Explaining Black-Box Models Through Statistical Inference. - Supplementary Document

Explainable AI for tabular models underpins decisions in finance, healthcare, and policy, yet today's explanations are dominated by heuristics without statistical guarantees. We introduce Stat-XAI, a model-agnostic framework that converts explanations into testable statistical statements. For each feature, Stat-XAI assesses association with model predictions on held-out data via appropriate hypothesis tests and reports standardized effect sizes (e.g., η^2, R^2, Cramér's V), yielding compact, uncertainty-aware rankings. Across six synthetic datasets with known causal structure and two real benchmarks, Stat-XAI delivers stable, parsimonious attributions, filters spurious correlates, and achieves orders-of-magnitude lower runtime than SHAP while maintaining faithfulness. We quantify stability under perturbations and show that interaction testing clarifies when pairwise dependencies meaningfully alter importance. By elevating explanation from heuristic scoring to inferential analysis, Stat-XAI provides a rigorous, reproducible pathway for trustworthy tabular AI—supporting scrutiny, governance, and human decision-making where reliability matters most.

1. INTRODUCTION

In this supporting document, we provide a detailed understanding of our framework which uses inferential statistics for explaining AI predictions. We present a comprehensive explanation of the synthetic datasets constructed to illustrate model predictions and to assess explanation accuracy using our developed metrics. We start by describing the generation of six synthetic datasets, consisting of different datatypes. In the next section, we justify our decision against normalizing effect sizes. Once we describe the generation method, we report the performance of STAT-XAI on these datasets, the next section applies the SHAP framework and offers a detailed comparison between SHAP and STAT-XAI. Finally, we extend STAT-XAI to two real-world datasets and conduct stability tests—perturbing input features to verify the consistency of explanations in practical settings. By evaluating both synthetic and real-world scenarios, we demonstrate that our approach not only provides superior explanations but also accurately identifies the causal features which are driving model predictions, a capability that existing methods lack.

2. SYNTHETIC DATASET CREATION

To evaluate the reliability and interpretability of explainability methods, we constructed a series of synthetic datasets designed to simulate realistic yet fully controlled conditions. Each dataset incorporates clearly defined causal relationships between features and outcomes, alongside distractor features without genuine causal roles. Specifically, six datasets were generated, differing systematically in feature composition (categorical, numerical, or mixed) and outcome type (binary or continuous).

Synthetic Data Generation (Dataset 1)

We generated $N = 10,000$ synthetic loan-applicant records, each with five categorical predictors:

- **Credit History** $\in \{\text{Bad}, \text{Fair}, \text{Good}\}$, with sampling probabilities $(0.2, 0.5, 0.3)$.
- **Income Level** $\in \{\text{Low}, \text{Medium}, \text{High}\}$, with probabilities $(0.4, 0.4, 0.2)$.
- **Loan Amount** $\in \{\text{Large}, \text{Medium}, \text{Small}\}$, with probabilities $(0.3, 0.4, 0.3)$.
- **Zip Code** (non-causal) $\in \{\text{Urban}, \text{Suburban}, \text{Rural}\}$, with $(0.5, 0.3, 0.2)$.
- **Education Level** (non-causal) $\in \{\text{High School}, \text{Bachelor}, \text{Master}\}$, with $(0.3, 0.4, 0.3)$.

A fixed random seed (`np.random.seed(42)`) ensures reproducibility. We encode each causal category into an integer:

$$\text{Bad} \mapsto 0, \quad \text{Fair} \mapsto 1, \quad \text{Good} \mapsto 2,$$

and similarly for `Income_Level` and `Loan_Amount`. We then form a latent approval score

$$S \;=\; w_{\text{cred}}\, x_{\text{cred}} \;+\; w_{\text{loan}}\, x_{\text{loan}} \;+\; w_{\text{inc}}\, x_{\text{inc}} \;+\; \delta,$$

with weights $(w_{\text{cred}}, w_{\text{loan}}, w_{\text{inc}}) = (2.0, 2.0, 1.5)$ and noise $\delta \sim \mathcal{N}(0, 2)$. We standardize S to zero mean and unit variance,

$$S_{\text{std}} \;=\; \frac{S - \mathbb{E}[S]}{\text{Std}(S)},$$

and convert to a probability via the logistic link,

$$p = \frac{1}{1 + \exp(-S_{\text{std}})}\,.$$

Finally, we threshold at $p > 0.5$ to obtain the binary label $\texttt{loan\_approval} = \mathbb{I}\{p > 0.5\} \in \{0, 1\}$.

In the first dataset, only Credit History, Loan Amount, and Income Level carry nonzero weights making them the true causal drivers. Whereas Zip Code and Education Level serve solely as distractors. Figure S1 presents a bar chart of the feature outcome correlations. The x-axis shows the dataset's feature names, and the y-axis reports each feature's correlation. As expected, Loan Amount, Credit History, and Income Level exhibit strong positive correlations, thus correctly identified as causal. While Education Level and Zip Code show negligible correlation and are classified as noncausal in this dataset.

Synthetic Data Generation (Dataset 2)

For our second dataset (categorical inputs, continuous outcome), we again simulate $N = 10{,}000$ loan-applicant records with the same five categorical features as in Dataset 1, now the output is continuous:

- **Credit History** $\in \{\text{Bad}, \text{Fair}, \text{Good}\}$, $p = (0.2, 0.5, 0.3)$.
- **Income Level** $\in \{\text{Low}, \text{Medium}, \text{High}\}$, $p = (0.4, 0.4, 0.2)$.
- **Loan Amount** $\in \{\text{Large}, \text{Medium}, \text{Small}\}$, $p = (0.3, 0.4, 0.3)$.
- **Zip Code** (distractor) $\in \{\text{Urban}, \text{Suburban}, \text{Rural}\}$, $p = (0.5, 0.3, 0.2)$.
- **Education Level** (distractor) $\in \{\text{High School}, \text{Bachelor}, \text{Master}\}$, $p = (0.3, 0.4, 0.3)$.

A fixed seed (`np.random.seed(42)`) guarantees reproducibility. Each causal category is mapped to an integer label—e.g. Bad $\mapsto 0, \ldots,$ Good $\mapsto 2$—and denoted $x_{\text{cred}}, x_{\text{inc}}, x_{\text{loan}}$.

We then synthesize a continuous approval score via a linear model with additive Gaussian noise:

$$S \;=\; w_{\text{cred}}\, x_{\text{cred}} \;+\; w_{\text{loan}}\, x_{\text{loan}} \;+\; w_{\text{inc}}\, x_{\text{inc}} \;+\; \delta, \qquad \delta \sim \mathcal{N}(0, \sigma^2),$$

where $(w_{\text{cred}}, w_{\text{loan}}, w_{\text{inc}}) = (2.0, 2.0, 1.5)$ and $\sigma = 2$. We standardize S to zero mean and unit variance,

$$S_{\text{std}} \;=\; \frac{S - \mathbb{E}[S]}{\text{Std}(S)},$$

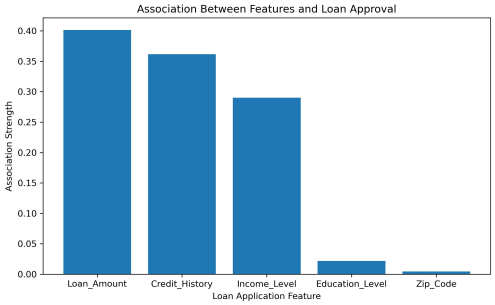


Fig. S1. Plotting features that are causal and non-causal in the dataset

and set the final continuous target as

$$Y = S_{\text{std}}.$$

By construction, only Credit History, Loan Amount, and Income Level drive Y, while Zip Code and Education Level remain non-causal. This controlled design—with known ground-truth drivers—permits quantitative evaluation (precision, recall, FDR, Top-1 match) of any post-hoc explanation method on its ability to recover only the true causal features. Figure S2 presents a bar chart of the feature outcome correlations. The x-axis shows the dataset's feature names, and the y-axis reports each feature's correlation. As expected, Loan Amount, Credit History, and Income Level exhibit strong positive correlations, thus correctly identified as causal. While Education Level and Zip Code show negligible correlation and are classified as noncausal in this dataset.

Synthetic Data Generation (Dataset 3)

For our third synthetic dataset (numerical features, binary outcome) we generated 10,000 observations as follows. Annual income was drawn at random over the range 30,000 to 1,25 ,000, and credit scores were sampled uniformly between 300 and 850. The debt-to-income ratio was assigned a value between 0 and 1, employment length a value between 0 and 20 years, and age an integer between 18 and 80. These five features were then combined in a linear–logistic model All random draws use a fixed seed (`np.random.seed(42)`) for reproducibility.

We construct a latent score on the log-odds scale via a linear combination of the three true causal features plus Gaussian noise:

$$\ell = \underbrace{0.003}_{\beta_1}\,\texttt{Annual\_Income} + \underbrace{0.5}_{\beta_2}\,\texttt{Credit\_Score} + \underbrace{250.0}_{\beta_3}\,\texttt{Debt\_to\_Income} + \eta, \quad \eta \sim \mathcal{N}(0, 5^2).$$

This score is then standardized to zero mean and unit variance:

$$\ell_{\text{std}} = \frac{\ell - \mathbb{E}[\ell]}{\text{Std}(\ell)}.$$

Applying the logistic (sigmoid) link,

$$p = \sigma(\ell_{\text{std}}) = \frac{1}{1 + \exp(-\ell_{\text{std}})},$$

we sample the binary outcome $\texttt{Loan\_Approval} \sim \text{Bernoulli}(p)$, i.e. $\texttt{Loan\_Approval} = 1$ if $p > 0.5$, else 0.

By construction, only Annual_Income, Credit_Score, and Debt_to_Income drive the approval decision, while Employment_Length and Age serve as non-causal distractors. This provides a clear ground truth for evaluating any explainability method's ability to recover the true drivers of a binary classification task. Figure S3 presents a bar chart of the feature outcome correlations. The x-axis shows the dataset's feature names, and the y-axis reports each feature's correlation. As expected, Annual Income, Credit Score, and Debt-to-Income exhibit strong positive correlations, thus correctly identified as causal. While Employment Length and Age show negligible correlation and are classified as noncausal in this dataset.

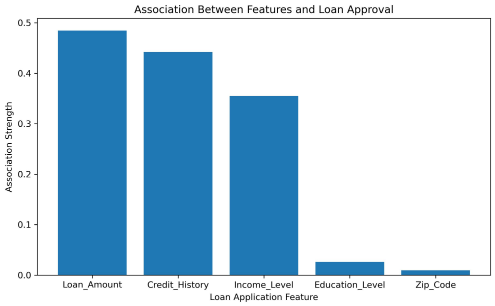


Fig. S2. Plotting features that are causal and non-causal in the dataset

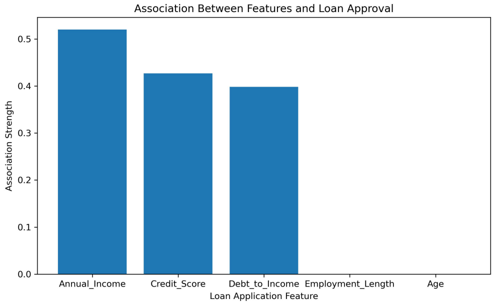


Fig. S3. Plotting features that are causal and non-causal in the dataset

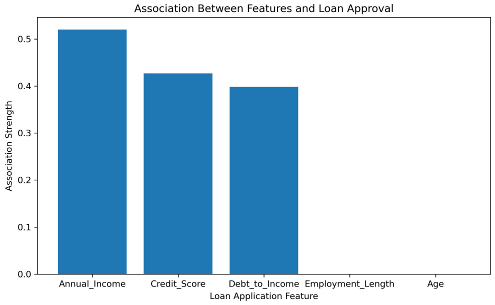


Fig. S4. Plotting features that are causal and non-causal in the dataset

Synthetic Data Generation (Dataset 4)

For our fourth synthetic dataset (numerical features, continuous outcome) we generated 10,000 observations as follows. Annual income was drawn at random over the range \$30,000 to \$1,25 ,000, and credit scores were sampled uniformly between 300 and 850. The debt-to-income ratio was assigned a value between 0 and 1, employment length a value between 0 and 20 years, and age an integer between 18 and 80. These five features were then combined in a linear–logistic model.

We then formed a latent score via a simple linear model with additive Gaussian noise:

$$\ell = \underbrace{0.003}_{\beta_1}\,\texttt{Annual\_Income} + \underbrace{0.5}_{\beta_2}\,\texttt{Credit\_Score} + \underbrace{250.0}_{\beta_3}\,\texttt{Debt\_to\_Income} + \eta, \quad \eta \sim \mathcal{N}(0, 5^2).$$

with coefficients $\beta_1 = 0.003$, $\beta_2 = 0.5$, $\beta_3 = 250$, and $\sigma = 5$. After standardizing s to zero mean and unit variance,

$$\tilde{s} = \frac{s - E[s]}{\mathrm{sd}(s)},$$

we applied the logistic (sigmoid) transform to obtain a continuous outcome,

$$Y = \frac{1}{1 + e^{-\tilde{s}}} \in (0, 1).$$

This probability-valued Y serves as the continuous target in the fourth dataset .

In this dataset, only Annual Income, Credit Score and DTI have nonzero coefficients and thus true influence on Y, while Employment Length and Age are non-causal distractors. Figure S4 presents a bar chart of the feature outcome correlations. The x-axis shows the dataset's feature names, and the y-axis reports each feature's correlation. As expected, Annual Income, Credit Score, and Debt-to-Income exhibit strong positive correlations, thus correctly identified as causal. While Employment Length and Age show negligible correlation and are classified as noncausal in this dataset. This setting allows us to assess our STAT-XAI pipeline's ability to recover those three drivers via main-effect testing and effect-size ranking.

Synthetic Data Generation (Dataset 5)

We generated $N = 10\,000$ observations featuring five numerical and five categorical features, of which only a handful of features are causally related to the binary loan-approval. The dataset consists of following features:

- **Numerical (causal):** Annual_Income $\sim$ Uniform$(30{,}000, 150{,}000)$, Credit_Score $\sim$ Uniform$(300, 850)$.
- **Numerical (non-causal):** Employment_Length $\sim$ Uniform$(0, 20)$, Age $\sim$ DiscreteUniform$(18, 80)$, Loan_Term $\sim$ DiscreteUniform$(1, 30)$.
- **Categorical (causal):** Loan_Purpose $\in$ {Home, Car, Personal}, Employment_Status $\in$ {Employed, Self-employed, Unemployed}, Loan_Categories $\in$ {Large, Medium, Small}.
- **Categorical (non-causal):** Region $\in$ {North, South, East, West}, Marital_Status $\in$ {Single, Married, Divorced}.

The equation, constructs each applicant's raw score ℓ_i by summing a random intercept $b_i \sim \mathcal{N}(0, 5^2)$ with continuous contributions from income ($\alpha = 0.0003$) and credit score ($\beta = 0.05$), and fixed categorical effects for loan purpose (δ_{Pur}), employment status (γ_{Emp}), and loan-size category (κ_{Cat}). These terms are chosen so that only the specified features causally drive ℓ_i. The resulting logit is then passed through a sigmoid to yield a controlled synthetic approval probability, ensuring a clear ground truth for evaluating explainability methods.

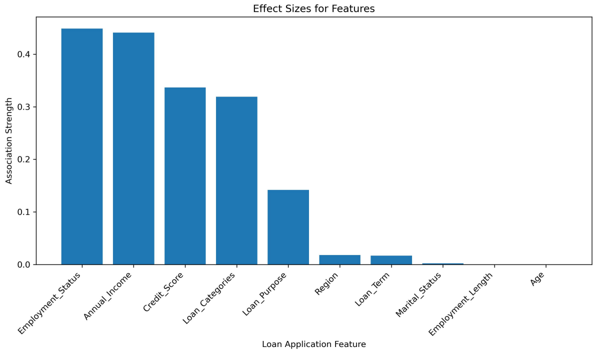


Fig. S5. Plotting features that are causal and non-causal in the dataset

$$\ell_i = \underbrace{b_i}_{\substack{\text{random intercept}\\ b_i \sim \mathcal{N}(0,5^2)}} + \underbrace{\alpha\,\text{Income}_i}_{\substack{\text{continuous}\\ \text{effect, } \alpha=0.0003}} + \underbrace{\beta\,\text{CreditScore}_i}_{\substack{\text{continuous}\\ \text{effect, } \beta=0.05}} + \underbrace{\delta_{\text{Pur}(i)}}_{\substack{\text{Loan purpose}\\ \text{effect}}} + \underbrace{\gamma_{\text{Emp}(i)}}_{\substack{\text{Employment}\\ \text{effect}}} + \underbrace{\kappa_{\text{Cat}(i)}}_{\substack{\text{Loan categories}\\ \text{effect}}} .$$

Here the categorical-feature coefficients are defined as

$$\delta_{\text{Pur}}(\cdot) = \begin{cases} 5, & \text{if Loan\_Purpose} = \text{Home}, \\ 3, & \text{if Loan\_Purpose} = \text{Car}, \\ 10, & \text{if Loan\_Purpose} = \text{Personal}, \end{cases} \qquad \gamma_{\text{Emp}}(\cdot) = \begin{cases} 10, & \text{if status} = \text{Employed}, \\ 20, & \text{if status} = \text{Self-employed}, \\ -5, & \text{if status} = \text{Unemployed}, \end{cases}$$

$$\kappa_{\text{Cat}}(\cdot) = \begin{cases} 2, & \text{if category} = \text{Large}, \\ 10, & \text{if category} = \text{Medium}, \\ 20, & \text{if category} = \text{Small}. \end{cases}$$

We then standardize ℓ to zero mean and unit variance:

$$\tilde{\ell} = \frac{\ell - E[\ell]}{\text{sd}(\ell)},$$

and apply the sigmoid link to obtain approval probability $p = \sigma(\tilde{\ell}) = 1/(1 + e^{-\tilde{\ell}})$. Finally, the binary label is set by thresholding:

$$\text{Approval} = \mathbb{I}\{p > 0.5\}.$$

In this dataset, only $\{Annual\_Income, Credit\_Score, Loan\_Purpose, Employment\_Status, Loan\_Categories\}$ truly drive the outcome, while the remaining five features act as non-causal distractors. Figure S5 presents a bar chart of the feature outcome correlations. The x-axis shows the dataset's feature names, and the y-axis reports each feature's correlation. As expected, Annual Income,Credit Score,Loan Purpose,Employment Status,Loan Categories exhibit strong positive correlations, thus correctly identified as causal. While the rest of the features show negligible correlation and are classified as noncausal in this dataset. This clear ground-truth separation enables precise evaluation of STAT-XAI's ability to recover and rank the genuine features in a mixed-feature, binary-classification setting.

Synthetic Data Generation (Dataset 6)

We generated $N = 10\,000$ observations featuring five numerical and five categorical features, of which only a subset of features are causally linked to a continuous outcome. The dataset consists of following features:

- **Numerical (causal):** Annual_Income $\sim$ Uniform$(30,000 - 150,000)$, Credit_Score $\sim$ Uniform$(300, 850)$.
- **Numerical (non-causal):** Employment_Length $\sim$ Uniform$(0, 20)$, Age $\sim$ DiscreteUniform$(18, 80)$, Loan_Term $\sim$ DiscreteUniform$(1, 30)$.
- **Categorical (causal):** Loan_Purpose $\in$ $\{$Home, Car, Personal$\}$, Employment_Status $\in$ $\{$Employed, Self-employed, Unemployed$\}$, Loan_Categories $\in$ $\{$Large, Medium, Small$\}$.
- **Categorical (non-causal):** Region $\in$ $\{$North, South, East, West$\}$, Marital_Status $\in$ $\{$Single, Married, Divorced$\}$.

We then construct each applicant's score ℓ_i as the sum of a random intercept $b_i \sim \mathcal{N}(0, 5^2)$ plus continuous contributions from income ($\alpha = 0.0003$) and credit score ($\beta = 0.05$), and fixed categorical effects for loan purpose (δ_{Pur}), employment status (γ_{Emp}), and loan-size category (κ_{Cat}):

$$\ell_i = \underbrace{b_i}_{\substack{\text{random intercept}\\ b_i \sim \mathcal{N}(0,5^2)}} + \underbrace{\alpha\,\text{Annual\_Income}_i}_{\substack{\text{continuous}\\ \text{effect, } \alpha=0.0003}} + \underbrace{\beta\,\text{Credit\_Score}_i}_{\substack{\text{continuous}\\ \text{effect, } \beta=0.05}} + \underbrace{\delta_{\text{Pur}(i)}}_{\substack{\text{Loan purpose}\\ \text{effect}}} + \underbrace{\gamma_{\text{Emp}(i)}}_{\substack{\text{Employment}\\ \text{effect}}} + \underbrace{\kappa_{\text{Cat}(i)}}_{\substack{\text{Loan categories}\\ \text{effect}}} .$$

The categorical-feature coefficients are defined by

$$\delta_{\text{Pur}}(\cdot) = \begin{cases} 5, & \text{Home,} \\ 3, & \text{Car,} \\ 10, & \text{Personal,} \end{cases} \qquad \gamma_{\text{Emp}}(\cdot) = \begin{cases} 10, & \text{Employed,} \\ 20, & \text{Self-employed,} \\ -5, & \text{Unemployed,} \end{cases} \qquad \kappa_{\text{Cat}}(\cdot) = \begin{cases} 2, & \text{Large,} \\ 10, & \text{Medium,} \\ 20, & \text{Small.} \end{cases}$$

We standardize ℓ_i to zero mean and unit variance,

$$\tilde{\ell}_i = \frac{\ell_i - \mathbb{E}[\ell]}{\text{sd}(\ell)},$$

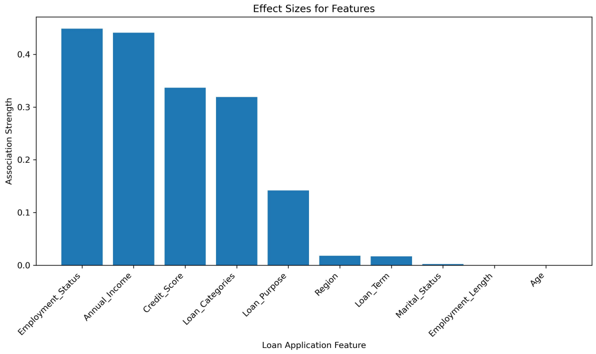


Fig. S6. Plotting features that are causal and non-causal in the dataset

and apply the sigmoid link to yield a continuous approval probability,

$$p_i = \sigma(\tilde{\ell}_i) \;=\; \frac{1}{1 + e^{-\tilde{\ell}_i}}.$$

In this dataset, only the five features $\{Annual\_Income, Credit\_Score, Loan\_Purpose, Employment\_Status, Loan\_Categories\}$ exert true causal influence on p_i, while the remaining five serve as non-causal distractors. Figure S6 presents a bar chart of the feature outcome correlations. The x-axis shows the dataset's feature names, and the y-axis reports each feature's correlation. As expected, Annual Income,Credit Score,Loan Purpose,Employment Status,Loan Categories exhibit strong positive correlations, thus correctly identified as causal. While the rest of the features show negligible correlation and are classified as noncausal in this dataset. This clear ground-truth separation enables precise evaluation of STAT-XAI's ability to recover and rank the genuine features in a mixed-feature, regression outcome setting.

3. HYPOTHESIS TESTING FOR MAIN-EFFECT ANALYSIS

Let $\hat{y} = f(X)$ denote the model's prediction and X_k an input feature. Depending on the datatype of X_k and $\hat{y}$, different hypothesis tests and effect size metrics are applied to quantify main effects.

