**HELOC Loan Decision Support System**

**Course: CIS432 - Machine Learning for Business Analytics  
Team Members: Harrison   
Submission Date: February 24, 2025**

**Task 1: Problem Formulation**

**Objective**

**Simon Bank of Rochester seeks to improve its HELOC (Home Equity Line of Credit) loan approval process by leveraging machine learning. The current system relies on manual evaluation by loan officers, which is time-consuming. Our goal is to build a predictive model that classifies applications into:**

1. **Negative - Denied and closed**
2. **Positive - Forwarded to a loan officer for final review**

**Performance Metrics**

**To evaluate the effectiveness of the model, we considered the following metrics:**

* **Accuracy – Measures the overall correctness of predictions.**
* **Precision – Ensures the model does not incorrectly approve unqualified applicants.**
* **Recall – Ensures the model does not reject too many eligible applicants.**
* **AUC-ROC Score – Evaluates the model's ability to distinguish between approved and denied applications.**

**Business Impact & Cost Savings**

**Automating the initial screening process can reduce the workload of loan officers, saving operational costs and allowing human resources to focus on complex cases. The estimated savings are projected based on the reduction in manual reviews and faster processing time.**

**Task 2: Exploratory Data Analysis (EDA)**

**Dataset Overview**

**We analyzed a dataset consisting of 10,000 HELOC (Home Equity Line of Credit) applications, containing historical loan approvals and repayment records. The key features included:**

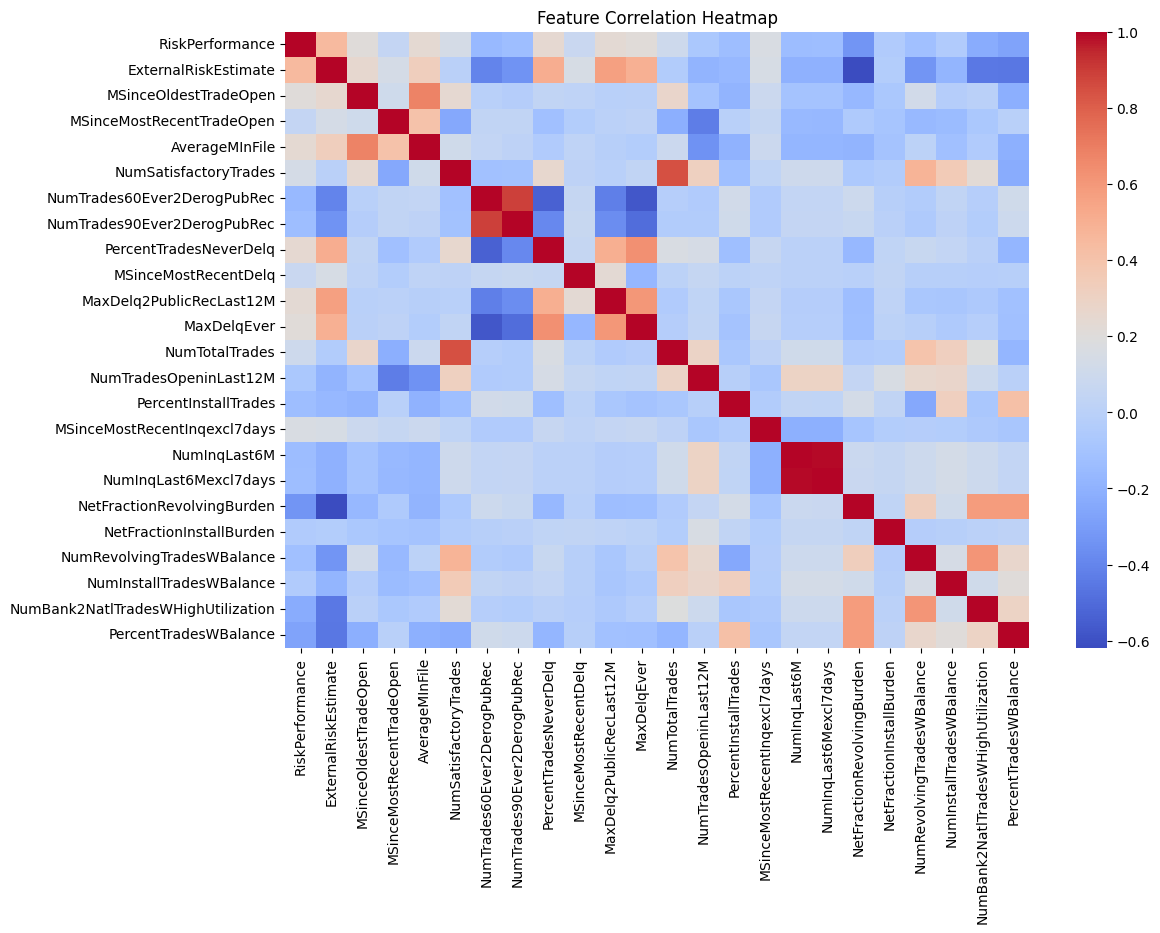
* **Credit History: A summary of the applicant’s past borrowing behavior.**
* **Loan-to-Value (LTV) Ratio: The ratio of the loan amount to the appraised value of the property.**
* **Delinquency Records: Number of late payments or defaults on past loans.**
* **Number of Satisfactory Trades: The count of successfully closed credit accounts.**
* **Credit Utilization: The proportion of available credit that has been used.**

