PROJECT REPORT

Machine Learning-Based Decision Support System for HELOC Approval

SIMON BUSINESS SCHOOL
UNIVERSITY of ROCHESTER

Team Members:
Irons, Matthew
Ikhanoba, Osigbemhe
Kumar, Devansh
Liu, Zhiying
Lan, Kai

INDEX

INTRODUCTION

A **Home Equity Line of Credit (HELOC)** is a financial product that allows homeowners to borrow against the equity of their property. Currently, the HELOC approval process at **Simon Bank of Rochester®** is manually conducted by expert loan officers, which is time-consuming and resource intensive. This project aims to develop a **machine learning-based decision support system** to automate the initial screening of HELOC applications, improving efficiency while ensuring transparency and regulatory compliance.
The proposed system will **classify applications into two categories:**

Negative (denied) and Positive (sent for further review). To maintain ethical lending practices, the system must generate clear explanations for denied applications and provide guidance for improving approval chances in future attempts.

Machine learning models require high-quality data for accurate predictions. Before model training, we will conduct an **exploratory data analysis (EDA)** to identify missing values, outliers, and feature importance. The dataset consists of 10,000 historical HELOC applications, including details on financial metrics such as credit scores, loan-to-value ratios, and debt-to-income ratios.
Since this is a classification task, key performance metrics include:

- **Accuracy –** Measures overall correctness of predictions.
- **Precision & Recall –** Ensures correct identification of valid approvals and rejections.
- **AUC-ROC Score –** Evaluates model effectiveness in distinguishing between approved and denied applicants.

A well-performing predictive model can lead to significant business improvements. By reducing the time required for manual reviews, the bank can lower operational costs while enhancing the customer experience. Additionally, ensuring **fair and unbiased lending decisions** supports regulatory compliance and strengthens customer trust.
The HELOC decision support system will be implemented as an **interactive prototype using Streamlit**, allowing applicants to enter financial details and receive real-time decisions. This tool will demonstrate how automation can optimize HELOC approvals while maintaining fairness and explainability.

DATA EXPLORATION AND PREPROCESSING

Our analysis revealed missing values in several financial attributes, particularly where special values like -7 and -9 were used to indicate missing or undefined data. To address this, we applied mean imputation for -7 values based on the average within each RiskPerformance category, while rows containing -9 values were removed to maintain data integrity. (See Figure 1 for the distribution of -7, -8, and -9 occurrences across features.)

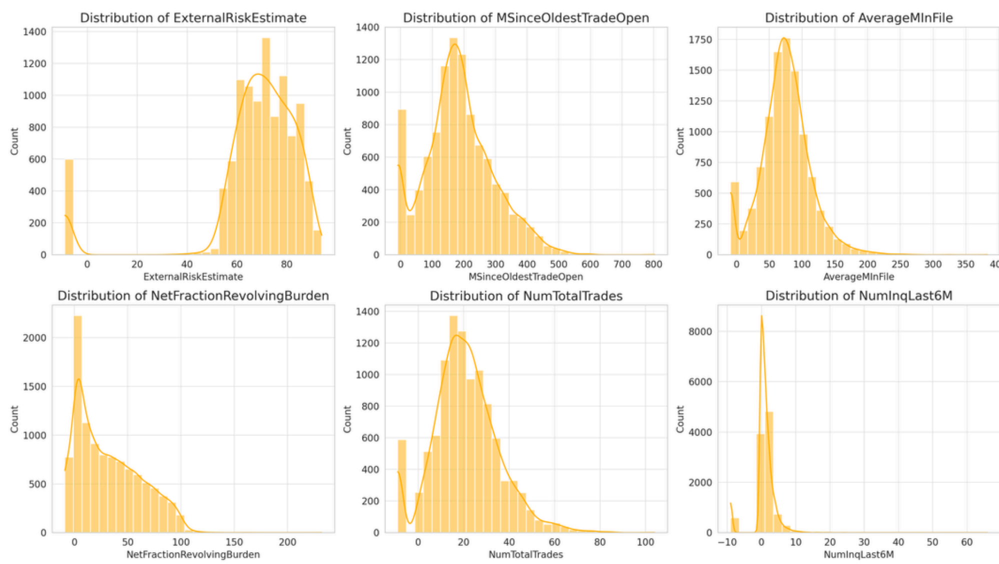


Fig. 1 : Occurrences of -7, -8, and -9 in dataset features.

