DATA 602 - Principles of Data Science

Fall Semester 2024

Targeted Risk Understanding & Scoring Technology (TRUST)

Instructor: Dr. Fardina Alam

Group Members: Arunbh Yashaswi, Swattik Maiti, Eniyan Ezhilan, Ajaykumar Balakannan, Ritik Pratap Singh

Final Tutorial:

`https://TRUSTRRKK.github.io/`

Project Source Code:

`https://github.com/kautilyaa/TRUST/`

Date of Submission: December 15, 2024

Contributions

Below are the contributions of each group member to the project, divided by tasks:

- **A: Project Idea**
  All group members contributed to brainstorming and finalizing the project idea for analyzing Home Credit Loan Default data. Each member participated in discussions and agreed on the approach.

- **B: Data Collection and Dataset Curation**
  Arunbh Yashaswi: Identified the dataset source, ensured data relevance, and curated the initial dataset.
  Swattik Maiti: Compiled data from multiple tables, ensuring consistency and completeness across sources.

- **C: Data Cleaning and Preprocessing**
  Swattik Maiti: Preprocessed the data, handled missing values, and implemented outlier detection and imputation strategies.
  Ritik Pratap Singh, Arunbh Yashaswi: Conducted feature scaling, normalization, and ensured final dataset readiness.

- **D: Exploratory Data Analysis (EDA)**
  Eniyan Ezhilan: Conducted comprehensive exploratory data analysis, generated summary statistics, and created insightful visualizations for feature understanding.
  Ajaykumar Balakannan: Supported EDA by analyzing key distributions, correlations, and identifying potential relationships among features.

- **E: Hypothesis Testing**
  Ajaykumar Balakannan: Formulated and tested hypotheses to validate relationships between features and target variables using statistical methods.
  Eniyan Ezhilan: Conducted ANOVA and Chi-Square tests to identify significant categorical and continuous variables.

- **F: Feature Creation, Engineering, and Feature Selection**
  Swattik Maiti: Designed and implemented key feature engineering strategies to create Amount Financed, Vintage, and Delinquency features.
  Swattik Maiti: Designed and implemented key feature engineering strategies to create Delinquency features.
  Ritik Pratap Singh: Performed feature engineering for Amount Financed features.
  Arunbh Yashaswi: Worked on creating Vintage features, including time-based aggregations and duration calculations.
  Ritik Pratap Singh: Performed recursive feature elimination (RFE), Lasso selection, and Random Forest feature importance for top feature selection.

- **G: Machine Learning Methodology and Architecture - Model Stacking**
  Swattik Maiti: Developed baseline models and ensemble stacking architectures to optimize prediction performance.
  Arunbh Yashaswi: Built and validated meta-models, combining predictions from multiple baseline learners.

- **H: Final Model Curation and Evaluation**
  Swattik Maiti: Trained final models on the curated 'ultimate_op_dataset', fine-tuned hyperparameters, and evaluated performance metrics.
  Arunbh Yashaswi: Conducted cross-validation and ensured robust evaluation across training, validation, and test sets.

- **I: Visualization, Result Analysis, and Conclusion**
  Arunbh Yashaswi: Created compelling visualizations to present results, trends, and insights.
  Ritik Pratap Singh: Interpreted results, derived actionable insights, and drafted the project conclusion connecting findings to the initial objectives.

- **J: Final Report Creation and Documentation**
  Swattik Maiti: Drafted key sections of the report, ensuring consistency and clarity.
  Ritik Pratap Singh: Reviewed and edited the final report for comprehensiveness and professional formatting.

- **K: Additional Contributions**
  Arunbh Yashaswi: Coordinated group activities, ensured timely progress, and managed overall project workflow.
  Ritik Pratap Singh: Assisted in resolving technical challenges, final data cleaning, and ensuring project reproducibility.

Introduction

In today's rapidly evolving financial landscape, accurately predicting credit risk is a significant challenge faced by financial institutions worldwide. Credit risk refers to the possibility of a borrower failing to meet their repayment obligations, leading to financial losses for lenders. To mitigate this risk, credit scoring systems have traditionally relied on historical credit data, such as repayment history and credit utilization. While these methods have proven effective for many borrowers, they fall short when evaluating first-time borrowers or individuals with limited credit histories. This limitation creates significant gaps in financial inclusion, leaving many deserving individuals without access to credit.

The purpose of this project is to analyze loan defaults using a comprehensive dataset provided by Home Credit Group. This dataset captures various financial, demographic, and behavioral attributes of borrowers, offering an opportunity to go beyond traditional credit scoring metrics. By leveraging advanced data science techniques, we aim to uncover meaningful insights and develop predictive models that address the limitations of existing credit risk frameworks.

This study is guided by the following key research questions:

- What are the most significant factors contributing to loan defaults? For example, how do income levels, employment stability, loan tenure, delinquency history, and past loan performance impact repayment behavior?
- How do different machine learning models compare in their ability to accurately predict loan defaults? Which models offer the best trade-off between predictive accuracy and computational efficiency?
- Can non-traditional data sources, such as behavioral and demographic information, improve the accuracy of credit risk predictions when combined with financial metrics?