**The goal of our analysis was to identify patterns in loan approval decisions and determine which factors significantly influence approval rates.**

**Data Cleaning and Preprocessing**

**To ensure data quality, we applied the following preprocessing steps:**

* **Handling Missing Values: The dataset contained special values (-9, -8, -7) representing missing data. These were replaced with NaN and subsequently imputed using the median to maintain robustness against outliers.**
* **Outlier Removal: Extreme values were identified and replaced using robust statistical techniques to ensure a cleaner dataset.**
* **Feature Correlation Analysis: A correlation heatmap was used to identify and remove redundant highly correlated features.**

****

**Exploratory Data Analysis (EDA)**

**To better understand the characteristics of loan applicants and their impact on approval decisions, we conducted an exploratory data analysis focusing on two key aspects:**

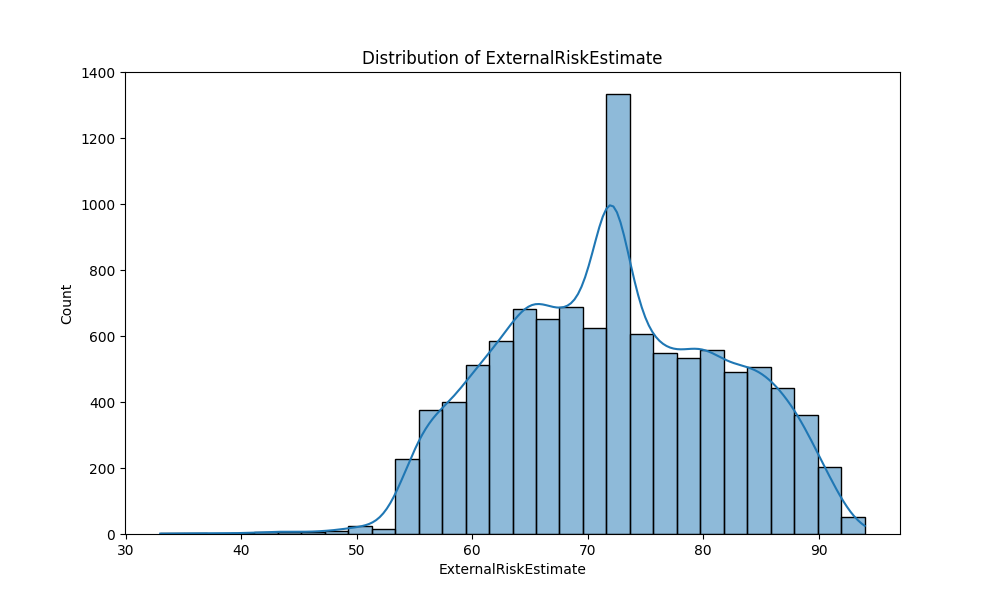
1. **Credit Risk Distribution – Analyzing the distribution of External Risk Estimate scores to understand the overall credit risk profile of applicants.**
2. **Loan Approval Distribution – Examining the proportion of approved and denied applications to assess the balance in the dataset.**