A critical challenge was class imbalance, where a significantly larger number of applicants were classified as "Bad" compared to "Good." This imbalance could lead to biased model training, making it difficult to accurately predict approvals. To mitigate this, we applied Synthetic Minority Over-sampling Technique (SMOTE) to generate synthetic samples for the minority class, ensuring a more balanced dataset. (See Figure 2 for class distribution before and after SMOTE.) Additionally, a PCA-based scatter plot visualizes how SMOTE generates synthetic instances within the feature space. (See Figure 3.)

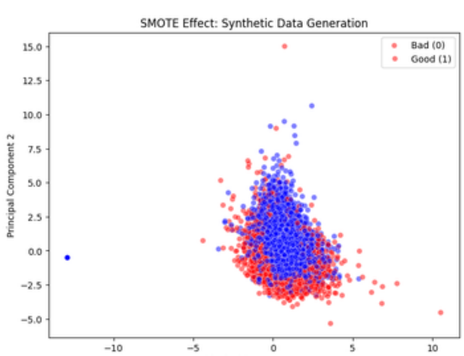


Figure 3: PCA-based visualization of synthetic data generation through SMOTE.

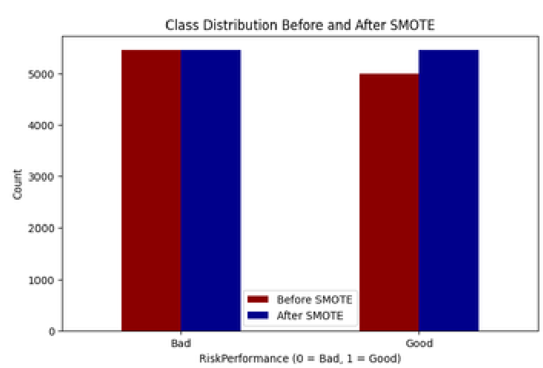


Figure 2: Class distribution before and after applying SMOTE.

MACHINE LEARNING MODEL DEVELOPMENT AND EXPLAINABILITY

To develop an accurate **HELOC approval model**, we implemented a structured workflow including **feature selection, model training, evaluation,** and **explainability analysis.** Using **Lasso Regression**, we identified **key predictive features** such as *MSinceMostRecentDelq, MaxDelqEver, and ExternalRiskEstimate* (Figure 1). This selection helped streamline the model by focusing on the most influential financial indicators.

S.No.	Model	Accuracy	Precision	Recall	F1-Score
0	XGBoost	0.880832	0.906467	0.835996	0.869806
1	Random Forest (Tuned)	0.876775	0.896355	0.838126	0.866263
2	Random Forest	0.868154	0.872667	0.846645	0.859459
3	Logistic Regression	0.715010	0.691760	0.724175	0.707596

Figure 3: Model performance comparison (Accuracy, Precision, Recall, F1-Score).

We tested multiple models, starting with **Logistic Regression**, which achieved **71.5% accuracy** but struggled with complex relationships. **Random Forest**, both in its default and tuned versions, performed better, with the tuned model reaching **87.6% accuracy.** However, **XGBoost emerged as the best model,** balancing **88.1% accuracy and 83.59% recall,** making it ideal for minimizing misclassifications. It handled **imbalanced data efficiently** with **SMOTE** and provided clear feature importance rankings (Figure 1).

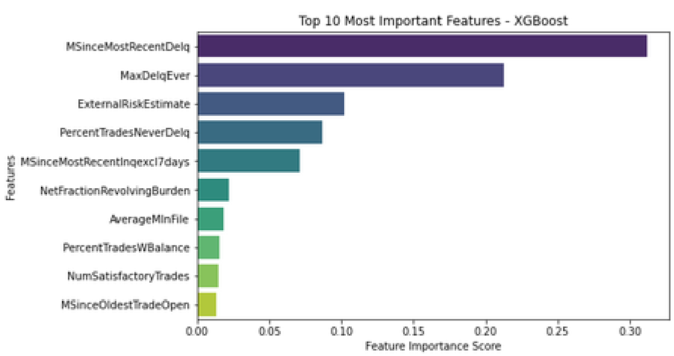


Figure 1: Top 10 most important features (XGBoost feature importance).

