Tech-GB 2336

Data Science for Business: Technical

by Prof. Chris Vollinsky

Home Credit Default Risk

Final Project Report

Kaartikeya Panjwani (kp3291)
Aditya Mittal (am13294)
Saurabh Paul (sp6296)
Teja Reddy (sa8238)

1 Business Understanding

1.1 Objective

The financial services sector faces a critical challenge in loan allocation, characterized by high rejection rates primarily stemming from insufficient credit histories. This systemic issue creates a significant barrier for potential borrowers, particularly those with limited or no formal credit documentation. Consequently, many individuals are compelled to seek financial assistance from unregulated and potentially predatory lenders, exposing themselves to exorbitant interest rates and precarious financial conditions.

The problem extends beyond individual hardship, representing a broader economic inefficiency. Traditional lending models rely heavily on conventional credit scoring mechanisms, which systematically exclude a substantial segment of the population with limited or alternative financial documentation. This approach not only constrains individual financial opportunities but also restricts economic mobility and financial inclusion.

Innovative solutions are emerging to address this challenge through advanced data analytics and comprehensive risk assessment methodologies. One such approach, exemplified by institutions like Home Credit, leverages expansive data ecosystems, including telecommunications and transactional data, to develop more nuanced and holistic credit evaluation frameworks.

The proposed system aims to create a transformative approach to loan underwriting that balances institutional risk management with expanded financial accessibility. Through sophisticated predictive modeling, lending institutions can more accurately assess an applicant's loan repayment capabilities, thereby creating a more equitable and efficient financial ecosystem.

1.2 Impact

The implementation of advanced data analytics in loan underwriting represents a transformative approach to financial inclusion, challenging traditional credit assessment methodologies. By leveraging sophisticated machine learning algorithms and comprehensive data analysis, financial institutions can now evaluate creditworthiness through a more nuanced and holistic lens that extends beyond conventional credit scoring.

The economic implications are profound, providing access to formal credit for individuals historically marginalized by traditional banking systems. This approach stimulates economic activity by enabling entrepreneurs and small businesses to access capital, potentially reducing systemic financial inequalities and creating pathways for economic mobility.

Technologically, the project represents a significant leap in applying data science to financial services. By integrating non-traditional data sources like telecommunications and transactional records, the model offers more comprehensive risk assessments that recognize individual potential beyond simplistic numerical credit scores. This innovative approach benefits both financial institutions and potential borrowers by creating a more transparent and equitable lending environment.

Despite its potential, the project must carefully navigate challenges such as data privacy, algorithmic fairness, and continuous model refinement. The ultimate goal is to balance institutional risk management with broader financial inclusion objectives, offering a more compassionate and data-driven approach to understanding individual economic potential.

This innovative model signals a fundamental reimagining of financial services—one that recognizes creditworthiness as a complex, multidimensional concept and prioritizes individual economic opportunity and participation.

1.3 **Target Stakeholders**

Primary Stakeholders:

Financial Institutions Lending organizations will be the primary beneficiaries of this innovative approach. Banks, microfinance institutions, credit unions, and digital lending platforms stand to gain significant advantages through more accurate risk assessment, reduced default rates, and expanded market reach. The data-driven methodology allows these institutions to make more informed lending decisions while minimizing financial risks.

Potential Borrowers Individuals with limited or non-traditional credit histories form the core beneficiary group. This includes:

- Young professionals without established credit
- Freelancers and gig economy workers
- Entrepreneurs in emerging markets
- Individuals from economically marginalized communities
- Migrants and those with limited financial documentation

Technology Providers, Data analytics firms, machine learning specialists, and financial technology (FinTech) companies will play a crucial role in developing and refining the predictive models. They will be instrumental in creating sophisticated algorithms that can process and interpret complex, multi-source financial data.

Regulatory Bodies Financial regulators and government agencies will be key stakeholders in ensuring the ethical implementation of these advanced credit assessment methodologies. They will be interested in how the project promotes financial inclusion while maintaining robust risk management and preventing potential discriminatory practices.

Economic Development Agencies Organizations focused on economic empowerment and financial inclusion will view this project as a potential breakthrough in expanding economic opportunities for underserved populations.

Secondary Stakeholders:

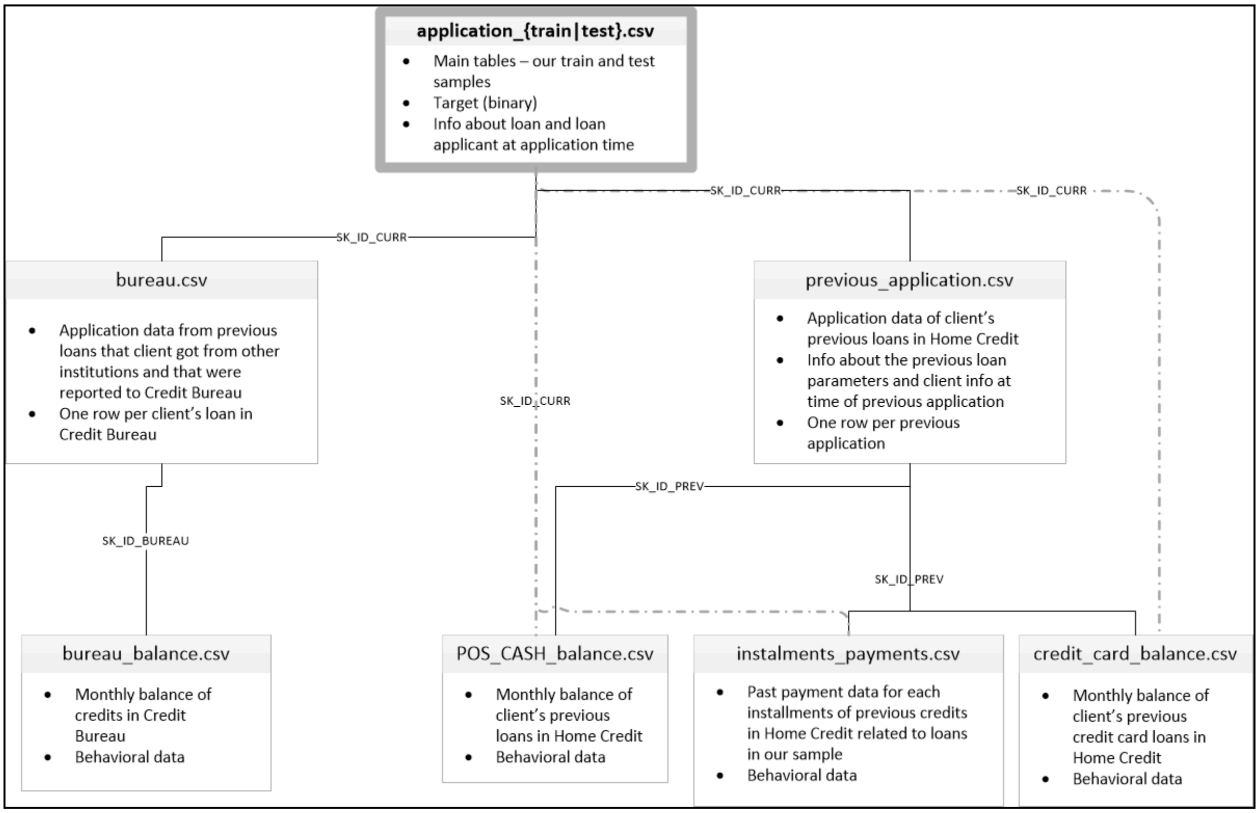
- Telecommunications companies (providers of alternative data)
- Credit bureaus
- Academic researchers in finance and data science
- Insurance companies
- Investment firms evaluating lending platforms

The project's comprehensive approach ensures a win-win scenario that balances institutional risk management with broader social and economic objectives, making it an attractive initiative for multiple stakeholder groups.

2 Data Understanding

2.1 Data Source

The dataset used for this analysis is taken from a Kaggle competition and is a robust and comprehensive collection, designed to provide a detailed view of loan applicants and their financial behaviors. It is structured into multiple interconnected files, similar to a relational database, with the SK_ID_CURR column acting as the primary key linking records across datasets. This relational design facilitates seamless integration of diverse data points, enabling in-depth analysis of an applicant's loan history and credit performance.



application_{train|test}.csv
- Main tables – our train and test samples
- Target (binary)
- Info about loan and loan applicant at application time

bureau.csv
- Application data from previous loans that client got from other institutions and that were reported to Credit Bureau
- One row per client's loan in Credit Bureau

previous_application.csv
- Application data of client's previous loans in Home Credit
- Info about the previous loan parameters and client info at time of previous application
- One row per previous application

bureau_balance.csv
- Monthly balance of credits in Credit Bureau
- Behavioral data

POS_CASH_balance.csv
- Monthly balance of client's previous loans in Home Credit
- Behavioral data

instalments_payments.csv
- Past payment data for each installments of previous credits in Home Credit related to loans in our sample
- Behavioral data

credit_card_balance.csv
- Monthly balance of client's previous credit card loans in Home Credit
- Behavioral data

Figure 2.1.1: Relationship between the different datasets used

The data supports a **binary classification problem** where the objective is to predict whether an applicant will default on a loan. The target variable (TARGET) is binary, with 1 indicating a predicted default and 0 indicating successful repayment. This design allows us to assess and model repayment risks effectively.

The datasets include various categories of information:

1. **Demographics**: Key features such as gender (CODE_GENDER), age (DAYS_BIRTH), and family details (NAME_FAMILY_STATUS) provide insight into the socioeconomic profile of borrowers.
2. **Financial Details**: Variables like income (AMT_INCOME_TOTAL), employment length (DAYS_EMPLOYED), and loan-specific details such as the amount of credit (AMT_CREDIT) and annuity (AMT_ANNUITY) capture applicants' financial stability.
3. **Credit History**: Previous credit performance is detailed through features like overdue payments (CREDIT_DAY_OVERDUE) and total outstanding debt (AMT_CREDIT_SUM_DEBT).

4. **Behavioral Insights**: Patterns in credit applications (DAYS_CREDIT) and registration updates (DAYS_REGISTRATION) highlight applicants' financial behaviors over time.