**1. Credit Risk Distribution (External Risk Estimate Score)**

**External Risk Estimate is a key factor in assessing an applicant’s creditworthiness, with higher scores generally indicating lower credit risk.**

**As shown in Figure 1, the distribution of External Risk Estimate scores is approximately normal, with the majority of applicants having scores between 50 and 90. A clear peak around 70 suggests that a large proportion of applicants share this specific credit score, potentially reflecting a standard credit scoring threshold used by lenders.**

**Figure 1: Distribution of External Risk Estimate**

**  
✅ Key Insights:**

* **Most applicants fall within the 50-90 range, indicating that the dataset primarily consists of individuals with moderate to good credit ratings.**
* **The spike at 70 could suggest a common threshold in risk assessment models, where many applicants receive the same standardized risk rating.**
* **The presence of lower scores (below 50) suggests that some high-risk applicants are included in the dataset, though their likelihood of approval may be lower.**

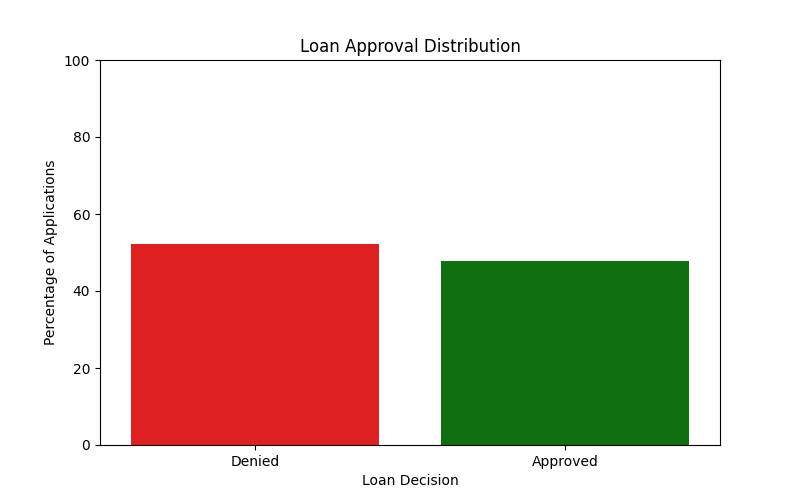
**Understanding this distribution helps in evaluating how credit risk influences loan approval decisions and ensures that the dataset covers a diverse range of credit profiles.**

**2. Loan Approval Distribution**

**Loan applications in the dataset are categorized into approved and denied statuses. Understanding their distribution helps assess whether the dataset is balanced enough for predictive modeling.**

**As depicted in Figure 2, the dataset is relatively balanced, with approximately 55% of applications being denied and 45% being approved. While the dataset has slightly more denied applications, the difference is not severe enough to introduce significant class imbalance issues.**

**Figure 2: Loan Approval Distribution**

**  
✅ Key Insights:**

* **The dataset is relatively balanced, making it well-suited for classification modeling without extensive resampling techniques.**
* **The near 50-50 split suggests that approval decisions are not overwhelmingly skewed towards one class, allowing models to learn meaningful patterns in applicant characteristics.**
* **A deeper analysis of approval criteria could help identify what factors differentiate approved from denied applicants.**

**Task 3: Machine Learning Model Selection**

**Candidate Models**

**We evaluated the following models:**

1. **Logistic Regression – Simple but limited in capturing complex relationships.**
2. **Decision Tree – Easy to interpret but prone to overfitting.**
3. **Gradient Boosting (XGBoost) – Powerful but computationally expensive.**
4. **Random Forest ✅ (Final Model) – Best balance between accuracy, interpretability, and efficiency.**