Answering these questions is critical for several reasons:

- **Financial Inclusion**: By developing more inclusive credit scoring models, financial institutions can expand access to credit for underserved populations, particularly those with limited or no credit history. This fosters greater economic opportunities and reduces inequality.
- **Risk Mitigation**: Improved predictive accuracy helps lenders identify high-risk borrowers more effectively, reducing default rates and minimizing financial losses.
- **Operational Efficiency**: Automation and accuracy in credit risk assessment streamline loan approval processes, saving time and resources for both lenders and borrowers.

The dataset for this study includes a wide range of tables, such as information on loan applications, credit card balances, previous loans, and repayment histories. These tables provide a holistic view of borrower profiles, enabling a thorough analysis of factors influencing loan performance. To ensure robust insights, we employ the full data science lifecycle, including data collection, preprocessing, exploratory analysis, feature engineering, and machine learning model development.

By combining domain knowledge with state-of-the-art analytics, this project seeks to contribute to the development of more accurate, inclusive, and efficient credit scoring systems. The findings from this study have the potential to benefit both financial institutions and borrowers, fostering trust, reducing risks, and promoting sustainable growth in the lending industry.

Dataset Overview and Structure

The dataset used in this analysis is the Home Credit Default Risk dataset, sourced from Kaggle (Home Credit Default Risk Dataset). It consists of multiple interconnected tables providing static, historical, and behavioral data for loan applicants. Figure 1 illustrates the relationships between these tables.

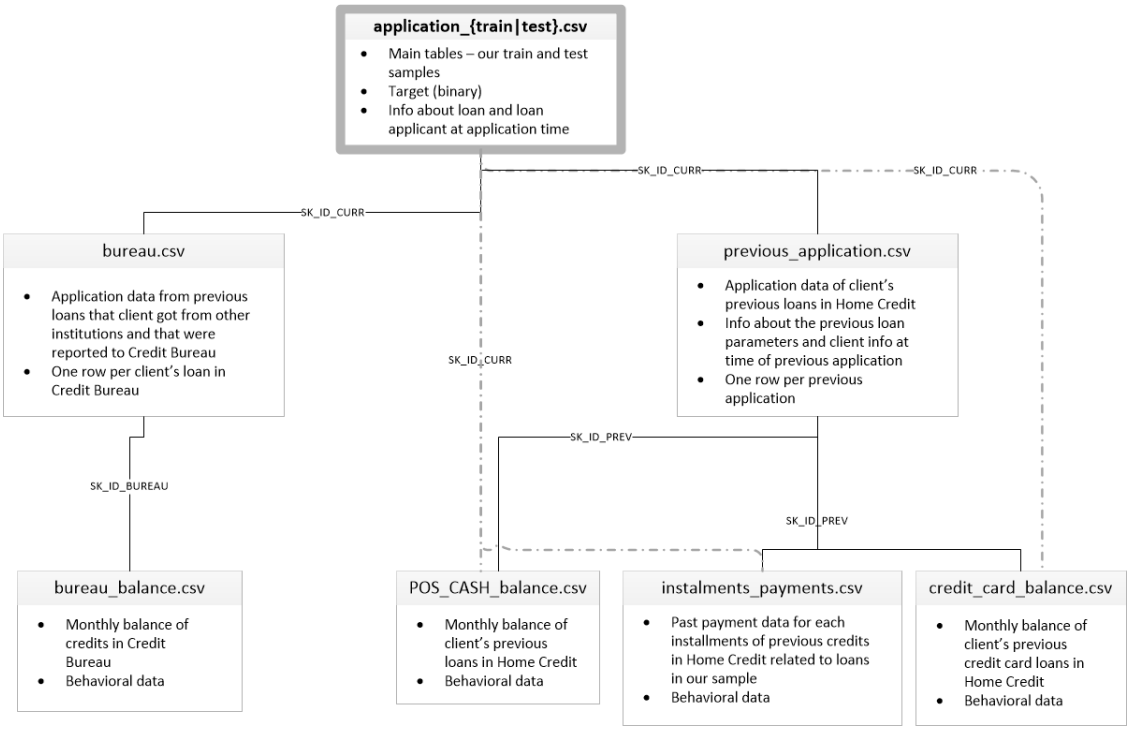


Figure 1: Dataset Structure and Relationships.

Description of Key Tables

- **application_{train—test}.csv:** Contains static data for all applications, split into Train (with TARGET column) and Test (without TARGET column). Each row represents one loan application.
- **bureau.csv:** Provides records of clients' previous credits reported to Credit Bureau. One row corresponds to one credit.
- **bureau_balance.csv:** Includes monthly balances for previous credits in Credit Bureau, with one row per month of history for each credit.
- **POS_CASH_balance.csv:** Contains monthly balance snapshots of previous POS (point of sale) and cash loans with Home Credit.
- **credit_card_balance.csv:** Offers monthly balance snapshots for clients' previous credit cards with Home Credit.
- **previous_application.csv:** Details all previous Home Credit loan applications by clients who have loans in our sample. Each row represents one previous application.

- **installments_payments.csv:** Provides repayment history for previously disbursed credits, including rows for payments made and missed.
- **HomeCredit_columns_description.csv:** A metadata file describing columns in all tables.

Data Curation and Cleaning

Low Fill Rate Columns: Columns with a fill rate less than 50% , such as OWN_CAR_AGE and EXT_SOURCE_1, were dropped. Retaining these columns would introduce unnecessary noise and bias during modeling.