This interconnected dataset, with over 120 features, provides a granular view of each applicant's profile, enabling the identification of key trends and critical variables essential for accurate loan default predictions.

2.2 Data Features

The dataset provided for this analysis is exceptionally detailed, capturing a wide range of information about loan applicants that collectively offers a comprehensive view of their financial profiles, behavioral patterns, and borrowing history. These features are critical for understanding applicant characteristics, identifying trends, and predicting loan default risks.

Demographic Information

The dataset includes key demographic variables that describe the socio economic profile of each applicant. Features such as CODE_GENDER (applicant's gender), CNT_CHILDREN (number of children), NAME_EDUCATION_TYPE (highest level of education achieved), and NAME_FAMILY_STATUS (marital status) help provide context about the applicant's personal background. These details are essential for understanding how socioeconomic factors correlate with repayment behavior. Additionally, attributes like FLAG_OWN_REALTY and FLAG_OWN_CAR indicate ownership of assets such as property or vehicles, which can reflect an applicant's financial stability.

Economic and Financial Details

Financial stability is a significant component of this dataset, captured through variables such as AMT_INCOME_TOTAL (total income of the applicant) and NAME_INCOME_TYPE (employment type, e.g., working, business, pensioner). These features help assess the applicant's earning capacity and income sources, which are direct indicators of their ability to repay loans. Other variables such as DAYS_EMPLOYED (days employed before application) and AMT_CREDIT (loan amount) provide insights into the applicant's employment stability and financial commitments.

Loan-related details such as AMT_ANNUITY (monthly loan repayment amount) and AMT_GOODS_PRICE (price of goods or services for which the loan is granted) highlight borrowing patterns. Together, these features offer a clear picture of the financial terms surrounding the loan, which can influence repayment behavior.

Credit History and External Risk Factors

A significant portion of the dataset captures applicants' past credit behavior through features such as CREDIT_TYPE (type of credit previously taken), AMT_CREDIT_SUM_DEBT (total outstanding debt), and CREDIT_DAY_OVERDUE (number of overdue days on past loans). These variables are instrumental in understanding the applicant's repayment history and any previous financial struggles.

Additionally, the dataset includes external credit scores (EXT_SOURCE_1, EXT_SOURCE_2, EXT_SOURCE_3) derived from third-party sources, which act as normalized risk indicators. These scores are essential for benchmarking the creditworthiness of applicants against broader industry standards.

Behavioral Insights

The dataset provides behavioral insights through variables that capture patterns over time, such as DAYS_BIRTH (age of the applicant in days), DAYS_REGISTRATION (time since the applicant’s last registration change), and DAYS_ID_PUBLISH (time since their identification documents were last updated). These features highlight temporal trends that might signal financial stability or instability. For instance, frequent updates to registration or ID documents could indicate volatility in the applicant's life circumstances.

Other variables, such as DAYS_CREDIT (days since the last credit application), provide insights into the frequency and timing of credit applications, which could signal financial stress if frequent applications are observed.

Categorical Insights

Several categorical features offer a segmented view of applicant behaviors and contexts. For example, NAME_INCOME_TYPE categorizes applicants by their income source, such as employed, pensioner, or business owner. Similarly, NAME_HOUSING_TYPE reflects their current housing situation (e.g., renting, owning a house). Features like OCCUPATION_TYPE identify the applicant's job profile, helping to segment borrowers based on employment type and income stability.

Relational Structure and Target Variable

The datasets are interconnected using the SK_ID_CURR identifier, which acts as a unifying key across files. This relational structure enables the combination of demographic, financial, and behavioral data into a single dataset for analysis. The target variable, TARGET, indicates whether an applicant defaulted (1) or successfully repaid their loan (0). This binary classification variable is critical for developing predictive models to minimize default risks.

Holistic View of the Dataset

With over 120 features, the dataset captures a multi-dimensional perspective on borrowers. Demographic and economic variables reveal their financial standing, while credit history and behavioral patterns offer insights into their borrowing and repayment habits. External risk indicators provide industry-wide benchmarks, and categorical features enable segmentation and targeted analysis. Together, these features allow for the development of robust models to predict loan defaults, optimize credit decisions, and gain a deeper understanding of financial behaviors across diverse applicant profiles.

2.3 Potential Bias

A critical aspect of working with any dataset is identifying and addressing potential biases that might affect the analysis and predictive modeling. The dataset used in this project, though comprehensive, is not immune to such challenges. Biases can arise from the data collection process, feature representation, and inherent imbalances within the dataset.

Imbalanced Target Variable

One of the most apparent biases in this dataset is the significant imbalance in the TARGET variable. Approximately 92% of the applicants are labeled as non-defaulters (TARGET = 0), while only about 8% are labeled as defaulters (TARGET = 1). This imbalance can lead to skewed model performance, where the algorithm favors predicting non-default outcomes due to their higher frequency. Without appropriate techniques like oversampling, undersampling, or applying weighted metrics, this imbalance could result in poor predictive performance for defaulters, who are the critical focus for risk assessment.

