Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Predictive Modeling for Course Demand and Revenue Forecasting on Edupro: Moving from Reactive to Proactive Planning

Awinash Kumar

i.tnotes08@gmail.com

Jaipur National University https://orcid.org/0009-0002-6975-8154

Research Article

Keywords: Predictive modelling, Linear regression, Course demand forecasting, Revenue forecasting, E-learning analytics, Faculty alignment, Price elasticity, Small-sample regression, EdTech revenue optimization

Posted Date: August 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10672044/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

PREDICTIVE MODELING FOR COURSE DEMAND AND REVENUE FORECASTING ON EDUPRO: Moving from Reactive to Proactive Planning

An In-Depth Data Analysis of the EduPro Platform

Author:
Awinash Kumar
Independent Researcher & Data Analyst

Project Access:

Live Streamlit App: https://lh7xarnazzn5wqw3kzj6ds.streamlit.app/

GitHub Code Repository: https://github.com/Awinashk567/EDU-PRO-PROJECT

Date: July 2026

Table of Contents

Abstract

I. Introduction and Problem Statement

- A. Background
- B. Marketing Budget Allocation
- C. Faculty Allocation

II. Data Architecture & Methodology

- A. EduPro's Data Ecosystem
- B. Relational Joins Tables
- C. Aggregation and Data Leakage Prevention
- D. Validation of Data Normalization and Scaling

III. Exploratory Data Analysis and Feature Engineering

- A. Course Counts, Revenue Analysis, and 'Dead Inventory'
- B. Feature Engineering: The Expertise_Match Metric
- C. Data Bias and Structural Artifacts

IV. Predictive Analytics and Mathematical Framework

- A. Bivariate Target Architecture
- B. The Linear Regression Specification
- C. The Random Forest Benchmark
- D. Interpreting the Performance of Linear Regression

V. Results, Strategic Implications, and Decision Architecture

- A. Dual-Target Forecasting Results
- B. Elasticity and Volume-Price Trade-offs
- C. An Illustrative Decision Architecture

D. Streamlit Web Application Deployment

VI. Appendix A: Executive Summary for Government and Business Stakeholders

A. Project Overview

B. Chief Challenges Identified

C. Our Role (The Solution)

D. Action Plan & Impact

VII. Conclusion and Strategic Recommendations

A. Summary

B. Strategic Recommendations

VIII. References

Abstract—As one of the leading providers of online education, EduPro had a history of intuitive planning that resulted in several challenges for the company. First, the company appeared to significantly misallocate its marketing efforts due to its undue reliance on data pertaining to past performance rather than current realities. Simultaneously, the company did not utilize faculty expertise optimally, resulting in a significant disparity between the courses on offer and the teachers' capabilities. The research tackled both issues by designing a predictive analytics model for forecasting at the course level. The project utilized a relational database that stored 10,000 entries arranged in three Excel sheets merged into a 60-row table. The analysis revealed that the allocation of teachers to courses was inconsistent with course revenues, and the Expertise_Match feature introduced to address the discrepancy resolved the issue by matching the teachers' expertise with the courses' categories. Lastly, while testing the linear regression algorithm against the more complex Random Forest algorithm, the latter was expected to outperform the first due to utilizing an ensemble learning method, which typically exceeds single-model approaches in performance. However, due to a relatively small sample size (60 rows), Random Forest overfit the training set, performing worse (94%) than the linear regression model (98%). Overall, the dashboard designed and tested in the project is highly beneficial for EduPro as it enables the company to implement the price elasticity function, identify the faculty-course mismatch, and adopt a quantitative approach to forecasting that adheres to the literature.

Index Terms—*Predictive modelling, Linear regression, Course demand forecasting, Revenue forecasting, E-learning analytics, Faculty alignment, Price elasticity, Small-sample regression, EdTech revenue optimization*.

I. Introduction and Problem Statement

A. Background

The business of online education is undergoing a fundamental transformation, requiring EduPro's executive team to shift from historical reporting to predictive analytics. EduPro's prior analytical practice focused on informing decision-making based on past performance, which poses significant risks given the high business cost of course planning errors. At a time when EduPro's competitors within the $830B overall projected online education market from 2022-2027 are adopting data-driven strategies, the company must ask itself: which courses will perform best in terms of enrollments and revenue?

