Predicting Credit Defaults Case Study

PRESENTED BY:

Monte B.

Agenda

Background

In the credit application process, lenders evaluate the details applicants submit to gauge their creditworthiness. Key to this evaluation is understanding default risk, which occurs when borrowers fail to fulfill their financial obligations.

By predicting the likelihood of default, lenders can reduce potential losses and make better-informed decisions, thus enhancing their risk management strategies and maintaining a healthy portfolio.

Back to Agenda

Back to Agenda

Requirements

MVP Perspective

The model is designed to comprehensively capture all risk indicators, ensuring no particular risk factor is emphasized over others. This balanced approach from the outset also sets the stage for ongoing model refinement.

Data

The analysis encompasses 1,000 entries, each detailed with 31 distinct data points, encompassing information on account status, car ownership, and the nature of employment. This thorough profiling underpins precise risk assessment.

Objective

The objective is to accurately predict the likelihood of credit default, providing insights that assist in making informed credit decisions and managing risks. By anticipating potential risks, lenders can take preemptive measures to mitigate them, optimizing their loan portfolios.

Back to Agenda

Analytical Considerations

Frequency of Predictions

Real-time predictions provide a substantial edge in risk management, offering lenders immediate insights into borrowers' financial statuses and any potential shifts in their creditworthiness. This approach contrasts sharply with batch processing, which can yield outdated or less pertinent data.

By utilizing real-time insights, lenders can make decisions grounded in the latest financial context, lowering the default risk and fostering a more dynamic and responsive approach to managing credit portfolios.

Dataset

From the initial dataset of 1,000 entries, 200 will be reserved to create a distinct test set, leaving 800 entries for training and validation. This separation ensures that an untouched dataset is available to evaluate the model's performance later in the process.

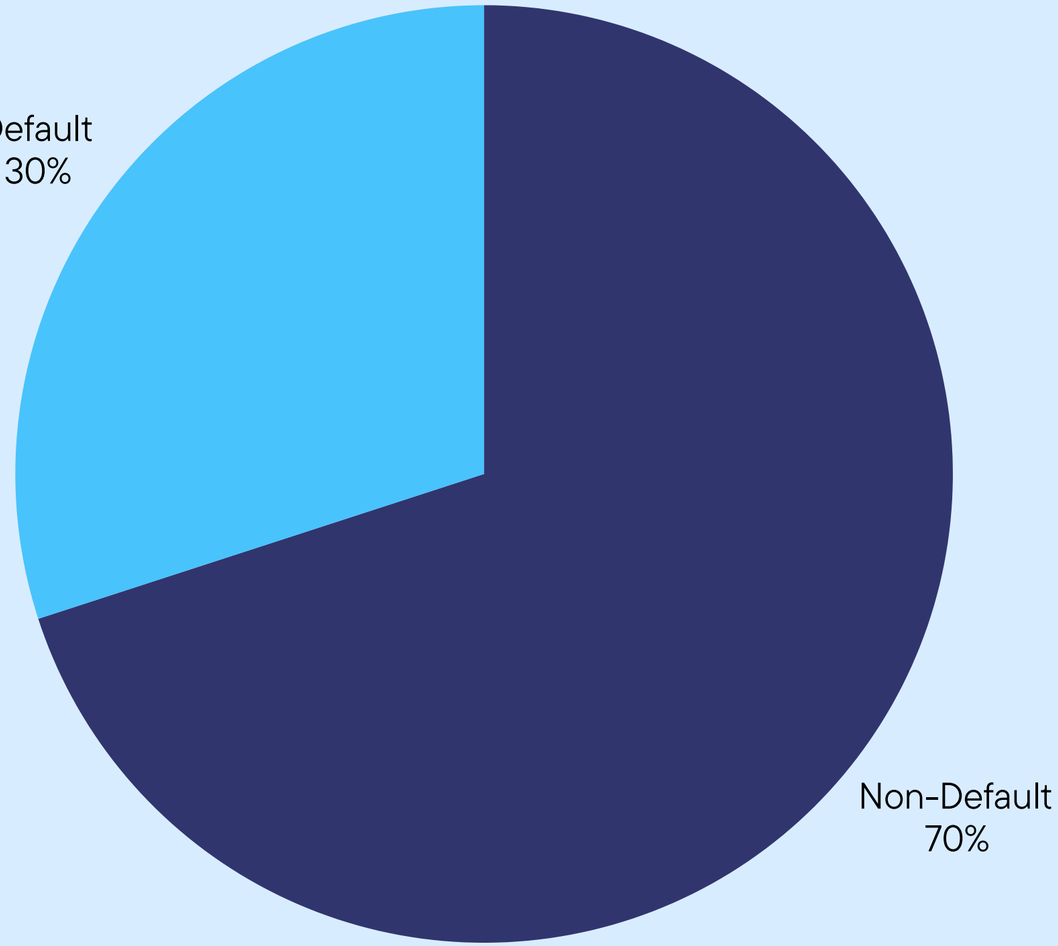
After the model is trained using the 800 entries, it will be tested against the reserved 200 entries. This testing phase is crucial to ascertaining the model's ability to generalize its learning to new, unseen data and verifying its effectiveness and applicability in real-world scenarios.