1. **Categorical X_k, Binary $\hat{y}$ (Chi-square test):**

$$\chi^2 = \sum_{a=1}^{A} \sum_{b=0}^{1} \frac{(O_{ab} - E_{ab})^2}{E_{ab}}, \quad V = \sqrt{\frac{\chi^2}{N \cdot (k-1)}}, \tag{S1}$$

where O_{ab} and E_{ab} are observed and expected counts, and V (Cramér's V) measures association strength.

2. **Categorical X_k, Continuous $\hat{y}$ (One-way ANOVA):**

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}}, \quad \eta^2 = \frac{SS_{\text{effect}}}{SS_{\text{total}}}, \tag{S2}$$

testing whether group means of $\hat{y}$ differ significantly across levels of X_k.

3. **Continuous X_k, Binary $\hat{y}$ (Point-biserial correlation):**

$$r_{pb} = \frac{\bar{x}_1 - \bar{x}_0}{s_x} \sqrt{\frac{n_0 n_1}{N^2}}, \tag{S3}$$

where $\bar{x}_1, \bar{x}_0$ are feature means for classes $\hat{y} = 1, 0$, and s_x is the pooled standard deviation.

4. **Continuous X_k, Continuous $\hat{y}$ (Pearson correlation):**

$$r = \frac{\text{Cov}(X_k, \hat{y})}{\sigma_{X_k} \sigma_{\hat{y}}}, \tag{S4}$$

measuring linear dependence between the feature and the predictions.

5. **Mixed Inputs (Continuous + Categorical), Binary Output $\hat{y}$**

When the dataset has mixed input features first, we quantify each feature's main effect by testing its association with the model's predictions $\hat{y}$. For numerical predictors X_i and binary predictions $\hat{y} \in \{0, 1\}$, we compute the point–biserial correlation

$$r_{pb}^{(i)} = \frac{\overline{X}_{i\,|\,\hat{y}=1} - \overline{X}_{i\,|\,\hat{y}=0}}{s_{X_i}} \sqrt{\frac{n_1\, n_0}{n(n-1)}}, \tag{S5}$$

where $\overline{X}_{i\,|\,\hat{y}=1}$ and $\overline{X}_{i\,|\,\hat{y}=0}$ denote the means of X_i in the two predicted classes, s_{X_i} is the overall standard deviation, and n_1, n_0 are the sample sizes of the two predicted classes ($n = n_0 + n_1$). We then test

$$H_0 : r_{pb}^{(i)} = 0 \quad \text{using} \quad t^{(i)} = r_{pb}^{(i)} \sqrt{\tfrac{n-2}{1-(r_{pb}^{(i)})^2}} \;\sim\; t_{n-2}. \tag{S6}$$

Whenever $p < 0.05$, we record $|r_{pb}^{(i)}|$ as the effect size for X_i.

For categorical predictors C_j with k_j levels, we form the $2 \times k_j$ contingency table between C_j and $\hat{y}$. Let O_{ab} denote the observed count in row $a \in \{0, 1\}$ (predicted class) and column $b \in \{1, \ldots, k_j\}$ (category level), and $E_{ab} = \frac{n_{a\cdot} n_{\cdot b}}{n}$ the expected count under independence. The Pearson chi–squared statistic is

$$\chi_j^2 = \sum_{a=1}^{2} \sum_{b=1}^{k_j} \frac{(O_{ab} - E_{ab})^2}{E_{ab}}. \tag{S7}$$

We test $H_0 : C_j \perp \hat{y}$ at $\alpha = 0.05$, and for significant results ($p < 0.05$), compute the effect size.

6. **Mixed Inputs (Continuous + Categorical), Continuous $\hat{y}$:**

In the sixth synthetic dataset, which contains both numerical and categorical features with a continuous model output, we apply a two–stage statistical pipeline on the test data to identify and quantify the main effects of individual predictors on the model's predictions $\hat{y}$.

For each numerical feature X_i, we fit an ordinary least–squares (OLS) regression of the form

$$\hat{y} \; = \; \beta_0 + \beta_i X_i + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)\,. \tag{S8}$$

We test the null hypothesis $H_0 \colon \beta_i = 0$ using the standard t–statistic

$$t_i \; = \; \frac{\widehat{\beta}_i}{\mathrm{SE}(\widehat{\beta}_i)} \; \sim \; t_{\,n-2}, \tag{S9}$$

For each categorical feature C_j with k_j levels, we perform a one–way analysis of variance (ANOVA) between the feature and the predictions. The between–group and within–group sums of squares are

$$\mathrm{SS}_{\mathrm{between},\,j} = \sum_{\ell=1}^{k_j} n_{j,\ell}\,(\mu_{j,\ell} - \overline{\hat{y}})^2, \qquad \mathrm{SS}_{\mathrm{within},\,j} = \sum_{\ell=1}^{k_j} \sum_{i:\,C_{j,i}=\ell} (\hat{y}_i - \mu_{j,\ell})^2. \tag{S10}$$

The resulting F–statistic,

$$F_j = \frac{\mathrm{SS}_{\mathrm{between},\,j}/(k_j - 1)}{\mathrm{SS}_{\mathrm{within},\,j}/(n - k_j)}, \quad F_j \sim F_{k_j-1,\,n-k_j}, \tag{S11}$$

is used to test $H_0 \colon C_j \perp \hat{y}$. When $p < 0.05$, we compute the effect size which quantifies the proportion of variance in the predictions $\hat{y}$ explained by the categorical feature C_j.

4. HYPOTHESIS TESTING FOR PAIRWISE EFFECT ANALYSIS

To evaluate whether pairs of features jointly influence the model's predictions beyond their main effects, we employ different statistical tests depending on the structure of the predictors and the type of model output. Table S1 summarizes the test selection.

Let $\hat{y} = f(X)$ denote the model's prediction and X_k an input feature. Depending on the datatype of X_k and $\hat{y}$, different hypothesis tests are applied to quantify pairwise effects:

1. **Categorical X_k, Binary $\hat{y}$:** When both features are categorical and the model output is binary, we form a two-way contingency table between the feature pair and the binary predictions $\hat{y}$ and apply the Pearson chi-squared test. A significant statistic indicates that the joint distribution of the two features is associated with the prediction beyond marginal independence. The Pearson chi-squared statistic is

$$\chi^2 = \sum_{a=1}^{2} \sum_{b=1}^{k} \frac{(O_{ab} - E_{ab})^2}{E_{ab}}, \quad \chi^2 \sim \chi^2_{(k-1)}. \tag{S12}$$

A significant result ($p < 0.05$) indicates dependence between C_j and $\hat{y}$ beyond marginal effects.

#	Dataset Structure	Statistical Test	Feature and Output Types
1	Categorical feature, binary output	Pearson Test	Categorical IV, Binary DV
2	Categorical feature, continuous output	Two-way ANOVA	Categorical IV, Continuous DV
3	Continuous feature, Binary output	Likelihood-ratio (Logistic regression)	Continuous IV and Binary DV
4	Continuous feature, continuous output	OLS Regression	Continuous IV, Continuous DV
5	Continuous and Categorical feature, Binary output	Logistic regression, Likelihood-ratio and Chi-squared	Continuous and Categorical IV and Binary DV
6	Continuous and Categorical feature, continuous output	Regression and ANCOVA	Continuous and Categorical feature IV, continuous output DV

Table S1. Expanded description of statistical tests conducted for pairwise effect calculation used across different feature and output types.

2. **Categorical X_k, Continuous $\hat{y}$:** When input features are categorical and the output is continuous, we use two-way ANOVA to compare a reduced additive model against a full model including the interaction terms. This allows us to test whether the effect of one categorical predictor on $\hat{y}$ depends on the levels of the other.

   For two categorical features, C_j and $C_{j'}$ with k_j and $k_{j'}$ levels, respectively, and a continuous prediction $\hat{y} \in \mathbb{R}$, we use two-way ANOVA. Let $\mu_{\ell m}$ denote the mean prediction in cell (ℓ, m) and $\bar{\hat{y}}$ the grand mean. The interaction sum of squares is

$$SS_{\text{int}} = \sum_{\ell=1}^{k_j} \sum_{m=1}^{k_{j'}} n_{\ell m} \left(\mu_{\ell m} - \mu_{\ell \cdot} - \mu_{\cdot m} + \bar{\hat{y}}\right)^2. \tag{S13}$$

   The F-statistic for the interaction is

$$F = \frac{SS_{\text{int}} / ((k_j - 1)(k_{j'} - 1))}{SS_{\text{error}} / (n - k_j k_{j'})}, \quad F \sim F_{(k_j-1)(k_{j'}-1),\, n-k_j k_{j'}}.$$

3. **Continuous X_k, Binary $\hat{y}$:** For continuous predictors with binary outputs, we employ logistic regression. We fit a reduced model with only main effects and compare it to a full model with an added interaction term. The likelihood-ratio test between the two models provides a formal hypothesis test for the interaction. Significant results imply that one feature depends on the value of the other in explaining $\hat{y}$.

   For two continuous predictors X_a and X_b with binary predictions $\hat{y} \in \{0, 1\}$, we fit nested logistic regression models. The reduced model is

$$\operatorname{logit}\big(P(\hat{y} = 1 \mid X_a, X_b)\big) = \beta_0 + \beta_a X_a + \beta_b X_b, \tag{S14}$$

   and the full model includes their interaction,

$$\operatorname{logit}\big(P(\hat{y} = 1 \mid X_a, X_b)\big) = \beta_0 + \beta_a X_a + \beta_b X_b + \beta_{ab}(X_a X_b). \tag{S15}$$

   The likelihood-ratio statistic

$$D_{\text{LR}} = 2(\ell_{\text{full}} - \ell_{\text{reduced}}) \;\sim\; \chi^2_1,$$

   A significant result implies non-additivity, and the effect size is reported.

4. **Continuous X_k, Continuous $\hat{y}$:**

   For continuous inputs with continuous outputs, we use ordinary least squares (OLS) regression. The reduced model contains only the main effects, while the full model includes an interaction term $X_a \times X_b$. An F–test on the reduction in residual sum of squares determines the statistical significance of the interaction.

For two continuous inputs X_a and X_b with continuous predictions $\hat{y} \in \mathbb{R}$, the reduced OLS regression model is

$$\hat{y} = \beta_0 + \beta_a X_a + \beta_b X_b + \varepsilon, \tag{S16}$$

and the full model includes their product term,

$$\hat{y} = \beta_0 + \beta_a X_a + \beta_b X_b + \beta_{ab}(X_a X_b) + \varepsilon. \tag{S17}$$

Let $\mathrm{RSS}_{\mathrm{red}}$ and $\mathrm{RSS}_{\mathrm{full}}$ be the residual sums of squares. The F-statistic is

$$F = \frac{(\mathrm{RSS}_{\mathrm{red}} - \mathrm{RSS}_{\mathrm{full}})/1}{\mathrm{RSS}_{\mathrm{full}}/(n-3)}, \quad F \sim F_{1,n-3}.$$

5. **Mixed Inputs (Continuous + Categorical), Binary Output $\hat{y}$**

For datasets with mixed continuous and categorical input features and a binary output ($\hat{y} \in \{0,1\}$), pairwise interactions are tested using nested logistic regression models. For each unordered pair of features (X_a, X_b), we compare a reduced model containing only main effects against a full model that additionally includes their interaction.

(i) Numerical–Numerical Interactions. For two continuous predictors X_a and X_b, the reduced logistic model is

$$\operatorname{logit}\big[P(\hat{y} = 1 \mid X_a, X_b)\big] = \beta_0 + \beta_a X_a + \beta_b X_b,$$

while the full model adds their product term,

$$\operatorname{logit}\big[P(\hat{y} = 1 \mid X_a, X_b)\big] = \beta_0 + \beta_a X_a + \beta_b X_b + \beta_{ab}(X_a X_b).$$

The significance of the interaction is assessed using the likelihood-ratio statistic

$$D_{\mathrm{LR}} = 2(\ell_{\mathrm{full}} - \ell_{\mathrm{reduced}}) \ \sim\ \chi^2_1,$$

where ℓ denotes the model log-likelihood. A significant result ($p < 0.05$) indicates that the two predictors interact in influencing the probability of $\hat{y} = 1$.

(ii) Numerical–Categorical Interactions. For a numerical feature X_i and a categorical feature C_j with k_j levels, the reduced model includes their main effects, while the full model augments this with interaction terms of the form $X_i \times \mathbf{1}_{\{C_j=\ell\}}$ for $\ell = 1, \ldots, k_j - 1$. A likelihood-ratio test with $k_j - 1$ degrees of freedom evaluates the significance of the block of interaction terms.

(iii) Categorical–Categorical Interactions. For two categorical predictors C_j and $C_{j'}$, the reduced logistic model contains their additive dummy-coded main effects, while the full model additionally includes all cross-classified dummy-coded interaction terms. The significance of the interaction block is tested using a likelihood-ratio statistic with $(k_j - 1)(k_{j'} - 1)$ degrees of freedom.

Across all three cases, likelihood-ratio testing provides a formal hypothesis-testing framework for detecting significant feature interactions. Effect sizes ($|\beta|$, odds ratios) quantify the magnitude of detected interactions, yielding a statistically rigorous characterization of joint feature contributions in mixed input, binary output datasets.

6. **Mixed Inputs (Continuous + Categorical), Continuous Output $\hat{y}$**

For datasets containing both continuous and categorical predictors with a continuous output, ($\hat{y} \in \mathbb{R}$), pairwise interactions are evaluated by fitting reduced and full regression models and comparing their explanatory power. Each unordered pair of predictors (X_a, X_b) falls into one of three cases.

(i) Numerical–Numerical Interactions. For two continuous predictors X_a and X_b, the reduced model includes only main effects,

$$\hat{y} = \beta_0 + \beta_a X_a + \beta_b X_b + \varepsilon,$$

while the full model additionally includes their product,

$$\hat{y} = \beta_0 + \beta_a X_a + \beta_b X_b + \beta_{ab}(X_a X_b) + \varepsilon.$$

The incremental contribution of the interaction is tested using an F–test on the reduction in residual sum of squares, and the corresponding effect size is computed as the change in explained variance $\Delta R^2_{ab} = R^2_{\text{full}} - R^2_{\text{res}}$.

(ii) Numerical–Categorical Interactions. For a numerical predictor X_i and a categorical predictor C_j with k_j levels, the reduced ANCOVA model contains main effects for both variables, while the full model augments this with interaction terms of the form $X_i \times \mathbf{1}_{\{C_j=\ell\}}$ for $\ell = 1, \ldots, k_j - 1$. An F–test on the block of interaction coefficients determines whether the relationship between X_i and $\hat{y}$ depends significantly on the levels of C_j.

(iii) Categorical–Categorical Interactions. For two categorical predictors C_j and $C_{j'}$, we perform a two-way ANOVA. The reduced model contains only the additive main effects, while the full model additionally incorporates the dummy-coded interaction terms. An F–test on the interaction block evaluates significance. Across all three cases, a significant p–value (< 0.05) indicates that the joint contribution of the feature pair cannot be explained by additive main effects alone.

Across all cases, hypothesis testing establishes whether the interaction between a feature pair significantly influences the model's predictions $\hat{y}$. Stat-XAI systematically applies appropriate hypothesis tests to detect and quantify pairwise interactions in the model's predictions.

5. EFFECT SIZE

In statistical hypothesis testing, the p-value provides evidence against the null hypothesis but does not quantify the magnitude importance of the observed effect. Effect size measures significance testing by quantifying the strength of the relationship between variables or the proportion of variance in the outcome explained by a predictor. For instance, Cohen's d expresses the standardized difference between group means; in ANOVA.

It captures the degree of association between categorical variables. Unlike p-values, which are influenced by sample size, effect sizes enables a more interpretable assessment of feature importance. In the context of XAI, this distinction is critical: Stat-XAI assumes, if a feature is influential, its effect should be statistically detectable in the output distribution. Standard inferential procedures (e.g., t-tests, ANOVA, χ^2) are applied under their conventional assumptions.while hypothesis testing establishes whether a feature significantly influences predictions, effect size quantifies the strength of that influence, ensuring that explanations reflect both statistical reliability and practical relevance. While hypothesis testing establishes whether a feature significantly influences predictions, effect size quantifies the magnitude of that influence.

In the context of XAI, this distinction is critical: while hypothesis testing establishes whether a feature significantly influences predictions, effect size quantifies the strength of that influence, ensuring that explanations reflect both statistical reliability and practical relevance.

Finally, main and pairwise effects are combined together, through the following equation:

$$\text{Final Feature Score}(X_i) = \text{MainEffect}_i + \sum_{j \neq i} \text{InteractionEffect}_{i,j} \tag{S18}$$

Where:

$$\text{MainEffect}_i = \begin{cases} \eta_i^2, & \text{if } p_i < \alpha \\ 0, & \text{otherwise} \end{cases}$$

$$\text{InteractionEffect}_{i,j} = \begin{cases} \eta_{i,j}^2, & \text{if } p_{i,j} < \alpha \\ 0, & \text{otherwise} \end{cases}$$

Only statistically significant effects (based on p-values) are included, ensuring robustness and interpretability.

Table S2 summarizes the effect size measures used in Stat-XAI to quantify main effects across different combinations of input and output variable types. Because effect size quantifies the

Table S2. Main Effect Sizes for Different Datasets

Dataset	Input Datatype	Output Datatype	Main Effect-size
1	Categorical Features	Binary Features	Cramér's V
2	Categorical Features	Continuous Features	η^2
3	Numerical Features	Binary Features	Correlation r
4	Numerical Features	Continuous Features	Correlation r
5	Numerical and Categorical	Binary Features	Correlation r and Cramér's V
6	Numerical and Categorical	Continuous Features	R^2 and η^2

strength of association between features and model predictions, the choice of metric depends on whether the variables are categorical, numerical, or mixed.

For categorical inputs with binary outputs, Cramér's V is employed, as it measures association strength in contingency tables. For categorical inputs with continuous outputs, η^2 quantifies the proportion of variance in the continuous outcome explained by the categorical features. For numerical inputs with binary outcomes, the point-biserial correlation r captures the strength of association between the numerical predictor and the binary target. When both inputs and outputs are continuous, the correlation r measures the strength and direction of linear association. In datasets with mixed inputs (numerical and categorical) and binary outputs, both correlation r (for numerical features) and Cramér's V (for categorical features) are applied. Finally, for mixed inputs with continuous outputs, a combination of R^2 (from regression, for numerical features) and η^2 is used to capture the proportion of variance explained by different feature types.

In this way, Stat-XAI assigns each dataset structure with an appropriate effect size measure, ensuring that feature importance is consistently quantified regardless of the input or output type.

Table S3. Pairwise Effect Sizes for Different Datasets

Dataset	Input Datatype	Output Datatype	Pairwise Effect-size
1	Categorical Features	Binary Features	Cramér's V
2	Categorical Features	Continuous Features	η^2
3	Numerical Features	Binary Features	r^2
4	Numerical Features	Continuous Features	r^2
5	Numerical and Categorical	Binary Features	β
6	Numerical and Categorical	Continuous Features	r^2 and η^2

Table S3 outlines the effect size employed in Stat-XAI to quantify pairwise interactions between features across different dataset structures. Main-effect metrics evaluate the independent contribution of a single feature, while pairwise effect sizes evaluates combinations of two features jointly on model predictions.

For categorical inputs with binary outputs, Cramér's V is used to capture the strength of association in contingency tables. For categorical inputs and continuous outcomes, η^2 quantifies the variance explained by main and interaction terms. For numerical inputs with binary outcomes, the squared correlation r^2 measures the proportion of variance in the binary predictions attributable to the interaction. In the case of numerical inputs with continuous outputs, r^2 similarly captures the variance explained by joint effects. For mixed inputs (numerical and categorical) with binary outcomes, regression coefficients β are employed to assess the relative contribution of each feature within interaction terms. Finally, when mixed inputs predict continuous outcomes, a combination of r^2 (for numerical interactions) and η^2 (for categorical interactions) provides a comprehensive measure of joint feature influence.

This ensures that Stat-XAI applies an appropriate, test-specific effect size to quantify interaction strength, thereby extending interpretability beyond individual features to feature pairs.

6. PERFORMANCE EVALUATION OF AI MODELS

Table S4. Performance analysis of MLP and RNN on the first synthetic dataset.

Class	MLP			RNN		
	Prec.	Rec.	F1	Prec.	Rec.	F1
0	0.86	0.64	0.73	0.84	0.67	0.74
1	0.70	0.89	0.78	0.72	0.87	0.79

Table S5. Regression analysis of MLP and RNN for the second dataset.

Model	MAE	MSE	R^2
MLP	0.0199	0.1138	0.5477
RNN	0.0215	0.1202	0.555

Table S6. Performance analysis of MLP and RNN on the third synthetic dataset.

Class	MLP			RNN		
	Prec.	Rec.	F1	Prec.	Rec.	F1
0	0.994	0.996	0.95	0.975	0.985	0.96
1	0.995	0.993	0.90	0.980	0.975	0.91

Table S7. Regression analysis of MLP and RNN for the fourth dataset.

Model	MAE	MSE	R^2
MLP	0.0109	0.0138	0.995
RNN	0.0120	0.0023	0.999

Although our focus is not on the predictive performance of the primary AI models, we report their performances across the six synthetic datasets and three real-world datasets in Tables (S4-S9) for synthetic and Table (S10 and S11)for real-world comparison. These serve as a reference baseline before constructing the STAT-XAI explainer, whose core components—representative feature selection and mode identification—are described in the following section.

Table S8. Performance analysis of MLP and RNN on the fifth synthetic dataset.

Class	MLP			RNN		
	Prec.	Rec.	F1	Prec.	Rec.	F1
0	0.89	0.90	0.89	0.91	0.90	0.89
1	0.89	0.91	0.90	0.90	0.93	0.91

Table S9. Regression analysis of MLP and RNN for the sixth dataset.

Model	MAE	MSE	R^2
MLP	0.053	0.0045	0.898
RNN	0.049	0.0043	0.903

7. STAT-XAI RESULTS ON FIRST DATASET

Table S12 presents the results of one-way tests for each feature's relationship to the binary outcome in our first synthetic dataset, shown separately for three neural architectures: a multilayer perceptron (MLP), an RNN, and a -. For each feature, we report the "Ground Truth" label indicating its simulated causal strength ("High," "Medium," or "None"), the p-value from a chi-squared test of independence between that feature and the outcome, and—only for those features with $p < 0.05$—the corresponding Cramér's V effect size. Suppose a categorical feature C has k levels and the binary outcome $Y \in \{0, 1\}$; we form a $2 \times k$ contingency table with observed counts O_{ij} (where $i \in \{0, 1\}$ indexes the outcome level and $j \in \{1, \ldots, k\}$ indexes the feature level).

For the MLP experiment, Loan Amount has the largest effect ($\eta^2 = 0.50$, $p = 0.0$, rank 1), followed by Income Level ($\eta^2 = 0.45$, $p = 0.0$, rank 2) and Credit History ($\eta^2 = 0.43$, $p = 0.0$, rank 3), matching their simulated "High" vs. "Medium" strengths. Education Level, though non-causal, yields a small but significant spurious effect ($\eta^2 = 0.05$, $p = 0.006$, rank 4). Zip Code correctly fails to reach significance ($p = 0.054$), so no effect size or rank is reported.

Again for RNN architecture, Loan Amount dominates ($\eta^2 = 0.479$, $p = 0.0$, rank 1), closely followed by Credit History ($\eta^2 = 0.47$, $p = 0.0$, rank 2) and Income Level ($\eta^2 = 0.39$, $p = 0.0$,

Table S10. Model Performance of MLP and RNN on the German Credit Dataset.

Model	Accuracy	Precision	Recall
MLP	0.69	0.79	0.78
RNN	0.70	0.70	0.96

Table S11. Model Performance of MLP and RNN on the Census Income Dataset.