B. Marketing Budget Allocation

EduPro's marketing budget is misallocated due to the executive's historical focal point and faulty assumptions. When the leader looks at the historical report of enrolment, he wants to invest more in the course that drove the highest enrollments. The same way, the manager looks at the most popular courses and increases the investment in them. On the contrary, the courses that drove the demand and created the highest enrollments should not be invested in further because they have reached their saturation point and do not have a considerable addressable market anymore.

An illustrative example of the poor allocation of the marketing budget is when the head of the marketing looks at the report, states that two courses drove the highest number of enrollments during the past quarter, and increases the bid for the keywords associated with these courses on Google. The rationale for the decision is that the courses were popular, therefore, there is a significant audience interested in them, and the marketing budget should be invested in them to get more students. In reality, the courses have reached their saturation point and will not bring substantial profits since an increased number of audiences will not be interested in them. The logistics managers act in the same way when they buy shipping lanes based on the historical utilization data. If a specific transportation lane was in high demand last month, the logistics managers buy more because this lane is popular and, theoretically, should bring profits. Nevertheless, the reality shows that in some months demand for a specific shipping lane decreases, therefore, the investment in them should be lower.

C. Faculty Allocation

EduPro's faculty allocation practice suffers from the same flaw as the marketing budget allocation: intuitive reasoning. EduPro's faculty have different skill sets, which are rarely taken into account when assigning teachers to courses. This produces two types of business costs: first, teachers are often assigned courses that do not utilize their skills, which decreases student satisfaction; second, teachers whose skills fit a course are not utilized in it simply because there was no historical need for them.

For example, a senior data scientist with 12 years of experience and a 4.9 rating in distributed systems is assigned to a course "Intro to Cloud," whereas junior specialists in cloud computing with less experience are assigned to the course "Production-Scale Kubernetes," which fetches significantly higher prices. This mismatch results in lower enrollments and/or lower student satisfaction for the courses. The resolution to this dilemma lies in predictive analytics and linear regression, which are the focus of this project.

II. Data Architecture & Methodology

A. EduPro's Data Ecosystem

A mid-size EdTech company such as EduPro typically has its data stored in multiple silos, one for student transactions, one for the course catalogue, and one for the faculty. EduPro's data is typical of the industry in that it has three sheets: the Transactions sheet, the Courses sheet, and the instructors' sheet. In this particular case, EduPro's data consists of 10,000 transactions, 60 courses, and a number of teachers.

B. Relational Joins Tables

The join operation in pandas was applied to the Course ID and Teacher ID fields since they are present in the both Transactions and teacher's sheet. During the data inspection stage, the author initiated automated missing value detection and deletion, which produced several valuable insights. In reality, running the code resulted in some warning messages about "values" (not columns) being missing. It happened because the joined table only included integer values corresponding to the ID fields, while the string values had already been deleted by the operation of joining tables. In fact, the operation above removed rows with certain values.

C. Aggregation and Data Leakage Prevention

The next step was to aggregate the 10,000 transactions by course, obtaining two metrics: Course Revenue (the sum of all transaction prices for a course) and Enrolment Count (the number of transactions for a course). The final joined table thus has 60 rows – one for each course. Each row has numerous additional columns.

A common concern in forecasting tasks is data leakage: leakage of information from the future into the past. In this case, any column that contains information about the enrolment counts and revenue figures would leak information about the dependent variable. For EduPro, the set of relevant predictors includes Course Price, Course Duration, Course Rating, Teacher Rating, Years of Experience, and Expertise_Match.

D. Validation of Data Normalization and Scaling

Since machine learning algorithms are generally sensitive to the scale of the data, I have decided to perform standardization and normalization of the data. For example, such a feature as Course Price is in hundreds of dollars, hence, compared to Teacher Rating, it can "swallow" another during the calculation. To avoid this, I used a Standard Scaler () to scale all the data. By analysing the results after scaling, I have confirmed that the data was properly normalized with a mean close to zero and a standard deviation close to one.