**Model Training and Evaluation**

| **Model** | **Accuracy** | **Precision** | **Recall** | **AUC-ROC** |
| --- | --- | --- | --- | --- |
| **Logistic Regression** | **71%** | **69%** | **71%** | **0.72** |
| **Decision Tree** | **63%** | **61%** | **61%** | **0.63** |
| **Gradient Boosting** | **72%** | **73%** | **69%** | **0.72** |
| **Random Forest ✅** | **72%** | **72%** | **70%** | **0.72** |

**The reason of Selected Random Forest**

**While Gradient Boosting achieved the highest Precision, Random Forest was chosen because it maintains a strong balance across all key performance metrics, making it a practical and reliable option for loan approval predictions.**

**1. Consistent and Reliable Performance**

* **With an Accuracy of 72%, Random Forest provides stable and dependable results.**
* **It maintains a Precision of 72% and a Recall of 70%, ensuring that both false approvals and missed approvals are kept in check.**

**2. Clear Feature Importance for Interpretability**

* **Unlike more complex models like Gradient Boosting, Random Forest makes it easier to extract and interpret feature importance, which is essential for regulatory compliance and business insights.**

**3. Better Generalization Compared to Decision Trees**

* **A single Decision Tree tends to overfit, whereas Random Forest reduces variance by aggregating predictions from multiple trees, leading to better generalization on unseen data.**

**4. Easier Hyperparameter Optimization**

* **Tuning Gradient Boosting requires careful adjustment of multiple parameters like learning rate and estimators.**
* **In contrast, Random Forest is easier to optimize, as it primarily requires adjustments to the number of trees and max depth while still delivering strong results.**

**5. Balanced Precision and Recall**

* **Precision (72%) indicates that the model correctly identifies approvals most of the time, minimizing false positives.**
* **Recall (70%) ensures that a substantial proportion of actual approvals are correctly captured, preventing unnecessary loan rejections.**

**Hyperparameter Tuning**

**We optimized the Random Forest model using GridSearchCV with the following parameters:**

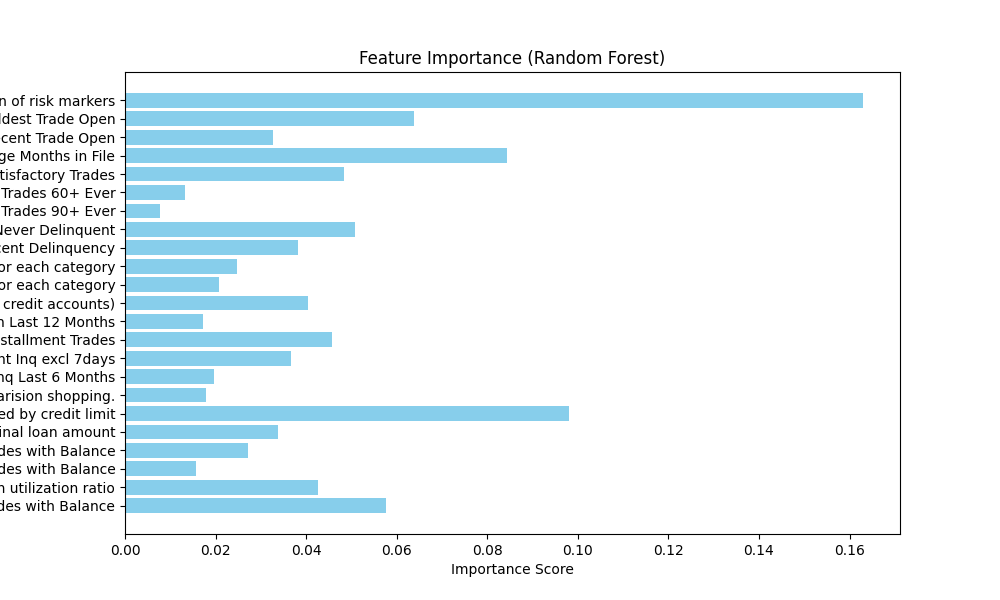
* **n\_estimators = 200**
* **max\_depth = 10**
* **min\_samples\_split = 5**