Demographic Bias

The demographic features, such as CODE_GENDER (gender), CNT_CHILDREN (number of children), and NAME_FAMILY_STATUS (marital status), might introduce bias if certain groups are over- or underrepresented. For instance, if a particular gender or family status is disproportionately prevalent in the dataset, the model might inadvertently learn patterns that reflect societal or sampling biases rather than actual risk factors.

Income and Employment Bias

Features like AMT_INCOME_TOTAL (total income) and NAME_INCOME_TYPE (income source) can also introduce bias if the dataset underrepresents applicants from specific income brackets or employment categories. For instance, if high-income earners are overrepresented, the model might generalize poorly for low-income applicants, potentially misjudging their creditworthiness.

Geographic and Regional Bias

The dataset includes variables like REGION_POPULATION_RELATIVE (population density of the applicant's region) and NAME_HOUSING_TYPE (housing situation). Bias could arise if certain regions or housing types are underrepresented, leading the model to perform better for applicants from urban areas while neglecting rural populations. This could limit the model's applicability to diverse geographic contexts.

Historical Bias in Credit Behavior

Variables capturing past credit behavior, such as CREDIT_TYPE (type of credit), CREDIT_DAY_OVERDUE (days past due), and AMT_CREDIT_SUM_DEBT (outstanding debt), may reflect systemic biases in how credit has been historically extended or managed. For example, individuals from marginalized groups may have faced unequal access to credit in the past, which could influence their credit history and lead to biased predictions

External Credit Scores

Features like EXT_SOURCE_1, EXT_SOURCE_2, and EXT_SOURCE_3 are derived from external credit scoring systems. These scores may already incorporate biases inherent in the external systems, such as favoring applicants with traditional credit histories while penalizing those with unconventional borrowing patterns.

Missing Data Bias

The dataset contains significant amounts of missing data in critical features, such as AMT_INCOME_TOTAL (income) and OWN_CAR_AGE (age of car). The strategies used to handle missing data, such as imputing with averages or removing incomplete records, could inadvertently introduce bias. For example, imputing missing income data with the mean might distort predictions for individuals whose actual incomes deviate significantly from the average.

Behavioral Bias

Behavioral variables, such as DAYS_BIRTH (age), DAYS_EMPLOYED (employment duration), and DAYS_CREDIT (time since the last credit application), could introduce bias if certain age groups or employment types dominate the dataset. For instance, younger individuals may be overrepresented, leading the model to underperform for older applicants.

2.4 Preliminary Observations

The dataset presents a wealth of information with over 120 interconnected features, offering deep insights into borrower demographics, financial profiles, and credit behavior. However, early exploration highlights key challenges and opportunities. A major issue is the imbalance in the TARGET variable, with only 8% of loans marked as defaults, necessitating resampling techniques or weighted metrics to ensure fair model performance. Significant correlations among variables, such as AMT_CREDIT (loan amount), AMT_ANNUITY (repayment), and AMT_GOODS_PRICE (goods price), point to potential redundancy that requires attention during feature selection. External credit scores (EXT_SOURCE_1, EXT_SOURCE_2, and EXT_SOURCE_3) emerge as strong predictors of default risk, offering reliable benchmarks. Behavioral trends, like higher default risks among younger borrowers and frequent credit applications indicating financial stress, provide actionable insights. Missing data in key features, such as income, and anomalies like implausible family sizes or mismatched income-to-loan ratios, underline the need for robust cleaning and imputation strategies. These observations guide the data preparation process to ensure reliable modeling and accurate loan default predictions.

3 Exploratory Data Analysis (EDA)

To better understand the underlying patterns and insights within the dataset, we performed extensive EDA. This analysis helped us in identifying the distribution of various features, their relationship to house prices, and any potential correlations that could influence predictive modeling. For this report, we focus on key findings from the EDA related to Income Sources, Family Status, Occupation, and Education Level. The complete and extensive EDA, which includes additional insights and visualizations, is available in the accompanying Python notebook.

3.1 Distribution of Target Variable

The target variable (TARGET) indicates the loan repayment status:

- 0: Loan was repaid.
- 1: Loan was not repaid (defaulted).

Key Insights

- The dataset is highly imbalanced, with 92% of loans repaid (TARGET=0) and 8% defaulted (TARGET=1).
- This imbalance suggests the need for techniques like class balancing or cost-sensitive learning during model development.

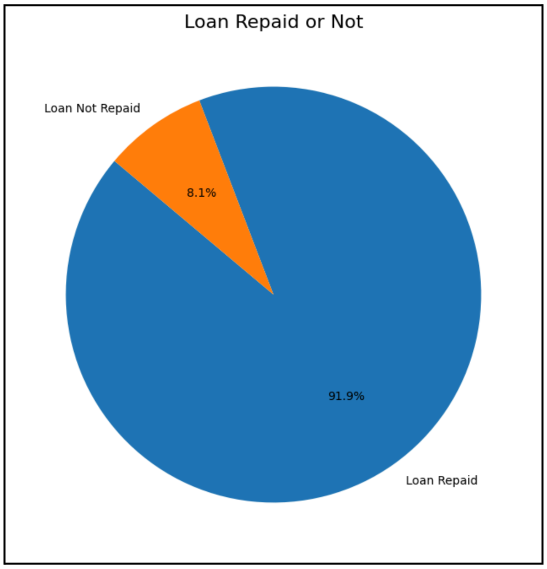


Fig. 3.1.1: Distribution of Target Variable

3.2 Income Sources

Application Demographics:

- Most applicants (51.6%) are working individuals, followed by commercial associates (23.3%) and pensioners (18%).
- Pensioners are a smaller group but show a strong repayment history.