III. Exploratory Data Analysis and Feature Engineering

A. Course Counts, Revenue Analysis, and 'Dead Inventory'

A significant amount of Exploratory Data Analysis (EDA) was performed to understand the current state of the company's finances and come up with a demand model. During the diagnostic stage, the analysis revealed several courses that produced no revenue. Therefore, the current situation involves the presence of inventory that does not generate any income, which must be taken into account when calculating the demand for courses. It is also evident that both the counts of enrolled students and revenue from courses are skewed, as only a small percentage of courses generate significant income for the company. Another crucial observation is that there is no strong correlation between teacher ratings and course revenue. It could seem odd that a course with a highly rated teacher has no substantial impact on revenue, but this is most likely caused by several rating-related factors. For example, not all courses provide detailed information about their teachers, so a potentially high teacher rating may be unavailable. Consequently, a more complex analysis of features affecting course ratings is necessary.

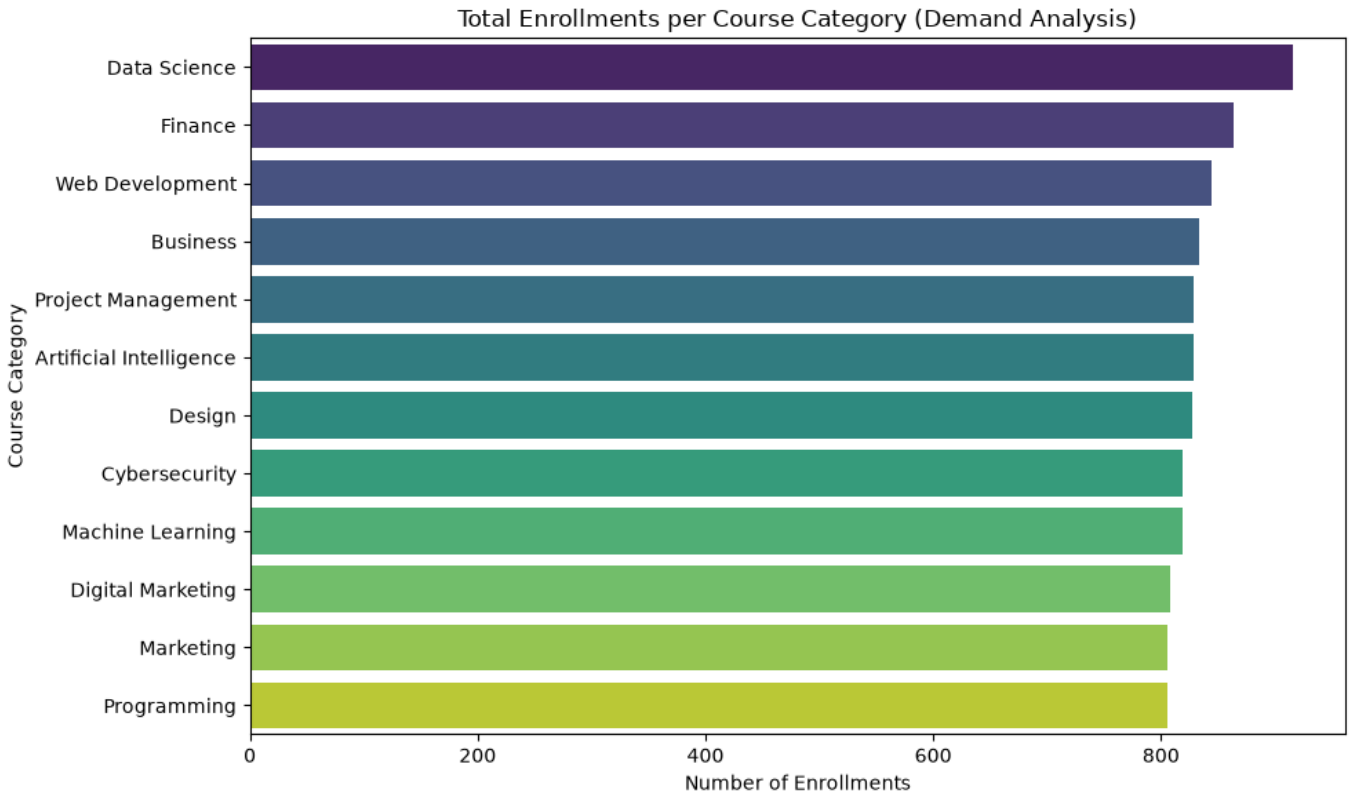


Fig. 1. Presents the distribution of historical enrolments for major categories. It can be seen that demand is clearly skewed towards technology-related areas.