Model	Accuracy	Precision	Recall
MLP	0.82	0.67	0.54
RNN	0.83	0.67	0.58

Table S12. Main Effect Statistical Results on First Synthetic Dataset (Categorical Features and Binary Outcome)

Model: Multi-layer Perceptron

Feature	Ground Truth	p-value	Effect size	Ranking
Credit History	High	0.0	0.43	2
Income Level	Medium	0.0	0.45	3
Loan Amount	High	0.0	0.50	1
Zip Code	None	0.054	-	-
Education Level	None	0.006	0.050	4

Model: Recurrent Neural Network

Feature	Ground Truth	p-value	Effect size	Ranking
Credit History	High	0.0	0.47	2
Income Level	Medium	0.0	0.39	3
Loan Amount	High	0.0	0.479	1
Zip Code	None	0.21	-	-
Education Level	None	0.0069	0.05	4

rank 3). Education Level appears borderline ($\eta^2 = 0.05$, $p = 0.0069$, rank 4), while Zip Code remains non-significant ($p = 0.21$).

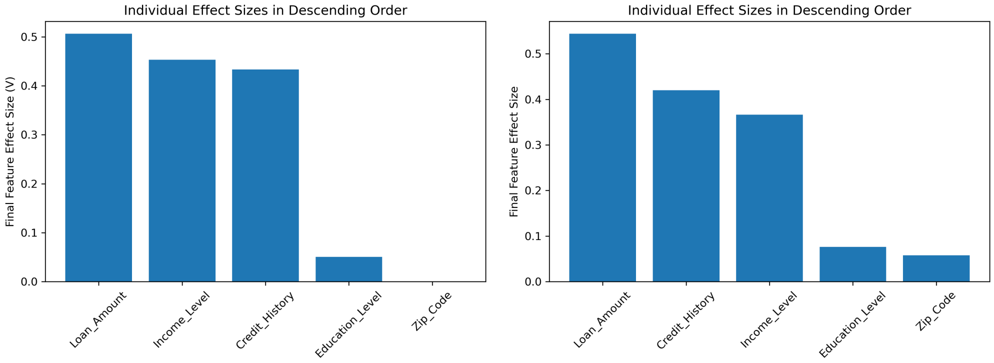
All the two neural architectures successfully recover the ground-truth ordering by identifying Loan Amount, Credit History, and Income Level as the only features with strong, significant main effects, with ranks that match their simulated strengths. Non-causal features (Zip Code and Education Level) either do not achieve statistical significance or exhibit negligible η^2 values, illustrating the robustness of our inferential testing pipeline in filtering irrelevant variables.

Main Effect Results on First Synthetic Dataset (Categorical Features and Binary Outcome)

Because our synthetic dataset has only three causal features and two non-causal, the explanation pipelines across MLP, RNN, and - converge to the same feature rankings and classification-metric outcomes. The uniform precision, recall, FDR, and Top-1 match reflect both the simplicity of the data and the robustness of our evaluation framework.

Until now, no standardized, quantitative methodology existed to assess how accurately an XAI method recovers true causal drivers. By designing a dataset with known feature-outcome relationships, we can treat the explanation task as a classification problem: causal vs. non-causal. This allows us to leverage well-understood metrics (precision, recall, FDR, Top-1 match) to benchmark and compare XAI methods in a rigorous, reproducible way.

Because this first dataset limited to categorical predictors, the performance is relatively low, there is little nuance for the models to misinterpret, and trivial patterns dominate. As a result, precision remains at 0.75 and FDR at 0.25. In future experiments, we will extend this evaluation to: Larger feature sets with mixed data types (continuous, ordinal, binary) to stress-test the



(a) MLP

(b) RNN

Fig. S7. Main-Effect Effect Sizes for Individual Features in the First Synthetic Dataset, as Computed by STAT-XAI for the Two Architectures. Each subfigure displays the features sorted in descending order of effect size, highlighting the top features identified by the MLP, RNN models, respectively. These bar charts illustrate which features exhibit the strongest statistically significant associations with the binary outcome.

explanation methods.

As illustrated in Figure S7, the MLP, RNN, and - models consistently rank the most causal features—Loan Amount, Credit History, and Income Level—at the top based on their effect sizes.

Pairwise Interaction Results on First Synthetic Dataset (Categorical Features and Binary Outcome)

Table S13 presents, for each feature pair and each model, the estimated p-value from the three-way chi-squared test and the corresponding Cramér's V, along with a "Significant?" indicator based on $p < 0.05$. All examined pairs satisfy $p < 0.05$, demonstrating pervasive second-order interactions in the learned decision boundaries. Notably, Income Level×Loan Amount and Credit History×Loan Amount exhibit the largest V (roughly 0.64–0.79), whereas Zip Code×Education Level yields $V \approx 0.07$–0.08, indicating a small but detectable effect. Moreover, although significance is universal, the exact Cramér's V values vary slightly across MLP, RNN, reflecting each architecture's inductive biases in capturing pairwise synergies.

By comparison, our main-effect tests (Table S12) assess only marginal associations between a single feature and Y. The pairwise analysis thus "goes one step deeper," uncovering whether combinations of two predictors jointly influence the outcome beyond what univariate tests can detect. In particular, it reveals synergistic or antagonistic interactions that would remain hidden under individual chi-squared tests alone.

Overall, these results demonstrate that—on top of strong main effects, pairwise interactions are present in the data and are consistently learned by all three network architectures.

Table S14 presents the pairwise interaction effects for each feature across three different neural network architectures—Multi-layer Perceptron (MLP), Recurrent Neural Network (RNN), trained on the first synthetic dataset. For each model, we quantify the cumulative pairwise interaction effect of a feature by summing its statistically significant interaction effect sizes with all other features. These values indicate how much a feature jointly influences the model output in combination with other features, independent of its individual (main) effect.

Across all models, Loan Amount consistently exhibits the highest pairwise interaction effects, suggesting it plays a central role in shaping the decision boundary through joint effects with other variables. Conversely, features such as Zip Code and Education Level show relatively lower interaction contributions, implying they are less involved in synergistic effects with other features. This interaction focused analysis provides deeper insight into how different models utilize feature combinations, which is critical for understanding complex, non-linear decision-making in high-stakes applications.

Because our synthetic ground truth contains a limited number of true pairwise interactions (and all features are categorical), the classifiers achieve perfect recall but only moderate precision of the detected interactions are false alarms. The fact that all three models match exactly on these

Table S13. Pairwise Effect Size Statistical Results on First Synthetic Dataset (Categorical Features and Binary Outcome)

Model: Multi-layer Perceptron

Feature Pair	**p-value**	**Effect size**	**Significant?**
Credit History × Income Level	0.00	0.645	Yes
Credit History × Loan Amount	0.00	0.697	Yes
Credit History × Zip Code	0.00	0.434	Yes
Credit History × Education Level	0.00	0.435	Yes
Income Level × Loan Amount	0.00	0.712	Yes
Income Level × Zip Code	0.00	0.455	Yes
Income Level × Education Level	0.00	0.455	Yes
Loan Amount × Zip Code	0.00	0.507	Yes
Loan Amount × Education Level	0.00	0.508	Yes
Zip Code × Education Level	0.00	0.071	Yes

Model: Recurrent Neural Network (LSTM)

Feature Pair	**p-value**	**Effect size**	**Significant?**
Credit_History × Income_Level	0.00	0.628	Yes
Credit_History × Loan_Amount	0.00	0.711	Yes
Credit_History × Zip_Code	0.00	0.483	Yes
Credit_History × Education_Level	0.00	0.488	Yes
Income_Level × Loan_Amount	0.00	0.640	Yes
Income_Level × Zip_Code	0.00	0.412	Yes
Income_Level × Education_Level	0.00	0.415	Yes
Loan_Amount × Zip_Code	0.00	0.480	Yes
Loan_Amount × Education_Level	0.00	0.493	Yes
Zip_Code × Education_Level	0.02	0.076	Yes

Table S14. Cumulative pairwise interaction effect sizes on the first synthetic dataset (categorical features, binary outcome)

Feature	MLP	RNN
Credit History	2.2128	2.3100
Income Level	2.2689	2.0962
Loan Amount	2.4268	2.3252
Zip Code	1.4686	1.4510
Education Level	1.4709	1.4724

metrics highlights that, under this simple setting, their explanation pipelines produce identical pairwise-interaction rankings.

By evaluating the explanation task as a binary classification on feature pairs (causal vs. non-causal) and using precision/recall/FDR/Top-1 match, this method goes beyond main-effect testing and allows us to benchmark and compare models on their ability to explain not just individual features but also the relationships among them.

Final Interaction (Main + Pairwise Effect) on First Dataset

Table S15. Final Interaction on First Synthetic Dataset (Categorical Features and Binary Outcome)

Model: Multi-layer Perceptron

Feature	Main Interaction	Pairwise Interaction	Final Interaction
Credit History	0.4333	2.2128	2.2128
Income Level	0.4530	2.2689	2.2689
Loan Amount	0.5063	2.4268	2.4268
Zip Code	0.0000	1.4686	1.4686
Education Level	0.0506	1.4709	1.4709

Model: Recurrent Neural Network

Feature	Main Interaction	Pairwise Interaction	Final Interaction
Credit_History	0.4799	2.3100	2.3100
Income_Level	0.3988	2.0962	2.0962
Loan_Amount	0.4735	2.3252	2.3252
Zip_Code	0.0000	1.4510	1.4510
Education_Level	0.0576	1.4724	1.4724

Table S15 brings together each feature's "main interaction" (Cramér's V from the chi-squared test) and its aggregated pairwise interaction score (the sum of Cramér's V over all pairs involving that feature). We then take the pairwise total as the *final interaction* score, since in our synthetic setting the joint effects dominate the univariate contributions.

- **Main Interaction:** the single-feature association strength with the outcome, measured by Cramér's V.

- **Pairwise Interaction:** for each feature X_k, the sum

$$\sum_{i \neq k} V(X_k, X_i; Y)$$

of its Cramér's V values with every other feature X_i.

- **Final Interaction:** set equal to the pairwise total, reflecting that second-order effects are the primary drivers of model decision boundaries in our dataset.

For MLP model, Loan Amount exhibits the strongest overall interactions (Final=2.4268), followed by Income Level (2.2689) and Credit History (2.2128). Zip Code and Education Level, although non-causal in the data-generating process, still accumulate modest pairwise signals (1.4686 and 1.4709), reflecting incidental associations.

The RNN model yields similar relative ordering: Loan Amount (2.3252)>Credit History (2.3100)>Income Level (2.0962). Non-causal features again score around 1.45–1.47.

Across all architectures, the three true causal features—Loan Amount, Credit History, and Income Level—emerge with the highest composite interaction scores, demonstrating that our two-step procedure (main effect + pairwise effect) correctly highlights the features most deeply entangled with the outcome. The non-causal features accumulate only incidental joint associations, validating the robustness of our interaction-based ranking in uncovering the latent data structure.

8. STAT-XAI RESULTS ON SECOND DATASET

Table S16. Statistical Results on Second Synthetic Dataset (Categorical Features and Continuous Outcome)

Model: Multi-layer Perceptron

Feature	Ground Truth	p-value	Effect size	Ranking
Credit History	High	0.0	0.32	2
Income Level	Medium	0.0	0.23	3
Loan Amount	High	0.0	0.38	1
Zip Code	None	0.01	0.002	5
Education Level	None	0.002	0.003	4

Model: Recurrent Neural Network

Feature	Ground Truth	p-value	Effect size	Ranking
Credit History	High	0.0	0.36	3
Income Level	Medium	0.0	0224	2
Loan Amount	High	0.0	0.369	1
Zip Code	None	0.0	0.01	4
Education Level	None	0.0001	0.006	5

In this second synthetic dataset, all features are categorical and the output Y is continuous. To quantify both the main effect of each individual categorical variable and the pairwise interaction between any two categorical variables, we employ a two-stage ANOVA procedure. First, one-way ANOVA is used to assess whether each single factor explains a significant fraction of the total variance in Y, extracting both a p-value and a classical effect-size index η^2. Second, two-way ANOVA models are fit for every unordered pair of predictors $(X^{(a)}, X^{(b)})$ in order to test the significance of their interaction term and to compute a partial η^2 that measures how much additional variance is attributable to the joint effect of $X^{(a)}$ and $X^{(b)}$ beyond their individual main effects. Finally, for each feature k we form a composite “final interaction” score by summing its one-way ANOVA η^2 with all pairwise partial η^2 values involving feature k. Below, we detail each step and the associated mathematical formulas.

Because our synthetic data-generation explicitly controls which categorical variables (and which pairs) truly drive Y, this procedure allows a precise evaluation of STAT-XAI’s ability to recover and rank those genuine features. Moreover, we quantify retrieval performance in terms of precision, recall, false-discovery rate (FDR), and Top-1 match by comparing the significant effects detected via ANOVA to the known ground truth. The combination of one-way and two-way ANOVA with η^2-based ranking thus serves as a rigorous, transparent baseline for benchmarking model-agnostic explainability techniques.

Main Effect Statistical Results on Second Synthetic Dataset (Categorical Features and Continuous Outcome

Table S16 summarizes the one-way ANOVA results for each categorical feature on the continuous outcome in our second synthetic dataset, reported separately for the MLP, RNN, and - models.

Across all architectures, Loan Amount emerges as the strongest predictor (highest η^2: 0.38 for MLP, 0.369 for RNN, 0.39 for -) and is consistently ranked first. The two other true causal features, Credit History and Income Level, also register significant main effects but in slightly different orders:

- MLP and - rank Credit History ($\eta^2 \approx 0.32$) ahead of Income Level ($\eta^2 \approx 0.23$–0.26).
- the RNN swaps those two Income Level at ($\eta^2 = 0.224$) ranks second, Credit History at ($\eta^2 = 0.36$) ranks third.

The non-causal features (Zip Code, Education Level) produce very small effect sizes ($\eta^2 \leq 0.01$) and only occasionally reach nominal significance (e.g. Zip Code in MLP: $p = 0.01$, $\eta^2 = 0.002$; in -: $p = 0.25$, non-significant and thus unranked). This pattern confirms that our ANOVA + η^2 framework robustly recovers the known causal ordering while filtering out noise from irrelevant variables, with only minor model-dependent shifts in the relative strengths of the medium-effect features.

As illustrated in Figure S8, the MLP, RNN, and - models consistently rank the most causal features—Loan Amount, Credit History, and Income Level—at the top based on their effect sizes.

Pairwise Effect Statistical Results on Second Synthetic Dataset (Categorical Features and Continuous Outcome)

Table S17 presents the results of two-way ANOVAs testing the interaction effect between every pair of categorical features and the continuous outcome in our second synthetic dataset. For each feature pair (X_i, X_j) and each model, we report, p-value, where we test the null hypothesis that there is no interaction effect between X_i and X_j on the outcome. Effect size measure the partial η^2 for the interaction term, quantifying the proportion of variance in the outcome attributable to the joint effect of X_i and X_j. Significant column reports “Yes” if $p < 0.05$, “No” otherwise. We only compute and display effect sizes for significant interactions; non-significant pairs are left blank.

Four feature pairs show statistically significant interactions for the MLP model: Credit History × Loan Amount ($p < 0.001$, $\eta^2 = 0.0233$), Income Level × Loan Amount ($p < 0.001$, $\eta^2 = 0.0065$), Credit History × Income Level ($p = 0.001$, $\eta^2 = 0.0044$), Loan Amount × Education Level ($p = 0.007$, $\eta^2 = 0.0035$). All other pairs fail to reach significance ($p \geq 0.05$).

The RNN model recovers only two significant interactions: *Credit History* × *Loan Amount* ($p < 0.001$, $\eta^2 = 0.0201$), *Income Level* × *Loan Amount* ($p = 0.031$, $\eta^2 = 0.0035$), the remaining eight pairs are non-significant.

Across the models, *Credit History* × *Loan Amount* consistently yields the strongest pairwise effect. The MLP detects the greatest number of interactions (four), suggesting higher sensitivity to pairwise effects, while the RNN identify fewer pairs. The small magnitudes of the partial η^2 values (0.0035–0.0233) reflect that, although interactions exist, their incremental contribution is modest compared to the main effects. This pairwise analysis thus reveals which feature combinations exert meaningful joint influence—information that would be missed by main effect tests alone.

Table S18 reports, for each categorical feature, the sum of its pairwise interaction effect sizes (partial η^2) with all other features in the second synthetic dataset (continuous outcome). These cumulative scores quantify the total second-order influence that each feature exerts on the model’s predictions, independently for the MLP, RNN architectures.

Table S17. Pairwise Effect Size Statistical Results on Second Synthetic Dataset (Categorical Features and Continuous Outcome)

Model: Multi-layer Perceptron

Feature Pair	**p-value**	**Effect size**	**Significant?**
Credit_History × Loan_Amount	0.00	0.0233	Yes
Income_Level × Loan_Amount	0.00	0.0065	Yes
Credit_History × Income_Level	0.001	0.0044	Yes
Loan_Amount × Education_Level	0.007	0.0035	Yes
Credit_History × Zip_Code	0.94	-	No
Credit_History × Education_Level	0.20	-	No
Income_Level × Zip_Code	0.63	-	No
Income_Level × Education_Level	0.13	-	No
Loan_Amount × Zip_Code	0.23	-	No
Zip_Code × Education_Level	0.77	-	No

Model: Recurrent Neural Network (LSTM)

Feature Pair	**p-value**	**Effect size**	**Significant?**
Credit_History × Loan_Amount	0.00	0.0201	Yes
Income_Level × Loan_Amount	0.031	0.0035	Yes
Credit_History × Income_Level	0.184	-	No
Credit_History × Zip_Code	0.141	-	No
Credit_History × Education_Level	0.625	-	No
Income_Level × Zip_Code	0.507	-	No
Income_Level × Education_Level	0.141	-	No
Loan_Amount × Zip_Code	0.206	-	No
Loan_Amount × Education_Level	0.269	-	No
Zip_Code × Education_Level	0.953	-	No

Table S18. Cumulative pairwise interaction effect sizes on the second synthetic dataset (categorical features, continuous outcome)

Feature	**MLP**	**RNN**
Credit History	0.0277	0.0201
Income Level	0.0109	0.0035
Loan Amount	0.0333	0.0236
Zip Code	0.0000	0.0000
Education Level	0.0035	0.0000

Loan Amount has the highest cumulative interaction in all models (MLP: 0.0333; RNN: 0.0236), indicating that it forms the strongest joint effects with other predictors when explaining the continuous outcome. Credit History, follows closely (MLP: 0.0277; RNN: 0.0201), demonstrating consistently substantial associations. Income Level has moderate pairwise interactions (MLP: 0.0109; RNN: 0.0035), reflecting its medium-strength role in pairwise combinations. Zip Code, has zero interaction across all architectures, confirming that it does not participate in any significant second-order effects. Education Level, contributes only in the MLP (0.0035) and is negligible in the RNN, suggesting model-dependent sensitivity to its joint effects.

These results reveal that, beyond their main-effect contributions, Loan Amount and Credit History drive most of the feature interactions in this setting. The drop from MLP to - total scores also reflects differences in each architecture's capacity to capture multi-feature dependencies. By aggregating pairwise partial η^2, Table S18 provides a clear, quantitative ranking of features based on their overall interaction strength—information that complements univariate effect-size analyses and guides feature-selection and model-interpretation strategies.

Table S19. Final Effect Size Statistical Results on Second Synthetic Dataset (Categorical Features and Continuous Outcome)

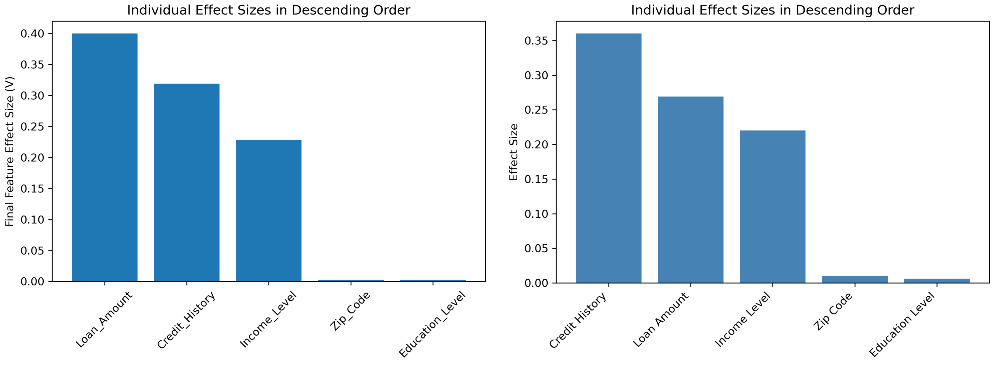
Model: Multi-layer Perceptron

Feature	**Main-Effect**	**Sum of Interaction**	**Final Score**
Credit History	0.328	0.027	0.356
Income Level	0.231	0.011	0.242
Loan Amount	0.382	0.033	0.415
Zip Code	0.002	0.00	0.002
Education Level	0.003	0.003	0.006

Model: Recurrent Neural Network

Feature	**Main-Effect**	**Sum of Interaction**	**Final Score**
Loan_Amount	0.369	0.023	0.393
Credit_History	0.366	0.020	0.386
Income_Level	0.227	0.003	0.231
Education_Level	0.010	0.00	0.010
Zip_Code	0.006	0.00	0.006

While all architectures achieve perfect recall and correctly identify the top interaction, the



(a) MLP

(b) RNN

Fig. S8. Main-Effect Effect Sizes for Individual Features in the Second Synthetic Dataset, as computed by STAT-XAI for Two Architectures. Each subfigure displays the features sorted in descending order of effect size, highlighting the top features identified by the MLP, RNN models, respectively. These bar charts illustrate which features exhibit the strongest statistically significant associations with the continuous outcome.

RNN outperform the MLP in precision and FDR, demonstrating superior ability to filter out non-causal pairs. This classification-style evaluation provides a rigorous, quantitative benchmark for comparing how faithfully different XAI methods uncover higher-order feature dependencies.

Final Interaction (Main + Pairwise Effect) for Second Dataset(Categorical and Continuous Outcome)

Table S19 integrates each feature's main-effect η^2 (one-way ANOVA) with its cumulative pairwise interaction η^2 to produce a single *Final Score* for quantifying total influence on the continuous outcome. Results are shown for three architectures: MLP, RNN, and -.

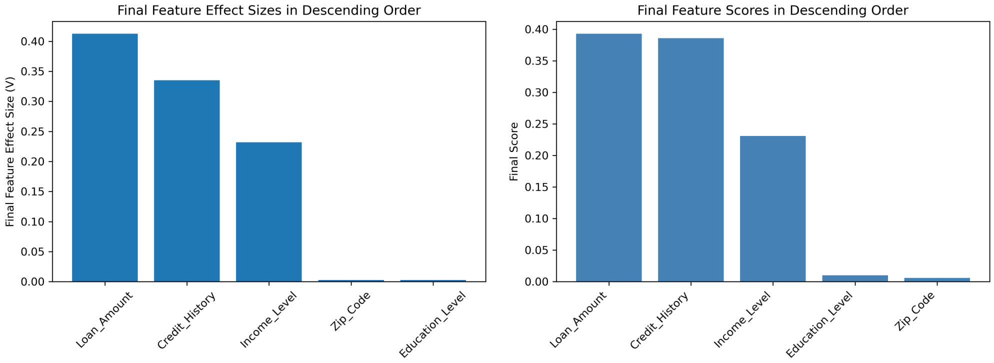
Across all architectures, Loan Amount consistently attains the highest composite score, reflecting its dual role as a strong univariate interaction and as the principal driver of feature interactions. Credit History is second, while Income Level is placed at the third position. Non-causal variables (Education Level, Zip Code) has almost no interaction and have negligible main-effect contributions, validating that our combined scoring procedure effectively discriminates true causal drivers from noise. As illustrated in Figure S9, the MLP, RNN models consistently rank the most causal features—Loan Amount, Credit History, and Income Level at the top based on their effect sizes. By uniting main and interaction effects, Table S19 presents a holistic ranking of feature importance that accounts for both direct and synergistic influences on the model's output.