Handling Missing Values: For columns with missing rates above 50% but less than 100%:

- **AMT_ANNUITY:** Imputed using predictive modeling with correlated features like AMT_CREDIT and AMT_GOODS_PRICE.
- **AMT_GOODS_PRICE:** Median imputation was employed, given the right-skewed distribution of values.
- **DAYS_LAST_PHONE_CHANGE:** Missing values were imputed using the mean of the column after observing that it was evenly distributed across groups.
- **EXT_SOURCE_2:** Imputed using the mean value, as it demonstrated strong correlations with the target variable and other features.

Outliers: Outliers, such as extreme values in AMT_INCOME_TOTAL (e.g., maximum of 117,000,000), were excluded to avoid distortions in analysis. Specifically:

- **AMT_CREDIT:** Values greater than the 99th percentile were capped to reduce the impact of extreme values.
- **DAYS_EMPLOYED:** Large positive values (indicating anomalies) were replaced with **NaN** and subsequently imputed with the median value.

Encoding Categorical Features: Categorical features were processed to ensure compatibility with machine learning models:

- **NAME_CONTRACT_TYPE, CODE_GENDER, NAME_EDUCATION_TYPE:** Label encoding was applied for binary variables.
- **OCCUPATION_TYPE:** One-hot encoding was used for multiclass categorical variables, as they provided meaningful separation during feature importance analysis.
- **ORGANIZATION_TYPE:** Categories were grouped into broader buckets based on frequency counts to reduce dimensionality.

Imputation Techniques

- **Predictive Imputation:**
  - LightGBM was used to estimate missing values in AMT_ANNUITY by leveraging correlated columns like AMT_CREDIT, DAYS_BIRTH, and AMT_GOODS_PRICE.
- **Median Imputation:**
  - Applied to AMT_GOODS_PRICE due to its highly skewed distribution.
  - Used for DAYS_EMPLOYED to handle anomalies flagged as outliers.
- **Mode Imputation:**
  - Categorical variables like DEF_30_CNT_SOCIAL_CIRCLE and NAME_FAMILY_STATUS were imputed with the most frequent value in their respective columns.
- **Feature Engineering for Imputation:**
  - New features like RATIO_AMT_CREDIT_TO_GOODS_PRICE were derived to simplify imputations and reduce collinearity.
  - DAYS_LAST_PHONE_CHANGE_RATIO: Created by dividing the DAYS_LAST_PHONE_CHANGE column by DAYS_BIRTH to normalize the values.

Handling Duplicates and Redundant Data:

- Duplicate rows in bureau.csv and previous_application.csv were identified and dropped to reduce data redundancy.
- Features with high multicollinearity (e.g., AMT_CREDIT and CREDIT_LIMIT) were checked, and one of the correlated columns was removed to simplify the model.

Database Setup

The cleaned dataset was organized into a structured format compatible with pandas DataFrames and SQL for querying and further analysis. Data validation, including distribution analysis and correlation checks, ensured the dataset's readiness for modeling.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to uncover insights about the dataset's distribution, identify patterns, and assess relationships between features and the target variable. This step is critical to prepare data for modeling and ensure that important underlying trends and issues (like skewness and bias) are addressed.

Feature Distributions and Skewness

Understanding the distribution of features was an essential part of the EDA. Several continuous variables in the dataset exhibited significant skewness:

- **AMT_INCOME_TOTAL**: The income data was highly right-skewed, with most values concentrated below 500,000 but outliers exceeding 10 million. Logarithmic transformation was applied to normalize this feature and make it suitable for linear models.
- **AMT_CREDIT**: Although less extreme than income, the credit amount showed slight positive skewness. Visualization confirmed a long tail that could impact model performance.
- **DAYS_EMPLOYED**: This column had anomalous values such as 365243, which distorted the distribution. These were treated as missing values and imputed with the median to reduce skewness.

Categorical features like NAME_CONTRACT_TYPE and NAME_EDUCATION_TYPE showed no skewness concerns, as their distributions were relatively balanced across categories.

Approaches to Address Skewness

- **Logarithmic Transformation:** Applied to AMT_INCOME_TOTAL and AMT_CREDIT to reduce their skewness and stabilize variance.
- **Winsorization:** Outliers in AMT_GOODS_PRICE were capped at the 99th percentile to prevent extreme values from influencing downstream models.

Bias-Variance Analysis

Bias-variance analysis was conducted to understand the trade-offs between underfitting and overfitting. Using validation curves and residual plots, we assessed model performance across key features:

- **EXT_SOURCE_2:** A highly predictive variable with a linear relationship to the target variable (TARGET). No transformation was necessary, but regularization was considered to prevent overfitting.
- **DAYS_BIRTH:** Exhibited a complex, nonlinear relationship with TARGET. Polynomial features were added to capture this nonlinearity, balancing bias and variance.
- **NAME_FAMILY_STATUS and OCCUPATION_TYPE**: High cardinality in OCCUPATION_TYPE led to overfitting when one-hot encoding was applied. Grouping categories into broader buckets (e.g., "Blue-Collar," "White-Collar") helped reduce model variance.

Validation Observations

Residual plots for logistic regression showed that features with high variance, such as DAYS_EMPLOYED, introduced noise. Conversely, using engineered features like RATIO_AMT_CREDIT_TO_GOODS_PRICE improved prediction consistency.