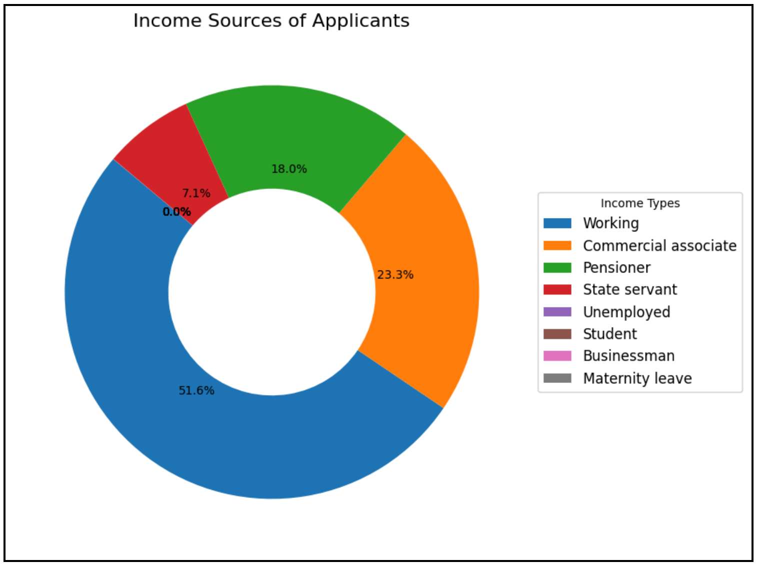


Fig. 3.2.1: Income Sources (Application Demographics)

Loan Repayment Trends:

- Pensioners exhibit the highest repayment rates, while working individuals show slightly higher default tendencies.
- Visualization: A stacked bar chart in the notebook illustrates repayment trends across income types.

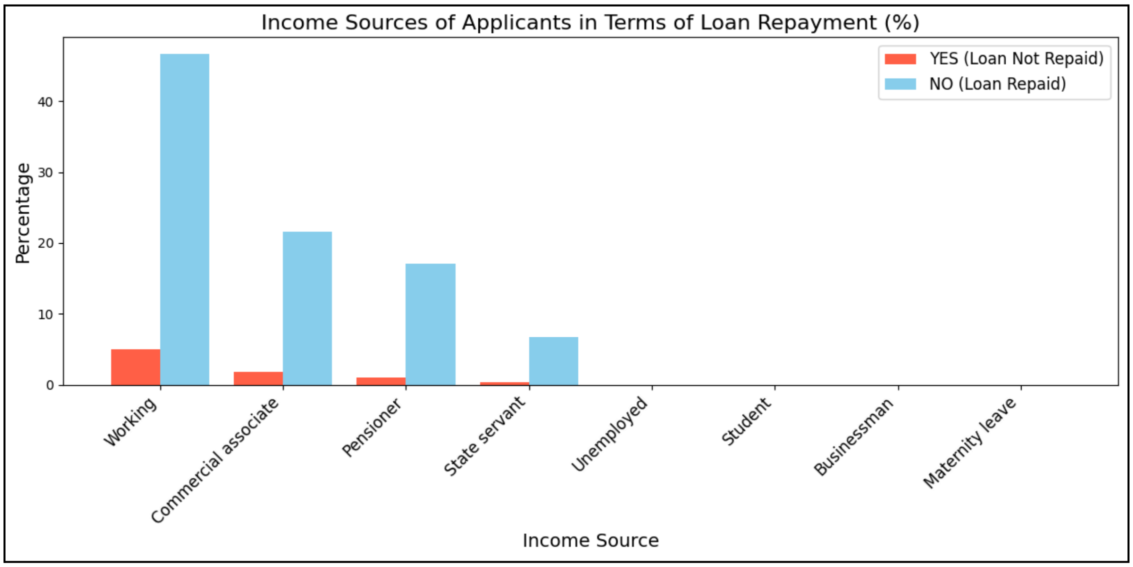


Fig. 3.2.2: Income Sources (Loan Repayment Trends)

3.3 Family Status

Application Demographics:

- Married individuals constitute the majority (63.9%), followed by single individuals (14.8%).
- Family status strongly correlates with repayment behavior.