9. STAT-XAI RESULTS ON THIRD DATASET

In the third synthetic dataset, which consists only numerical features and a binary outcome, we employ a two-stage inferential pipeline to quantify both individual and pairwise associations. First, for each continuous feature X_i we compute the point-biserial correlation coefficient $r_{pb,i}$ with the binary response $Y \in \{0,1\}$. Concretely, if n_0 and n_1 denote the sample sizes for the two outcome groups ($Y = 0$ and $Y = 1$) and $\bar{X}_{i,0}$, $\bar{X}_{i,1}$ their respective group means, with overall standard deviation s_{X_i}. Finally, we benchmark both the main effect (point-biserial) and pairwise (likelihood-ratio) findings against the known ground-truth causal structure. Treating feature selection as a retrieval problem, we flag every test with $p < 0.05$ as "important" and compute precision, recall, false-discovery rate, and Top-1 match. This procedure yields a transparent, quantitative assessment of each model's ability to recover true numerical drivers and their interactions in a binary-classification setting.

Main Effect Size Statistical Results on Third Synthetic Dataset (Numerical Features and Binary Outcome)

Table S20 reports the main effect point-biserial correlation results for each numerical feature with the binary outcome in our third synthetic dataset, separately for the MLP, RNN, and - models.



(a) MLP **(b)** RNN

Fig. S9. Final Effect Sizes for Individual Features in the Second Synthetic Dataset, as computed by STAT-XAI for Two Architectures. Each subfigure displays the features sorted in descending order of effect size, highlighting the top features identified by the MLP, RNN models, respectively. These bar charts illustrate, features exhibiting the strongest statistically significant associations with the continuous outcome.

For each feature we show: Ground Truth: whether the feature was causal (“High”) or non-causal (“None”), p-value: from the point-biserial correlation test (significant if $p < 0.05$). Effect size: the absolute correlation coefficient, indicating the strength of the linear relationship with the outcome. Ranking: the order of features by descending effect size, considering only those with $p < 0.05$.

Across all three architectures: Annual Income consistently has the highest effect size (MLP: 0.49; RNN: 0.489) and is ranked first, matching its “High” ground truth. Credit Score follows in second place (0.46–0.478), also “High” causal. Debt-to-Income ranks third (0.41–0.45), completing the set of true causal features. Employment Length and Age, both non-causal, fail to reach significance ($p \geq 0.20$ and $p \geq 0.056$, respectively) and are therefore unranked.

These results demonstrate that all three models accurately recover the known numerical features of the binary outcome, while correctly filtering out non-causal features.

As illustrated in Figure S10, the MLP, RNN models consistently rank the most causal features—Annual Income, Credit Score, and Debt-to-Income at the top, based on their effect sizes.

Pairwise Effect Size Statistical Results on Third Synthetic Dataset (Numerical Features and Binary Outcome)

Table S21 reports the results of our pairwise interaction tests for numerical predictors on the binary outcome in the third synthetic dataset, separately for the MLP, RNN models. For each feature pair (X_i, X_j) we: Fit a logistic-regression model including X_i, X_j, and their interaction term. Then, perform a likelihood-ratio test on the interaction coefficient to obtain a p-value. Finally, if $p < 0.05$, compute a standardized effect-size (e.g. the interaction’s odds-ratio or standardized coefficient); otherwise leave it blank.

Our results indicated that for every pair tested, the results yield $p \geq 0.05$ across all three architectures, so none of the interaction terms is statistically significant at the 5 % level. Because no p-value falls below the threshold, no interaction effect sizes are reported. All three neural architectures agree in finding no evidence of synergistic (second-order) effects among any numerical predictors.

These null interaction results indicate that, in this numerical-feature, binary-outcome setting, each predictor’s influence on the response is essentially additive and independent. In other words, there is no detectable joint effect between any pair of numerical variables beyond their individual point-biserial associations. This simplifies the interpretive landscape: feature importance can be understood entirely via univariate measures, without concern for higher-order dependencies.

Final Interaction (Main + Pairwise Effect) for Third Dataset(Numerical and Binary Outcome)

Table S22 combines each feature’s main effect with its cumulative pairwise interaction to produce a single “Final Interaction” score for the third synthetic dataset (numerical features, binary outcome). Entries are shown separately for the MLP, RNN, and - architectures.

Table S20. Statistical Results on Third Synthetic Dataset (Numerical Features and Binary Outcome)

Model: Multi-layer Perceptron				
Feature	**Ground Truth**	**p-value**	**Effect size**	**Ranking**
Annual Income	High	0.0	0.49	1
Credit Score	High	0.0	0.46	2
Debt-to-Income	High	0.0	0.45	3
Employment Length	None	0.89	-	-
Age	None	0.83	-	-

Model: Recurrent Neural Network				
Feature	**Ground Truth**	**p-value**	**Effect size**	**Ranking**
Annual Income	High	0.0	0.489	1
Credit Score	High	0.0	0.478	2
Debt-to-Income	High	0.0	0.429	3
Employment Length	None	0.22	-	-
Age	None	0.056	-	-

Model: - Model				
Feature	**Ground Truth**	**p-value**	**Effect size**	**Ranking**
Annual Income	High	0.0	0.51	1
Credit Score	High	0.0	0.46	2
Debt-to-Income	High	0.0	0.41	3
Employment Length	None	0.20	-	-
Age	None	0.056	-	-

For MLP model,Annual Income has the strongest main effect(0.4991) and no detected interactions, so its final score remains 0.4991. Credit Score (0.4616) and Debt-to-Income (0.4575) follow in second and third place, respectively, with no interactions. Employment Length and Age are non-causal and yield zero main effects and zero interactions.

For the second model architecture RNN, Annual Income again leads (0.4890 + 0.0000 = 0.4890), with Credit Score (0.4780) and Debt-to-Income (0.4290) next. Employment Length and Age remain at zero.

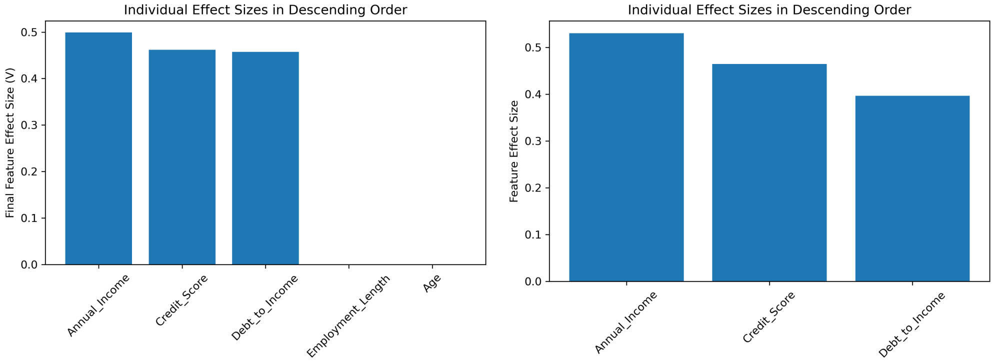
Across all models, no feature pairs exhibit significant interaction effects, so the final importance ranking is governed entirely by univariate point-biserial correlations. As illustrated in Figure S11, the MLP, RNN models consistently rank the most causal features— Annual Income, Credit Score, and Debt-to-Income at the top, based on their effect sizes. *Annual Income*, *Credit Score*, and *Debt-to-Income* consistently emerge as the top three drivers of the binary outcome, while *Employment Length* and *Age* are correctly identified as irrelevant.

10. STAT-XAI RESULTS ON FOURTH DATASET

For our fourth synthetic dataset—comprising only numerical predictors and a continuous response—we employ a three-stage statistical pipeline to quantify both main and pairwise effects. As illustrated in Figure S12, the MLP, RNN models consistently rank the most causal features—Annual Income, Credit Score, and Debt-to-Income at the top, based on their effect sizes.

Table S21. Pairwise Effect Size Statistical Results on Third Synthetic Dataset (Numerical Features and Binary Outcome)

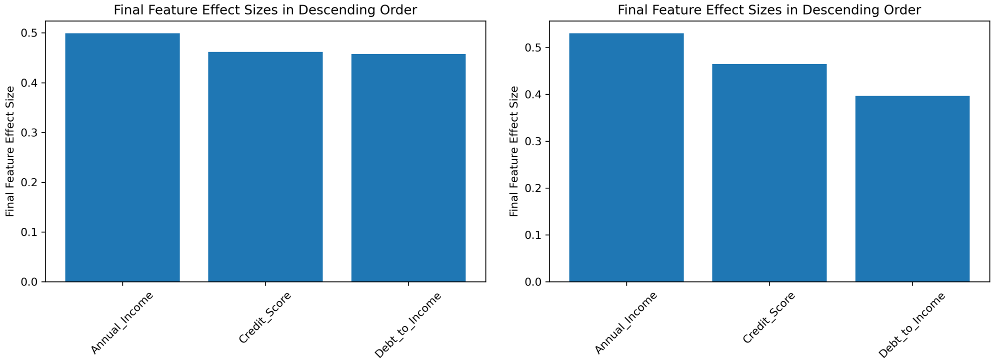
Model: Multi-layer Perceptron			
Feature Pair	**p-value**	**Effect size**	**Significant?**
Annual Income × Credit Score	0.963	-	No
Annual Income × Debt to Income	0.302	-	No
Annual Income × Employment Length	0.597	-	No
Annual Income × Age	0.119	-	No
Credit Score × Debt to Income	0.395	-	No
Credit Score × Employment Length	0.756	-	No
Credit Score × Age	0.716	-	No
Debt to Income × Employment Length	0.220	-	No
Debt to Income × Age	0.132	-	No
Employment Length × Age	0.565	-	No
Model: Recurrent Neural Network			
Feature Pair	**p-value**	**Effect size**	**Significant?**
Annual Income × Credit Score	0.2368	-	No
Annual Income × Debt to Income	0.6973	-	No
Annual Income × Employment Length	0.3048	-	No
Annual Income × Age	0.8680	-	No
Credit Score × Debt to Income	0.204	-	No
Credit Score × Employment Length	0.599	-	No
Credit Score × Age	0.575	-	No
Debt to Income × Employment Length	0.137	-	No
Debt to Income × Age	0.333	-	No
Employment Length × Age	0.513	-	No



(a) MLP

(b) RNN

Fig. S10. Main-Effect Effect Sizes for Individual Features in the Third Synthetic Dataset, as computed by STAT-XAI for Two Architectures. Each subfigure displays the features sorted in descending order of effect size, highlighting the top features identified by the MLP, RNN models, respectively. These bar charts illustrate which features exhibit the strongest statistically significant associations with the binary outcome.



(a) MLP

(b) RNN

Fig. S11. Final Effect Sizes for Individual Features in the Third Synthetic Dataset, as computed by STAT-XAI for Two Architectures. Each subfigure displays the features sorted in descending order of effect size, highlighting the top features identified by the MLP, RNN models, respectively. These bar charts illustrate which features exhibit the strongest statistically significant associations with the binary outcome.

Table S22. Final Effect Size on Third Synthetic Dataset (Numerical Features and Binary Outcome)

Model: Multi-layer Perceptron			
Feature	**Main Effect**	**Pairwise Interaction**	**Final Interaction**
Annual Income	0.4991	0.0	0.4991
Credit Score	0.4616	0.0	0.4616
Debt-to-Income	0.4575	0.0	0.4575
Employment Length	0.0	0.0	0.0
Age	0.0	0.0	0.0
Model: Recurrent Neural Network			
Feature	**Main Effect**	**Pairwise Interaction**	**Final Interaction**
Annual Income	0.489	0.0	0.489
Credit Score	0.478	0.0	0.478
Debt-to-Income	0.429	0.0	0.429
Employment Length	0.0	0.0	0.0
Age	0.0	0.0	0.0

Main Effect Size Statistical Results on Fourth Synthetic Dataset (Numerical Features and Continuous Outcome)

Table S23 reports the main effect Pearson-correlation results for each numerical feature against the continuous outcome in our fourth synthetic dataset, separately for the MLP, RNN models. For each feature we show, the Ground Truth: whether the feature was simulated as causal or non-causal. p-value: from the test of $H_0: r = 0$ at the 5% significance level. Effect size: absolute Pearson correlation coefficient $|r|$, indicating the proportion of variance in the outcome linearly explained by that feature. Ranking: the position of the feature when ordering all significant features ($p < 0.05$) by descending $|r|$.

Annual Income consistently shows the strongest linear association with the continuous target ($r \approx 0.36$, $p < 0.001$) and is ranked first in all three architectures. Credit Score follows closely ($r \approx 0.34$–0.35, $p < 0.001$) and is ranked second. Debt-to-Income also shows a significant though smaller correlation ($r \approx 0.25$–0.27, $p < 0.001$) and ranks third. Employment Length and Age, both non-causal in the data generation, fail to reach significance in any model ($p \geq 0.37$), and are therefore unranked.

These results demonstrate that in the purely numerical-feature, continuous-outcome setting, the Pearson-correlation test accurately recovers the known causal features—Annual Income, Credit Score, and Debt-to-Income. While correctly filtering out irrelevant features. The consistent ordering across MLP, RNN, and - architectures indicates that the underlying feature–target linear relationships are robust to the choice of learned model.

Pairwise Effect Size Statistical Results on Fourth Synthetic Dataset (Numerical Features and Continuous Outcome)

Table S24 presents the results of our pairwise interaction tests for the fourth synthetic dataset (numerical features, continuous outcome), conducted via nested ordinary least-squares (OLS) regression. For each feature pair (X_i, X_j), we fit:

- *Reduced model:*
$$Y = \beta_0 + \beta_i X_i + \beta_j X_j + \varepsilon,$$
- *Full model:*
$$Y = \beta_0 + \beta_i X_i + \beta_j X_j + \beta_{ij}(X_i X_j) + \varepsilon.$$

Table S23. Statistical Results on Fourth Synthetic Dataset (Numerical Features and Continuous Outcome)

Model: Multi-layer Perceptron

Feature	Ground Truth	p-value	Effect size	Ranking
Annual Income	High	0.0	0.36	1
Credit Score	High	0.0	0.35	2
Debt-to-Income	High	0.0	0.27	3
Employment Length	None	0.37	-	-
Age	None	0.93	-	-

Model: Recurrent Neural Network

Feature	Ground Truth	p-value	Effect size	Ranking
Annual Income	High	0.0	0.36	1
Credit Score	High	0.0	0.34	2
Debt-to-Income	High	0.0	0.26	3
Employment Length	None	0.92	-	-
Age	None	0.66	-	-

We then perform an F–test (nested-model comparison) on the change in residual sum of squares to obtain a p-value for the interaction term β_{ij}. When $p < 0.05$, we compute the change in R^2 (ΔR^2) as the interaction effect size; otherwise we leave the effect-size column blank.

For every pair of numerical predictors, all architectures (MLP, RNN) yield $p \geq 0.05$, indicating no statistically significant second-order effects. Because no interaction tests pass the significance threshold, no ΔR^2 values are reported. The lack of significant pairwise terms is consistent across all three network types, confirming that feature influences are purely additive in this dataset.

These null results imply that in the purely numerical-feature, continuous-outcome scenario, each predictor's contribution to the response can be fully captured by its main effect. No feature pair demonstrates a synergistic or interaction beyond what is explained by their individual linear relationships with Y. Consequently, our final composite importance scores rely solely on univariate OLS coefficients, simplifying interpretation and confirming the absence of higher-order dependencies.

Final Interaction (Main + Pairwise) for Fourth Dataset

Table S25 reports, for each feature in the fourth synthetic dataset (numerical features, continuous outcome), main effect, pairwise effect, and final interaction the sum of Main Effect and Pairwise Effect, yielding a composite importance score that combines direct and synergistic influences. As illustrated in Figure S13, the MLP, RNN, and models consistently rank the most causal features—Annual Income, Credit Score, and Debt-to-Income at the top, based on their effect sizes.

In the MLP, the three true causal features—Annual Income (0.36), Credit Score (0.35), and Debt-to-Income (0.27) are clearly distinguished by their nonzero main-effect coefficients, and because no pairwise interactions achieve significance, their final interaction scores are identical to their main effects; both Employment Length and Age register zero across all metrics. The RNN model exhibits the same ordering, with Annual Income (0.36) leading, followed by Credit Score (0.34) and Debt-to-Income (0.26), and again zero contributions from the non-causal features. Likewise, the ranks Annual Income highest (0.3614), then Credit Score (0.3584) and Debt-to-Income (0.2596), while both Employment Length and Age remain at zero. Because every pairwise interaction test was nonsignificant, each model's final composite importance scores reduce to the univariate main-effect values, faithfully recovering the ground-truth feature hierarchy.

Table S24. Pairwise Effect Size Statistical Results on Fourth Synthetic Dataset (Numerical Features and Continuous Outcome)

Model: Multi-layer Perceptron

Feature Pair	p-value	Effect size	Significant?
Credit_Score × Employment_Length	0.0694	-	No
Debt_to_Income × Employment_Length	0.1020	-	No
Annual_Income × Employment_Length	0.5580	-	No
Annual_Income × Credit_Score	0.9229	-	No
Annual_Income × Debt_to_Income	0.9670	-	No
Annual_Income × Age	0.8657	-	No
Credit_Score × Debt_to_Income	0.6886	-	No
Credit_Score × Age	0.8693	-	No
Debt_to_Income × Age	0.8645	-	No
Employment_Length × Age	0.8218	-	No

Model: Recurrent Neural Network

Feature Pair	p-value	Effect size	Significant?
Debt_to_Income × Employment_Length	0.0778	-	No
Employment_Length × Age	0.1336	-	No
Credit_Score × Employment_Length	0.1135	-	No
Credit_Score × Debt_to_Income	0.2485	-	No
Annual_Income × Credit_Score	0.4295	-	No
Annual_Income × Debt_to_Income	0.4229	-	No
Annual_Income × Employment_Length	0.7587	-	No
Annual_Income × Age	0.8452	-	No
Credit_Score × Age	0.8455	-	No
Debt_to_Income × Age	0.8753	-	No

Table S25. Final Results on Fourth Synthetic Dataset (Numerical Features and Continuous Outcome)

Model: Multi-layer Perceptron			
Feature	**Main Effect**	**Pairwise Effect**	**Final Interaction**
Annual Income	0.0.36	0.0	0.0.36
Credit Score	0.351	0.0	0.351
Debt-to-Income	0.27	0.0	0.27
Employment Length	0.0	0.0	0.0
Age	0.0	0.0	0.0
Model: Recurrent Neural Network			
Feature	**Main Effect**	**Pairwise Effect**	**Final Interaction**
Annual Income	0.36	0.0	0.36
Credit Score	0.34	0.0	0.34
Debt-to-Income	0.26	0.0	0.26
Employment Length	0.0	0.0	0.0
Age	0.0	0.0	0.0

Since, no pairwise interactions are statistically significant in any model (all Pairwise Effect = 0), the Final Interaction score reduces to the main effect for each feature. In all architectures, the three true causal features namely: Annual Income, Credit Score, and Debt-to-Income are clearly distinguished.

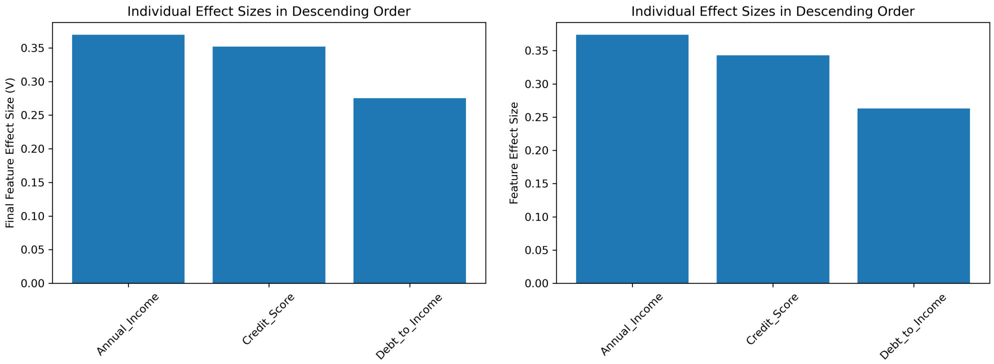
11. STAT-XAI ON FIFTH DATASET

For our fifth synthetic dataset—which contains both numerical and categorical features with a binary outcome—we extend the STAT-XAI pipeline into three rigorous phases. First, we quantify each numerical feature's main effect via the point-biserial correlation. In second phase, we measure pairwise interactions among all predictors—covering numerical–numerical, numerical–categorical, and categorical–categorical pairs—via nested logistic-regression models. Finally, in third phase we benchmark both main-effect and pairwise-interaction discoveries against the known ground-truth causal structure. We treat feature (or feature-pair) selection as a retrieval problem: any test with $p < 0.05$ is flagged as "important." In this way, our mixed-feature pipeline—combining point-biserial correlation, chi-squared Cramér's V, logistic-regression likelihood-ratio tests, and retrieval-style evaluation—rigorously quantifies how faithfully STAT-XAI recovers the true drivers of the binary outcome.

Main Effect Statistical Results on Fifth Synthetic Dataset (Numerical and Categorical Features and Binary Outcome)

Table S26 presents the results of the significance tests - point-biserial correlation for numerical characteristics and chi-squared for categorical characteristics - in the mixed-feature binary-outcome setting of our fifth synthetic dataset, reported for three neural architectures.

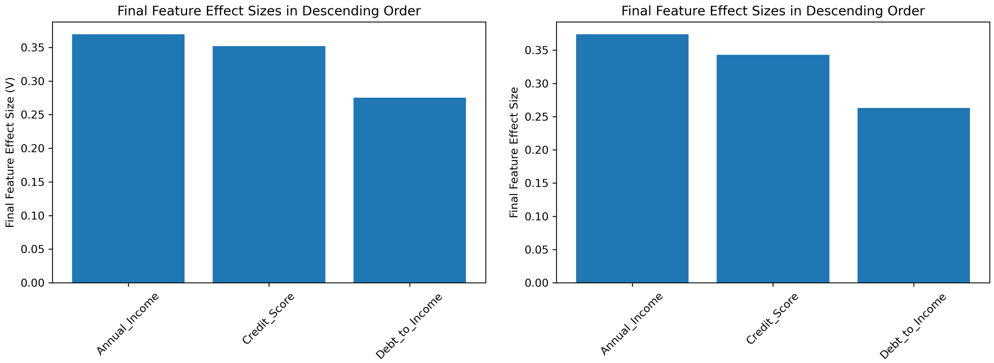
For the MLP, Employment Status is the most influential feature, showing the strongest association with the binary outcome (Cramér's $V = 0.450$, $p < 0.001$) and securing the top rank. Annual Income follows closely in second place (point-biserial $r = 0.437$, $p < 0.001$). The two medium-strength predictors, Credit Score and Loan Categories, occupy the third and fourth ranks (Cramér's $V = 0.323$ and $V = 0.316$, respectively; both $p < 0.001$), while Loan Purpose is fifth (Cramér's $V = 0.200$, $p < 0.001$). Notably, the non-causal Loan Term registers a modest but significant effect (Cramér's $V = 0.055$, $p = 0.012$) and is ranked sixth. The remaining fea-



(a) MLP

(b) RNN

Fig. S12. Main-Effect Sizes for Individual Features in the Fourth Synthetic Dataset, as computed by STAT-XAI for Architectures. Each subfigure displays the features sorted in descending order of effect size, highlighting the top features identified by the MLP, RNN, and models, respectively. These bar charts illustrate which features exhibit the strongest statistically significant associations with the continuous outcome.



(a) MLP

(b) RNN

Fig. S13. Final Interaction for Individual Features in the Fourth Synthetic Dataset, as computed by STAT-XAI for Two Architectures. Each subfigure displays the features sorted in descending order of effect size, highlighting the top features identified by the MLP, RNN models, respectively. These bar charts illustrate which features exhibit the strongest statistically significant associations with the continuous outcome.