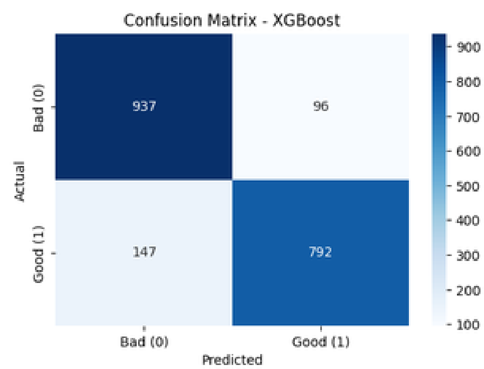
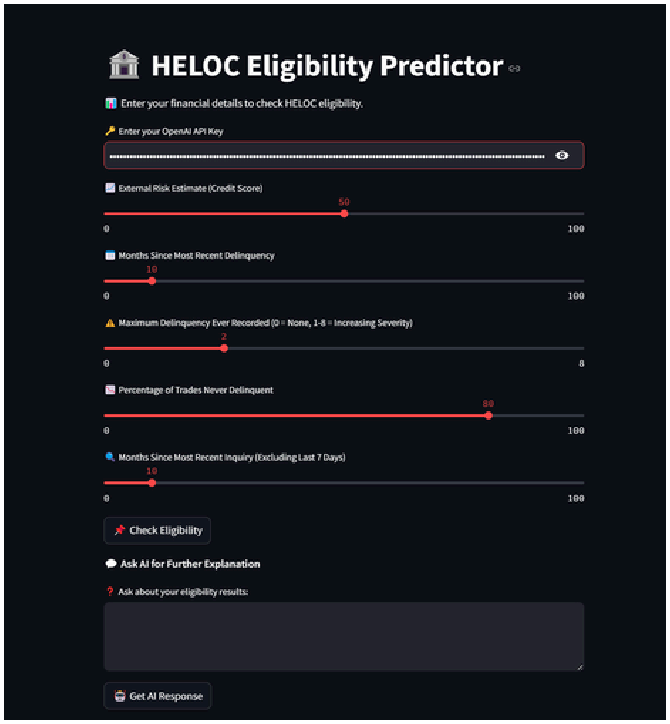


Figure 2: Confusion Matrix for XGBoost predictions.

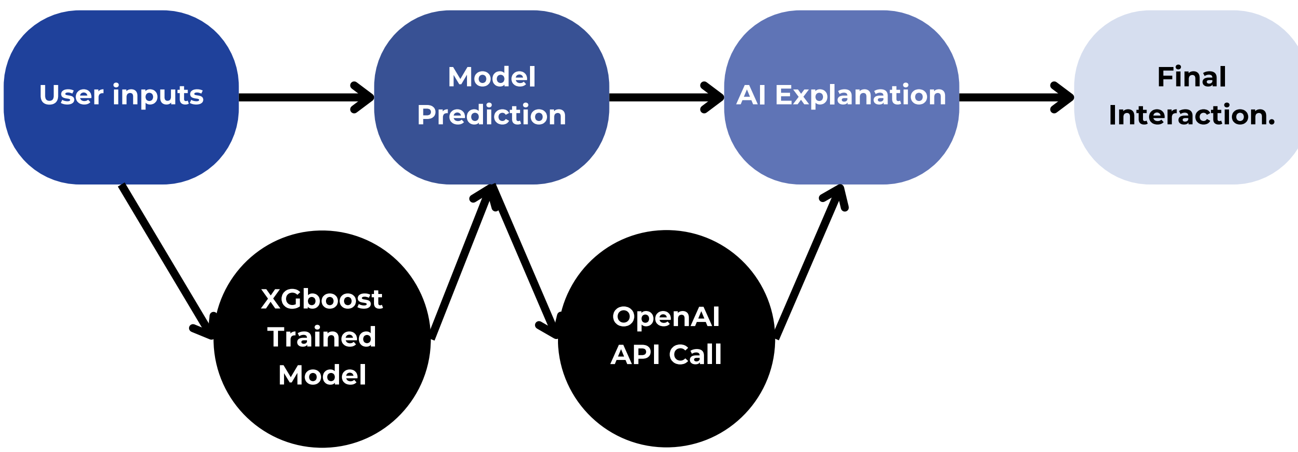
The **confusion matrix** (Figure 2) confirmed that XGBoost effectively distinguished between **approved and denied applications,** reducing false positives and negatives. A **comparative analysis** (Figure 3) showed that while Random Forest performed similarly, XGBoost was computationally efficient and more interpretable. To ensure transparency, we utilized **feature importance rankings and model evaluation metrics**, providing a clear understanding of how credit decisions are made.

PROTOTYPE DEVELOPMENT AND IMPLEMENTATION

To bring our HELOC approval model to life, we developed an interactive web-based prototype using Streamlit, enabling users to check their loan eligibility in real time. The prototype seamlessly integrates our trained XGBoost model with a user-friendly interface, making the credit assessment process more accessible and interpretable.