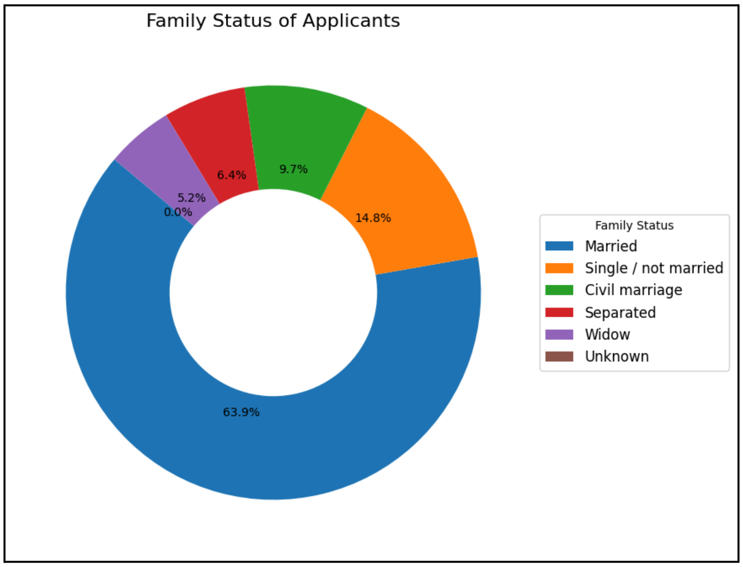


Fig. 3.3.1: Family Status (Application Demographics)

Loan Repayment Trends:

- Married applicants have higher repayment rates compared to single or separated individuals.
- Visualization: A stacked bar chart in the notebook highlights repayment patterns by family status.

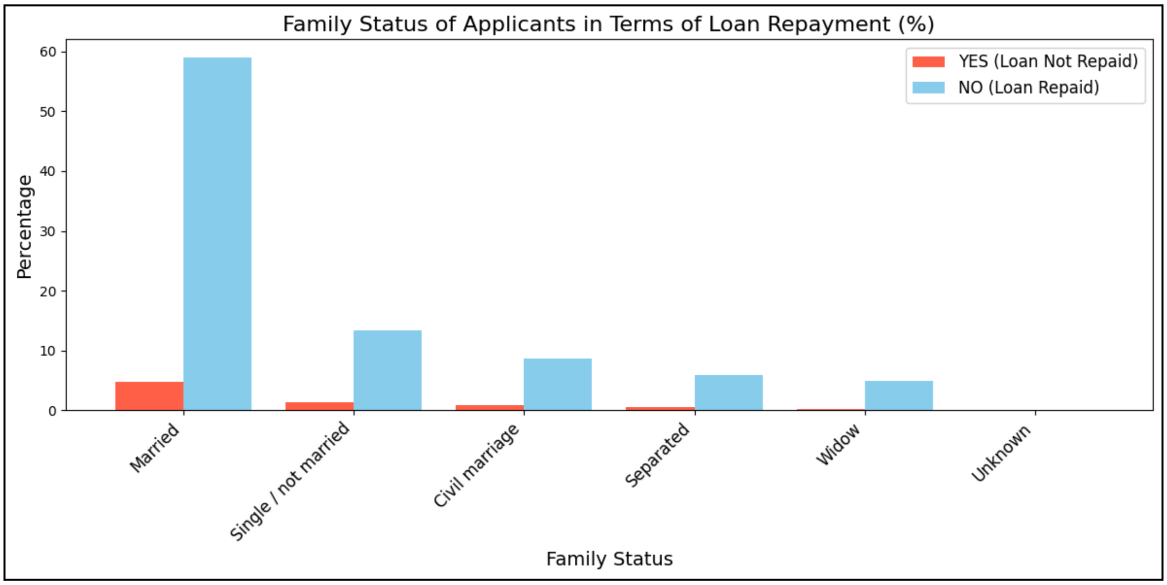


Fig. 3.3.2: Family Status (Loan Repayment Trends)

3.4 Occupation

Application Demographics:

- Laborers (55K), sales staff (32K), and core staff (28K) are the top occupations among applicants.
- Laborers represent a large portion but have a higher tendency to default as seen in the last visualization..