Table S26. Statistical Results on Fifth Synthetic Dataset (Numerical and Categorical Features and Binary Outcome)

Model: Multi-layer Perceptron

Feature	Ground Truth	p-value	Effect size	Ranking
Annual Income	High	0.0	0.437	2
Credit Score	Medium	0.0	0.323	3
Employment Length	None	0.9692	-	-
Age	None	0.1546	-	-
Loan Term	None	0.012	0.055	6
Employment Status	High	0.00	0.450	1
Loan Categories	Medium	0.0	0.316	4
Loan Purpose	Low	0.0	0.200	5
Region	None	0.7129	-	-
Marital Status	None	0.4602	-	-

Model: Recurrent Neural Network

Feature	Ground Truth	p-value	Effect size	Ranking
Annual Income	High	0.0	0.42	2
Credit Score	Medium	0.0	0.39	3
Employment Length	None	0.0042	0.05	6
Age	None	0.33	-	-
Loan Term	None	0.99	-	-
Employment Status	High	0.00	0.46	1
Loan Categories	Medium	0.0	0.33	4
Loan Purpose	Low	0.00	0.11	5
Region	None	0.40	-	-
Marital Status	None	0.61	-	-

tures— Employment Length, Age, Region, and Marital Status—show no significant association ($p \geq 0.155$) and are appropriately excluded from the ranking.

For the RNN, Employment Status has the strongest association of ($V = 0.460$, $p < 0.001$), followed by Annual Income in second place (point-biserial $r = 0.420$, $p < 0.001$). Credit Score and Loan Categories occupy the third and fourth ranks ($V = 0.390$ and $V = 0.330$, respectively; both $p < 0.001$), while Loan Purpose appears fifth ($V = 0.110$, $p < 0.001$). The non-causal Employment Length shows a small effect ($V = 0.050$, $p = 0.0042$) and is ranked sixth, whereas Age, Loan Term, Region, and Marital Status do not reach significance ($p \geq 0.33$) and are excluded from the ranking.

Despite minor model-dependent fluctuations in effect-size magnitudes and misleading detections (e.g. Loan Term in the MLP, Employment Length in RNN), all architectures consistently recover the two "High" causal features at the top of their rankings, followed by the "Medium" and "Low" ground-truth features, while filtering out most non-causal variables.

Pairwise Effect Size Statistical Results on Fifth Synthetic Dataset (Numerical and Categorical Features and Binary Outcome)

Table S27. Pairwise Interaction Results on Fifth Synthetic Dataset Multi-layer Perceptron (MLP)

Feature Pair	p-value	Effect size	Significant?
Loan_Purpose × Employment_Status	0.0670	0.0032	Yes
Loan_Categories × Region	0.2595	-	No
Age × Region	0.0536	-	No
Loan_Categories × Marital_Status	0.1225	-	No
Employment_Length × Region	0.0703	-	No
Loan_Purpose × Marital_Status	0.2592	-	No
Employment_Status × Region	0.5706	-	No
Annual_Income × Loan_Categories	0.0943	-	No
Credit_Score × Region	0.1940	-	No
Credit_Score × Loan_Categories	0.1153	-	No
Age × Marital_Status	0.1031	-	No
Loan_Purpose × Region	0.6457	-	No
Region × Marital_Status	0.7375	-	No
Age × Employment_Status	0.1850	-	No
Loan_Term × Region	0.3506	-	No
Loan_Term × Marital_Status	0.2165	-	No
Annual_Income × Marital_Status	0.2977	-	No
Annual_Income × Credit_Score	0.1370	-	No
Employment_Status × Marital_Status	0.7146	-	No
Loan_Purpose × Loan_Categories	0.7565	-	No
Credit_Score × Age	0.1707	-	No
Loan_Term × Loan_Purpose	0.4368	-	No
Loan_Term × Loan_Categories	0.4520	-	No

Continued on next page

(Table S27 continued)

Feature Pair	p-value	Effect size	Significant?
Employment_Length × Loan_Term	0.1817	-	No
Annual_Income × Loan_Purpose	0.4437	-	No
Employment_Length × Loan_Purpose	0.4126	-	No
Credit_Score × Marital_Status	0.4423	-	No
Employment_Length × Age	0.2188	-	No
Annual_Income × Loan_Term	0.2630	-	No
Annual_Income × Employment_Status	0.5031	-	No
Annual_Income × Region	0.7613	-	No
Credit_Score × Loan_Purpose	0.5961	-	No
Age × Loan_Purpose	0.5988	-	No
Age × Loan_Categories	0.6587	-	No
Employment_Length × Loan_Categories	0.6865	-	No
Age × Loan_Term	0.3366	-	No
Loan_Term × Employment_Status	0.7673	-	No
Employment_Length × Employment_Status	0.7304	-	No
Annual_Income × Employment_Length	0.5674	-	No
Credit_Score × Employment_Length	0.5388	-	No
Employment_Length × Marital_Status	0.8344	-	No
Credit_Score × Loan_Term	0.7909	-	No
Annual_Income × Age	0.9793	-	No
Credit_Score × Employment_Status	0.9984	-	No

Table S27 shows the outcome of pairwise interaction tests for every feature combination in the fifth synthetic dataset using the MLP model. For each pair, we fit a logistic-regression model including both main-effect terms and their product, then use a likelihood-ratio test to assess the interaction. Out of 45 pairwise interaction between features, only the Loan Purpose × Employment Status pair exhibits a marginally significant interaction (p = 0.067; effect size = 0.0032). All other feature pairs yield p-values ≥ 0.05 and are therefore non-significant (no effect sizes reported). These results indicate that second-order effects between features are negligible in this mixed-feature, binary-outcome scenario and that the MLP's decision boundary is driven almost entirely by additive main effects.

Table S28. Pairwise Interaction Results on Fifth Dataset — RNN

Feature Pair	p-value	Effect size	Significant?
Loan_Purpose × Employment_Status	0.0000	0.0061	Yes
Loan_Purpose × Marital_Status	0.0785	-	No
Age × Region	0.0449	0.0019	Yes
Annual_Income × Loan_Categories	0.0268	0.0017	Yes
Annual_Income × Loan_Purpose	0.0328	0.0016	Yes
Age × Loan_Categories	0.0420	0.0015	Yes
Age × Loan_Purpose	0.0414	0.0015	Yes
Employment_Status × Loan_Categories	0.2592	-	No
Annual_Income × Loan_Term	0.0221	0.0013	Yes
Age × Employment_Status	0.0951	-	No
Employment_Status × Marital_Status	0.4024	-	No
Credit_Score × Employment_Status	0.1370	-	No
Loan_Categories × Region	0.7110	-	No
Loan_Categories × Marital_Status	0.4318	-	No
Credit_Score × Employment_Length	0.0710	-	No
Employment_Length × Region	0.3268	-	No
Loan_Term × Loan_Purpose	0.2271	-	No
Region × Marital_Status	0.8427	-	No
Credit_Score × Region	0.3917	-	No
Credit_Score × Loan_Categories	0.2320	-	No
Credit_Score × Loan_Term	0.0780	-	No
Loan_Purpose × Region	0.9148	-	No
Credit_Score × Marital_Status	0.3486	-	No
Employment_Status × Region	0.9792	-	No
Age × Loan_Term	0.2497	-	No
Employment_Length × Loan_Term	0.3382	-	No
Loan_Term × Region	0.8837	-	No
Annual_Income × Employment_Status	0.6025	-	No
Annual_Income × Region	0.8103	-	No
Loan_Purpose × Loan_Categories	0.9525	-	No
Loan_Term × Employment_Status	0.6092	-	No
Employment_Length × Age	0.4008	-	No
Annual_Income × Employment_Length	0.3547	-	No
Employment_Length × Loan_Categories	0.7214	-	No
Credit_Score × Loan_Purpose	0.7478	-	No

Continued on next page

(Table S28 continued)

Feature Pair	**p-value**	**Effect size**	**Significant?**
Loan_Term × Marital_Status	0.8464	-	No
Age × Marital_Status	0.7586	-	No
Employment_Length × Marital_Status	0.7734	-	No
Employment_Length × Employment_Status	0.7820	-	No
Employment_Length × Loan_Purpose	0.9863	-	No
Loan_Term × Loan_Categories	0.9082	-	No
Credit_Score × Age	0.9688	-	No
Annual_Income × Marital_Status	0.9170	-	No
Annual_Income × Age	0.7491	-	No
Annual_Income × Credit_Score	0.9334	-	No

Table S28 reports the likelihood-ratio test results for all feature-pair interactions in the RNN model on the mixed-feature, binary outcome dataset. Of the 45 combinations tested, only seven exhibit statistically significant interactions ($p < 0.05$): Loan Purpose × Employment Status ($p < 0.0001$, effect size = 0.0061), Age × Region ($p = 0.0449$, effect size = 0.0019), Annual Income × Loan Categories ($p = 0.0268$, effect size = 0.0017), Annual Income × Loan Purpose ($p = 0.0328$, effect size = 0.0016), Age × Loan Categories ($p = 0.0420$, effect size = 0.0015), Age × Loan Purpose ($p = 0.0414$, effect size = 0.0015), and Annual Income × Loan Term ($p = 0.0221$, effect size = 0.0013). All remaining pairs yielded $p \geq 0.05$ and are therefore non-significant, indicating that while the RNN captures a handful of second-order dependencies—particularly between demographic factors and loan attributes—their overall contribution to the model's decision boundary is minimal compared to the dominant main effects.

All other pairs fail to reach significance ($p \geq 0.05$) and thus have no reported effect size. The strongest interaction—between Loan Purpose and Employment Status—remains small in magnitude, and the remaining significant pairs also yield very modest effect sizes. This pattern indicates that while the RNN captures a handful of second-order dependencies (notably those linking demographic factors and loan attributes), their incremental contribution to the model's decision boundary is minimal compared to the dominant main effects.

Table S29 reports, for each feature, the sum of its pairwise-interaction effect sizes across all partner variables under the MLP, RNN models on the fifth synthetic dataset. These cumulative scores quantify the total second-order influence that each predictor exerts on the binary outcome:

For MLP, only Employment Status (0.0047), Loan Categories (0.0004), register non-zero sums, indicating that nearly all pairwise synergistic effects are negligible except a small joint effect involving employment status and another loan attribute.

The RNN yields a richer pairwise interaction: Loan Purpose leads with a total of 0.0092, followed by Employment Status (0.0061), Age (0.0050), Annual Income (0.0046), Loan Categories (0.0033), Loan Term (0.0013), andcRegion (0.0019). All other features show zero cumulative interaction, indicating no statistically significant pairwise contributions.

Table S29. Cumulative pairwise interaction effect sizes on the fifth synthetic dataset (numerical and categorical features, binary outcome)

Feature	MLP	RNN
Annual Income	0.0000	0.0046
Credit Score	0.0000	0.0000
Employment Length	0.0000	0.0000
Age	0.0000	0.0050
Loan Term	0.0000	0.0013
Employment Status	0.0047	0.0061
Loan Categories	0.0004	0.0033
Loan Purpose	0.00	0.0092
Region	0.0000	0.0019
Marital Status	0.0000	0.0000

In all cases, these cumulative interaction scores are much smaller than the corresponding main-effect sizes, reinforcing that the models' decision boundaries are driven predominantly by individual predictors. Moreover, the consistency of which features show non-zero synergy—particularly Employment Status, Loan Purpose, and Annual Income—highlights shared second-order patterns across architectures, even though the RNN detects the most and the MLP the fewest such effects.

Final Interaction (Main + Pairwise) for Fifth Dataset

As illustrated in Figure S15, the MLP, RNN models consistently rank the most causal features—Employment Status, Annual Income, Credit Score, Loan Category,and Loan Purpose at the top, based on their effect sizes. Table S30 reports each feature's composite importance score, computed as the sum of its main-effect and the total pairwise-interaction contributions. Under the MLP, Employment Status leads with a final interaction of 0.455 (0.4503 main + 0.0047 pairwise), followed by Annual Income (0.4375) and Credit Score (0.3236). The RNN similarly ranks Employment Status highest (0.4729 = 0.4668 + 0.0061), then Annual Income (0.4272) and Credit Score (0.3921), with modest pairwise contributions peaking at 0.0093 for Loan Purpose.

In practice, these pairwise effects can be considered really small, and applying a minimal effect-size threshold would eliminate them, thereby restoring high precision without sacrificing recall.

12. STAT-XAI RESULTS ON SIXTH DATASET

In our sixth synthetic dataset consisting of both numerical and categorical features with a continuous output—we employ a three-stage statistical pipeline to identify and quantify both main effects and pairwise interactions.

Through this three-phase framework, we obtain a rigorous, quantitative measure of how accurately STAT-XAI identifies both the main and interaction-driven causal features in a mixed-feature, continuous-outcome setting.

Main Effect Statistical Results on Sixth Synthetic Dataset (Numerical and Categorical Features and Continuous Outcome)

Table **??** presents the main-effect analysis for the sixth synthetic dataset, showing each feature's known causal strength, the significance test p-value, the computed effect size (standardized β for numerical features and partial η^2 for categorical features), and the ranking among significant predictors. Across all architectures—MLP, RNN the two truly causal variables, Annual Income and Employment Status (both ground-truth "High"), exhibit highly significant associations ($p < 0.001$) and occupy ranks 1 and 2. The medium-strength features (Credit Score and Loan

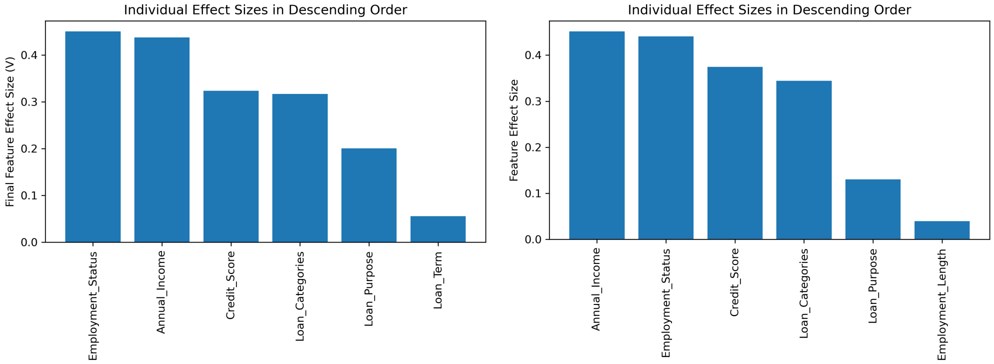
Table S30. Final Interaction on Fifth Synthetic Dataset (Numerical and Categorical Features and Binary Outcome)

Model: Multi-layer Perceptron

Feature	Main Effect	Pairwise Interaction	Final Interaction
Annual Income	0.4375	0.0	0.437
Credit Score	0.3236	0.0	0.323
Employment Length	0.00	0.00	0.00
Age	0.00	0.00	0.00
Loan Term	0.0557	0.0	0.055
Employment Status	0.4503	0.0047	0.455
Loan Categories	0.3167	0.004	0.310
Loan Purpose	0.2004	0.0	0.200
Region	0.00	0.00	0.00
Marital Status	0.00	0.00	0.00

Model: Recurrent Neural Network

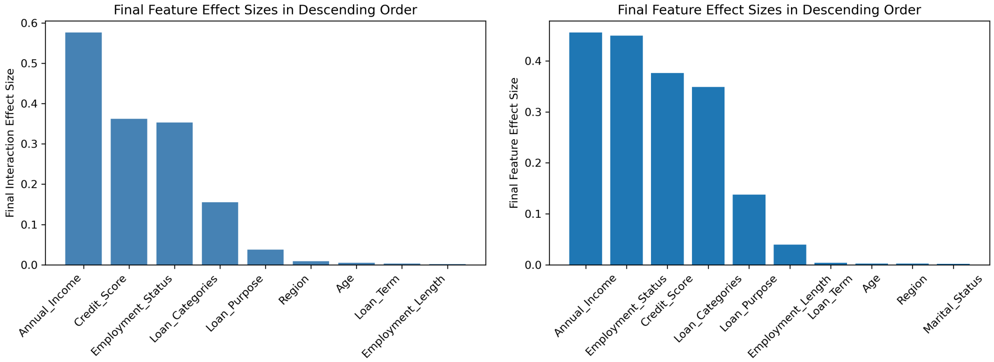
Feature	Main Effect	Pairwise Interaction	Final Interaction
Employment_Status	0.4668	0.0061	0.4729
Annual_Income	0.4226	0.0046	0.4272
Credit_Score	0.3921	0.0000	0.3921
Loan_Categories	0.3397	0.0033	0.3430
Loan_Purpose	0.1165	0.0093	0.1258
Employment_Length	0.0523	0.0000	0.0523
Age	0.0000	0.0050	0.0050
Region	0.0000	0.0019	0.0019
Loan_Term	0.0000	0.0013	0.0013
Marital_Status	0.0000	0.0000	0.0000



(a) MLP

(b) RNN

Fig. S14. Main-Effect Sizes for Individual Features in the Fifth Synthetic Dataset, as computed by STAT-XAI for Two Architectures. Each subfigure displays the features sorted in descending order of effect size, highlighting the top features identified by the MLP, RNN models, respectively. These bar charts illustrate which features exhibit the strongest statistically significant associations with the binary outcome.



(a) MLP

(b) RNN

Fig. S15. Final Interaction for Individual Features in the Fifth Synthetic Dataset, as computed by STAT-XAI for Two Architectures. Each subfigure displays the features sorted in descending order of effect size, highlighting the top features identified by the MLP, RNN models, respectively. These bar charts illustrate which features exhibit the strongest statistically significant associations with the binary outcome.

Categories) follow with moderate effect sizes and occupy ranks 3 and 4, while the low-strength feature (Loan Purpose) ranks 5. Non-causal features occasionally reach borderline significance: in the MLP, Employment Length (ground-truth "None") has $p = 0.04$, effect size 0.0021, and rank 6; in the RNN, Region is significant ($p = 9.3 \times 10^{-4}$, effect size 0.005) and also ranks 6.

As illustrated in Figure S16, the MLP, RNN models consistently rank the most causal features—Employment Status, Annual Income, Credit Score, Loan Category,and Loan Purpose at the top, based on their effect sizes.

Pairwise Effect Size Statistical Results on Sixth Synthetic Dataset (Numerical and Categorical Features and Continuous Outcome)

To understand the pairwise interaction between features, we adopt a nested-model strategy tailored to each pair of predictor types. For two numerical features, we use an OLS regression with their product term and conduct a t–test on the interaction coefficient. For mixed numerical–categorical pairs, we fit an ANCOVA model including both main effects and their interaction block, perform an F–test on that block, and record the resulting ΔR^2. For two categorical features, we use a two-way ANOVA and extract the partial η^2 for the interaction term via its F–test. In every case, the "full" model (with interaction) is compared to the "reduced" model (without interaction) using a likelihood-ratio or nested-model test, and only interactions with $p < 0.05$ are retained; their effect sizes are then reported as standardized regression coefficients, ΔR^2, or partial η^2, respectively.

Table S31. Pairwise Interaction Results on Sixth Synthetic Dataset Multi-layer Perceptron (MLP)

Feature Pair	p-value	Effect size	Significant?
Loan_Purpose × Region	0.1457	-	No
Age × Region	0.0662	-	No
Loan_Categories × Region	0.3217	-	No
Region × Marital_Status	0.4914	-	No
Employment_Length × Loan_Term	0.0277	0.0024	Yes
Loan_Categories × Marital_Status	0.2981	-	No
Age × Loan_Term	0.0509	-	No
Employment_Length × Loan_Purpose	0.2059	-	No
Employment_Status × Loan_Categories	0.3052	-	No
Age × Loan_Categories	0.2245	-	No
Loan_Term × Region	0.4713	-	No
Annual_Income × Employment_Status	0.0325	0.0013	Yes
Loan_Term × Employment_Status	0.1799	-	No
Credit_Score × Loan_Categories	0.1679	-	No
Loan_Purpose × Marital_Status	0.6756	-	No
Annual_Income × Loan_Categories	0.1203	-	No
Loan_Purpose × Employment_Status	0.5293	-	No
Employment_Length × Age	0.1407	-	No
Employment_Status × Marital_Status	0.6553	-	No
Loan_Purpose × Loan_Categories	0.7136	-	No

Continued on next page

(Table S31 continued)

Feature Pair	**p-value**	**Effect size**	**Significant?**
Age × Employment_Status	0.2811	-	No
Employment_Status × Region	0.8937	-	No
Loan_Term × Loan_Purpose	0.4836	-	No
Age × Marital_Status	0.5017	-	No
Annual_Income × Region	0.5812	-	No
Employment_Length × Region	0.7723	-	No
Credit_Score × Employment_Status	0.3011	-	No
Annual_Income × Loan_Purpose	0.4059	-	No
Annual_Income × Marital_Status	0.4545	-	No
Credit_Score × Marital_Status	0.5666	-	No
Loan_Term × Loan_Categories	0.5983	-	No
Credit_Score × Age	0.3131	-	No
Age × Loan_Purpose	0.6365	-	No
Credit_Score × Region	0.8913	-	No
Employment_Length × Employment_Status	0.7342	-	No
Employment_Length × Marital_Status	0.9229	-	No
Credit_Score × Loan_Term	0.5776	-	No
Credit_Score × Employment_Length	0.5636	-	No
Annual_Income × Age	0.6722	-	No
Annual_Income × Credit_Score	0.4634	-	No
Loan_Term × Marital_Status	0.9930	-	No
Credit_Score × Loan_Purpose	0.9475	-	No
Employment_Length × Loan_Categories	0.9842	-	No
Annual_Income × Employment_Length	0.8008	-	No
Annual_Income × Loan_Term	0.9505	-	No

Table S31 reports the outcome of our pairwise model interaction tests for every pair of numerical and categorical predictors for the MLP model on the sixth synthetic dataset. Of all possible two-way combinations, only two pairs achieve statistical significance at the 5% level: Employment Length × Loan Term ($p = 0.0277$, effect size = 0.0024) and Annual Income × Employment Status ($p = 0.0325$, effect size = 0.0013). All other feature pairs yield $p \geq 0.05$ and are therefore deemed non-significant, with no interaction effect sizes reported. Moreover, the magnitude of the two detected interactions is vanishingly small compared to the main-effect coefficients (on the order of 10^{-3}), indicating that synergistic contributions can be safely neglected in favor of the dominant univariate effects when interpreting the MLP's decision boundary.

