

Joint Estimation of Credit Risk in Consumer Lending: A Multi-Gate Mixture-of-Experts Approach

Wenlu Zhang

wenlu.zhang@csulb.edu

California State University, Long Beach

Fang Fang

California State University Los Angeles

Lusi Li

California State University Los Angeles

Liuliu Fu

California State University Los Angeles

Research Article

Keywords: Credit risk, consumer lending, expected-loss calibration, multitask learning, mixture-of-experts

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11066647/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Joint Estimation of Credit Risk in Consumer Lending: A Multi-Gate Mixture-of-Experts Approach

Abstract

Credit risk management requires estimates of probability of default (PD) and loss given default (LGD) to calculate expected loss. Standard approaches estimate these components separately, potentially underutilizing shared borrower information and allowing component errors to affect the combined estimate. This study investigates whether joint estimation with selective parameter sharing improves expected-loss calibration. We develop a multitask framework based on the Multi-Gate Mixture-of-Experts (MMoE) architecture and evaluate it using Lending Club data. MMoE uses a shared pool of expert networks and task-specific gates, allowing PD and LGD to share information selectively. The multitask models substantially outperform conventional sequential approaches in expected-loss calibration. Compared with the hard-sharing benchmark, MMoE further reduces expected-loss calibration root mean squared error (EL-Cal RMSE) by 9.0%. These improvements are concentrated in the middle and high risk borrower groups. Learned gate weights also differ systematically across tasks and credit-risk segments, providing descriptive evidence of selective expert utilization.

Keywords: Credit risk, consumer lending, expected-loss calibration, multitask learning, mixture-of-experts

1 Introduction

Consumer lending has grown rapidly over the past several decades. As of April 2026, total U.S. consumer credit outstanding reached approximately \$5.10 trillion (Board of Governors of the Federal Reserve System, 2026). In a data-rich lending environment, lenders rely heavily on statistical and machine-learning models to evaluate borrower credit quality, price risk, and manage portfolio losses. A central challenge is to estimate not only whether a borrower will default, but also how much will be lost if default occurs. These two quantities are commonly referred to as probability of default (PD) and loss given default (LGD). Together with exposure at default, they determine expected loss (EL), which is a key input in credit-risk management, regulatory capital assessment, provisioning, and risk-based pricing (Basel Committee on Banking Supervision, 2006).

In the literature, PD and LGD are typically estimated using separate statistical or machine-learning models, including logistic regression, linear regression, and gradient-boosted trees (Lessmann et al, 2015; Loterman et al, 2012; Chen and Guestrin, 2016). Expected loss is then calculated by multiplying the predicted components. Although this sequential approach is straightforward, it treats PD and LGD as independent modeling problems even though the two outcomes may depend on overlapping borrower and loan characteristics. Information related to affordability, leverage, repayment burden, and credit history may therefore be underutilized when the tasks are estimated

separately. In addition, because EL is formed from the product of predicted PD and LGD, errors in the two component models may compound in the final expected-loss estimate. These limitations motivate a joint modeling framework that can exploit shared information while preserving differences between the two prediction tasks.

This study examines whether joint estimation with selective parameter sharing improves expected-loss calibration. We develop a multitask framework based on the Multi-Gate Mixture-of-Experts (MMoE) architecture and compare it with three baseline models. The baselines consist of sequential XGBoost and LightGBM models and a hard-sharing multitask model. A key challenge in LGD estimation is right censoring caused by incomplete observation periods. To address this issue, we use the public Lending Club loan dataset and restrict the sample to loans that had reached contractual maturity by the observation date. We evaluate performance using separate metrics for PD and LGD, together with expected loss calibration, which measures the agreement between predicted and realized loss rates across the portfolio. The joint estimation models substantially reduce EL-Cal RMSE relative to the sequential baselines. MMoE further reduces EL-Cal RMSE by approximately 9.0% relative to the hard-sharing benchmark. Its advantage is concentrated in the middle and high risk loan segments, whereas hard sharing performs slightly better among the safest borrowers. Gate weights vary systematically across tasks and borrower segments, showing that the MMoE model uses different experts mixtures for different prediction tasks and borrower groups.

This study makes three contributions. First, it centers the empirical evaluation of joint PD and LGD models on the calibration of the combined expected-loss estimate rather than relying only on separate component metrics. Second, it adapts the MMoE architecture to joint PD-LGD estimation, enabling task-specific routing through a

shared pool of experts. Third, it provides portfolio-level and segment-level evidence on when selective routing yields better expected-loss calibration than the hard-sharing multitask model.

The remainder of the paper is organized as follows. Section 2 reviews the related literature and identifies the research gap. Section 3 describes the data, target construction, model architecture, baselines, and evaluation design. Section 4 reports the main results, segment-level analyses, gate-weight diagnostics, expert-count robustness, and economic interpretation. Section 5 concludes, discusses limitations, and outlines directions for future research.

2 Literature Review

This paper connects the consumer credit-risk literature with research on multitask learning and mixture-of-experts architectures. The following subsections synthesize the most relevant work and identify the gap addressed by the present study.