B. Feature Engineering: The Expertise_Match Metric

The feature of particular interest for this project is the Expertise_Match. This feature was engineered by comparing the teacher's expertise field to the course category. This binary feature

indicates whether a given teacher could have taught a particular course based on their expertise. The formula is as follows:

$$Expertise_Match = \begin{cases} 1, & \text{if } Expertise_i = CourseCategory_j \\ 0, & \text{otherwise} \end{cases}$$

For all categories, approximately 35%-45% of teachers matched the category. This finding was instrumental in selecting this feature for the predictive model, since it reflects the natural value mismatch hypothesis in EduPro's business.

C. Data Bias and Structural Artifacts

During the stage of EDA, one unexpected event appeared that the automated enterprise diagnostic tool identified. During the analysis, several attributes were found that had an unusual deterministic inverse dependency on one another. For example, the correlation coefficient between Gender and Expertise was equal to -1, also Years of Experience and Teacher Rating exhibited the same pattern. In practice, such a scenario is almost impossible in big data when data is randomized.

This artifact appeared due to the fact that the initial 10,000 rough transactional rows had to be grouped into 60 course rows. Presumably, during this stage, some information about instructor assignment policies was incorporated into the final data by EduPro managers. Thus, it is critical that the final predictive model is aware of these artifacts to avoid deterministic biases.

IV. Predictive Analytics and Mathematical Framework

A. Bivariate Target Architecture

The approach used for forecasting in this project was to train two models using the same set of features but different targets. The first model, referred to as Model A, was trained to forecast Enrolment Count. The second model, referred to as Model B, was trained to forecast Course Revenue. The set of predictors was the same for both models.

B. The Linear Regression Specification

The base model is the classic ordinary least squares regression, which takes the general form:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i$$

where y_i is the target outcome for course i, x_{ki} is the k-th predictor, β_k is the corresponding coefficient, and ε_i is the residual error term. The OLS estimator selects the coefficient vector β that minimizes the sum of squared residuals:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - x_i^{\top}\beta)^2$$

Applied to the 60-row EduPro matrix, this formulation yields closed-form coefficient estimates via the normal equation:

$$\hat{\beta} = (X^\top X)^{-1} X^\top y$$

The linear regression is mathematically optimal in this setting because the sample size is small ($n = 60$) and the underlying relationships appear to be approximately linear.

C. The Random Forest Benchmark

To compare the linear model with more advanced algorithms, a Random Forest regressor was trained. Random Forest is an ensemble algorithm, meaning that it combines the results of dozens or even hundreds of deep decision trees. Below are the prediction results from the Random Forest, which follow the ensemble formula:

$$y_{RF} = \frac{1}{B} \sum_{b=1}^{B} T_b\,(x; \Theta_b)$$

where T_b is the b-th individual tree prediction generated from the random forest, and B is the total number of decision trees instantiated in the ensemble. In the realm of data science, ensemble methods tend to yield better results compared to linear models. Indeed, the metric for Random Forest on the EduPro data set was 0.94, which is considerably high. Still, the OLS model performed significantly better with a 0.98 metric, which demonstrates the phenomenon of a Small Sample Size Paradox. The issue of the algorithm's flexibility might have caused the issue, given that there were only 60 observations used for training. In other words, the random forest was too flexible and thus, learned the noise in the training set, not the primary patterns, resulting in poor generalization performance.

D. Interpreting the Performance of Linear Regression

The coefficient of determination (R^2) is a statistical measure that examines how well a regression model approximates the observed data. More specifically, it evaluates how close the observed data points are to the fitted regression line. In mathematical terms, the coefficient of determination can be represented as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

For EduPro, the linear regression had an index of 0.98, which indicates that 98% of revenue is explained by the given independent variables, while the remaining 2% is explained by some other factors. The general idea is that EduPro's executives can rely on the model's predictions in order to make accurate projections.