Back to Agenda

EDA: Imbalance

The dataset's imbalance, with 70% non-defaults and 30% defaults, could bias the model toward predicting non-defaults, potentially overlooking critical default risks. To address this, the Synthetic Minority Over-sampling Technique (SMOTE) will be used. SMOTE generates synthetic default cases, enhancing the minority class's representation to balance the dataset.

This method ensures that the model can accurately generalize across both classes, improving its predictive performance and reducing bias, essential for reliable risk assessment.



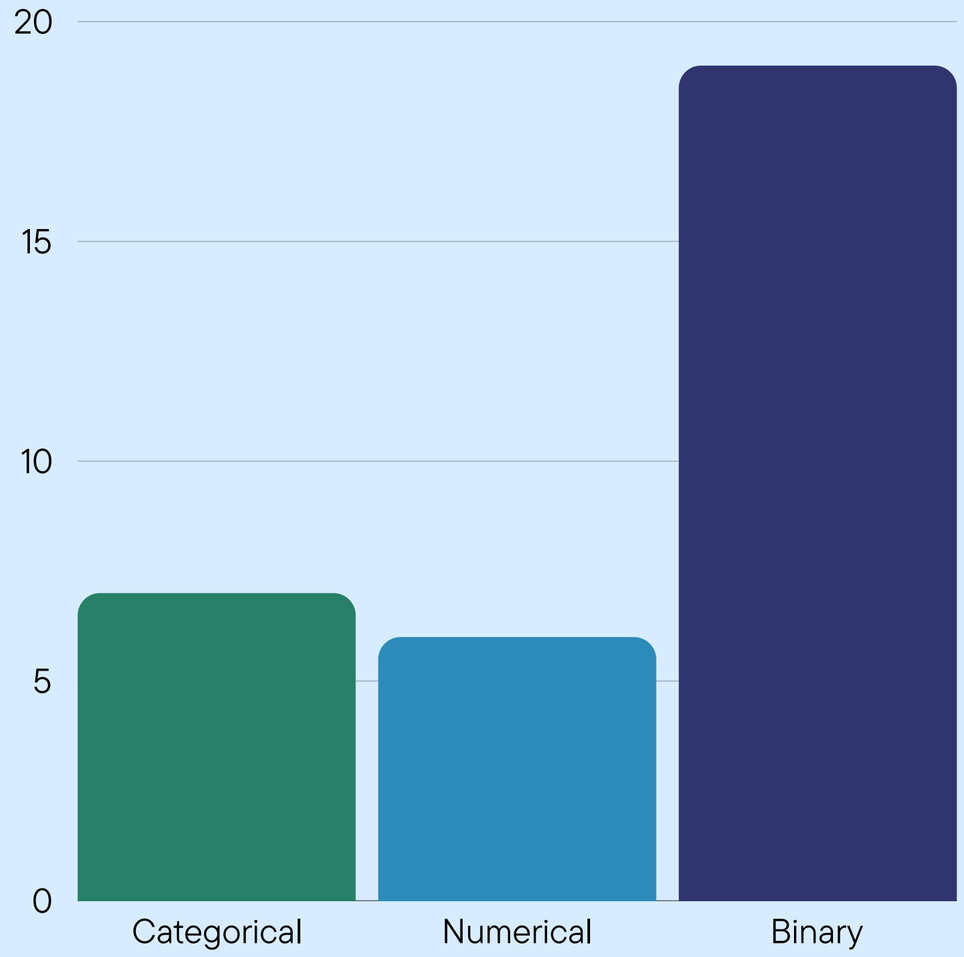
Back to Agenda

EDA: Mixed Data Types

The dataset features 19 binary, seven categorical, and six numerical columns, necessitating specific preprocessing techniques to accurately capture each data type's unique characteristics.

Failing to apply the appropriate preprocessing methods can create substantial challenges in predictive analysis, potentially compromising the model's accuracy and effectiveness.

It is imperative to preprocess each data type correctly to uphold the integrity and reliability of the predictive analysis, thereby enabling the model to generate accurate and valuable predictions.



Back to Agenda

EDA: Solutions

01

Data Preprocessing

Preprocessing is tailored for categorical, numerical, and binary data to enhance model utility. Binary data undergoes minimal preprocessing to maintain its dichotomy. Categorical data is converted to numerical formats, like one-hot encoding, to avoid artificial ordinal interpretations. Numerical data benefits from normalization or standardization, ensuring no variable disproportionately influences the model due to scale differences.

02

Feature Engineering

After preprocessing, logistic regression models were independently applied to numerical, binary, and categorical variables to identify significant predictors within each category. Significant features from all three types were combined into a comprehensive dataset for further analysis, followed by oversampling techniques such as SMOTE to address class imbalance. This comprehensive approach to feature engineering optimizes the predictive model's performance and reliability.

03

Feature Selection

Based on the results from feature engineering, a set of crucial features for predicting the target variable was finalized. This systematic approach ensures that only the most relevant features are retained, enhancing the model's predictive power while reducing complexity and computational overhead.