2.1 *Credit Risk Estimation*

Probability-of-default modeling in consumer lending has long been studied in the credit-scoring literature. Early work relied on linear discriminant analysis and logistic regression because these methods are interpretable and relatively easy to implement in retail credit settings (Altman, 1968; Thomas et al, 2002). More recently, tree-based ensemble methods such as XGBoost (Chen and Guestrin, 2016) and LightGBM (Ke

et al, 2017) have become strong empirical benchmarks because they capture non-linearities, handle missing values, and perform well on heterogeneous tabular data. Accordingly, boosted trees are widely used as benchmark models for PD prediction (Lessmann et al, 2015). Recent studies have also combined machine-learning credit scoring with explainability methods. For example, Bussmann et al (2021) apply SHAP-based explainable machine learning to peer-to-peer credit risk assessment and show that post-hoc interpretability tools can identify economically meaningful drivers of default risk. Using Lending Club data, Monje et al (2025) evaluate machine-learning and explainable-AI methods for early loan default prediction. These studies demonstrate the value of machine learning and interpretability tools for consumer-credit PD modeling. Loss given default is commonly modeled as a regression problem conditional on default (Loterman et al, 2012). A major challenge in LGD modeling is right-censoring, because many loans remain active and their recovery outcomes are not yet fully observed. For this reason, our study combines terminal-status filtering with a contractual-maturity restriction when constructing the LGD target. Deep learning methods have also been applied to credit risk, including convolutional models for mortgage default prediction (Kvamme et al, 2018). Despite these advances, prior studies generally optimize PD and LGD independently and evaluate them using task-specific metrics. Relatively little attention has been paid to whether joint estimation can improve the calibration of the combined expected-loss estimate, which is ultimately the quantity used in many lending decisions.

2.2 Multitask Learning and Multi-Gate Mixture-of-Experts

Multitask learning is based on the idea that related tasks can benefit from shared representations (Caruana, 1997; Zhang and Yang, 2022). A classic implementation is

hard parameter sharing, in which several tasks share the same backbone network while using separate output heads. In finance, multitask learning has been applied to related prediction problems such as the joint modeling of mortgage delinquency, foreclosure, and prepayment risk (Sadhvani et al, 2021). Although hard-sharing architectures are computationally efficient, they can be restrictive when tasks are only partially related or generate conflicting optimization signals (Yu et al, 2020).

To relax this restriction, the multitask-learning literature has proposed more flexible forms of parameter sharing. Examples include cross-stitch networks (Misra et al, 2016) and attention-based multitask models (Liu et al, 2019), which allow tasks to share information selectively rather than through a single fully common representation. Another related architecture is the MMoE model, in which multiple tasks share a common pool of experts but use separate gating networks (Ma et al, 2018). Each task therefore learns its own weighted combination of experts, creating a form of selective sharing that lies between fully separate models and fully shared models. This flexibility makes MMoE well suited to credit-risk estimation, where PD and LGD are related but not identical learning problems. Nevertheless, whether selective multitask learning improves expected-loss calibration beyond both sequential estimation and hard-sharing multitask learning remains largely unexplored.

3 Methods

3.1 *Data*

The analysis is based on public Lending Club loan-level records covering loans originated between 2013 and 2016. We construct the estimation sample in two stages. First,

we retain only loans with a terminal status of either *Fully Paid* or *Charged Off*, thereby excluding loans that were still active or otherwise unresolved at the time of the data observation. Second, we apply a maturity filter and retain only loans whose contractual maturity date falls on or before the observation date. The contractual maturity date is defined as the loan issue date plus the original loan term. This restriction ensures that every loan included in the analysis had reached the end of its scheduled term by the cutoff date.

The maturity restriction is important because terminal status alone does not make repayment and recovery histories comparable across origination cohorts. For example, a recently originated loan that defaults after only a few payments has had substantially less time to repay principal or generate post-default recoveries than an otherwise similar loan observed over its full contractual term. Including such loans may therefore place disproportionate weight on early defaults with high observed loss rates and distort the empirical distribution of LGD. The maturity filter creates more comparable contractual observation windows across origination cohorts and reduces the right-censoring problem associated with incompletely observed repayment and recovery histories.

To preserve the temporal ordering that a lender would face in practice, we divide the sample chronologically. Loans originated in 2013 and 2014 form the training set, loans originated in 2015 form the validation set, and loans originated in January and February 2016 form the test set. After applying the terminal-status and maturity restrictions, the final sample contains 305,755 training loans, 283,026 validation loans, and 50,695 test loans. The default rates are 14.88%, 14.89%, and 14.84% in the training, validation, and test sets, respectively. The corresponding average LGD values are

53.39%, 52.09%, and 54.76%. The relative stability of these outcome distributions suggests that the maturity restriction reduces differences in cohort composition caused by incomplete loan resolution.

To prevent information leakage, we retain only features that would have been observable at the time of loan origination. The final feature set comprises 26 variables capturing borrower characteristics, contractual loan terms, and credit history. These variables include FICO score, annual income, employment length, home-ownership status, delinquency history, and other origination-time measures.

We define the target variables as follows. For the PD target, the default indicator equals 1 for Charged Off loans and 0 for Fully Paid loans. For defaulted loans, the LGD target is defined as

$$LGD_i = 1 - \frac{\text{total_rec_prncp}_i + \text{recoveries}_i - \text{collection_recovery_fee}_i}{\text{funded_amnt}_i}, \quad (1)$$

where *total_rec_prncp* is cumulative principal recovered during the loan term, *recoveries* are payments collected after the loan has been written off, and *collection_recovery_fee* is the cost of recovery collection. Equation (1) therefore captures the net economic loss rate on defaulted loans. LGD is undefined for non-defaulted loans. Therefore we assign these observations a sentinel value of -1 and exclude them from the LGD objective during training.

Because exposure at default is not directly available in the Lending Club dataset, this study evaluates the expected-loss rate rather than the expected-loss amount. The

predicted expected-loss rate for loan i is defined as

$$\widehat{ELRate}_i = \widehat{PD}_i \times \widehat{LGD}_i. \quad (2)$$

3.2 Architecture

We use an MMoE architecture to jointly estimate PD and LGD. Figure 1 illustrates the overall framework.