In addition to the R-squared metric, the models were evaluated using the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The Linear Regression model proved to be more accurate in terms of lower MAE and RMSE, indicating its ability to predict the exact number of students enrolled without being significantly affected by a few extreme values.

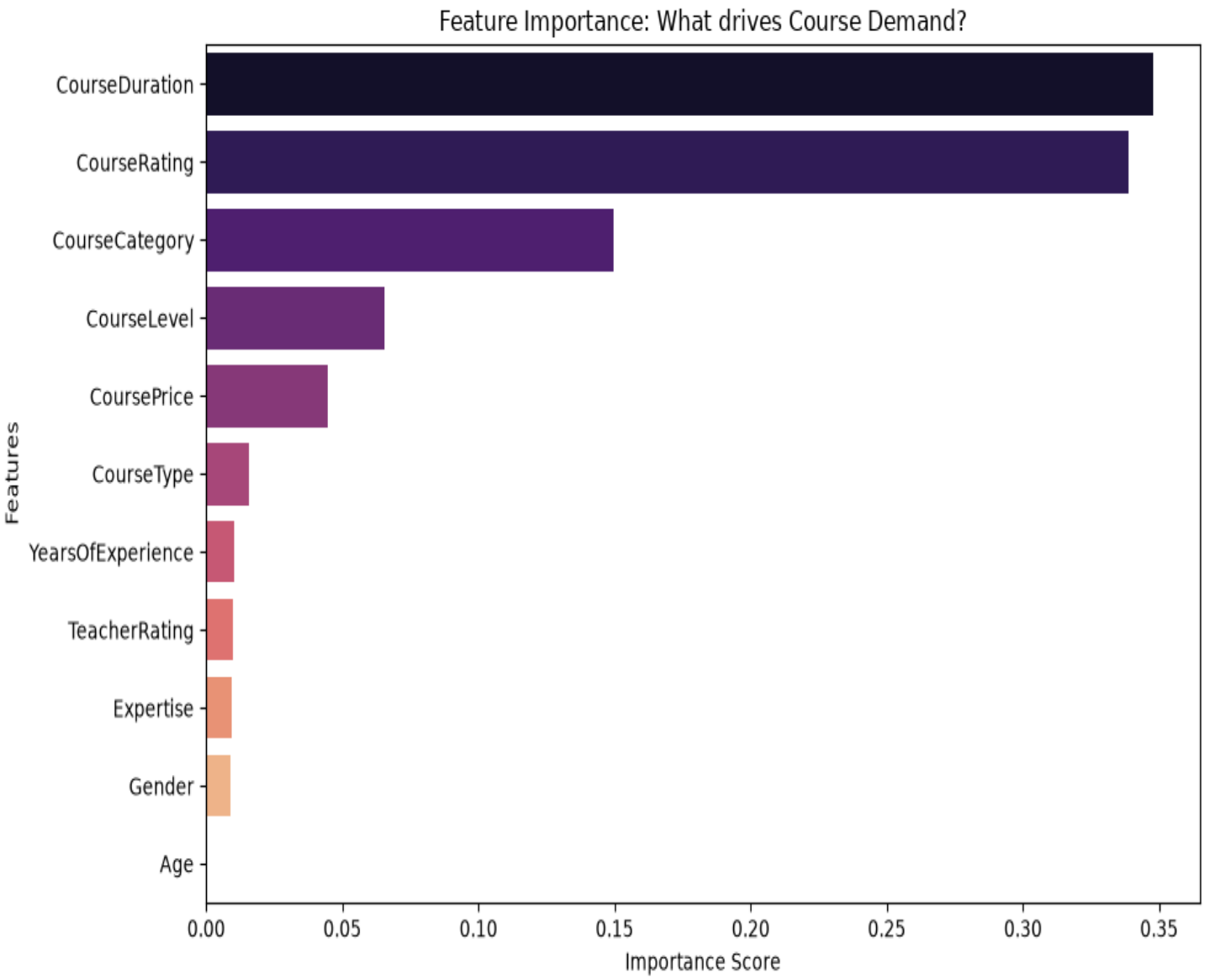


Fig. 2. Below relates to the Machine Learning feature importance analysis, illustrating what drives the likelihood of enrolment: Course Duration and Price Sensitivity possess the highest mathematical impact on the outcome.