Back to Agenda

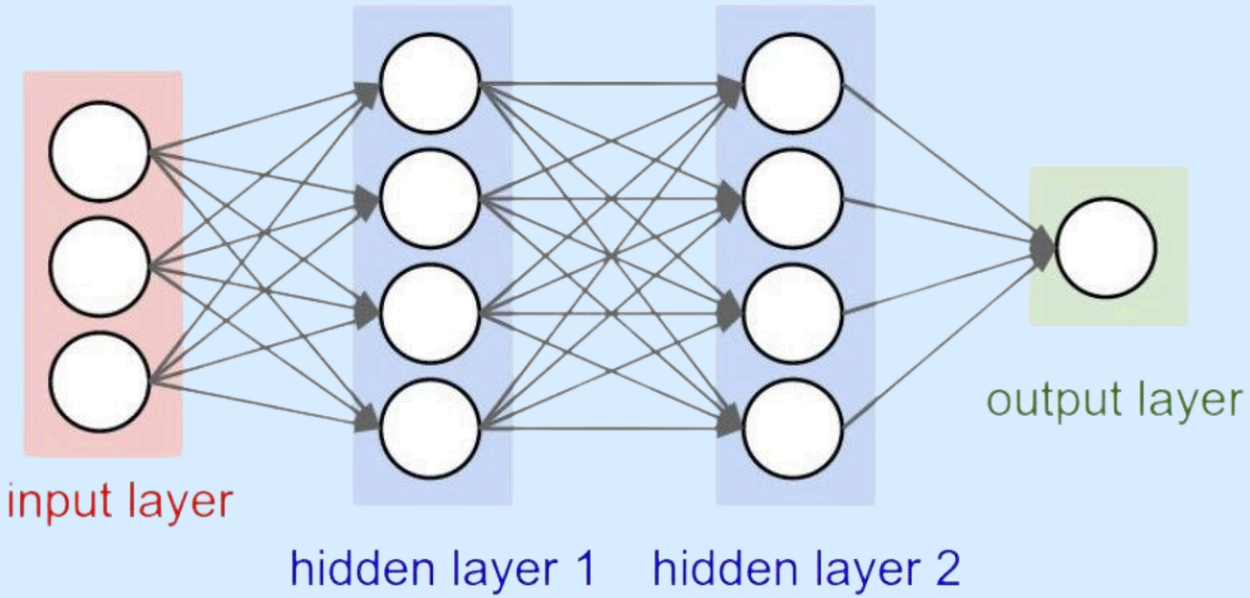
Modeling: Options

	Linear Regression	Random Forest	Keras
Pro	Offers excellent interpretability and speed, ideal for linear relationships in sizable datasets.	Effectively manages non-linear data and diverse variables, providing insightful feature importance.	Keras excels in capturing intricate, non-linear patterns and scales effectively with large datasets and feature sets.
Con	Struggles with complex, non-linear relationships and is sensitive to outliers, potentially oversimplifying data.	It can be computationally demanding and prone to overfitting, with lower interpretability compared to simpler models.	It requires substantial computational power and presents challenges in interpretability and potential overfitting.