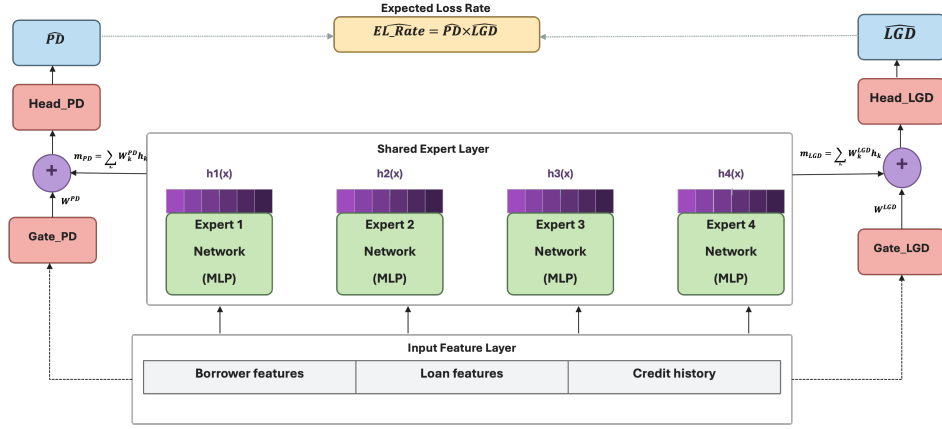


Fig. 1 Multi-Gate Mixture-of-Experts architecture for joint credit-risk estimation

Let x denote the feature vector observed at loan origination. The model consists of K expert networks, each of which maps the input vector x to a latent representation $h_k(x)$. All experts share the same architecture but have independent parameters. Specifically, each expert is implemented as a multilayer perceptron with two hidden layers of sizes 128 and 64. Each hidden layer uses batch normalization, ReLU activation, and dropout with a rate of 0.2. For each task $t \in \{PD, LGD\}$, a task-specific

gating network computes a soft weighting over the shared experts:

$$w^t(x) = \text{softmax}(W^t x + b^t), \quad (3)$$

where W^t and b^t are the gating parameters for task t . The gate weights determine how much each expert contributes to the representation for task t . The task-specific mixture representation is then given by

$$m^t(x) = \sum_{k=1}^K w_k^t(x) h_k(x). \quad (4)$$

Each task is assigned a dedicated prediction head with a sigmoid activation. The PD head produces a default probability, while the LGD head produces a bounded loss-severity estimate.

Because PD and LGD are optimized jointly, each task contributes its own loss term. The PD objective is binary cross-entropy:

$$\mathcal{L}_{PD} = \text{BCE}(\hat{y}^{PD}, y^{PD}). \quad (5)$$

The LGD objective is mean squared error evaluated exclusively on the defaulted subset:

$$\mathcal{L}_{LGD} = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} (\hat{y}_i^{LGD} - y_i^{LGD})^2, \quad (6)$$

where $\mathcal{D}$ denotes the set of observations with realized default.

Rather than choosing task weights manually, we combine the two objectives using the homoscedastic uncertainty-weighting approach of Kendall et al (2018):

$$\mathcal{L} = \exp(-s_{PD})\mathcal{L}_{PD} + \exp(-s_{LGD})\mathcal{L}_{LGD} + s_{PD} + s_{LGD}, \quad (7)$$

where s_{PD} and s_{LGD} are learned log-variance parameters. This formulation allows the model to adaptively balance the contributions of the two objectives during training.

3.3 *Baselines and Evaluation Metrics*

For PD, we report AUC-ROC. For LGD, we report mean absolute error (MAE). Beyond these task-specific metrics, we evaluate the quality of the joint expected-loss estimate using a calibration metric, EL-Cal RMSE, defined as

$$\text{EL-Cal RMSE} = \sqrt{\frac{1}{B} \sum_{b=1}^B \left(\bar{r}_b^{pred} - \bar{r}_b^{real} \right)^2}. \quad (8)$$

Loans are first ranked by predicted expected-loss rate and partitioned into $B = 10$ decile buckets. For each bucket b , we compute the mean predicted EL rate and the mean realized EL rate. The realized EL rate for loan i equals the observed net loss divided by funded amount. Lower EL-Cal RMSE indicates closer agreement between predicted and realized loss rates across the risk distribution.

We construct three baselines for comparison. The first is a sequential XGBoost benchmark, in which an XGBoost classifier estimates PD using the full sample and a separate XGBoost regressor estimates LGD using only defaulted loans. The second is

a sequential LightGBM benchmark constructed in the same manner. The third is a hard-sharing multitask model with a shared multilayer perceptron trunk containing hidden layers of sizes 256, 128, and 64, followed by task-specific heads for PD and LGD. For the MMoE model, each expert is a two-layer multilayer perceptron with hidden-layer sizes 128 and 64, and four experts. Separate softmax gating networks determine the expert mixture for the PD and LGD tasks. All neural models are trained using the Adam optimizer with a learning rate of 10^{-3} , weight decay of 10^{-4} , and a batch size of 512. Training uses early stopping with a maximum of 100 epochs, and task-loss weights are learned using the homoscedastic uncertainty approach of Kendall et al (2018). The tree-based models use standard gradient-boosting configurations, with the same specification applied separately to the PD classifier and LGD regressor. Detailed hyperparameter settings are reported in Appendix A.