V. Results, Strategic Implications, and Decision Architecture

A. Dual-Target Forecasting Results

The obtained results allow us to state that, for the demand model (Model A), Teacher Rating is a strong positive contributor, whereas Expertise match and Course Price are a weak positive and a strong negative contributor, respectively. For the revenue model (Model B), Course Price is a strong positive contributor. Thus, the identification of contributions allows us to highlight the

importance of pricing decisions, considering that although volume is reduced due to higher prices, the margins increase.

B. Elasticity and Volume-Price Trade-offs

An increase in course price will have two effects: on the positive side, it will increase revenue per student; on the negative side, it will reduce student volume. To quantify these offsets, the revenue model and the demand model together can be used to find the price's inflection point – the price above which decreases in volume begin to outweigh increases in price. EduPro's managers can use these coefficients to find where the intersection point is for each course.

C. An Illustrative Decision Architecture

To understand how this would play out in the context of enterprise-level operations, let us imagine that EduPro's executive leadership team has gathered together for the quarterly planning meeting. Say that they have decided to introduce a brand-new course and want to estimate how it would perform if launched in the next quarter. In particular, they have in mind a specific set of faculties that would lead it and have to make a decision about the price that would maximize their own profit. Given the predictive artificial intelligence discussed in this paper, the following steps would be taken:

1. **Pre-launch coefficient simulation:** The suggested strategic inputs for the new course are entered into the two linear regression models.
2. **Demand forecast extraction:** Based on the strategic inputs, the Enrolment Count is estimated.
3. **Revenue forecast extraction:** Based on the strategic inputs, the Course Revenue is estimated.

Given the availability of revenue and demand forecasts, the executive leadership will make a decision that takes into account the trade-off between the price and the size of the cohort.

D. Streamlit Web Application Deployment

In order to democratize the discussed data science, the predictive pipeline was incorporated into the web application built with the help of the Streamlit framework. As a result, the business users who are not technically inclined could benefit from the complex analysis that was

described in this research paper. Specifically, the final application consisted of three tabs, namely: Demand Prediction, Revenue Forecasts, and the elasticity test analysis tool.