Hypothesis Testing

To validate the significance of relationships between features and the target variable, several hypothesis tests were conducted:

- **Chi-Square Test for Independence:**
  - Categorical variables like NAME_CONTRACT_TYPE and CODE_GENDER were tested for independence with the target variable.
  - Results indicated that both variables were statistically significant ($p < 0.01$).
- **Kolmogorov-Smirnov Test:**
  - Applied to EXT_SOURCE_1, EXT_SOURCE_2, and EXT_SOURCE_3, which had high correlation with TARGET.
  - The test confirmed that the distributions of these variables were significantly different between default and non-default groups.
- **T-Test for Means:**
  - Conducted on continuous variables like AMT_INCOME_TOTAL and AMT_CREDIT.
  - Results showed significant differences in means between default and non-default groups, confirming their predictive power.
- **ANOVA (Analysis of Variance):**
  - Performed on OCCUPATION_TYPE to analyze the variation in default rates across different occupations. The results showed significant differences, indicating that certain occupations are more likely to default.

Feature Correlations

A heatmap of Pearson correlations between numerical features revealed the following insights:

- **EXT_SOURCE_1, EXT_SOURCE_2, and EXT_SOURCE_3**: Strong negative correlations with the target variable (TARGET), indicating that higher scores are associated with lower default risk.
- **AMT_GOODS_PRICE and AMT_CREDIT**: Moderately positively correlated, reflecting a logical relationship between loan size and the price of goods.
- Multicollinearity was observed between AMT_ANNUITY and AMT_CREDIT, which was addressed through dimensionality reduction techniques.

Conclusions from EDA

The EDA process revealed key insights and actionable steps:

- Addressing skewness and handling outliers ensured that continuous variables were suitable for modeling.
- Categorical encoding strategies, combined with dimensionality reduction, reduced noise and overfitting risks.
- Hypothesis testing validated the predictive power of features like EXT_SOURCE_2 and DAYS_BIRTH, guiding feature selection.

These findings informed subsequent feature engineering and model-building steps.

Feature Creation and Feature Engineering

Feature engineering is a critical step in machine learning and data analysis, where raw data is transformed into meaningful features to improve model performance. The goal is to create features that capture the underlying patterns in the data and provide the model with relevant inputs for accurate predictions.

Why Feature Engineering?

Raw data often contains noise, redundancy, and missing values. Effective feature engineering helps to:

- Improve model interpretability by creating features that represent meaningful relationships.
- Enhance predictive power by reducing noise and emphasizing critical attributes.
- Address issues like missing data and inconsistencies, ensuring the data is ready for modeling.

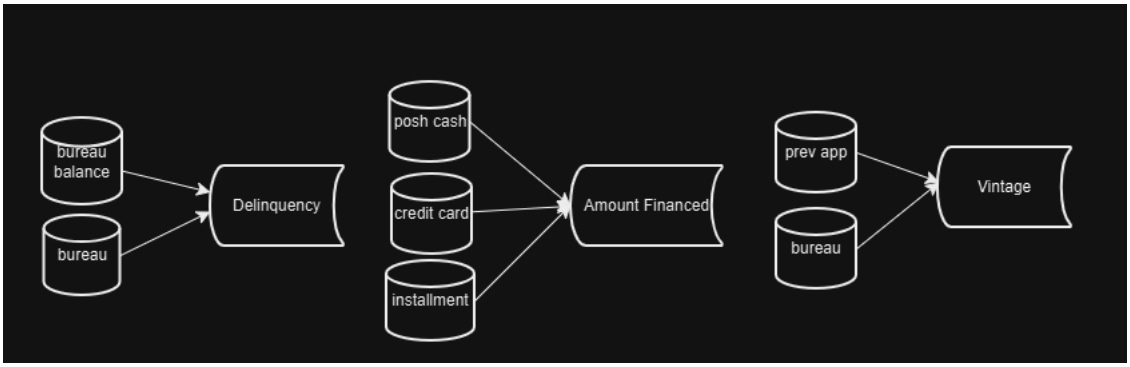
In this project, feature engineering played a pivotal role in consolidating data from multiple interconnected tables and generating key attributes that summarize clients' financial and behavioral characteristics. These features were engineered at the `SK_ID_CURR` level, ensuring that each client was represented by a single row in the final dataset.

How Feature Engineering Helped

By aggregating data from different sources (e.g., bureau, POS cash, credit cards), feature engineering helped us:

- Capture critical metrics like the total amount financed, repayment behaviors, and historical loan performance.
- Create time-based features (vintage) to understand the duration of financial activities and credit history.
- Highlight past delinquencies, providing insights into clients' risk profiles.
- Prepare a clean and unified dataset with informative features for training the predictive model.

Implementation of Feature Engineering



We created three primary categories of features: **Amount Financed**, **Vintage (Time Duration)**, and **Delinquency**. Features were aggregated from multiple tables using functions like `max`, `min`, `mean`, and `sum`.**In the end we had over 600+ custom aggregated features**.

1. Amount Financed

This category focuses on the actual amount of money the clients borrowed across various loans throughout his credit history, we created approx 120 features and below are few examples of them:

- **POS_CASH_balance.csv:** Aggregated monthly balances of POS(point of sale) and cash loans:
  - `TOTAL_POS_AMT`: Sum of all POS cash balances.
  - `AVG_POS_AMT`: Mean balance across months.
- **credit_card_balance.csv:** Captured monthly balances of credit card loans:
  - `MAX_CREDIT_CARD_AMT`: Maximum credit card balance.
  - `RATIO_CREDIT_USED`: Ratio of credit used to credit limit.
- **installments_payments.csv:** Summarized repayment amounts:
  - `SUM_INSTALLMENT_PAID`: Total repayment amount.
  - `AVG_INSTALLMENT_PAID`: Average installment payment.