4 Results

4.1 Main Performance Comparison

Table 1 Out-of-sample performance of sequential and multitask models

Model	AUC-ROC	LGD MAE	EL-Cal RMSE
XGBoost (PD + LGD, sequential)	0.6971	0.1878	0.1114
LightGBM (PD + LGD, sequential)	0.6904	0.1901	0.1016
Hard-sharing MTL	0.6977	0.1841	0.0222
MMoE ($K = 4$)	0.7006	0.1832	0.0202

Table 1 compares the out-of-sample performance of the sequential and multitask models. Performance is broadly similar on the individual PD and LGD metrics. PD AUC ranges from 0.6904 to 0.7006, while LGD MAE ranges from 0.1832 to 0.1901. MMoE

achieves the highest PD AUC and the lowest LGD MAE, although its advantages on these component metrics are modest. The differences in expected-loss calibration are much larger. The sequential XGBoost and LightGBM models produce EL-Cal RMSE values of 0.1114 and 0.1016, respectively. The hard-sharing model reduces EL-Cal RMSE to 0.0222, and MMoE lowers it further to 0.0202. These results show that similar performance on the individual components does not imply similar calibration of the combined expected loss estimate. A model may rank borrowers accurately by default risk and estimate LGD with a competitive error while still producing a poorly calibrated product of predicted PD and LGD. The comparison between the two multitask models demonstrates the additional value of selective expert routing. MMoE reduces EL-Cal RMSE by approximately 9.0% relative to hard sharing. Overall, the results indicate that joint estimation accounts for most of the improvement in expected loss calibration, while task-specific expert routing provides a further gain.

4.2 Credit Quality Analysis by Lending Club Grade

The portfolio-level EL-Cal RMSE reported in Table 1 summarizes calibration performance across the full test sample. However, this aggregate measure may hide differences across borrower groups. We therefore examine model performance within groups defined by Lending Club’s internally assigned loan grades. These grades reflect the platform’s overall assessment of borrower and loan risk.

We divide the test sample into three groups. The low risk group contains grades A–B, the middle risk group contains grades C–D, and the high risk group contains grades E–G. For each group, we report PD AUC, LGD MAE, and EL-Cal RMSE. To calculate EL-Cal RMSE, loans are ranked and divided into five equal-frequency buckets within

each grade group. This procedure measures calibration among loans with relatively similar credit quality and limits the influence of differences in grade composition. Table 2 and Figure 2 report the results.

Table 2 Test set performance by Lending Club grade group

Risk group	XGBoost	LightGBM	Hard-sharing MTL	MMoE ($K = 4$)
<i>EL-Cal RMSE</i>				
Low risk (A–B)	0.0886	0.0799	0.0098	0.0103
Middle risk (C–D)	0.1397	0.1293	0.0322	0.0263
High risk (E–G)	0.1461	0.1343	0.0665	0.0642
<i>PD AUC</i>				
Low risk (A–B)	0.6468	0.6401	0.6494	0.6505
Middle risk (C–D)	0.5950	0.5870	0.5931	0.5989
High risk (E–G)	0.6051	0.5944	0.5998	0.5995
<i>LGD MAE</i>				
Low risk (A–B)	0.1898	0.1915	0.1882	0.1872
Middle risk (C–D)	0.1866	0.1894	0.1825	0.1811
High risk (E–G)	0.1864	0.1883	0.1769	0.1783

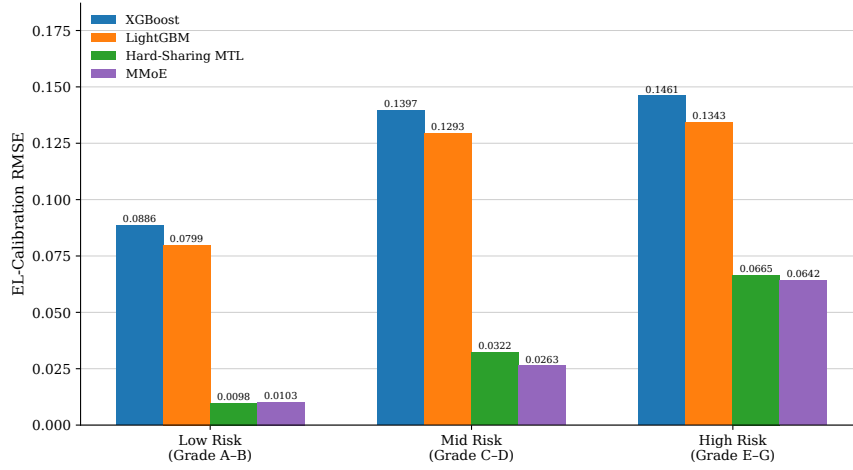


Fig. 2 EL-Cal RMSE by Lending Club grade group

The results show that both multitask models achieve substantially better expected loss calibration than the sequential models in every grade group. EL-Cal RMSE ranges

from 0.0799 to 0.1461 for the sequential models and from 0.0098 to 0.0665 for the multitask models. The calibration advantage of joint estimation therefore remains after loans are compared within narrower grade groups. The relative performance of MMoE and hard sharing varies across grade groups. In the low risk group, hard sharing produces a slightly lower EL-Cal RMSE than MMoE, with values of 0.0098 and 0.0103, respectively. In the middle risk group, MMoE reduces EL-Cal RMSE from 0.0322 to 0.0263, which represents an improvement of approximately 18.3%. In the high risk group, MMoE reduces EL-Cal RMSE from 0.0665 to 0.0642, an improvement of approximately 3.5%. The additional calibration benefit of selective expert routing is therefore largest in the middle risk group and remains positive but smaller in the high risk group.