Back to Agenda

Modeling: Choice

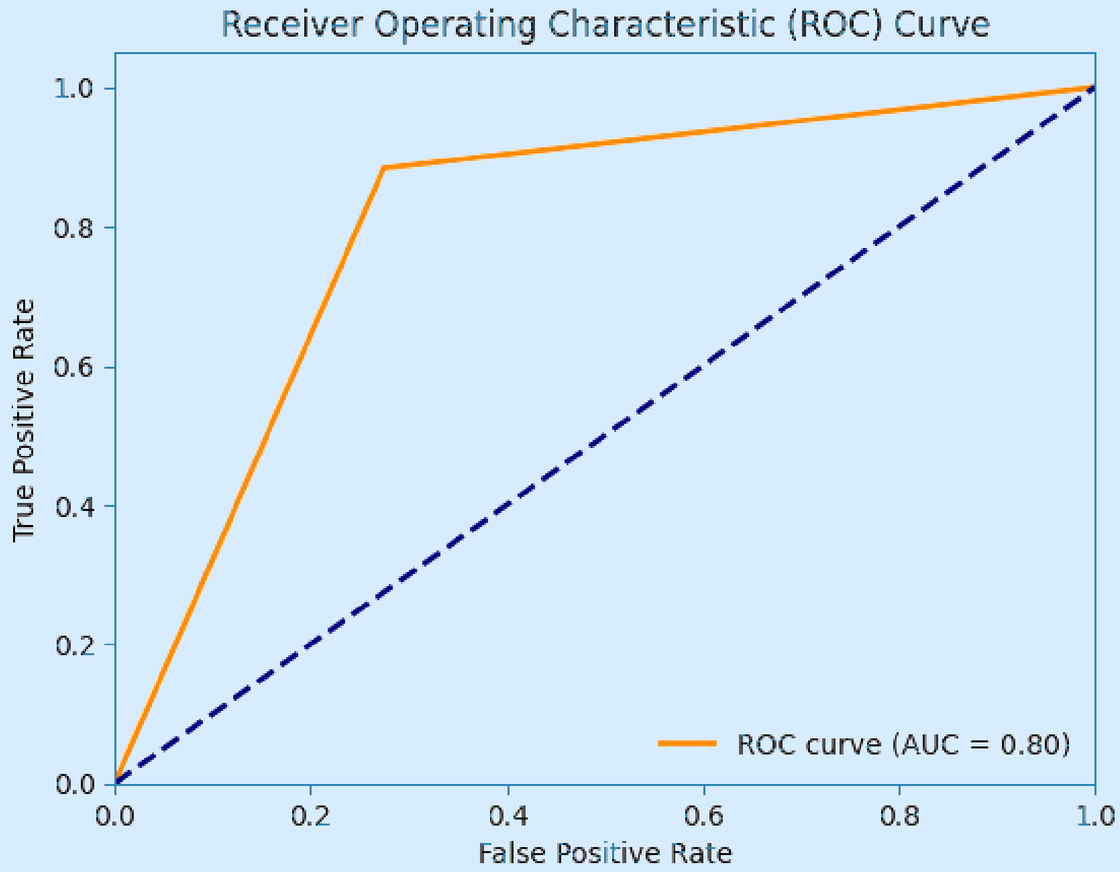
Considering the context, Keras was selected due to its suitability for the dataset's characteristics and alignment with the project's objectives. Its capability to model intricate relationships and accommodate diverse data types surpasses its drawbacks, particularly in the pursuit of leveraging advanced analytics for accurate credit default prediction. Although random forest and linear regression present their own advantages, the data's depth and complexity warrant a more robust model, such as the one provided by Keras.