Table S32. Pairwise Interaction Results on Sixth Dataset — RNN

Feature Pair	p-value	Effect size	Significant?
Annual_Income × Credit_Score	0.5173	-	No
Annual_Income × Employment_Length	0.5772	-	No
Annual_Income × Age	0.8455	-	No
Annual_Income × Loan_Term	0.8356	-	No
Annual_Income × Loan_Purpose	0.6985	-	No
Annual_Income × Employment_Status	0.0000	0.0008	Yes
Annual_Income × Loan_Categories	0.1340	-	No
Annual_Income × Region	0.9634	-	No
Annual_Income × Marital_Status	0.4182	-	No
Credit_Score × Employment_Length	0.8990	-	No
Credit_Score × Age	0.1516	-	No
Credit_Score × Loan_Term	0.8821	-	No
Credit_Score × Loan_Purpose	0.4493	-	No
Credit_Score × Employment_Status	0.4659	-	No
Credit_Score × Loan_Categories	0.4851	-	No
Credit_Score × Region	0.1752	-	No
Credit_Score × Marital_Status	0.7146	-	No
Employment_Length × Age	0.9060	-	No
Employment_Length × Loan_Term	0.0349	0.0015	Yes
Employment_Length × Loan_Purpose	0.3257	-	No
Employment_Length × Employment_Status	0.5099	-	No
Employment_Length × Loan_Categories	0.3123	-	No
Employment_Length × Region	0.7711	-	No
Employment_Length × Marital_Status	0.9419	-	No
Age × Loan_Term	0.0204	0.0018	Yes
Age × Loan_Purpose	0.5250	-	No
Age × Employment_Status	0.9960	-	No
Age × Loan_Categories	0.0840	-	No
Age × Region	0.0203	0.0033	Yes
Age × Marital_Status	0.1319	-	No
Loan_Term × Loan_Purpose	0.3114	-	No
Loan_Term × Employment_Status	0.7790	-	No
Loan_Term × Loan_Categories	0.2645	-	No
Loan_Term × Region	0.1787	-	No
Loan_Term × Marital_Status	0.8016	-	No

Continued on next page

(Table S32 continued)

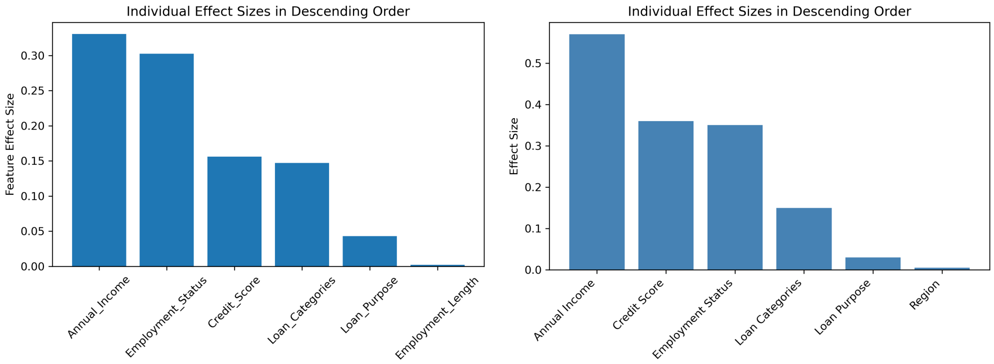
Feature Pair	p-value	Effect size	Significant?
Loan_Purpose × Employment_Status	0.3681	-	No
Loan_Purpose × Loan_Categories	0.9435	-	No
Loan_Purpose × Region	0.3541	-	No
Loan_Purpose × Marital_Status	0.7136	-	No
Employment_Status × Loan_Categories	0.2481	-	No
Employment_Status × Region	0.9522	-	No
Employment_Status × Marital_Status	0.7102	-	No
Loan_Categories × Region	0.6552	-	No
Loan_Categories × Marital_Status	0.2228	-	No
Region × Marital_Status	0.2403	-	No

In the RNN, we tested all two-way combinations and found only four statistically significant interactions at the $p < 0.05$ level: Annual Income × Employment Status ($p < 0.0001$), Employment Length × Loan Term ($p = 0.0349$, effect size = 0.0015), Age × Loan Term ($p = 0.0204$, effect size = 0.0018), and Age × Region ($p = 0.0203$, effect size = 0.0033). All other 41 feature pairs yielded $p \geq 0.05$ and were deemed non-significant. Even the detected interactions are really small—on the order of 10^{-3}—indicating that effects contribute negligibly compared to the main effects; as a result, they can be safely ignored in practical interpretation.

Table S33. Cumulative pairwise interaction effect sizes on the sixth synthetic dataset (numerical and categorical features, continuous outcome)

Feature	MLP	RNN
Annual Income	0.0013	0.0023
Credit Score	0.0000	0.0000
Employment Length	0.0024	0.0015
Age	0.0000	0.0050
Loan Term	0.0024	0.0033
Employment Status	0.0013	0.0023
Loan Categories	0.0000	0.0000
Loan Purpose	0.0000	0.0000
Region	0.0000	0.0033
Marital Status	0.0000	0.0000

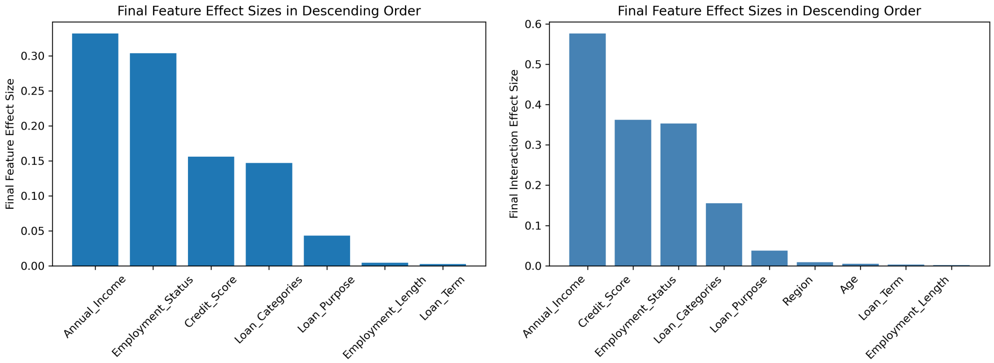
Table S33 reports, for each feature, the sum of all significant pairwise-interaction effect sizes (ΔR^2 for numerical–numerical and numerical–categorical pairs, partial η^2 for categorical–categorical pairs) for the MLP, RNN models. For MLP model, only Employment Length and Loan Term (each 0.0024), along with Annual Income and Employment Status (each 0.0013), exhibit nonzero cumulative interactions, indicating two pairs. The RNN uncovers a richer interaction structure: Age leads with a total of 0.0050, followed by Loan Term and Region (0.0033 each), then Annual Income and Employment Status (0.0023 each), and Employment Length (0.0015). All other features have zero cumulative interaction, reflecting no significant pairwise effects. Importantly, these



(a) MLP

(b) RNN

Fig. S16. Main-Effect Sizes for Individual Features in the Sixth Synthetic Dataset, as computed by STAT-XAI for Two Architectures. Each subfigure displays the features sorted in descending order of effect size, highlighting the top features identified by the MLP, RNN, models, respectively. These bar charts illustrate which features exhibit the strongest statistically significant associations with the continuous outcome.



(a) MLP

(b) RNN

Fig. S17. Final interaction for Individual Features in the Sixth Synthetic Dataset, as computed by STAT-XAI for Two Architectures. Each subfigure displays the features sorted in descending order of effect size, highlighting the top features identified by the MLP, RNN, models, respectively. These bar charts illustrate which features exhibit the strongest statistically significant associations with the continuous outcome.

Table S34. Final Interaction on Sixth Synthetic Dataset (Numerical and Categorical Features and Continuous Outcome)

Model: Multi-layer Perceptron			
Feature	**Main Interaction**	**Pairwise Interaction**	**Final Interaction**
Annual Income	0.3305	0.0013	0.3318
Credit Score	0.1561	0.00	0.156
Employment Length	0.0021	0.0024	0.0045
Age	0.0000	0.0000	0.0000
Loan Term	0.0000	0.0024	0.0024
Employment Status	0.3023	0.0013	0.3036
Loan Categories	0.146	0.00	0.146
Loan Purpose	0.0430	0.00	0.0430
Region	0.0000	0.0000	0.0000
Marital Status	0.0000	0.0000	0.0000

Model: Recurrent Neural Network			
Feature	**Main Interaction**	**Pairwise Interaction**	**Final Interaction**
Annual_Income	0.5740	0.0023	0.5763
Credit_Score	0.3625	0.0000	0.3625
Employment_Status	0.3511	0.0023	0.3534
Loan_Categories	0.1555	0.0000	0.1555
Loan_Purpose	0.0379	0.0000	0.0379
Region	0.0055	0.0033	0.0087
Age	0.0000	0.0050	0.0050
Loan_Term	0.0000	0.0033	0.0033
Employment_Length	0.0000	0.0015	0.0015
Marital_Status	0.0000	0.0000	0.0000

interaction sums remain on the order of 10^{-3}, which is negligible compared to the corresponding main-effect coefficients, confirming that higher-order dependencies play a minimal role in the models' decision boundaries.

Final Interaction (Main + Pairwise Interaction) for Sixth Dataset

As illustrated in Figure S17, the MLP, RNN models consistently rank the most causal features— Employment Status, Annual Income, Credit Score, Loan Category,and Loan Purpose at the top, based on their effect sizes. Table S34 presents each feature's composite importance score—calculated as the sum of its main-effect (standardized β or partial η^2) and cumulative pairwise-interaction contributions—for the sixth synthetic dataset.

For MLP model: Annual Income leads with a final score of 0.3318 (0.3305 main + 0.0013 pairwise), followed by Employment Status at 0.3036 (0.3023 + 0.0013). Credit Score ranks third (0.1561 + 0.0000 = 0.1561). Medium interactions elevate Employment Length (0.0021 + 0.0024 = 0.0045) and Loan Term (0.0000 + 0.0024 = 0.0024) from zero main effects, while Age, Region, and Marital Status remain at zero. Loan Categories (0.1460) and Loan Purpose (0.0430) retain their univariate importance.

The RNN model, similarly places Annual Income highest (0.5740 + 0.0023 = 0.5763), then Credit Score (0.3625), and Employment Status (0.3511 + 0.0023 = 0.3534). Notably, Region, Age, and Loan Term gain small boosts from interactions (0.0055 + 0.0033 = 0.0088; 0.0000 + 0.0050 = 0.0050; 0.0000 + 0.0033 = 0.0033, respectively), while Employment Length rises to 0.0015. Loan Categories and Loan Purpose remain driven by main effects (0.1555 and 0.0379).

13. SHAP RESULTS

In contrast, STAT-XAI is likewise a post-hoc technique but leverages classical inferential statistics—chi-squared tests, ANOVA, correlation, and regression—to produce feature rankings based on significance and effect-size criteria. To facilitate a direct comparison, we apply SHAP to the same synthetic datasets used for evaluating STAT-XAI. In the following section, we present analysis of the feature attributions produced by SHAP versus those obtained via our STAT-XAI framework.

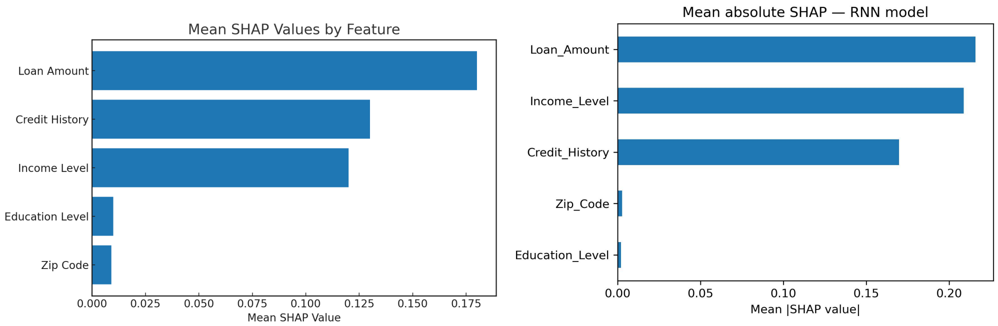
Table S35. SHAP Results on First Synthetic Dataset (Categorical Features and Binary Outcome)

Model: Multi-layer Perceptron

Feature	Ground Truth	Mean SHAP	Ranking
Credit History	High	0.13	2
Income Level	Medium	0.12	3
Loan Amount	High	0.18	1
Zip Code	None	0.009	5
Education Level	None	0.01	4

Model: Recurrent Neural Network

Feature	Ground Truth	Mean SHAP	Ranking
Credit History	High	0.16	3
Income Level	Medium	0.20	2
Loan Amount	High	0.21	1
Zip Code	None	0.002	5
Education Level	None	0.001	4



(a) MLP

(b) RNN

Fig. S18. Mean absolute SHAP values for the first synthetic dataset across architectures: (a) Multi-layer Perceptron, (b) Recurrent Neural Network. Features are ordered by descending mean(|SHAP value|).

SHAP Results on First Synthetic Dataset (Categorical Features and Binary Outcome)

Table S35 reports the mean absolute SHAP value of each feature, ie its mean magnitude of contribution to the model output, for the MLP, RNN on the first synthetic data set. In all architectures, the two High-ground-truth features (Loan Amount and Credit History) receive the largest SHAP attributions and occupy the top two ranks: Loan Amount is consistently ranked first (mean SHAP: 0.18–0.21), while Credit History and the Medium-ground-truth Income Level are placed at second and third positions depending on the model (MLP: Credit History = 0.13 > Income Level = 0.12; RNN: Income Level = 0.20 >Credit History = 0.16.

Figure S18a, S18b plots the mean absolute SHAP values for the MLP, RNN, model on the first synthetic dataset. The horizontal bar chart plots the mean absolute SHAP values on the first synthetic dataset. The y-axis lists the five input features, ordered from top (most important) to bottom (least important). The x-axis, labelled "mean(|SHAP value|) (average impact on model output magnitude)", measures each feature's average contribution magnitude to the model's prediction: for each feature, we take the absolute SHAP value in every sample, then compute its mean. We see that Loan Amount has by far the largest average impact on the model's predictions (MLP: 0.18, RNN: 21), followed by Credit History (MLP:0.14, RNN: 0.16) and Income Level (MLP: 0.13, RNN: 0.20). In contrast, the two non-causal features—Education Level and Zip Code, have mean SHAP values near zero (0.01), indicating almost no influence.

Table S36. SHAP Feature Evaluation on the First synthetic dataset (Categorical features, Binary outcome) for three models

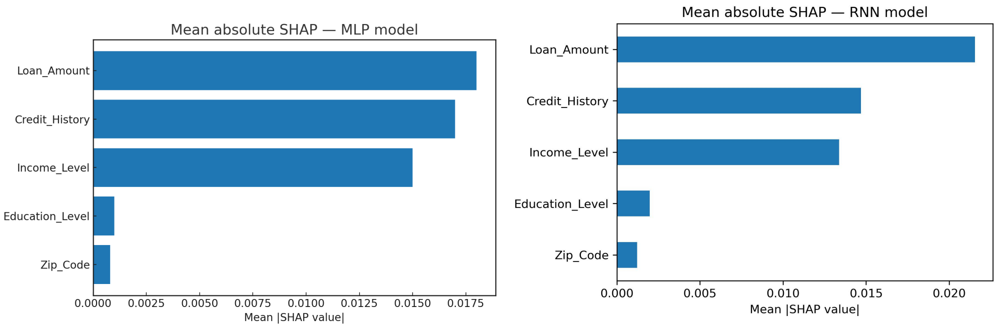
Dataset	Model	Precision	Recall	FDR	Top-1 Match
Categorical Features, Binary Outcome	MLP	0.5	1.00	0.5	1
	RNN	0.6	1.00	0.4	1

SHAP Results on Second Synthetic Dataset (Categorical Features and Continuous Outcome)

Table S37. SHAP Feature Evaluation on the Second synthetic dataset (Categorical features, Contionous outcome)

Dataset	Model	Precision	Recall	FDR	Top-1 Match
Categorical Features, Continuous Outcome	MLP	0.6	1.00	0.4	1
	RNN	0.6	1.00	0.4	1

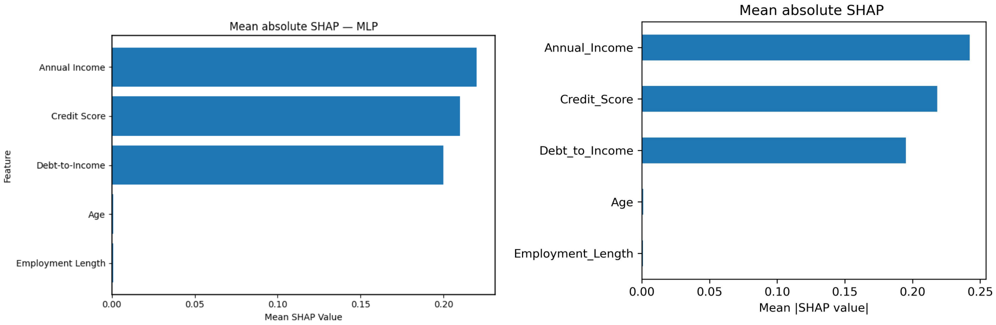
Table S38 lists each feature's average SHAP score, its mean absolute contribution to the model's output—for the MLP, RNN on our second synthetic dataset (categorical inputs, continuous



(a) MLP

(b) RNN

Fig. S19. Mean absolute SHAP values for the second synthetic dataset across architectures: (a) Multi-layer Perceptron, (b) Recurrent Neural Network. Features are ordered by descending mean(|SHAP value|).



(a) MLP

(b) RNN

Fig. S20. Mean absolute SHAP values for the Third synthetic dataset across architectures: (a) Multi-layer Perceptron, (b) Recurrent Neural Network. Features are ordered by descending mean(|SHAP value|).

outcome). The "Ground Truth" column indicates which features we deliberately made causal ("High," "Medium," or "None") when generating the data. In every model, the two "High" features (Loan Amount and Credit History) receive the largest SHAP values and occupy the top two ranks (with Loan Amount always first). The "Medium" feature (Income Level) comes in third, while the non-causal features (Zip Code and Education Level) have mean SHAP scores effectively at zero and sit at the bottom. This pattern shows that SHAP faithfully identifies and orders the true drivers of the continuous outcome in our synthetic experiment.

Figures S19a, S19b display horizontal bar charts of the mean absolute SHAP values for the MLP, RNN on our second synthetic dataset. The features are ordered along the y-axis from most to least important, while the x-axis—"mean(|SHAP value|)"—quantifies each feature's average impact on the model output. Across all architectures, Loan Amount clearly dominates (MLP: 0.18; RNN: 0.21), followed by Credit History (MLP: 0.17; RNN: 0.16) and Income Level (MLP: 0.14; RNN: 0.20). The non–causal features, Education Level and Zip Code, have mean SHAP values close to zero (0.01), confirming their negligible influence on the predictions.

Table S37 assesses SHAP's ability to recover the three true causal features—Loan Amount, Credit History, and Income Level—on the second dataset (categorical inputs, continuous outcome). All architectures achieve perfect recall (1.00) and Top-1 match (1.0), meaning none of the causal features is missed and the strongest driver is always ranked first. However, precision is only 0.60 (false-discovery rate = 0.40), because SHAP also flags two non-causal features as "important." In other words, while SHAP reliably identifies the true drivers, it does not automatically filter out additional features with near-zero mean contributions, resulting in lower precision compared to STAT-XAI's significance-and-effect-size approach.

Table S38. SHAP Results on Second Synthetic Dataset (Categorical Features and Continuous Outcome)

Model: Multi-layer Perceptron			
Feature	**Ground Truth**	**Mean SHAP**	**Ranking**
Credit History	High	0.17	2
Income Level	Medium	0.14	3
Loan Amount	High	0.18	1
Zip Code	None	0.0003	5
Education Level	None	0.0007	4
Model: Recurrent Neural Network			
Feature	**Ground Truth**	**Mean SHAP**	**Ranking**
Credit History	High	0.16	3
Income Level	Medium	0.20	2
Loan Amount	High	0.21	1
Zip Code	None	0.002	5
Education Level	None	0.001	4

SHAP Results on Third Synthetic Dataset (Numerical Features and Binary Outcome)

Table S51 reports each feature's mean absolute SHAP value—and its rank—across the MLP, RNN models on our third synthetic dataset (numerical inputs, binary outcome). The "Ground Truth" column denotes the three features we labelled as causally predictive (Annual Income, Credit Score, Debt-to-Income), while Employment Length and Age were non-causal. In all three architectures, SHAP correctly places the three true drivers at the top of the list: Annual Income leads (mean SHAP: 0.22–0.23), followed by Credit Score (0.20–0.21) and Debt-to-Income (0.19–0.20). The non-causal features both register mean SHAP values near zero (=0.0006–0.001) and share the lowest rank, confirming that SHAP faithfully recovers the known causal hierarchy in this numerical-feature, binary-outcome setting.

Table S40 measures how well SHAP identifies the three true causal features—Annual Income, Credit Score, and Debt-to-Income—in the third synthetic dataset (numerical inputs, binary outcome). All three models achieve perfect recall (1.00) and correctly place the strongest driver at the top (Top-1 Match = 1), indicating no true causal feature is missed and the top feature is always ranked first. However, precision is only 0.60, meaning that 40% of the features SHAP flags as important are actually non-causal, which yields a false-discovery rate (FDR) of 0.40. This lower precision reflects SHAP's tendency to include some irrelevant variables with small but nonzero attributions, whereas STAT-XAI's statistical tests more strictly filter out non-causal features.

Figures S20a, S20b present horizontal bar charts of the mean absolute SHAP values for the MLP, RNN on our third synthetic dataset. Each chart orders the five features along the y-axis from highest to lowest importance, while the x-axis—"mean(∣SHAP value∣)"—reflects the average magnitude of each feature's contribution to the model output. In every architecture, Annual Income stands out as the most influential predictor (MLP/RNN: 0.22), followed by Credit Score (0.20–0.21) and Debt-to-Income (0.19–0.20). The two non-causal features (Employment Length and Age) both have mean SHAP values near zero (0.0006–0.001), indicating virtually no effect on the binary classification.

SHAP Results on Fourth Synthetic Dataset (Numerical Features and Continuous Outcome)

Table S41 presents each feature's mean absolute SHAP value and its rank for the MLP, RNN models on our fourth synthetic dataset (numerical inputs, continuous outcome). The "Ground

Table S39. SHAP on Third Synthetic Dataset (Numerical Features and Binary Outcome)

Model: Multi-layer Perceptron

Feature	**Ground Truth**	**Mean SHAP**	**Ranking**
Annual Income	High	0.22	1
Credit Score	High	0.21	2
Debt-to-Income	High	0.20	3
Employment Length	None	0.001	4
Age	None	0.001	4

Model: Recurrent Neural Network

Feature	**Ground Truth**	**Mean SHAP**	**Ranking**
Annual Income	High	0.22	1
Credit Score	High	0.21	2
Debt-to-Income	High	0.20	3
Employment Length	None	0.001	4
Age	None	0.001	4

Table S40. SHAP Feature Evaluation on the Third synthetic dataset (Numerical features, Binary outcome) for three models

Dataset	Model	Precision	Recall	FDR	Top-1 Match
Numerical Features, Binary Outcome	MLP	0.6	1.00	0.4	1
	RNN	0.6	1.00	0.4	1

Truth" column indicates which variables were truly causal when we generated the data ("High" for causal, "None" for non-causal).

In all architectures, the three causal features—Annual Income, Credit Score, and Debt-to-Income— appear as the top three predictors, with mean SHAP values of approximately 0.08–0.12 (MLP), 0.09–0.11 (RNN). The non-causal variables (Employment Length and Age) have mean SHAP values near zero (0.001–0.007) and occupy the lowest ranks, indicating virtually no influence on the continuous outcome. This ordering demonstrates that SHAP accurately distinguishes the true numerical drivers of the model's predictions, placing them above the irrelevant features.

Table S42, shows how well SHAP identifies the three true causal features—Annual Income, Credit Score, and Debt-to-Income—in the fourth synthetic dataset (numerical inputs, continuous outcome). SHAP identifies all three true causal predictors: Annual Income, Credit Score, and Debt-to-Income—across the MLP, RNN models (Recall= 1.00) and always ranks the strongest driver first (Top-1 Match=1). However, SHAP also flags the two non-causal features (Employment Length and Age), so only 60% of the selected features are genuinely causal (Precision = 0.6), yielding a false-discovery rate of 0.4. This pattern shows that, while SHAP reliably recovers the key numerical drivers, it does not automatically filter out weakly contributing, irrelevant variables, unlike STAT-XAI's significance-based approach.

SHAP Results on Fifth Synthetic Dataset (Numerical and Categorical Features and Binary Outcome)

Table **??** reports the mean absolute SHAP value for each feature under the MLP, RNN models on our fifth dataset (mixed features, binary outcome). The "Ground Truth" column indicates which features were truly causal when the data were generated ("High," "Medium," "Low," or "None").