The component metrics do not produce a uniform model ranking. MMoE achieves the highest PD AUC and the lowest LGD MAE in the low and middle risk groups. In the high risk group, XGBoost achieves the highest PD AUC, hard sharing produces the lowest LGD MAE, and MMoE provides the lowest EL-Cal RMSE. These differences show that strong performance on an individual component does not necessarily lead to the best calibration of the combined expected loss estimate.

4.3 Gate Weight Analysis

The analysis by Lending Club grade shows that the relative performance of MMoE and hard sharing varies across credit-quality groups. To better understand how the fitted MMoE uses the shared experts differently across tasks, we examine the task-specific gate weights. We first describe their overall distributions, then relate each expert’s weight to the input features to characterize the routing profiles of the PD

and LGD gates, and finally examine how these routing patterns vary across Lending Club grades.

4.3.1 Gate Weight Distributions

Figure 3 presents the distribution of gate weights by task and expert across the full test set. For the PD task, the four experts have broadly similar weight distributions, and no single expert dominates across most loans. In contrast, the LGD weights are more concentrated among a smaller subset of experts, indicating more selective routing. Thus, the two task-specific gates use the shared expert pool differently at the aggregate level. The PD gate tends to blend information from several experts, whereas the LGD gate places greater emphasis on particular experts. The next section examines which borrower and loan characteristics are associated with these routing patterns.

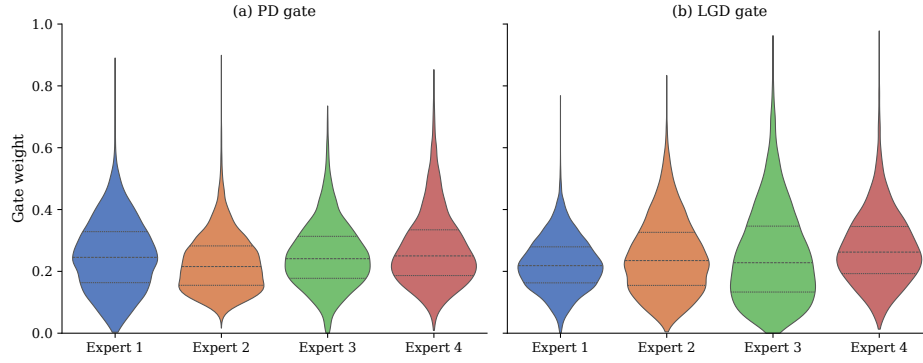


Fig. 3 Distribution of MMoE gate weights by task and expert across the test set

4.3.2 Routing Profile Analysis

To characterize how each gate allocates loans across experts, we calculate the Pearson correlation between each expert’s gate weight and each input feature across the test set. Figures 4 and 5 report the results for the PD and LGD gates. Because the gating network in Equation (3) applies a softmax function to a linear combination of the input features, these correlations describe routing patterns rather than the internal representations produced by the expert networks.

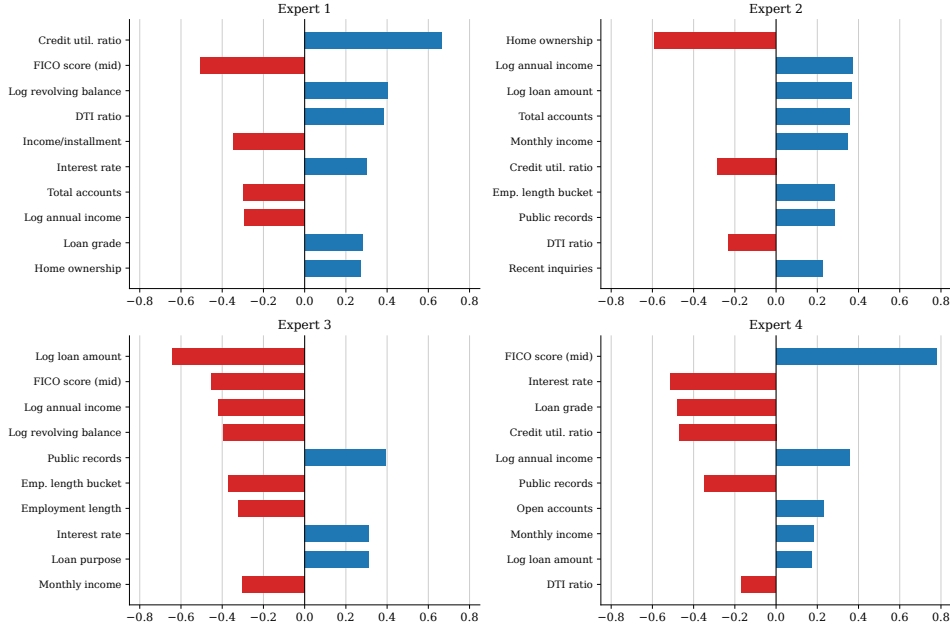


Fig. 4 Pearson correlations between PD gate weights and input features

For the PD gate, the routing profiles are organized primarily along a credit-quality dimension, with the clearest contrast appearing between Experts 4 and 1. The weight assigned to Expert 4 exhibits a high-credit-quality profile. It is positively correlated with FICO score and negatively correlated with interest rate, loan grade, and credit

utilization. By contrast, the weight assigned to Expert 1 exhibits a credit-stress profile, increasing with credit utilization, log revolving balance, and the debt-to-income ratio, and decreasing with FICO score and income per installment. Expert 3 is associated with a weaker-borrower, smaller-loan profile. Its weight is negatively correlated with log loan amount, FICO score, and log annual income, and positively correlated with public records, interest rate, and loan purpose. Expert 2 is less directly aligned with credit quality. The weight is associated primarily with income, loan size, and account depth, showing positive correlations with log annual income, log loan amount, and total accounts, and negative correlations with home ownership and credit utilization.