2. Vintage (Time Duration)

Vintage features represent performance metrics of loans, accounts, or customers over time, segmented by their origination period (e.g., month or year), to track and compare risk, delinquency, or behavioral trends across cohorts, we created over 200+ features and below are few examples of them:

- **bureau.csv:** Duration of loans from external institutions:
  - `MAX_LOAN_DURATION`: Maximum duration of external loans.
  - `AVG_LOAN_DURATION`: Mean duration of all loans.
- **previous_application.csv:** Time gaps and durations of past Home Credit applications:
  - `MIN_TIME_SINCE_LAST_APPLICATION`: Minimum time gap between previous loans.
  - `AVG_LOAN_VINTAGE`: Average duration of previous loans.

3. Delinquency

Delinquency features capture past performance(success and failures) in repaying loans, we created over 360 features and below are few examples of them:

- **bureau.csv:** Extracted delinquency patterns:
  - `COUNT_LOANS_DELINQUENT`: Total number of delinquent loans.
  - `MAX_DELINQUENCY_DAYS`: Maximum days past due for any loan.

- **bureau_balance.csv:** Monthly bureau balance data:
  - `SUM_DELINQUENT_MONTHS`: Total delinquent months.
  - `RATIO_DELINQUENT_MONTHS`: Ratio of delinquent months to total credit history.

Aggregation at `SK_ID_CURR` Level

To ensure each client had a single row summarizing their financial and behavioral data, features were aggregated at the `SK_ID_CURR` level using:

- **Aggregation Functions:** `max`, `min`, `mean`, and `sum`.
- **Ratios:** Derived relationships such as `RATIO_AMT_CREDIT_TO_INCOME`.
- **Time-Based Features:** Calculated differences between key dates to derive durations.

Benefits of Feature Engineering

The engineered features provided critical insights into client risk profiles:

- High delinquency ratios were strong indicators of default risk.
- Time-based features revealed clients' financial history and stability.
- Consolidating financial activities across tables ensured a holistic view of each client.

Feature engineering ensured the dataset was comprehensive and ready for predictive modeling, forming the backbone of our analysis.

Approach and Machine Learning Analysis

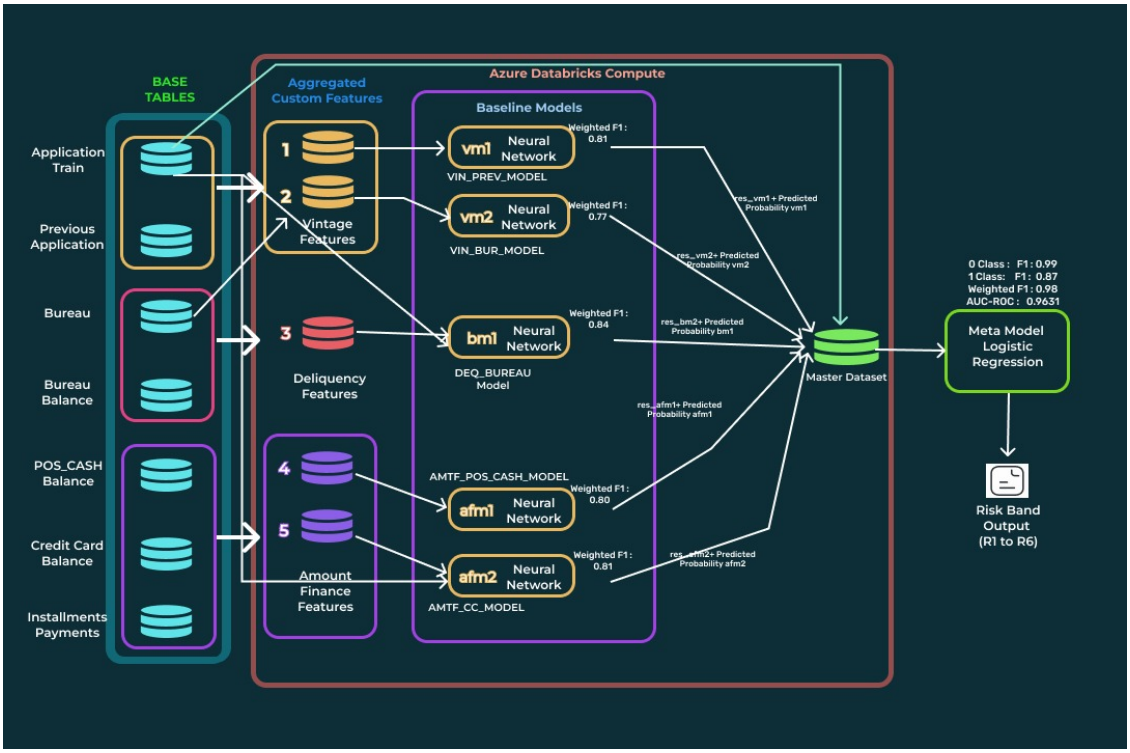
In this section, we outline the primary machine learning approach used to address the objectives defined in the introduction. Based on the insights gained from exploratory data analysis and feature engineering, we adopted a robust and systematic methodology for prediction tasks, leveraging the power of model stacking and neural networks to handle the complexity and imbalance in the dataset.