Model: Evaluation

Back to Agenda

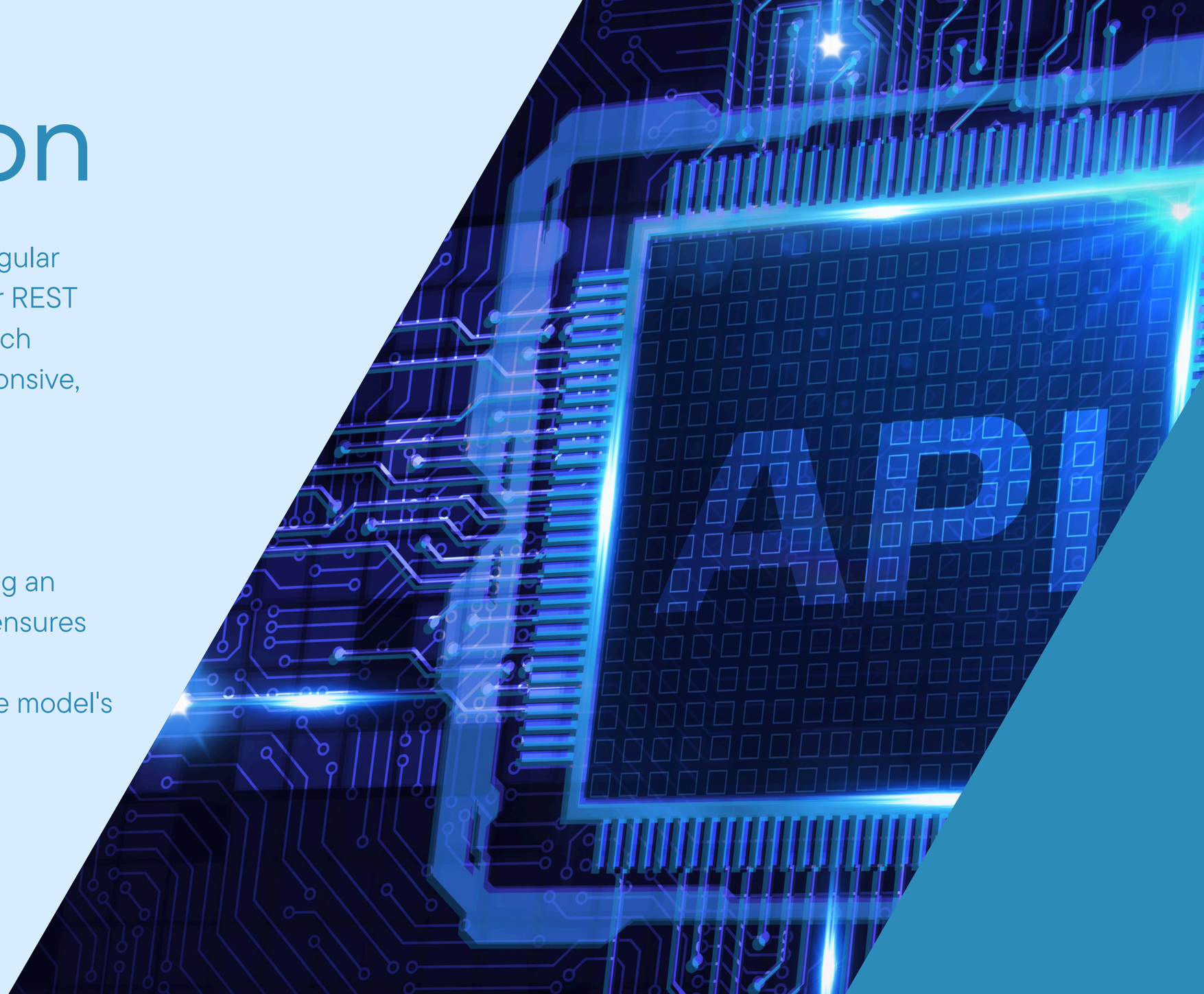
Integrating K-fold cross-validation and ROC curve analysis offers a robust evaluation framework. K-fold cross-validation partitions the dataset, training the model iteratively on various subsets. This reduces overfitting risk and improves generalization. ROC curve analysis examines the model's discriminatory power across thresholds, revealing sensitivity-specificity trade-offs. The area under the ROC curve quantifies predictive performance, with higher AUC indicating better discrimination.