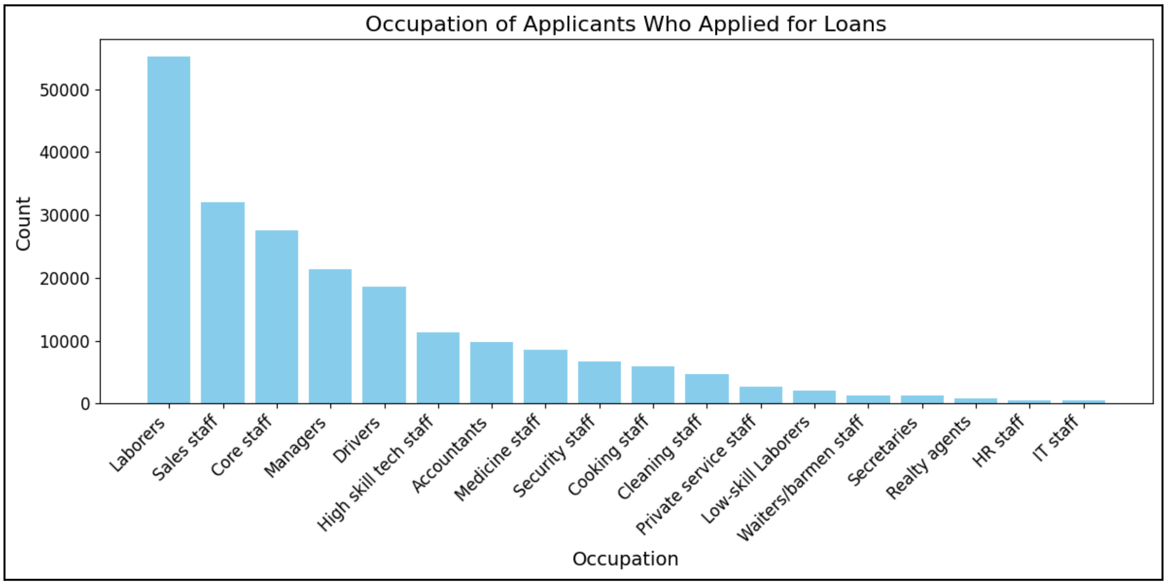


Fig. 3.4.1: Occupation (Application Demographics)

Loan Repayment Trends:

- Laborers show higher default rates, while managers and core staff exhibit better repayment behavior.
- Visualization: The notebook includes a stacked bar chart for default rates by occupation.

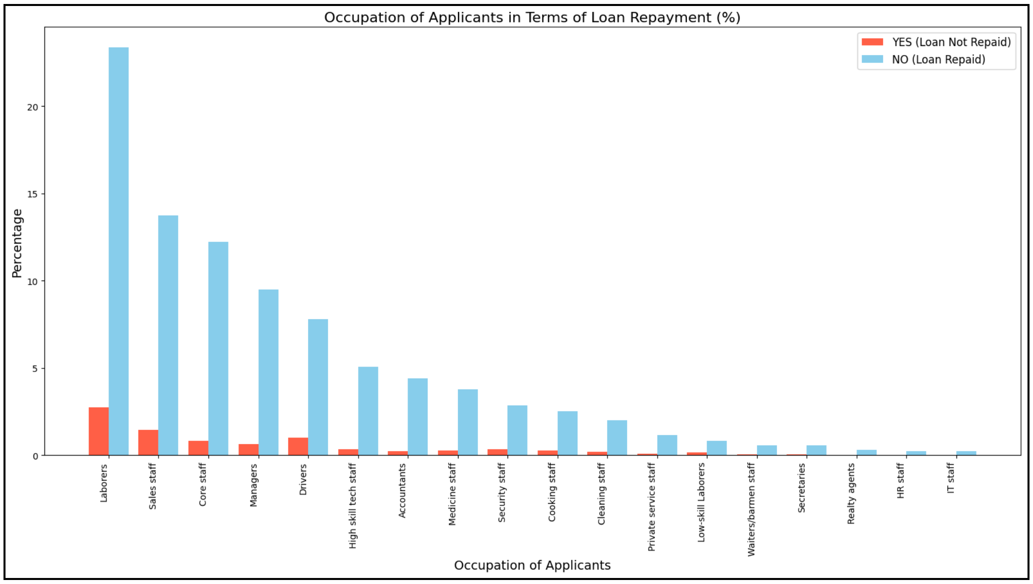


Fig. 3.4.2: Occupation (Loan Repayment Trends)

3.4 Education Level

Application Demographics:

- Most applicants (71%) have secondary education, with 24.3% holding higher education qualifications.
  - Higher education correlates with lower default rates.

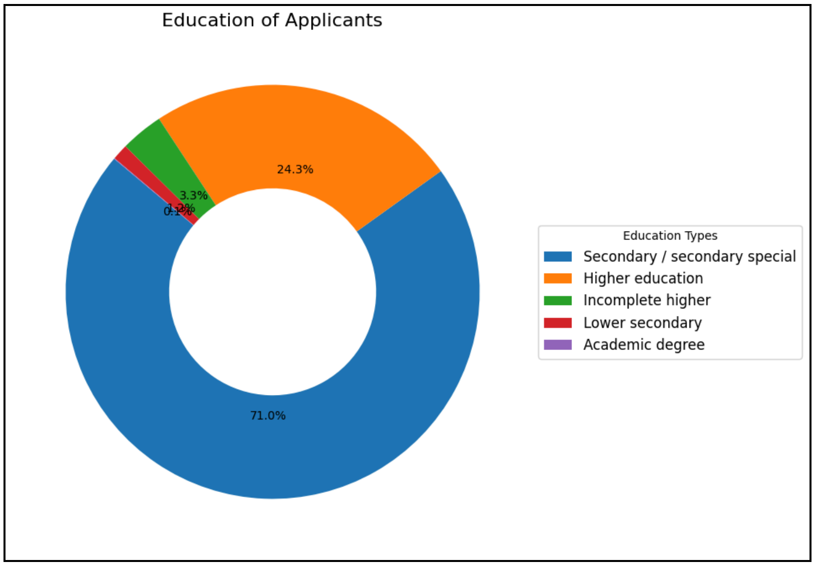


Fig. 3.5.1: Education Level (Application Demographics)

Loan Repayment Trends:

- Applicants with higher education are more likely to repay loans compared to those with secondary education or less.
- Visualization: A bar chart in the notebook compares repayment rates across education levels.