Algorithm Selection

Given the highly imbalanced nature of the target variable (loan default), our primary focus was on implementing techniques that effectively handle such data distribution while avoiding overfitting. After experimenting with various machine learning algorithms, we concluded that a **model stacking approach(custom ensemble of models)**, combined with neural networks, would provide the best results. This decision was based on the following considerations:

- **Handling Imbalanced Data:** Traditional machine learning models often struggle with imbalanced data. To address this, we tested algorithms like Gradient Boosting (e.g., LightGBM, XGBoost), Random Forest, and Logistic Regression, which incorporate mechanisms for imbalance handling.
- **Avoiding Multicollinearity:** Since our features were derived from multiple related tables, multicollinearity could skew the model's predictions. By creating distinct baseline models for each feature group, we minimized feature interactions that could introduce noise.
- **Stacking for Improved Generalization:** Model stacking allowed us to combine the predictive power of multiple baseline models, ensuring robust predictions and reducing overfitting by leveraging the strengths of each model.

Our Approach



The workflow of our approach can be summarized as follows:

1. **Feature Groups and Baseline Models:** We divided the engineered features into three distinct groups based on their origin and type:

   - **Amount Financed Features:** Derived from `POS_CASH_balance`, `credit_card_balance`, and `installments_payments` tables.
   - **Delinquency Features:** Derived from `bureau` and `bureau_balance` tables.
   - **Vintage Features:** Derived from `previous_application` and `bureau` tables.

   For each feature group, we built separate baseline models, experimenting with various algorithms such as LightGBM, Random Forest, Logistic Regression, Support Vector Machines and Neural Network. **Finally we decided to proceed with Neural Network for baseline model creation**.

2. **Baseline Models Creation and Evaluation:** After Creating the Baseline models we got the following results:

   - **Amount_Financed_Features:** Derived from `POS_CASH_balance`, `credit_card_balance`, and `installments_payments` tables.

     Amount Finance Credit Card:

Model: "sequential"

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 256)	16,384
dense_1 (Dense)	(None, 128)	32,896
dropout (Dropout)	(None, 128)	0
dense_2 (Dense)	(None, 64)	8,256
dropout_1 (Dropout)	(None, 64)	0
dense_3 (Dense)	(None, 1)	65

Total params: 57,601 (225.00 KB)

Trainable params: 57,601 (225.00 KB)

Non-trainable params: 0 (0.00 B)

	precision	recall	f1-score	support
0	0.94	0.80	0.86	15874
1	0.18	0.49	0.27	1507
accuracy			0.77	17381
macro avg	0.56	0.64	0.57	17381
weighted avg	0.88	0.77	0.81	17381

Amount Finance Pos Cash:

Model: "sequential_1"

Layer (type)	Output Shape	Param #
dense_4 (Dense)	(None, 256)	14,080
dense_5 (Dense)	(None, 128)	32,896
dropout_2 (Dropout)	(None, 128)	0
dense_6 (Dense)	(None, 64)	8,256
dropout_3 (Dropout)	(None, 64)	0
dense_7 (Dense)	(None, 1)	65

Total params: 55,297 (216.00 KB)

Trainable params: 55,297 (216.00 KB)

Non-trainable params: 0 (0.00 B)

	precision	recall	f1-score	support
0	0.94	0.80	0.87	53164
1	0.15	0.40	0.22	4724
accuracy			0.77	57888
macro avg	0.55	0.60	0.54	57888
weighted avg	0.87	0.77	0.81	57888

- **Delinquency Features:** Derived from `bureau` and `bureau_balance` tables.
  Delinquency:

Model: "sequential"

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 256)	65,792
dense_1 (Dense)	(None, 128)	32,896
dropout (Dropout)	(None, 128)	0
dense_2 (Dense)	(None, 64)	8,256
dropout_1 (Dropout)	(None, 64)	0
dense_3 (Dense)	(None, 1)	65

Total params: 107,009 (418.00 KB)

Trainable params: 107,009 (418.00 KB)

Non-trainable params: 0 (0.00 B)

	precision	recall	f1-score	support
0	0.94	0.85	0.89	48625
1	0.16	0.35	0.23	4074
accuracy			0.81	52699
macro avg	0.55	0.60	0.56	52699
weighted avg	0.88	0.81	0.84	52699

- **Vintage Features:** Derived from `previous_application` and `bureau` tables.

  Vintage Bureau:

	precision	recall	f1-score	support
0.0	0.94	0.73	0.82	48648
1.0	0.12	0.42	0.18	4051
accuracy			0.71	52699
macro avg	0.53	0.58	0.50	52699
weighted avg	0.87	0.71	0.77	52699

  Vintage Prev:

	precision	recall	f1-score	support
0.0	0.93	0.81	0.87	53432
1.0	0.12	0.30	0.18	4780
accuracy			0.77	58212
macro avg	0.53	0.56	0.52	58212
weighted avg	0.86	0.77	0.81	58212

3. **Stacking and Integration:** The residual and Predicted Probabilities from each baseline gives model is used as an input to train our final model(meta model).We also used the top features from every feature group in training of our final model. These six columns (two each from amount financed, delinquency, and vintage models) were appended to the original `application_train` dataset to create the master dataset.