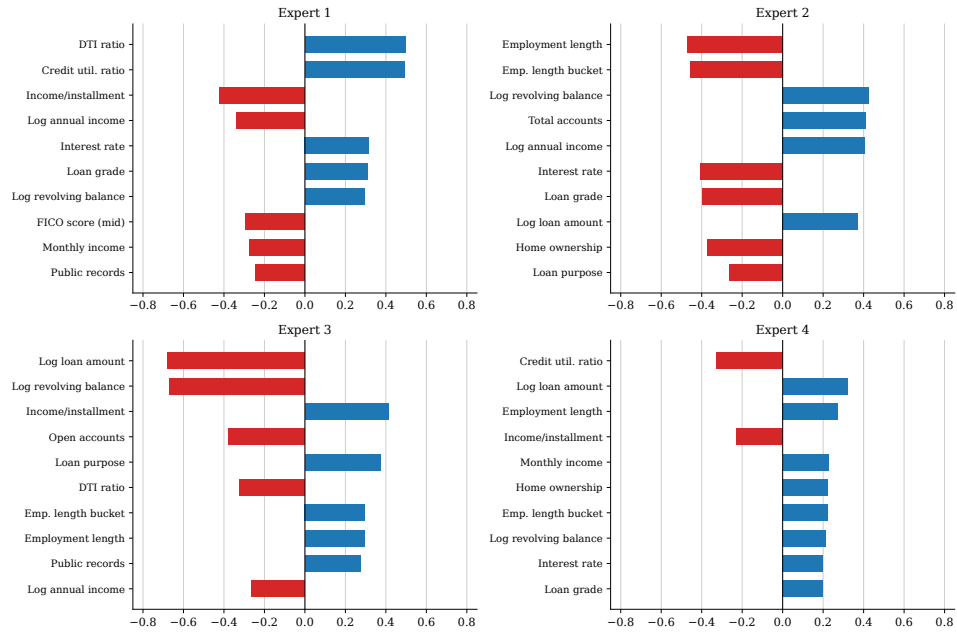


Fig. 5 Pearson correlations between LGD gate weights and input features

For the LGD gate, the routing profiles combine credit-risk information with variables more directly related to loss severity, including exposure size, revolving balances, repayment capacity, and employment history. Expert 1 again carries a credit-stress

profile, with weight positively correlated with DTI ratio, credit utilization, and interest rate, and negatively correlated with income per installment, log annual income, and FICO score. Expert 2 is associated with larger balances and greater account depth, Expert 3 with smaller loan and revolving balances and higher income per installment, and Expert 4 with loan size and employment history combined with lower utilization.

Taken together, the two sets of correlations provide descriptive evidence that the PD and LGD gates use the same expert pool but organize the shared representations differently. The PD gate distinguishes borrowers primarily along a credit-quality dimension. The LGD gate retains some of the same risk information, most visibly in the routing profile associated with Expert 1, but its routing patterns are more strongly associated with exposure size, balances, repayment capacity, and employment characteristics, the variables closely related to loss severity conditional on default. This task-specific mixing distinguishes MMoE from the hard-sharing model, in which both tasks must rely on a single shared representation.

4.3.3 Routing by Lending Club Grade

Figure 6 reports mean gate weights by Lending Club grade. The figure uses five grade groups, A, B, C, D, and E–G, to show how routing changes across the credit-quality distribution.

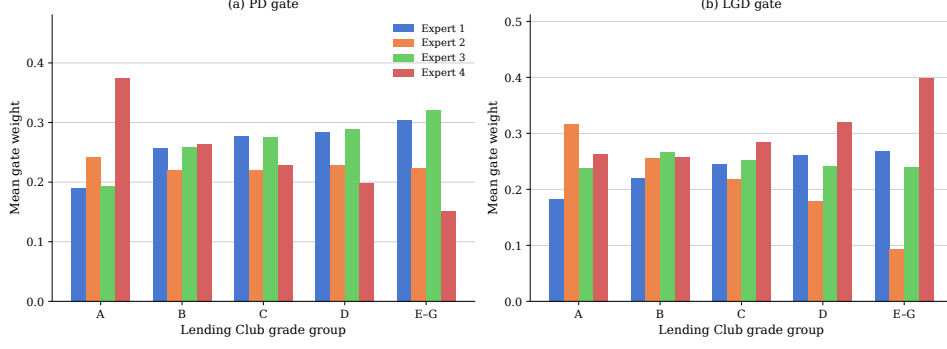


Fig. 6 Mean gate weights by Lending Club grade for the PD and LGD tasks

For the PD task, routing changes in the direction indicated by the feature-level profiles in Section 4.3.2. The weight assigned to Expert 4, which exhibits the clearest high-credit-quality routing profile, is highest among grade A loans at approximately 0.38 and declines steadily to about 0.15 for grades E–G. By contrast, the weights assigned to Experts 1 and 3, whose routing profiles are associated with credit stress and weaker borrower characteristics, increase as grade deteriorates, rising from approximately 0.19 each in grade A to about 0.31 and 0.32 in grades E–G. Expert 2 remains comparatively stable across grades, with a mean weight of approximately 0.22–0.24.