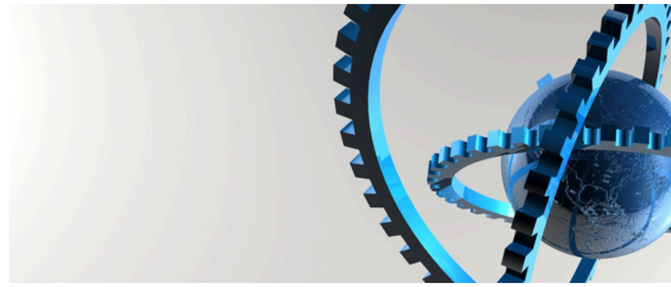
The front-end interface allows users to input their financial details, such as credit score, delinquency history, and trade performance. These values are passed to the trained XGBoost model, which predicts HELOC eligibility with a corresponding approval probability. The system provides immediate feedback, indicating whether an applicant qualifies for the loan or not. To enhance transparency, the prototype integrates OpenAI's GPT-4 for AI-driven explanations. When a user receives a loan rejection, the system generates personalized insights explaining the decision, along with recommendations to improve future eligibility. Users can also chat with the AI to get further clarifications on their credit assessment.



For deployment, we utilized Ngrok to create a secure public URL, making the application easily accessible without requiring extensive server setup. This ensures that financial institutions and users can test the system remotely.

IMPACT AND APPLICATIVE ASPECTS OF AUTOMATION

Operational Impact and Cost Savings
A bank's financial viability depends on accurate credit risk assessments. The cost savings and revenue impact of this model can be evaluated in the following ways:



Reduction in Manual Processing Costs:
Automating HELOC approvals can reduce the need for manual review by loan officers, leading to significant cost savings on labor.

Lower Default Rates: By leveraging machine learning to identify high-risk applicants, banks can reduce non-performing loans (NPLs) and associated default costs.

Increased Loan Approvals: A high-performing model ensures that creditworthy applicants are not mistakenly rejected, leading to increased revenue generation.

Operational Efficiency: Faster and automated loan processing improves customer experience and enhances the bank's competitive edge.

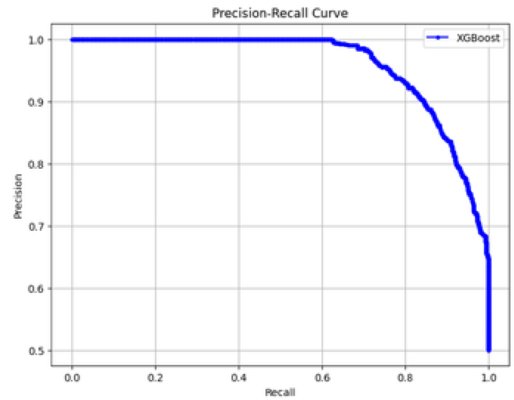
A well-optimized HELOC prediction model can potentially save millions of dollars annually by reducing default rates, optimizing loan portfolios, and minimizing human resource expenses. The expected cost savings scale directly with model accuracy—a 1% improvement in recall could mean thousands of additional successful loans without increasing financial risk.

APPENDIX

MODEL EVALUATION

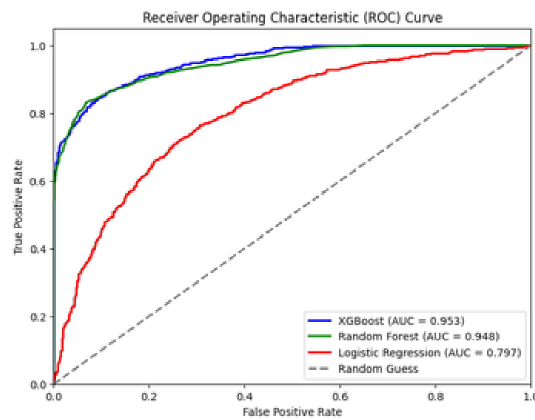
Precision-Recall Curve

The Precision-Recall Curve evaluates the trade-off between precision and recall across different classification thresholds. A high area under this curve indicates that the model maintains strong precision and recall even when adjusting decision boundaries. The curve shows that the XGBoost model effectively balances false positives and false negatives, making it highly reliable in predicting HELOC eligibility.



Receiver Operating Characteristic (ROC)

The ROC Curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR), illustrating the model's ability to distinguish between approved and rejected applicants. The AUC (Area Under the Curve) score is high, indicating that the XGBoost model significantly outperforms a random classifier. The steeper the curve, the better the model is at minimizing misclassification.



Learning Curve - Overfitting Analysis

The Learning Curve compares training and validation scores as the model is exposed to increasing amounts of data. The red curve (training score) and green curve (cross-validation score) gradually converge, suggesting that the model generalizes well and is not overfitting. If a large gap were present between these curves, it would indicate overfitting, where the model memorizes training data instead of learning general patterns.

These evaluations confirm that the XGBoost model is well-optimized, robust, and capable of accurate HELOC eligibility predictions.