4. **Top Features Selection:** For selecting top features from baseline models we have used the following techniques:

   - **Lasso Feature Selection**
   - **Recursive Feature elimination**

- **Random forest feature importance**

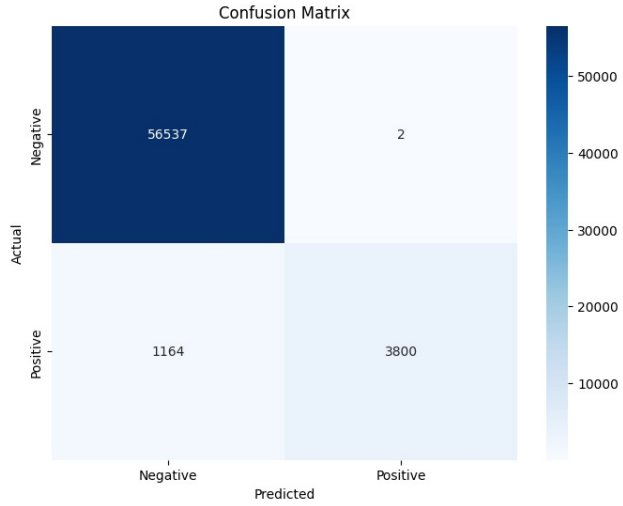
5. **Logistic Regression for Final Predictions:** The augmented dataset, now containing demographic features,top features from each aggregated feature group, the target variable, and the six prediction columns, was fed into a Logistic Regression model. Logistic Regression model is a simple ML model which help us to get coefficients against each variable which help us identify the most important features contributing to the target prediction.This enhances the explain ability if the model for business stakeholders.

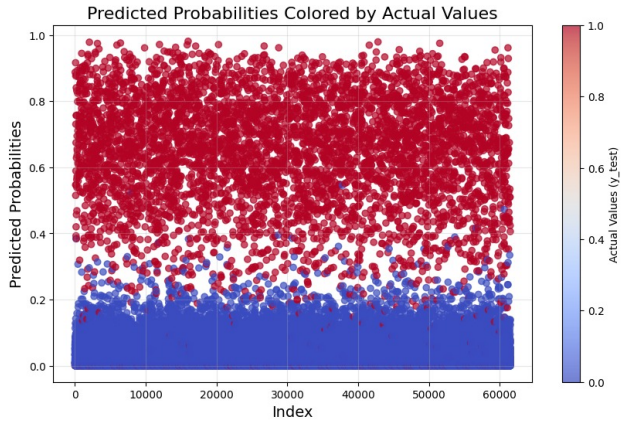
6. **Final Model Evaluation and Visualization:** Performance metrics like Area Under the Curve (AUC), F1-score, and Precision-Recall curves were used to evaluate the models. The logistic Regression, enhanced by stacking, demonstrated the best performance across all metrics.

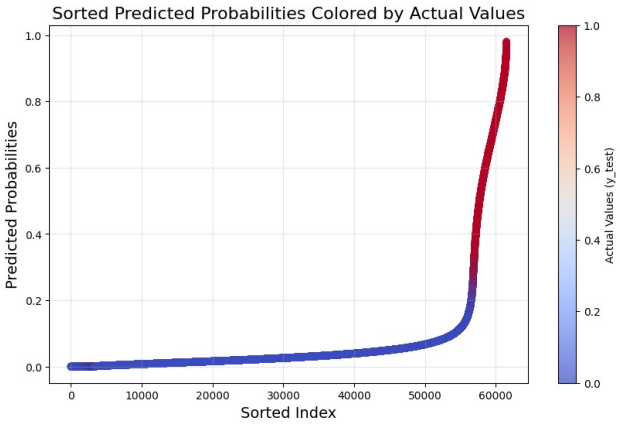
   Below are the evalution results for logistic regression meta model:

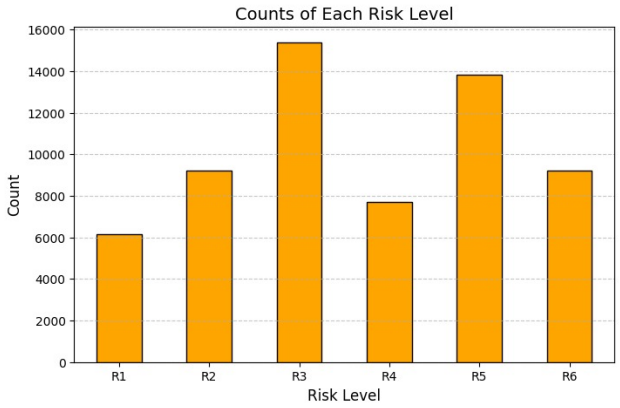
	precision	recall	f1-score	support
0	0.98	1.00	0.99	56539
1	1.00	0.77	0.87	4964
accuracy			0.98	61503
macro avg	0.99	0.88	0.93	61503
weighted avg	0.98	0.98	0.98	61503

AUC-ROC: 0.963172198110055









Summary

Our approach effectively leveraged the strengths of different machine learning techniques. By creating separate baseline models for distinct feature groups, we ensured better interpretability and avoided potential collinearity issues. Stacking these models and feeding the results into a Logistic Regression allowed us to achieve a high-performing solution capable of handling complex data relationships and imbalances in the dataset.

Insights and Conclusions

This project aimed to predict loan default probabilities using the Home Credit Default Risk dataset, a challenging task due to the highly imbalanced nature of the data. Through careful feature creation, thoughtful model selection, and advanced techniques like model stacking, we successfully tackled the complexities of this dataset and gained meaningful insights.