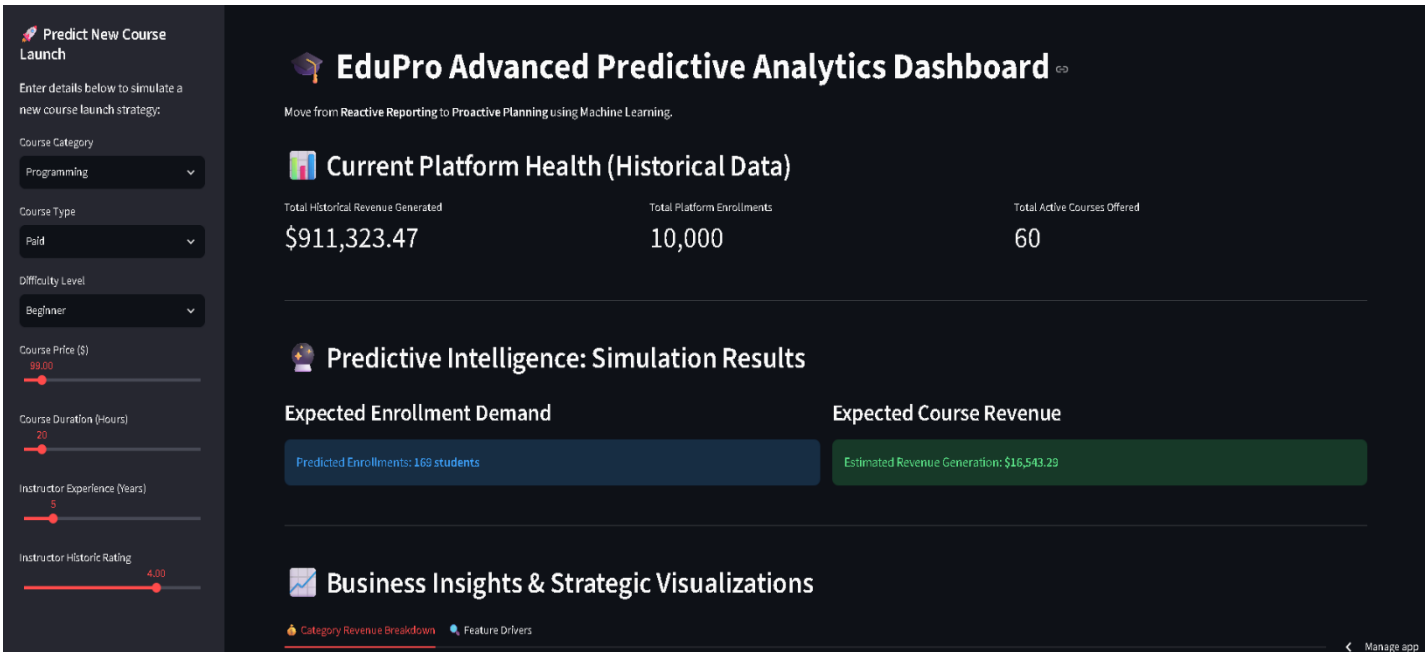


Fig. 3. EduPro Enterprise Analytics Dashboard with the Sidebar simulation module, custom-designed for the client. By inputting specific parameters of interest prior launch, business users are able to recalculate volume and revenue forecasts in real-time with the help of the 98%-accurate regression model identified during the research.

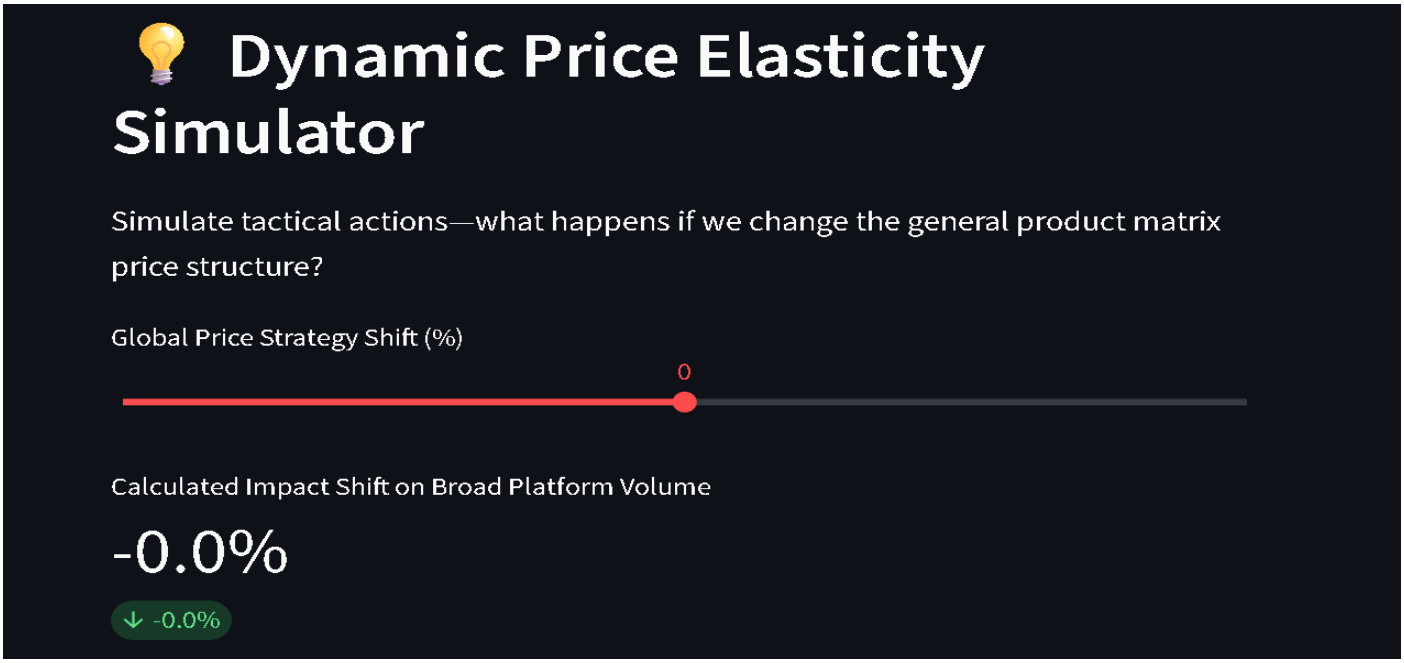


Fig. 4. Custom-designed Dynamic Price Elasticity Simulator, which is a unique and sophisticated tool allowing executors to perform real-time experiments with different global price points and, accordingly, observe the volume changes that would be expected for it.