Production

Our deployment strategy prioritizes regular model updates and caching within our REST API framework. This structured approach keeps the model up-to-date and responsive, which is crucial in the dynamic credit landscape.

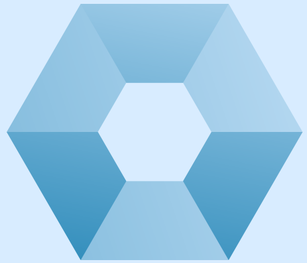
Additionally, proactive monitoring of performance metrics and implementing an alerting mechanism are integral. This ensures swift detection and resolution of any performance decline, safeguarding the model's accuracy and reliability over time.



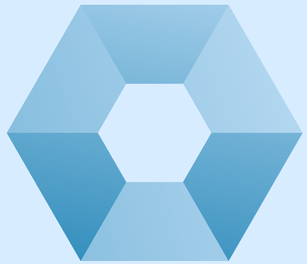
Back to Agenda

Back to Agenda

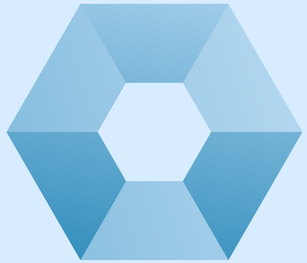
Limitations



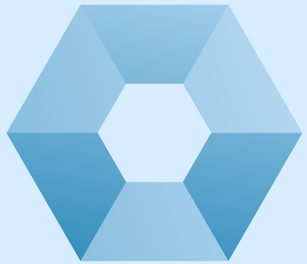
Small Training Set
Limited data necessitates acquiring a larger dataset for training.



Data Imbalance
Severe data imbalance limits analysis for unbiased results.



Multicollinearity Concerns
Multicollinearity is overlooked for a minimally viable product.



Hyperparameter
Hyperparameter changes were deliberately kept to a minimum.

Thank You