The Story of Our Approach

The key to our success lay in the structured feature creation process and the use of model stacking. By engineering features from multiple interconnected tables, we captured the nuances of each applicant's financial behavior. For instance:

- **Amount Financed:** Features derived from POS cash, credit card balances, and installment payments reflected the applicant's borrowing and repayment patterns.
- **Delinquency:** Features from bureau and bureau balance tables highlighted the applicant's past performance with external loans.
- **Vintage (Time Duration):** Historical trends from previous loan applications and bureau data revealed the applicant's experience and consistency in borrowing over time.

To deal with the data's imbalance, we created three independent feature groups (amount financed, delinquency, and vintage) and built separate baseline models for each. This allowed us to capture relationships specific to each feature type while avoiding the noise and collinearity that arise when all features are used together.

The outputs from these baseline models, representing probabilities for each feature group, were then stacked together with demographic data from the main table. This enriched dataset was fed into a logistic Regression, which captured non-linear relationships and refined the predictions further. This multi-stage approach allowed us to leverage the strengths of different algorithms while mitigating their individual weaknesses.

How It Helped and What We Learned

The feature creation and model stacking process significantly improved our results. Breaking down the data into logical feature groups allowed for better interpretability and reduced the risk of overfitting. By stacking models, we combined the predictive power of different algorithms, leading to a robust solution.

Key takeaways include:

- **Impact of Feature Engineering:** Thoughtful feature engineering from raw data tables enabled us to extract meaningful patterns that directly improved model performance.
- **Handling Imbalance:** The stacking approach, combined with the neural network, effectively addressed the challenge of imbalanced data, ensuring that minority class predictions were not overlooked.
- **Generalizability:** The final logistic Regression, trained on enriched data, provided consistent and reliable predictions across the test dataset.

Concluding Thoughts

This project demonstrates the power of combining domain knowledge with advanced machine learning techniques. An uninformed reader can appreciate how structured data processing, careful feature engineering, and a multi-stage modeling approach can unravel insights from even the most complex datasets. For those familiar with the topic, this project highlights the benefits of model stacking and the importance of leveraging data relationships to overcome challenges like imbalance.

In conclusion, our approach not only provided accurate predictions but also shed light on the importance of financial behavior, past performance, and historical trends in assessing credit risk. These insights can be valuable for institutions seeking to make informed decisions about lending and risk management.

Data Science Ethics

When working with sensitive financial data such as the Home Credit Default Risk dataset, it is essential to consider the ethical implications of the analysis to ensure fairness, transparency, and the avoidance of harm to individuals. Below, we address the potential ethical concerns related to our project and the steps we took to mitigate them.

Potential Ethical Concerns

1. Bias in Data Collection The dataset may reflect biases in historical loan approvals and defaults. The dataset consist of only 8% of people defaulting which is target variable 1.Also in other instance, certain demographic groups may have been disproportionately represented in loan applications or repayment histories, leading to potential biases in the data. This can result in models that inadvertently perpetuate or amplify these biases.

2. Discrimination in Predictions If the model assigns higher default probabilities to certain groups based on sensitive attributes like age, gender, or education level, it could lead to unfair treatment. Such discrimination could reinforce stereotypes and disadvantage specific populations.

3. Imbalance in Data The dataset's imbalance between default and non-default cases poses a risk of favoring the majority class (non-defaults), potentially ignoring the patterns and characteristics of the minority class (defaults). This could result in a model that fails to detect high-risk cases accurately.

4. Transparency and Interpretability Black-box models, such as neural networks, can make it difficult to interpret predictions and understand why certain decisions were made. This lack of transparency can reduce trust in the system.

Mitigation Strategies

1. Addressing Bias in Data To minimize potential biases, we carefully analyzed the dataset to ensure that sensitive demographic features (e.g., gender, age, marital status) were not directly used in the modeling process. Instead, we focused on behavioral and financial features, such as repayment history and credit utilization, that are more directly related to credit risk.

2. Ensuring Fairness in Predictions We conducted fairness checks to ensure that the model's predictions were not disproportionately skewed against specific groups. By analyzing the performance of our model across different demographic segments, we verified that the model treated individuals fairly regardless of sensitive attributes.

3. Handling Imbalanced Data To address the imbalance in the dataset, we employed techniques such as stratified sampling and cost-sensitive learning. Additionally, model stacking allowed us to incorporate diverse algorithms, ensuring that the minority class was adequately represented and learned.

4. Promoting Transparency While neural networks were used in our final model, we mitigated the transparency issue by providing feature-level insights using baseline models. These simpler models (e.g., logistic regression, decision trees) offered interpretable intermediate outputs that complemented the neural network's predictions. This two-stage approach enhanced trust in the results.

Ethical Responsibility in Loan Predictions

Recognizing the potential societal impact of credit risk models, we adhered to the principles of fairness, accountability, and transparency throughout this project. By prioritizing behavioral and financial features over sensitive demographic attributes, we aimed to ensure that the model's predictions were objective and equitable.

Concluding Thoughts

Ethical considerations are central to responsible data science practices. This project highlights the importance of recognizing and addressing biases, ensuring fairness in predictions, and maintaining transparency in machine learning applications. By taking these steps, we not only improved the quality of our analysis but also ensured that our approach aligns with the broader goal of promoting fairness and equity in financial decision-making.