The LGD gate exhibits a different pattern of reallocation. Expert 2 receives the largest mean weight among grade A loans, at approximately 0.32, but its weight declines sharply as credit quality deteriorates, reaching about 0.09 in grades E–G. Expert 4 moves in the opposite direction, with its mean weight increasing from approximately 0.26 in grade A to about 0.40 in grades E–G, making it the most heavily weighted LGD expert in the highest-risk grade group. Expert 1 increases more moderately across grades, whereas Expert 3 remains relatively stable at approximately 0.24–0.27. These grade-level patterns corroborate the feature-level evidence that the two task-specific gates do not simply mirror one another. As credit quality deteriorates, the PD gate

reallocates weight primarily toward Experts 1 and 3, whereas the LGD gate reallocates weight primarily from Expert 2 toward Expert 4.

Taken together, the gate-weight results show that MMoE makes substantive use of its task-specific routing mechanism. The overall weight distributions indicate that the gates do not collapse onto a single expert, the feature correlations reveal distinct routing profiles for PD and LGD, and the grade-level analysis shows that these profiles vary systematically with borrower credit quality. These patterns provide descriptive context for the performance differences reported in Table 2. In the middle and high risk segments, the PD and LGD gates form different expert mixtures, and MMoE reduces EL-Cal RMSE relative to hard sharing by approximately 18.3% and 3.5%, respectively. In the low-risk segment, although the PD and LGD gates assign their largest mean weights to different experts, realized loss rates are relatively low and homogeneous. This limited task-specific heterogeneity offers less scope for selective routing to improve calibration, which is consistent with the slightly lower EL-Cal RMSE achieved by hard sharing among the safest borrowers.

4.4 Expert-Count Robustness

To assess the sensitivity of results to the number of experts, we train 2-task MMoE models with $K \in \{2, 4, 6, 8\}$ and report results alongside the hard-sharing MTL baseline. Table 3 reports the results.

The results show that the main conclusion is not tied to a single expert count, but additional experts do not improve performance monotonically. The $K = 4$ model achieves the best value for each reported metric. The $K = 2$ and $K = 6$ specifications

Table 3 Performance metrics across expert counts

Model	AUC-ROC	LGD MAE	EL-Cal RMSE
Hard-sharing MTL	0.6977	0.1841	0.0222
MMoE $K = 2$	0.6986	0.1846	0.0218
MMoE $K = 4$	0.7006	0.1832	0.0202
MMoE $K = 6$	0.6993	0.1843	0.0202
MMoE $K = 8$	0.7003	0.1843	0.0263

also outperform hard sharing on EL-Cal RMSE, whereas the $K = 8$ model performs worse, suggesting that excess capacity can make the routing problem less stable. This pattern indicates that a moderately sized expert pool is sufficient for the present data, and the ablation therefore supports $K = 4$ as the preferred empirical specification.

4.5 *Economic Interpretation*

The empirical results have two important implications for credit-risk practice. First, model selection based only on component-level performance can be suboptimal when the operational target is expected loss. The sequential tree-based models remain competitive in terms of PD AUC and LGD MAE, but their substantially higher EL-Cal RMSE shows that strong performance on the individual components does not necessarily translate into an accurate expected-loss estimate. By estimating PD and LGD jointly, the multitask models produce more coherent component predictions and substantially improve the calibration of the combined expected-loss estimate. Second, selective parameter sharing provides additional value beyond joint estimation. At the portfolio level, MMoE reduces EL-Cal RMSE by 9.0% relative to hard sharing. Its advantage is particularly evident in the middle and high risk borrower segments, where accurate loss assessment is especially important for risk pricing, reserve planning, and portfolio monitoring. These results indicate that the additional benefit of MMoE is

associated with task-specific routing rather than joint estimation alone. By allowing PD and LGD to draw on different combinations of shared representations, MMoE can accommodate variation in the relationship between default occurrence and loss severity across the credit-risk distribution.

MMoE provides a flexible framework for modeling two related but distinct credit-risk outcomes. PD and LGD are influenced by overlapping borrower and loan characteristics, but the relevance of this information may differ between default occurrence and loss severity. MMoE captures this structure by combining a shared pool of expert networks with task-specific routing. The observed differences in gate weights across tasks and credit-quality groups show that the model adjusts its use of shared information across the risk distribution. This capacity to represent both common and task-specific relationships is central to the economic relevance and empirical performance of the MMoE framework.

5 Conclusion and Discussion

This paper examines whether PD and LGD can be estimated more effectively as related tasks when the primary objective is expected-loss calibration. Using maturity-filtered Lending Club data and a chronological train-validation-test design, we compare sequential tree-based models, a hard-sharing multitask network, and an MMoE architecture with task-specific gates. The results show that joint estimation substantially improves expected-loss calibration relative to separate PD and LGD modeling. MMoE provides a further 9.0% reduction in EL-Cal RMSE relative to hard sharing, with its strongest advantages observed outside the lowest-risk borrower groups.

These findings highlight the importance of aligning model development and evaluation with the final decision objective. Models with similar PD AUC and LGD MAE can produce substantially different expected-loss calibration, indicating that component-level metrics alone do not provide a complete assessment of credit-risk performance. The gate-weight patterns further indicate that the information shared between PD and LGD varies across tasks and credit-quality groups. By allowing each task to use a different combination of shared representations, MMoE provides a flexible framework for capturing this variation and improving expected-loss assessment across heterogeneous loan portfolios.

The study can be extended in several ways. Future studies could evaluate the framework using data from other lenders, more recent origination cohorts, and periods of greater macroeconomic stress. Further research could also compare MMoE with alternative selective-sharing architectures, including cross-stitch networks and task-conditioned attention. These extensions would help establish the broader applicability of joint PD–LGD estimation across different lending environments and identify the conditions under which flexible parameter sharing provides the greatest value.