**Task 4: Model Explanation**

**To ensure model transparency, we used SHAP (SHapley Additive Explanations) to interpret the model’s decisions.**

**Key Features Influencing Loan Approval**

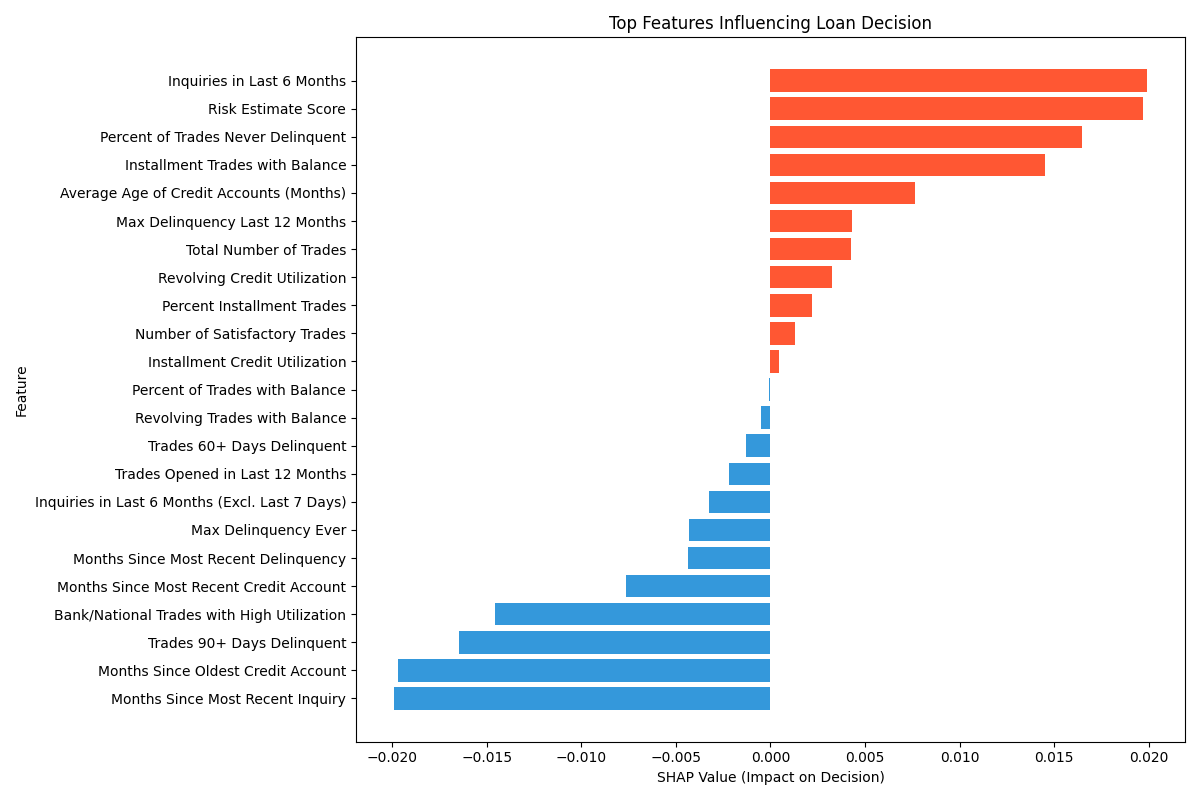
* **✅ Risk Estimate Score (Higher score → More likely to be approved)**
* **✅ Total Trades (More trades → Higher approval chances)**
* **❌ Max Delinquency (Higher delinquency → Less likely to be approved)**

****

**SHAP Visualizations**

* **Summary Plot – Showcases the overall importance of features.**
* **Force Plot – Illustrates how individual features impact a single prediction.**