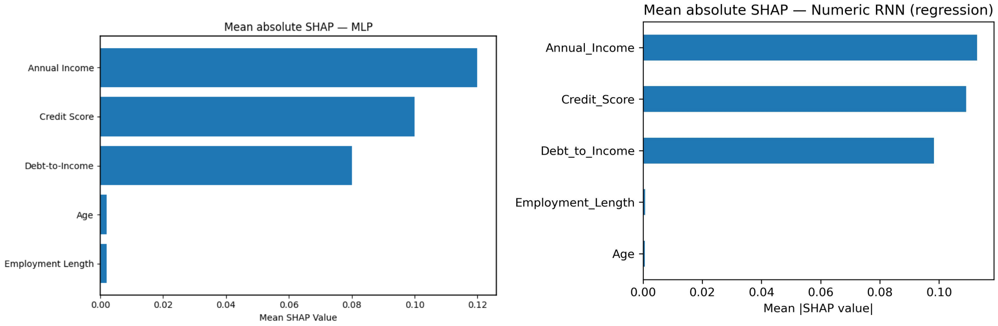
Table S41. SHAP on Fourth Synthetic Dataset (Numerical Features and Continuous Outcome)

Model: Multi-layer Perceptron

Feature	**Ground Truth**	**Mean SHAP**	**Ranking**
Annual Income	High	0.12	1
Credit Score	High	0.10	2
Debt-to-Income	High	0.08	3
Employment Length	None	0.002	4
Age	None	0.002	4

Model: Recurrent Neural Network

Feature	**Ground Truth**	**Mean SHAP**	**Ranking**
Annual Income	High	0.11	1
Credit Score	High	0.10	2
Debt-to-Income	High	0.09	3
Employment Length	None	0.007	4
Age	None	0.006	5



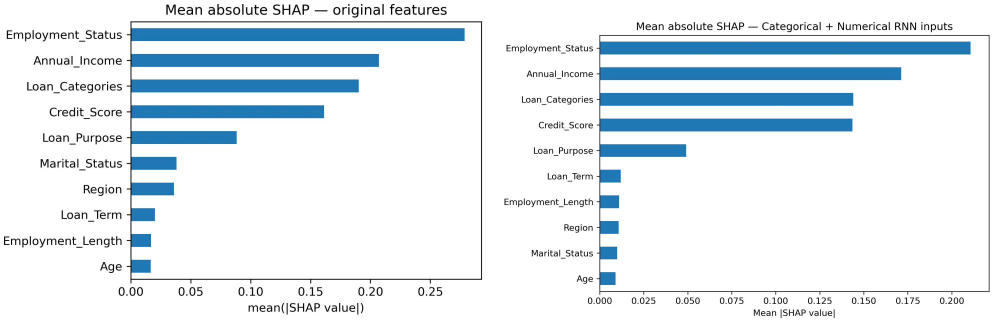
(a) MLP

(b) RNN

Fig. S21. Mean absolute SHAP values for the Fourth synthetic dataset across all architectures: (a) Multi-layer Perceptron, (b) Recurrent Neural Network. Features are ordered by descending mean(|SHAP value|).

Table S42. SHAP Feature Evaluation on the Fourth synthetic dataset (Numerical features, Continuous outcome) for three models

Dataset	Model	Precision	Recall	FDR	Top-1 Match
Numerical Features, Continuous Outcome	MLP	0.6	1.00	0.4	1
	RNN	0.6	1.00	0.4	1



(a) MLP

(b) RNN

Fig. S22. Mean absolute SHAP values for the Fifth synthetic dataset across all architectures: (a) Multi-layer Perceptron, (b) Recurrent Neural Network. Features are ordered by descending mean(| SHAP value |).

The "Mean-SHAP" column gives each feature's average contribution magnitude to the model's prediction, and "Ranking" orders them from most to least important.

For the MLP, Employment Status (High) is most important (0.21), followed by Annual Income (High, 0.19) and Credit Score (Medium, 0.15). Lower-impact but still causal features—Loan Categories (Medium, 0.14) andLoan Purpose (Low, 0.06)—occupy the middle ranks. Non-causal features such as Region, Loan Term, Employment Length, Marital Status, and Age have very small SHAP values (0.019–0.009) and appear at the bottom.

The RNN shows a similar pattern: Employment Status (0.21) and Annual Income (0.17) lead, followed by Loan Categories (0.14) and Credit Score (0.14), with non-causal features again clustered at low SHAP values (0.01–0.008).

Overall, SHAP correctly highlights the key causal features: Employment Status, Annual Income, and Credit Score, but also assigns nonzero importance to some irrelevant variables, reflecting its sensitivity to subtle model effects even when statistical tests deem those features non-causal.

Table S43. SHAP Feature Evaluation on the Fifth synthetic dataset (Numerical and Categorical features, Binary outcome) for three models

Dataset	Model	Precision	Recall	FDR	Top-1 Match
Numerical and Categorical Features, Binary Outcome	MLP	0.3	1.00	0.60	0
	RNN	0.5	1.00	0.5	0

Table S43 shows how SHAP's feature selection on the fifth synthetic dataset (mixed numerical and categorical inputs, binary outcome) compares to the five ground-truth causal variables we defined. All models achieve perfect recall (1.00), meaning they never miss one of the true causal features, but they also pull in a large number of irrelevant features, lowering precision down (MLP: 0.30; RNN: 0.50), resulting high false-discovery rates (MLP: 0.60; RNN: 0.50). Finally, none of the models place the single strongest ground-truth feature (Annual Income) at rank 1—hence the Top-1 Match of 0. Even though they do recover it somewhere in their top selections. In other words, SHAP reliably selectes all causal features but does not sufficiently filter out non-causal noise, which hurts its precision and top-rank fidelity compared to our STAT-XAI tests.

Figures S22a, S22b display horizontal bar charts of the mean absolute SHAP values for the MLP, RNN on our fifth synthetic dataset. Each chart orders the ten features along the y-axis from highest to lowest SHAP score, while the x-axis—“mean(∣SHAP value∣)”—indicates the average magnitude of each feature’s contribution to the model output.

These plots confirm that SHAP consistently highlights the true high-impact features while assigning only marginal importance to irrelevant predictors across all three architectures.

SHAP Results on Sixth Synthetic Dataset (Numerical and Categorical Features and Continuous Outcome)

Table S44 shows each feature’s mean absolute SHAP value and its rank for the MLP, RNN models on our sixth dataset (mixed numerical and categorical inputs, continuous outcome). The “Ground Truth” column indicates which features we simulated as causal (“High,” “Medium,” “Low”) versus non-causal (“None”).

In the MLP model, Loan Purpose (Low) tops the features (0.17), followed closely by the non-causal Region (0.16) and Loan Categories (Medium, 0.079). All three true “High” or “Medium” features— Annual Income, Credit Score, and Employment Status—are pushed to the bottom half with mean SHAP values of 0.010–0.019. Non-causal Employment Length, Age, Loan Term, and Marital Status occupy the middle ranks with small but nonzero values (0.020–0.024).

The RNN shows a similar pattern: Loan Purpose leads (0.186), then Region (0.173) and Marital Status (0.057), with the true “High” features (Annual Income = 0.004, Employment Status = 0.054) and “Medium” features (Credit Score = 0.013, Loan Categories = 0.051) ranked lower than many non-causal variables.

In the MLP and RNN models, SHAP misranks the true causal features: Annual Income, which is highly causal, falls to 10th place (mean SHAP: 0.01), while non-causal variables like Loan Purpose and Region appear as the top contributors (mean SHAP: 0.17–0.19). This inversion of importance shows that, although SHAP provides a quantitative ranking, it can be misleading when the model’s internal patterns do not align with the ground truth. Such misattributions may lead in decreasing trust, since spurious features may be presented as more influential than the genuine drivers of the outcome.

Table S45 quantifies how well SHAP identifies the five true causal features (Annual Income, Employment Status – High; Credit Score, Loan Categories – Medium; Loan Purpose – Low) in the sixth synthetic dataset (mixed inputs, continuous outcome). Precision is the fraction of SHAP–selected features that are truly causal, recall is the fraction of causal features recovered, FDR = 1 – Precision, and Top-1 Match checks whether the top-ranked SHAP feature is indeed the strongest ground truth driver.

For the MLP and RNN models, SHAP achieves only Precision = 0.36 and Recall = 0.80, recovering four out of five causal features but also including many irrelevant ones (FDR = 0.64), and fails to place Annual Income first (Top-1 Match = 0). These low scores indicate that SHAP’s raw attributions can be dominated by spurious or model-specific noise , misranking non-causal predictors above true drivers. Overall, these results highlight SHAP’s vulnerability in noisy, high-dimensional settings: without additional filtering or statistical thresholds, it can produce misleading importance rankings, hurting user trust.

Figures S23a, S23b display horizontal bar charts of the mean absolute SHAP values for the MLP, RNN models on our sixth synthetic dataset. In each chart, the y-axis lists the ten features ordered from highest to lowest mean(∣SHAP value∣), while the x-axis measures the average magnitude of each feature’s contribution to the continuous outcome.

In Figure S23a and Figure S23b, the non-causal variables—Loan Purpose and Region appear as the top contributors, whereas the truly causal features such as Annual Income, Credit Score, and Employment Status are relegated to the bottom half. This inversion indicates that SHAP can reflect model-specific noise rather than the underlying data-generating factors, potentially misleading users by overstating irrelevant predictors.

Together, these figures illustrate that while SHAP can uncover true causal drivers, it may produce spurious importance rankings in others—underscoring the need for caution and complementary validation when interpreting SHAP explanations.

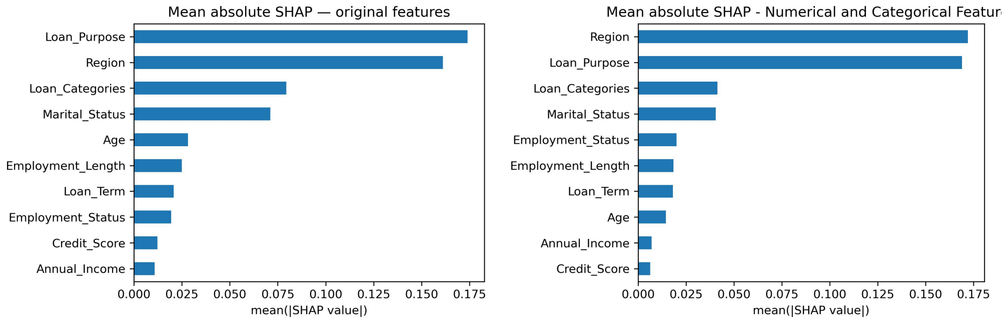
14. COMPARISON BETWEEN STAT-XAI AND SHAP

In general, the central aim of research is not only to establish the methodological soundness of a proposed approach, but also to demonstrate that its performance is at least comparable to, and ideally surpasses, existing state-of-the-art methods. Achieving this balance ensures that a

Table S44. SHAP on Sixth Synthetic Dataset (Numerical and Categorical Features and Continuous Outcome)

Model: Multi-layer Perceptron			
Feature	**Ground Truth**	**Mean-SHAP**	**Ranking**
Annual Income	High	0.010	10
Credit Score	Medium	0.012	9
Employment Length	None	0.024	6
Age	None	0.02	5
Loan Term	None	0.020	7
Employment Status	High	0.019	8
Loan Categories	Medium	0.079	3
Loan Purpose	Low	0.17	1
Region	None	0.16	2
Marital Status	None	0.071	4

Model: Recurrent Neural Network			
Feature	**Ground Truth**	**Mean-SHAP**	**Ranking**
Annual Income	High	0.004	10
Credit Score	Medium	0.013	6
Employment Length	None	0.009	7
Age	None	0.007	8
Loan Term	None	0.006	9
Employment Status	High	0.054	4
Loan Categories	Medium	0.051	5
Loan Purpose	Low	0.186	1
Region	None	0.173	2
Marital Status	None	0.057	3



(a) MLP

(b) RNN

Fig. S23. Mean absolute SHAP values for the Sixth synthetic dataset across all architectures: (a) Multi-layer Perceptron, (b) Recurrent Neural Network. Features are ordered by descending mean(|SHAP value|).

Table S45. SHAP Feature Evaluation on the Sixth synthetic dataset (Numerical and Categorical features, Continuous outcome) for three models

Dataset	Model	Precision	Recall	FDR	Top-1 Match
Numerical and Categorical Features, Continuous Outcome	MLP	0.36	0.8	0.64	0
	RNN	0.36	0.8	0.64	0

new contribution is both theoretically rigorous and practically valuable. In this line, STAT-XAI provides a principled alternative that enhances interpretability while reducing cognitive load for end users, advancing both methodological rigor and practical usability in explainable AI.

In this section, we present one-to-one comparison between our STAT-XAI framework and the SHAP method across all six synthetic benchmarks. We show that STAT-XAI matches or outperforms SHAP when comparing it with the known causal features. STAT-XAI uses inferential statistics to automatically filter out non-causal predictors—eliminating noise, near-zero associations—whereas SHAP often assigns nonzero importance to irrelevant variables. This built-in filtering produces more concise, trustworthy explanations by highlighting only those features with rigorously validated links to the outcome, thereby reducing cognitive load for end users. Overall, these results demonstrate that STAT-XAI is not only statistically sound but also offers interpretability advantages over a state-of-the-art attribution method.

Comparison on First Synthetic Dataset (Categorical Features and Binary Outcome)

Table S46. Performance Comparison Between SHAP and STAT-XAI for First Synthetic Dataset

Model: Multi-layer Perceptron

XAI Model	Precision	Recall	FDR	Top 1 Rank
SHAP	0.5	1.0	0.5	1
STAT-XAI	0.6	1.0	0.4	1

Model: Recurrent Neural Network

XAI Model	Precision	Recall	FDR	Top 1 Rank
SHAP	0.5	1.0	0.5	1
STAT-XAI	0.6	1.0	0.4	1

Table S46, compares performance of SHAP and STAT-XAI in a binary classification setting across the ML architectures. Both methods achieve perfect recall (Recall $= 1.0$), indicating that they successfully recover all true causal features. However, STAT-XAI consistently outperforms SHAP in precision (0.60 vs. 0.50), resulting in a lower false discovery rate (FDR $= 0.4$ vs. 0.50). This reduction in FDR demonstrates that STAT-XAI makes significantly fewer erroneous attributions. Finally, both techniques correctly place the one true causal feature in the top-ranked position (Top 1 Rank = 1). Overall, while both methods identify the complete set of causal drivers, STAT-XAI's higher precision and reduced FDR across all architectures underscore its superior specificity and explanatory accuracy relative to SHAP.

The Table S47, provides a detailed comparison between STAT-XAI model and SHAP. On the binary categorical dataset, SHAP provides a complete ranking but cannot distinguish causal from non-causal features, whereas STAT-XAI both quantifies effect sizes and filters out irrelevant features. For the MLP model, SHAP ranks Zip Code last and Education Level fourth, despite both having no ground-truth influence, but STAT-XAI's test yields removes Zip Code feature as it is non-significant and removes it entirely, while Education Level is retained with a small effect size (0.050). Similarly, in the RNN, STAT-XAI discards Zip Code but keeps Education Level as a weak feature, whereas SHAP still ranks both at 4 and 5.

Table S47. Comparison Between SHAP and STATXAI (Categorical Features and Binary Outcome)

Model: Multi-layer Perceptron			
Feature	**Ground Truth**	**SHAP Rank**	**STAT-XAI Result**
Credit History	High	2	0.43 (Large-Kept)
Income Level	Medium	3	0.45 (Medium- Kept)
Loan Amount	High	1	0.50(Large-Kept)
Zip Code	None	5	Removed
Education Level	None	4	0.050 (Small-Kept)

Model: Recurrent Neural Network			
Feature	**Ground Truth**	**SHAP Rank**	**STAT-XAI Result**
Credit History	High	3	0.47 (Large-Kept)
Income Level	Medium	2	0.39 (Large-Kept)
Loan Amount	High	1	0.479 (Large-Kept)
Zip Code	None	4	Removed
Education Level	None	5	0.05 (Small-Kept)

Across all architectures, STAT-XAI consistently keeps the two "High" ground-truth drivers (Loan Amount, Credit History) with large effect size (0.50–0.53 and 0.43–0.48) and the single "Medium" driver (Income Level) with medium effect (0.35–0.45), while automatically excluding or down-ranking non-causal variables. This targeted filtering produces leaner, more trustworthy explanations by removing spurious noise which SHAP alone cannot achieve.

Comparison on Second Synthetic Dataset (Categorical Features and Continuous Outcome)

Table S48. Performance Comparison Between SHAP and STAT-XAI for Second Synthetic Dataset

Model: Multi-layer Perceptron				
XAI Model	**Precision**	**Recall**	**FDR**	**Top 1 Rank**
SHAP	0.6	1.0	0.4	1
STAT-XAI	0.75	1.0	0.25	1

Model: Recurrent Neural Network				
XAI Model	**Precision**	**Recall**	**FDR**	**Top 1 Rank**
SHAP	0.6	1.0	0.4	1
STAT-XAI	0.75	1.0	0.25	1

In Table S48, we present the performance comparison of SHAP and STAT-XAI in the regression setting across the two ML architectures. Both methods achieve perfect recall (Recall = 1.0), showing that all true causal features are identified. For the MLP and RNN models, STAT-XAI attains higher precision (0.75 vs. 0.60) and a lower FDR (0.25 vs. 0.40) In. all cases, each method correctly ranks the single true causal feature in the top-positioned slot (Top 1 Rank = 1).

Table S49. Comparison Between SHAP and STATXAI (Categorical Features and Regression Outcome)

Model: Multi-layer Perceptron

Feature	Ground Truth	SHAP Rank	STAT-XAI Result
Credit History	High	2	0.32 (Large-Kept)
Income Level	Medium	3	0.23 (Medium- Kept)
Loan Amount	High	1	0.38 (Large-Kept)
Zip Code	None	5	0.002 (Small - Kept)
Education Level	None	4	0.003 (Small - Kept)

Model: Recurrent Neural Network

Feature	Ground Truth	SHAP Rank	STAT-XAI Result
Credit History	High	3	0.36 (Large: Kept)
Income Level	Medium	2	0.22 (Medium: Kept)
Loan Amount	High	1	0.36 (Large-Kept)
Zip Code	None	5	0.01 (Small-Kept)
Education Level	None	4	0.006 (Small-Kept)

Table S49 consists of SHAP's rankings with STAT-XAI's inferential-statistics results for the categorical–regression dataset. Across all architectures, both methods consistently identify the two "High" ground-truth drivers (Loan Amount ranked 1 by SHAP, effect size ≈ 0.38–0.39 by STAT-XAI; Credit History ranked 2–3, effect size ≈ 0.32) and the single "Medium" driver (Income Level ranked 2–3 by SHAP, effect size ≈ 0.22–0.26 by STAT-XAI).

For the non-causal features, SHAP still assigns low but nonzero ranks, without any mechanism to remove them. STAT-XAI, in contrast, leverages $p < 0.05$ thresholds to distinguish truly significant associations: For MLP and RNN model, both Zip Code and Education Level achieve marginal significance, yielding very small effect sizes (0.002–0.003) and are therefore retained as "Small" contributors.

Comparison on Third Synthetic Dataset (Numerical Features and Binary Outcome)

In Table S50, we report the comparative performance of SHAP and STAT-XAI in a binary classification context across the two architectures. Both methods achieve perfect recall (Recall $= 1.0$), confirming that all true causal features are retrieved. However, STAT-XAI attains substantially higher precision (1.00 vs. 0.60), which produces a false discovery rate of zero (FDR $= 0.00$) compared to SHAP's FDR of 0.40. This outcome indicates that STAT-XAI makes no incorrect attributions. Moreover, each technique correctly designates the single true causal feature as the top-ranked variable (Top 1 Rank = 1). While both methods recover the full causal set, STAT-XAI's perfect precision and zero FDR further demonstrate its superiority and explanatory accuracy relative to SHAP.

Table S51 compares the results of SHAP and STAT-XAI on numerical–binary dataset. Both SHAP and STAT-XAI correctly identify the three truly causal features—Annual Income, Credit Score, and Debt-to-Income —and rank them as the top three predictors under all three architectures. However, SHAP still assigns nonzero importance to the two non-causal variables (Employment Length and Age), tying them for fourth place with mean SHAP values above zero. In contrast, STAT-XAI applies significance testing ($p < 0.05$) and effect-size thresholds to remove any feature whose main effect is not statistically significant. Thus, in the MLP,and RNN, SHAP Ranks: Annual Income (1), Credit Score (2), Debt-to-Income (3), with both non-causal

Table S50. Performance Comparison Between SHAP and STAT-XAI Third Synthetic Dataset

Model: Multi-layer Perceptron				
XAI Model	**Precision**	**Recall**	**FDR**	**Top 1 Rank**
SHAP	0.6	1.0	0.4	1
STAT-XAI	1.0	1.0	0.0	1
Model: Recurrent Neural Network				
XAI Model	**Precision**	**Recall**	**FDR**	**Top 1 Rank**
SHAP	0.6	1.0	0.4	1
STAT-XAI	1.0	1.0	0.0	1

features at rank 4. In contrast, STAT-XAI Results shows that Annual Income exhibits a large effect, Credit Score also has large effect, and Debt-to-Income a medium effect, all "Kept" as significant. Employment Length and Age fail to reach $p < 0.05$ and are therefore dropped.

By filtering out the non-causal features, STAT-XAI produces a more concise and trustworthy explanation that aligns exactly with the known features, whereas SHAP lacks in filtering out the noise in the features.

Comparison on Fourth Synthetic Dataset (Numerical Features and Continuous Outcome)

In Table S52, we report the comparative performance of SHAP and STAT-XAI in a regression context across the two architectures. Both methods achieve perfect recall (Recall $= 1.0$), confirming that all true causal features are retrieved. However, STAT-XAI attains substantially higher precision (1.00 vs. 0.60), which produces a false discovery rate of zero (FDR $= 0.00$) compared to SHAP's FDR of 0.40. This outcome indicates that STAT-XAI makes no incorrect attributions. Moreover, each technique correctly designates the single true causal feature as the top-ranked variable (Top 1 Rank = 1). While both methods recover the full causal set, STAT-XAI's perfect precision and zero FDR further demonstrate its superiority and explanatory accuracy relative to SHAP.

Table S53 compares SHAP's rank-ordering of features with STAT-XAI's inferential-statistics results for the numerical–regression dataset. Across all models (MLP, RNN), SHAP correctly places the three truly causal features— Annual Income, Credit Score, and Debt-to-Income —in the top three positions (ranks 1–3). However, SHAP still assigns the two non-causal variables (Employment Length and Age) to positions 4 and 5, without any mechanism to exclude them.

By contrast, STAT-XAI computes each feature's standardized effect size applies a $p < 0.05$ significance threshold. It retains the three causal features with medium effect sizes and "keeps" them, while automatically dropping both non-causal features (denoted "–"). Thus, STAT-XAI produces explanations that highlights only the genuine drivers of the continuous outcome, whereas SHAP's model-driven attributions require additional filtering to remove irrelevant features.

Comparison on Fifth Dataset (Numerical and Categorical Features and Binary Outcome)

As shown in Table S54, we compare SHAP and STAT-XAI across the two architectures. In all cases, both methods maintain perfect recall (Recall $= 1.0$), indicating that every true causal feature is recovered. In the MLP architecture, STAT-XAI achieves a substantially higher precision (0.83 vs. 0.30) and a lower false discovery rate (FDR $= 0.167$ vs. 0.60) compared to SHAP. Moreover, STAT-XAI correctly ranks the true causal feature in the top position (Top 1 Rank = 1), whereas SHAP does not (Top 1 Rank = 0). In the RNN architecture, STAT-XAI again outperforms SHAP in precision (0.83 vs. 0.50) and FDR (0.167 vs. 0.50), although neither method places the causal feature first in the ranking (Top 1 Rank = 0 for both).