Appendix A Model Hyperparameters

Table A1 summarizes the hyperparameter configurations used for the neural and tree-based models. For XGBoost and LightGBM, the same specification is applied separately to the PD classifier and the LGD regressor. Neural-model settings are held fixed.

Table A1 Hyperparameter configurations for all models.

Component	Configuration
<i>MMoE 2-task</i>	
Expert networks	2-layer MLP: [128, 64], BatchNorm, ReLU, Dropout 0.2
Number of experts K	4 (ablation: $K \in \{2, 4, 6, 8\}$)
Gating networks	Linear($d_{\text{in}} \rightarrow K$), Softmax—one per task
Task heads	Linear $\rightarrow$ Sigmoid (PD); Linear $\rightarrow$ Sigmoid (LGD)
<i>Hard-sharing MTL</i>	
Shared trunk	3-layer MLP: [256, 128, 64], BatchNorm, ReLU, Dropout 0.2
Task heads	Linear $\rightarrow$ Sigmoid (PD); Linear $\rightarrow$ Sigmoid (LGD)
<i>Shared neural training settings</i>	
Optimizer	Adam, lr = 10^{-3} , weight decay = 10^{-4}
Batch size	512
Max epochs	100 (early stopping: patience = 15)
LR scheduler	ReduceLROnPlateau (factor 0.5, patience 5, min 10^{-5})
Gradient clipping	Max ℓ_2 norm = 5.0
Task loss weighting	Homoscedastic uncertainty (Kendall et al, 2018)
<i>XGBoost (PD classifier and LGD regressor, sequential)</i>	
n_estimators	500
max_depth	6
learning_rate	0.05
subsample	0.8
colsample_bytree	0.8
<i>LightGBM (PD classifier and LGD regressor, sequential)</i>	
n_estimators	1000
num_leaves	63
learning_rate	0.05
subsample	0.8
colsample_bytree	0.8
min_child_samples	50

References

- Altman EI (1968) Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance* 23(4):589–609. <https://doi.org/10.1111/j.1540-6261.1968.tb00843.x>
- Basel Committee on Banking Supervision (2006) International convergence of capital measurement and capital standards: A revised framework. Tech. rep., Bank for International Settlements
- Board of Governors of the Federal Reserve System (2026) Consumer credit—G.19. Statistical release, July 8, 2026, URL <https://www.federalreserve.gov/releases/g19/current/default.htm>
- Bussmann N, Giudici P, Marinelli D, et al (2021) Explainable machine learning in credit risk management. *Computational Economics* 57(1):203–216. <https://doi.org/10.1007/s10614-020-10042-0>
- Caruana R (1997) Multitask learning. *Machine Learning* 28(1):41–75. <https://doi.org/10.1023/A:1007379606734>
- Chen T, Guestrin C (2016) XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp 785–794, <https://doi.org/10.1145/2939672.2939785>
- Ke G, Meng Q, Finley T, et al (2017) LightGBM: A highly efficient gradient boosting decision tree. In: *Advances in Neural Information Processing Systems* 30, pp 3146–3154

- Kendall A, Gal Y, Cipolla R (2018) Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 7482–7491, <https://doi.org/10.1109/CVPR.2018.00781>
- Kvamme H, Sellereite N, Aas K, et al (2018) Predicting mortgage default using convolutional neural networks. *Expert Systems with Applications* 102:207–217. <https://doi.org/10.1016/j.eswa.2018.02.029>
- Lessmann S, Baesens B, Seow HV, et al (2015) Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research* 247(1):124–136. <https://doi.org/10.1016/j.ejor.2015.05.030>
- Liu S, Johns E, Davison AJ (2019) End-to-end multi-task learning with attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 1871–1880, <https://doi.org/10.1109/CVPR.2019.00197>
- Loterman G, Brown I, Martens D, et al (2012) Benchmarking regression algorithms for loss given default modeling. *International Journal of Forecasting* 28(1):161–170. <https://doi.org/10.1016/j.ijforecast.2011.01.006>
- Ma J, Zhao Z, Yi X, et al (2018) Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 1930–1939, <https://doi.org/10.1145/3219819.3220007>
- Misra I, Shrivastava A, Gupta A, et al (2016) Cross-stitch networks for multi-task learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3994–4003, <https://doi.org/10.1109/CVPR.2016.433>

- Monje L, Carrasco RA, Sánchez-Montañés M (2025) Machine learning XAI for early loan default prediction. *Computational Economics* 67(5):4033–4062. <https://doi.org/10.1007/s10614-025-10962-9>
- Sadhwani A, Giesecke K, Sirignano J (2021) Deep learning for mortgage risk. *Journal of Financial Econometrics* 19(2):313–368. <https://doi.org/10.1093/jjfinc/nbaa025>
- Thomas LC, Edelman DB, Crook JN (2002) Credit Scoring and Its Applications. Society for Industrial and Applied Mathematics, <https://doi.org/10.1137/1.9780898718317>
- Yu T, Kumar S, Gupta A, et al (2020) Gradient surgery for multi-task learning. In: *Advances in Neural Information Processing Systems* 33, pp 5824–5836
- Zhang Y, Yang Q (2022) A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering* 34(12):5586–5609. <https://doi.org/10.1109/TKDE.2021.3070203>

Statements and Declarations

Funding. The authors received no specific funding for this work.

Competing interests. The authors declare that they have no competing interests.

Data availability. The Lending Club loan data analyzed in this study are publicly available. The processed data required to reproduce the reported results are available from the corresponding author upon request.

Code availability. The code used for data processing, model estimation, and evaluation is available from the corresponding author upon request.

Author contributions. All authors contributed equally to this work.