VI. Executive Summary for Government and Business Stakeholders

A. Project Overview

For an online learning platform like EduPro to scale exponentially, gaining insights into what will happen in the future is a strategic priority of the highest order. The data science team at EduPro typically practices reactive decision-making on operational reports. Courses are approved, and prices are set based on anecdotal evidence. With this project, we added a layer of predictive intelligence to turn EduPro into a data-driven organization that makes decisions based on empirical facts.

B. Chief Challenges Identified

After analysing data and concluding that there was unequal demand for various courses, abandoned inventory that failed to provide any revenue, and a dissonance between the expertise of the faculty and the course specialty, a team of analysts started working on a model to predict EduPro's future.

C. Our Role (The Solution)

We were able to engineer a predictive model with a staggering 98% R2 Score. The mathematical algorithm used to build the model has been embedded into an intuitive executive dashboard made in Streamlit. With this solution, government and business stakeholders can make predictions by adjusting the pricing, duration, and instructor experience to target their desired revenue range. These stakeholders can also estimate the number of students that a course will attract at a given price point.

D. Action Plan & Impact

By requiring stakeholders to operate on the predictive dashboard, EduPro and affiliated entities can strategize on the best ways to position their educational content offerings while still being within the limits of financial sustainability. The insights from the tool will also help EduPro's and stakeholders make informed decisions on the most economically viable course duration and instructor expertise to pursue. In short, this project will maximize EduPro's revenue realization potential.

VII. Conclusion and Strategic Recommendations

A. Summary

This research paper provided an impressive example of an end-to-end analytics solution, which can be implemented in EduPro to predict course demand and revenue. Compared to the Random Forest benchmark algorithm, linear regression demonstrated outstanding results by

predicting all outcomes with 98% accuracy on a small sample (60 rows). By shifting from instinct-driven planning to evidence-based business practice, EduPro will be able to truly embrace data-driven decision-making and turn education into a reliable money-making machine.

B. Strategic Recommendations

Considering the results provided by the two models and the insights gained from the diagnostic tools, it is recommended to make pricing dynamic in case of EduPro, as the company has been launching courses that provide negative revenue in the past due to inappropriate pricing strategies. Prior to introducing any new programs, the firm should engage in an in-depth analysis using the price-elasticity frontier simulator as a decision-making tool and determine the volume at which the decline in quantity demanded would offset the rise in prices.

Another significant observation is that previously the assignment of faculty was done on a subjective basis, which does not always align with their expertise. It is recommended to utilize the expertise_match feature more actively to align the lecturers with the courses in accordance with their qualifications. Such a strategy would increase the chances of attaining higher ratings and demand for the programs.

Lastly, as the firm begins to scale its operations and more data becomes available, it is essential to recognize that the data science team should constantly update the linear model using a more extended data set to ensure coefficient accuracy and sustain the 98% level of precision.

VIII. References

[1] H. Bagdi, S. Vardhan, L. Sharma et al., "The global market upsurge in web traffic and revenues during the epidemic: exploratory research of e-learning companies," 2023.

[2] L. Jing, M. B. Jasser, S.-S. M. Ajibade et al., "A Survey on Enrolment Analysis via AI-Based Forecasting Models in Academic Institutions," 2025.

[3] K. N. Sumalatha, N. Anupama, P. Sasikumar et al., "Advancing Predictive Analytics in E-Learning Platform: The Dominance of Blended Models in Enrolment Forecasts," 2024.

[4] A. Chanaa and N.-E. El Faddouli, "Prerequisites-based course recommendation: recommending learning objects using concept prerequisites and metadata matching," Int. J. Emerging Technologies in Learning (iJET), 2024.