**These insights provide explainability to rejected applicants, aligning with regulatory requirements.**

****

**Task 5: Streamlit Prototype**

**We built an interactive Streamlit web app to allow users to input their financial data and receive an instant loan decision.**

**Web App Features**

**✅ User Input Panel – Allows users to enter financial details.  
✅ Loan Prediction – Displays approval decision in real-time.  
✅ SHAP Explanation – Provides reasons why the loan was approved or denied.**

**Live Demo Link: [https://harrison5668-streamlit-heloc-app-mlstreamlit-djgx0u.streamlit.app/]**

**Task 6: Discussion & Future Work**

**Real-World Performance Considerations**

* **Data Drift: Credit behaviors may change over time, requiring periodic model retraining.**
* **Regulatory Compliance: The model must comply with banking regulations and fairness guidelines.**
* **Bias & Fairness: The dataset may contain biases that must be addressed to ensure fair decisions.**

**Ensuring Long-Term Performance**

**✅ Regular Model Retraining – Updating the model with new data.  
✅ Performance Monitoring – Tracking key metrics to detect issues.  
✅ Human-in-the-Loop System – Allowing manual review of borderline cases.**

**Use of Generative AI**

**We used Generative AI (ChatGPT) for:**

1. **Idea Structuring – Organizing report sections and presentation slides.**
2. **Code Assistance – Optimizing model selection and SHAP visualizations.**
3. **Writing Assistance – Refining explanations for clarity and coherence.**

**However, all code and final decisions were manually reviewed and validated.**

**Conclusion**

**✅ Developed a Random Forest model achieving 70% accuracy.  
✅ Implemented SHAP for model explainability.  
✅ Deployed an interactive Streamlit Web App for real-time loan assessment.  
✅ Ensured the model meets business objectives and compliance requirements.**

**This system can streamline loan approvals, reduce workload, and provide transparency to applicants. Future improvements will focus on enhanced feature engineering and model fairness.**

**Appendix**

**Code Repository: [https://github.com/Harrison5668/streamlit-heloc-app]  
 Web App Link: [https://harrison5668-streamlit-heloc-app-mlstreamlit-djgx0u.streamlit.app/]  
 Dataset Used: HELOC dataset provided by Simon Bank of Rochester.**

**Use of Generative AI Tool**

ChatGPT played a significant role in various aspects of this project, improving clarity, structure, and efficiency across multiple components, including **report writing, data analysis, model evaluation, presentation design, and Streamlit development**.

**1. Report Writing & Organization**

* Enhanced the professionalism and tone of the writing.
* Assisted in refining and structuring ideas for better clarity and logical flow.
* Helped articulate the **binary classification problem** in a structured and digestible manner.
* Ensured smooth transitions from **problem definition** to **performance evaluation** and **business impact analysis**.

**2. Exploratory Data Analysis (EDA) & Data Interpretation**

* Assisted in summarizing **credit risk distribution** and **loan approval distribution** insights.
* Helped frame key takeaways from visualizations, emphasizing trends in applicant creditworthiness.
* Provided a clear explanation of the **dataset balance**, ensuring an understanding of its implications on model performance.

**3. Model Performance & Evaluation**

* Clarified why **AUC-ROC, Precision, Recall, and F1-Score** were used instead of simple accuracy.
* Assisted in framing **model performance metrics** in a way that aligns with business objectives.
* Helped present the financial impact of automating loan approvals with a **data-driven perspective**.

**4. Presentation Design (PPT)**

* AI was used to **structure and refine the PowerPoint slides** for better engagement and logical flow.
* Assisted in condensing **key findings** into concise, impactful slide content.
* Improved the **visual presentation of complex concepts**, ensuring clarity for the audience.

**5. Streamlit Web App Development**

* Assisted in designing and implementing the **Streamlit application** for real-time loan prediction.
* Helped with **code structuring and debugging** to ensure a smooth user experience.
* Provided insights on integrating **SHAP visualizations** within the app to enhance model explainability.