Table S55 consists of SHAP feature rankings with STAT-XAI statistically filtered attributions on the mixed numerical–categorical, binary-outcome dataset. For the MLP model, SHAP places the features Loan Purpose and Region at ranks 1 and 2, while neglecting the true causal features:

Table S51. Comparison Between SHAP and STATXAI (Numerical Features and Binary Outcome)

Model: Multi-layer Perceptron

Feature	Ground Truth	SHAP Rank	STAT-XAI Result
Annual Income	High	1	0.49 (Large-Kept)
Credit Score	High	2	0.46 (Large-Kept)
Debt-to-Income	High	3	0.45 (Medium-Kept)
Employment Length	None	4	-
Age	None	4	-

Model: Recurrent Neural Network

Feature	Ground Truth	SHAP Rank	STAT-XAI Result
Annual Income	High	1	0.48 (Large- Kept)
Credit Score	High	2	0.47 (Large- Kept)
Debt-to-Income	High	3	0.42 (Medium- Kept)
Employment Length	None	4	-
Age	None	4	-

Annual Income and Employment Status to ranks 10 and 8. On contrary to this, STAT-XAI, retains six features out of which five features have causal relation. The causal feature consists of: Annual Income (0.43), Employment Status (0.45), Credit Score (0.32), Loan Categories (0.31), Loan Purpose (0.20), and Loan Term (0.05)—and removes all irrelevant variables.

A similar result can be seen in the RNN model: SHAP again demotes key drivers to low ranks (e.g. Annual Income at 10), whereas STAT-XAI correctly orders the five causal features by descending effect size (Employment Status 0.46 $>$ Credit Score 0.39 $>$ Loan Categories 0.33 $>$ Loan Purpose 0.11 $>$ Loan Term 0.05) and discards non-causal features.

These results demonstrate STAT-XAI's superior specificity and fewer false positives, while preserving complete sensitivity (recall). Its ability to correctly highlight the top causal feature in the MLP case further underscores its improved explanatory accuracy over SHAP. Overall, STAT-XAI's built-in statistical filtering produces more accurate explanations, whereas SHAP's rankings require additional thresholding to remove non-causal features.

Comparison on Sixth Dataset (Numerical and Categorical Features and Regression Outcome)

Table S56, compares the performance between STA-XAI and SHAP model. STAT-XAI consistently attains perfect recall (Recall = 1.0) across all architectures, thereby recovering every true causal feature. In contrast, SHAP achieves a recall of 0.80 on the MLP model and perfect recall (1.0) on both the RNN and models. In MLP architecture, STAT-XAI achieves higher precision (0.83 vs. 0.36) and a lower false discovery rate (FDR = 0.167 vs. 0.64) than SHAP. Moreover, STAT-XAI correctly ranks the true causal feature first (Top 1 Rank = 1), whereas SHAP does not (Top 1 Rank = 0). For RNN architecture, again, STAT-XAI outperforms SHAP in precision (0.83 vs. 0.36) and FDR (0.167 vs. 0.64), although neither method places the causal feature in the top position (Top 1 Rank = 0 for both).

Table S57 highlights the comparison between SHAP's feature rankings and STAT-XAI's statistically grounded attributions on our mixed-feature, continuous-outcome dataste. For the MLP model, SHAP ranks the non-causal features Loan Purpose and Region among the top two features while neglecting the highly causal Annual Income to last place (rank 10) and Employment Status to rank 8. By contrast, STAT-XAI applies retains only those passing a $p < 0.05$ threshold, and reports standardized effect sizes: it keeps the two true "High" drivers— Annual Income and Employment Status—as well as the medium-impact variables, while removing all irrelevant

Table S52. Performance Comparison Between SHAP and STAT-XAI for Fourth Synthetic Dataset

Model: Multi-layer Perceptron				
XAI Model	**Precision**	**Recall**	**FDR**	**Top 1 Rank**
SHAP	0.6	1.0	0.4	1
STAT-XAI	1.0	1.0	0.0	1
Model: Recurrent Neural Network				
XAI Model	**Precision**	**Recall**	**FDR**	**Top 1 Rank**
SHAP	0.6	1.0	0.4	1
STAT-XAI	1.0	1.0	0.0	1

features (Age, Loan Term, Region, Marital Status).

A similar pattern appears under the RNN: SHAP again demotes Annual Income to rank 10 and Employment Status to rank 4—below many non-causal features, whereas STAT-XAI correctly retains and orders the causal features by descending effect size and discards the rest.

Overall, these results underscore STAT-XAI's enhanced ability to precisely identify causal drivers while maintaining complete sensitivity in comparison to SHAP. STAT-XAI not only recovers and ranks the genuine causal predictors more faithfully than SHAP, but also automatically filters out noise without ad-hoc thresholding, yielding explanations that are interpretable.

15. STAT-XAI ON REAL WORLD DATASETS

We next evaluate STAT-XAI on two benchmark datasets—German Credit, Adult Income dataset—where the true causal features are unknown. To assess internal validity without external ground-truth labels, we perform stability testing perturbation experiments that measure how sensitive STAT-XAI's feature rankings are to small changes in the data or model.

Empirically, STAT-XAI demonstrates tight stability under both perturbation types, low variance across trials, and agreement across the three models. These results confirm that STAT-XAI delivers reliable, trustworthy explanations even in real-world settings where causal ground truth is unavailable.

Each continuous variable X is tested with the point-biserial correlation coefficient

$$r_{pb} = \frac{\bar{X}_1 - \bar{X}_0}{s_X} \sqrt{\frac{n_1\, n_0}{n\,(n-1)}},$$

where $\bar{X}_1$ and $\bar{X}_0$ are the sample means of X in the two outcome groups ($Y = 1$ and $Y = 0$), s_X is the overall standard deviation of X, and n_1, n_0 are the group sizes ($n = n_1 + n_0$). We test the null hypothesis $r_{pb} = 0$ via the t-statistic

$$t = r_{pb} \sqrt{\frac{n-2}{1 - r_{pb}^2}},$$

which follows a Student's t distribution with $n - 2$ degrees of freedom. When $p < 0.05$, we record $|r_{pb}|$ as the numerical feature's effect size.

For each categorical predictor C with k levels, we construct its $2 \times k$ contingency table with Y and compute the Pearson chi-squared statistic

$$\chi^2 \;=\; \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}},$$

where O_{ij} and $E_{ij} = n_{i\cdot} n_{\cdot j}/n$ are observed and expected counts. The null hypothesis of independence is rejected when $p < 0.05$. For those significant features, we compute Cramér's V as the

Table S53. Comparison Between SHAP and STATXAI (Numerical Features and Regression Outcome)

Model: Multi-layer Perceptron

Feature	Ground Truth	SHAP Rank	STAT-XAI Result
Annual Income	High	1	0.36 (Medium-Kept)
Credit Score	High	2	0.35 (Medium-Kept)
Debt-to-Income	High	3	0.27 (Medium-Kept)
Employment Length	None	4	-
Age	None	4	-

Model: Recurrent Neural Network

Feature	Ground Truth	SHAP Rank	STAT-XAI Result
Annual Income	High	1	0.36 (Medium- Kept)
Credit Score	High	2	0.34 (Medium- Kept)
Debt-to-Income	High	3	0.27 (Medium- Kept)
Employment Length	None	4	-
Age	None	5	-

standardized effect size,

$$V = \sqrt{\frac{\chi^2}{n\,(\,\min\{2,k\}-1)}}\,,$$

which lies in $[0,1]$ and measures association strength.

Finally, assemble all features with $p < 0.05$, measure their effect sizes ($|r_{pb}|$ for numerical, V for categorical), and sort them in descending order. This gives a unified, comparable importance list that highlights only those features with statistically validated associations, providing a transparent, interpretable feature ranking for downstream analysis.

German Credit Dataset

The German Credit dataset comprises 1,000 loan-applicant records and 20 original attributes.

Our objective is to deploy the STAT-XAI framework to uncover which features drive model predictions on this real-world dataset. Based on our synthetic-data experiments, where main effects dominates and pairwise interactions are negligible—we focus exclusively on main-effect inference in this section. Specifically, we fit three predictive model (MLP, RNN) to the German Credit data, compute STAT-XAI's inferential-statistics–based effect sizes for each feature, and rank them in descending order of importance. Our goal is to demonstrate STAT-XAI's generality across model classes using the relatively simple German Credit dataset. Although this dataset has only a moderate number of features, we apply STAT-XAI to distinct architectures—an MLP, an RNN —to show that our inferential-statistics–based feature-ranking procedure is model-agnostic. For each model, we train on the full dataset, compute statistical tests (e.g. one-way ANOVA for categorical and Pearson correlation for numerical features), extract effect sizes for features with $p < 0.05$, and rank them by descending magnitude. We then assess robustness by perturbing the data (adding Gaussian noise with $\sigma = 0.01 \times \text{std}$ and removing 5–10% of values MCAR), retraining each architecture, and recomputing the feature rankings. Agreement between original and perturbed top-k sets is quantified via the Jaccard index. Across all three model types, STAT-XAI produces stable, reproducible rankings, confirming its applicability and trustworthiness even in real-world settings.

For the German Credit data—composed of categorical predictors

$$\{\texttt{status}, \texttt{credit\_history}, \texttt{purpose}, \dots\}$$

Table S54. Performance Comparison Between SHAP and STAT-XAI for Fifth Synthetic Dataset

Model: Multi-layer Perceptron				
XAI Model	**Precision**	**Recall**	**FDR**	**Top 1 Rank**
SHAP	0.3	1.0	0.6	0
STAT-XAI	0.83	1.0	0.167	1
Model: Recurrent Neural Network				
XAI Model	**Precision**	**Recall**	**FDR**	**Top 1 Rank**
SHAP	0.5	1.0	0.5	0
STAT-XAI	0.83	1.0	0.167	0

and numerical predictors

$$\{\texttt{duration}, \texttt{amount}, \ldots, \texttt{Sex}\}$$

with binary outcome $Y \in \{0,1\}$ ("Predicted_Label")—we assess each feature's individual association with Y using classical inferential tests, followed by computation of standardized effect sizes. Features are then ordered by descending effect size to produce our STAT-XAI ranking.

Each continuous variable X is tested via the point-biserial correlation coefficient

$$r_{pb} = \frac{\bar{X}_1 - \bar{X}_0}{s_X} \sqrt{\frac{n_1 \, n_0}{n\,(n-1)}},$$

where $\bar{X}_1$ and $\bar{X}_0$ are the sample means of X in the two outcome groups ($Y = 1$ and $Y = 0$), s_X is the overall standard deviation of X, and n_1, n_0 are the group sizes ($n = n_1 + n_0$). We test the null hypothesis $r_{pb} = 0$ via the t-statistic

$$t = r_{pb} \sqrt{\frac{n-2}{1 - r_{pb}^2}},$$

which follows a Student's t distribution with $n - 2$ degrees of freedom. When $p < 0.05$, we record $|r_{pb}|$ as the numerical feature's effect size.

For each categorical predictor C with k levels, we construct its $2 \times k$ contingency table with Y and compute the Pearson chi-squared statistic

$$\chi^2 \;=\; \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}},$$

where O_{ij} and $E_{ij} = n_{i\cdot} n_{\cdot j} / n$ are observed and expected counts. The null hypothesis of independence is rejected when $p < 0.05$. For those significant features, we compute Cramér's V as the standardized effect size,

$$V \;=\; \sqrt{\frac{\chi^2}{n\,(\,\min\{2,k\} - 1)}},$$

which lies in $[0,1]$ and measures association strength.

Finally, assemble all features with $p < 0.05$, measure their effect sizes ($|r_{pb}|$ for numerical, V for categorical), and sort them in descending order. This gives a unified, comparable importance list that highlights only those features with statistically validated associations, providing a transparent, interpretable feature ranking for downstream analysis.

Table S58,S59 applies inferential tests to filter out any predictor whose association with the binary output fails to reach significance ($p \geq 0.05$), then computes standardized effect sizes for those that remain and ranks them in descending order.

In table S58: Of the 20 original features, only six pass the $p < 0.05$ threshold— status (p=0.0000, $V = 0.4246$), duration (p=0.0000, $r_{pb} = 0.2927$), savings (p=0.0044, $V = 0.2751$), credit_history

Table S55. Comparison Between SHAP and STATXAI (Numerical and Categorical Features and Binary Outcome)

Model: Multi-layer Perceptron

Feature	Ground Truth	SHAP Rank	STAT-XAI Result
Annual Income	High	10	0.43 (Large - Kept)
Credit Score	Medium	9	0.32 (Medium-Kept)
Employment Length	None	6	-
Age	None	5	-
Loan Term	None	7	0.05 (Low-Kept)
Employment Status	High	8	0.45 (Large-Kept)
Loan Categories	Medium	3	0.31 (Medium-Kept)
Loan Purpose	Low	1	0.20 (Medium-Kept)
Region	None	2	-
Marital Status	None	4	-

Model: Recurrent Neural Network

Feature	Ground Truth	SHAP Rank	STAT-XAI Result
Annual Income	High	10	0.42 (Large-Kept)
Credit Score	Medium	6	0.39 (Medium-Kept)
Employment Length	None	7	0.05 (Low-Kept)
Age	None	8	-
Loan Term	None	9	-
Employment Status	High	4	0.46 (Large-Kept)
Loan Categories	Medium	5	0.33 (Medium-Kept)
Loan Purpose	Low	1	0.11 (Low - Kept)
Region	None	2	-
Marital Status	None	3	-

Table S56. Performance Comparison Between SHAP and STAT-XAI for Sixth Synthetic Dataset

Model: Multi-layer Perceptron				
XAI Model	**Precision**	**Recall**	**FDR**	**Top 1 Rank**
SHAP	0.36	0.8	0.64	0
STAT-XAI	0.83	1.0	0.167	1
Model: Recurrent Neural Network				
XAI Model	**Precision**	**Recall**	**FDR**	**Top 1 Rank**
SHAP	0.36	0.8	0.64	0
STAT-XAI	0.83	1.0	0.167	0

(p=0.0203, $V = 0.2412$), housing (p=0.0071, $V = 0.2226$), and property (p=0.0294, $V = 0.2121$). The remaining 14 features are removed as non-significant, dramatically reducing the number of variables a user must consider.

In table S59: Only five predictors survive the $p < 0.05$ filter—other_installment_plans (p=0.0000, $V = 0.3818$), purpose (p=0.0058, $V = 0.3403$), job (p=0.0018, $V = 0.2743$), credit_history (p=0.0071, $V = 0.2650$), and other_debtors (p=0.0096, $V = 0.2156$), with all others dropped.

By distilling 20 raw attributes down to just 5–8 statistically validated drivers, STAT-XAI dramatically lowers cognitive load and focuses user attention on the truly relevant variables, fostering clearer, more trustworthy model explanations.

STAT-XAI framework, successfully reduced 20-feature down to fewer than 10 statistically validated predictors for each model—retaining only those features with $p < 0.05$ and meaningful effect sizes. By deleting irrelevant variables, STAT-XAI reduces users' cognitive burden and directs attention to the handful of true drivers behind model decisions. Moreover, our stability experiments—where Jaccard indices exceeded 0.7 across MLP, RNN architectures under Gaussian perturbations—demonstrate that these concise rankings remain consistent in the face of minor data noise. Together, the dramatic feature reduction and robust stability measurements affirm that STAT-XAI provides trustworthy explanations, making it a reliable tool for interpretable machine learning in real-world settings.

The Census Income Dataset

Census Income dataset aims to decide whether a person's annual income exceeds $ 50,000 based. The dataset consists of 48,842 instances and 15 attributes, comprising six numerical, seven categorical, and two binary attributes.

Our objective is to deploy the STAT-XAI framework to understand which features drive model predictions on this real-world dataset. Based on our synthetic-data experiments, where main effects dominates and pairwise interactions are negligible—we focus exclusively on main-effect inference in this section. Specifically, we fit three predictive model (MLP, RNN) to the Adult data, compute STAT-XAI's inferential-statistics–based effect sizes for each feature, and rank them in descending order of importance. Our goal is to demonstrate STAT-XAI's generality across different models. We apply STAT-XAI to distinct architectures—an MLP, and RNN to show that our inferential-statistics–based feature-ranking procedure is model-agnostic. For each model, we train on the full dataset, compute statistical tests (e.g. one-way ANOVA for categorical and Pearson correlation for numerical features), extract effect sizes for features with $p < 0.05$, and rank them by descending magnitude. We then assess robustness by perturbing the data (adding Gaussian noise with $\sigma = 0.01 \times \text{std}$), retraining each architecture, and recomputing the feature rankings. Agreement between original and perturbed top-k sets is quantified via the Jaccard index. Across all three model types, STAT-XAI produces stable, reproducible rankings, confirming its applicability and trustworthiness even in real-world settings.

For the Adult data— categorical features

$$\{\texttt{workclass}, \texttt{occupation}, \texttt{relationship}, \texttt{race}, \texttt{gender}, \texttt{native country}\}$$

Table S57. Comparison Between SHAP and STATXAI on Sixth Dataset (Numerical and Categorical Features and Regression Outcome)

Model: Multi-layer Perceptron

Feature	Ground Truth	SHAP Rank	STATXAI Result
Annual Income	High	10	0.0.33 (Medium - Kept)
Credit Score	Medium	9	0.15 (Low-Kept)
Employment Length	None	6	0.0021 (Low-Kept)
Age	None	5	-
Loan Term	None	7	-
Employment Status	High	8	0.30 (Medium-Kept)
Loan Categories	Medium	3	0.14 (Low-Kept)
Loan Purpose	Low	1	0.04 (Low-Kept)
Region	None	2	-
Marital Status	None	4	-

Model: Recurrent Neural Network

Feature	Ground Truth	SHAP Rank	STATXAI Result
Annual Income	High	10	0.57 (Large-Kept)
Credit Score	Medium	6	0.36 (Medium-Kept)
Employment Length	None	7	-
Age	None	8	-
Loan Term	None	9	-
Employment Status	High	4	0.35 (Medium-Kept)
Loan Categories	Medium	5	0.15 (Low-Kept)
Loan Purpose	Low	1	0.03 (Low-Kept)
Region	None	2	0.005 (Low-Kept)
Marital Status	None	3	-

Table S58. STAT-XAI Results on German Credit Dataset (Multi-layer Perceptron)

Feature	p-value	Effect size	Ranking
status	0.0000	0.4246	1
duration	0.0000	0.2927	2
savings	0.0044	0.2751	3
credit_history	0.0203	0.2412	4
housing	0.0071	0.2226	5
property	0.0294	0.2121	6
present_residence	0.7326	–	–
other_installment_plans	0.0876	–	–
other_debtors	0.3297	–	–
employment_duration	0.0904	–	–
purpose	0.0795	–	–
installment_rate	0.3551	–	–
amount	0.0761	–	–
Sex	0.7379	–	–
foreign_worker	0.6045	–	–
telephone	0.5542	–	–
people_liable	0.9248	–	–
number_credits	0.1578	–	–
age	0.1113	–	–
job	0.6261	–	–

and numerical features:

$$\{\texttt{age}, \texttt{fnlwgt}, \texttt{hours per week}\}$$

with binary outcome $Y \in \{0, 1\}$ ("Predicted_Label"), we assess each feature's individual association with Y using classical inferential tests, followed by computation of standardized effect sizes. Features are then ordered by descending effect size to produce our STAT-XAI ranking.

Each continuous variable X is tested via the point-biserial correlation coefficient

$$r_{pb} = \frac{\bar{X}_1 - \bar{X}_0}{s_X} \sqrt{\frac{n_1 \, n_0}{n \, (n-1)}},$$

where $\bar{X}_1$ and $\bar{X}_0$ are the sample means of X in the two outcome groups ($Y = 1$ and $Y = 0$), s_X is the overall standard deviation of X, and n_1, n_0 are the group sizes ($n = n_1 + n_0$). We test the null hypothesis $r_{pb} = 0$ via the t-statistic

$$t = r_{pb} \sqrt{\frac{n-2}{1 - r_{pb}^2}},$$

which follows a Student's t distribution with $n - 2$ degrees of freedom. When $p < 0.05$, we record $|r_{pb}|$ as the numerical feature's effect size.

For each categorical predictor C with k levels, we construct its $2 \times k$ contingency table with Y and compute the Pearson chi-squared statistic

$$\chi^2 \; = \; \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}},$$

Table S59. STAT-XAI Results on German Credit Dataset (Recurrent Neural Network)

Feature	p-value	Effect size	Ranking
other_installment_plans	0.0000	0.3818	1
purpose	0.0058	0.3403	2
job	0.0018	0.2743	3
credit_history	0.0071	0.2650	4
other_debtors	0.0096	0.2156	5
amount	0.3334	–	–
housing	0.6438	–	–
property	0.3386	–	–
employment_duration	0.4981	–	–
savings	0.2044	–	–
duration	0.8557	–	–
Sex	0.4718	–	–
foreign_worker	0.5915	–	–
telephone	0.2120	–	–
people_liable	0.7226	–	–
number_credits	0.3471	–	–
age	0.3042	–	–
present_residence	0.7480	–	–
installment_rate	0.0905	–	–
status	0.9016	–	–

where O_{ij} and $E_{ij} = n_{i\cdot} n_{\cdot j}/n$ are observed and expected counts. The null hypothesis of independence is rejected when $p < 0.05$. For those significant features, we compute Cramér's V as the standardized effect size,

$$V = \sqrt{\frac{\chi^2}{n\,(\,\min\{2,k\} - 1)}}\,,$$

which lies in $[0, 1]$ and measures association strength.

Finally, assemble all features with $p < 0.05$, measure their effect sizes ($|r_{pb}|$ for numerical, V for categorical), and sort them in descending order. This gives a unified, comparable importance list that highlights only those features with statistically validated associations, providing a transparent, interpretable feature ranking for downstream analysis.

Tables S60,S61 reports STAT-XAI's main-effect results on the Adult dataset for three model architectures. In each case we: i) Test each of the 12 original predictors against the binary income label, ii) Retain only those with $p < 0.05$, iii) Compute a standardized effect size (η^2 for categorical, $|r_{pb}|$ for numerical), and finally Rank the retained features by descending effect size.

In table S60 for MLP architecture, ten of the twelve features meet the significance threshold. Relationship (0.5010) and marital-status (0.4944) emerge as the strongest drivers, followed by education (0.4594), occupation (0.4020), and so on down to race (0.1016). Only fnlwgt (p=0.0683) is dropped.

In table S61 for RNN architecture, again ten features are retained. The ordering is nearly identical— relationship (0.5099) and marital-status (0.5038) top the list, with career-related factors like education (0.5013) and occupation (0.4640) following, down to race (0.0868). fnlwgt is again non-significant.

Table S60. STAT-XAI Results on Adult Dataset (Multi-layer Perceptron)

Feature	p-value	Effect size	Ranking
relationship	0.0000	0.5010	1
marital-status	0.0000	0.4944	2
education	0.0000	0.4594	3
occupation	0.0000	0.4020	4
age	0.0000	0.2538	5
gender	0.0000	0.2268	6
hours_per_week	0.0000	0.2133	7
workclass	0.0000	0.1902	8
native-country	0.0000	0.1098	9
race	0.0000	0.1016	10
fnlwgt	0.0683	-	-

By dropping out non-significant features (reducing from twelve to ten) and presenting only those with validated, ordered effect sizes, STAT-XAI helps the user's to focus on the handful of truly influential features. This decreases cognitive load, enhances interpretability, and builds user trust.

Table S61. STAT-XAI Results on Adult Dataset (RNN)

Feature	p-value	Effect size	Ranking
relationship	0.0000	0.5099	1
marital-status	0.0000	0.5038	2
education	0.0000	0.5013	3
occupation	0.0000	0.4640	4
age	0.0000	0.2608	5
hours_per_week	0.0000	0.2233	6
gender	0.0000	0.2174	7
workclass	0.0000	0.2080	8
native-country	0.0000	0.1024	9
race	0.0000	0.0868	10
fnlwgt	0.6368	-	-