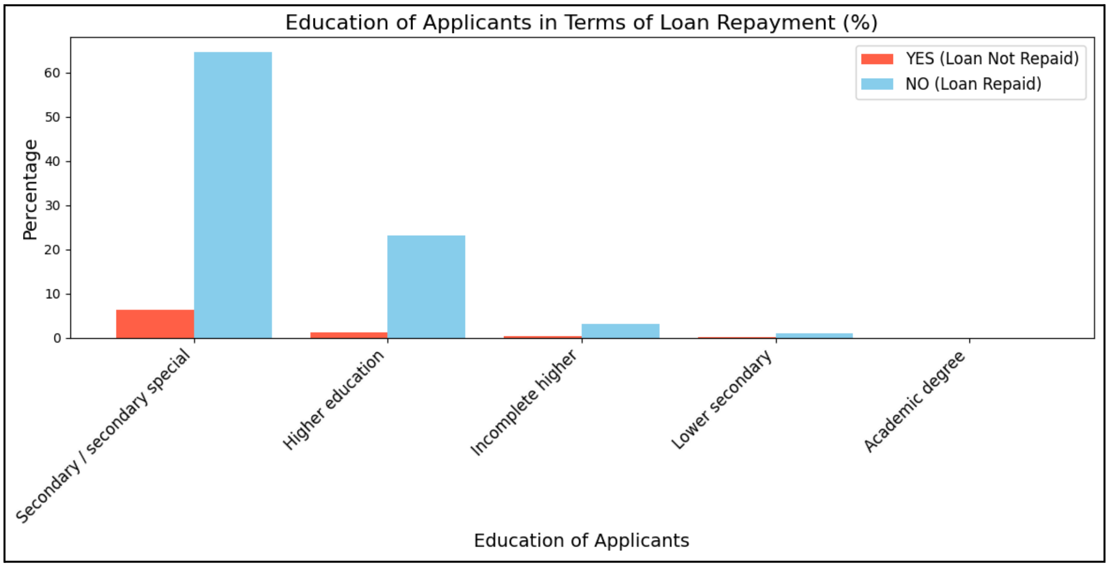


Fig. 3.5.2: Education Level (Loan Repayment Trends)

4 Data Preparation

4.1 Cleaning

Data cleaning is a crucial step in preparing the dataset for robust analysis and modeling. It ensures the removal of noise, redundancy, and inconsistencies, thereby enhancing the reliability of the results. This process involved addressing missing values, outliers, logical inconsistencies, and multicollinearity while preserving the integrity of the data.

Handling Missing Values

The dataset exhibited a significant proportion of missing values across multiple features, necessitating a systematic approach to imputation. Features with more than 75% missing values, such as OWN_CAR_AGE, were removed due to their limited informational value. For the remaining features, appropriate imputation techniques were employed. For example:

- Numerical features: Missing values in AMT_INCOME_TOTAL were replaced with the median value to prevent the influence of outliers.

- Categorical features: Missing entries in OCCUPATION_TYPE were imputed with the mode, preserving the overall distribution of categories.

Outlier Detection and Treatment

Outliers were identified using visual inspections, such as box plots, and statistical measures like the interquartile range (IQR). For instance, extreme values in income, such as applicants reporting over $1 million in annual income, were capped at the 99th percentile using winsorization. This approach ensured that these anomalies did not skew the analysis while retaining valuable insights from less extreme data points.

Addressing Logical Inconsistencies

Logical inconsistencies, particularly in time-based variables, were detected and corrected. Examples include negative values in DAYS_EMPLOYED due to data entry errors, which were replaced with valid positive values. Additionally, discrepancies in DAYS_BIRTH and DAYS_REGISTRATION were cross-referenced to ensure chronological accuracy.

Multicollinearity Reduction

Multicollinearity among numerical features was assessed through a correlation matrix. Highly correlated variables (correlation coefficient > 0.85) were identified and addressed. For example, redundant features like AMT_GOODS_PRICE were removed in favor of AMT_CREDIT, which had a stronger predictive relationship with the target variable.

Scaling and Transformation

To standardize the numerical features, z-score normalization was applied, ensuring that all features were on a comparable scale. This step was essential for algorithms sensitive to feature magnitudes, such as support vector machines. Furthermore, skewed distributions, such as DAYS_EMPLOYED_PERCENT, were log-transformed to approximate normality.

Summary of Cleaning Process

The data cleaning process resulted in a refined dataset with improved integrity and usability. Features with excessive missing values were eliminated, while outliers were managed to retain their informative aspects. Logical errors were resolved, and redundant features were removed to mitigate multicollinearity. These steps collectively ensured that the dataset was well-prepared for subsequent stages of analysis and modeling.

This systematic cleaning approach preserved the informational richness of the data while addressing inconsistencies and redundancies, laying a solid foundation for reliable and interpretable results.

4.2 Feature Engineering

Feature engineering is a critical step in preparing a dataset for predictive modeling, as it transforms raw data into meaningful features that enhance the model's predictive capabilities. For the given dataset, extensive feature engineering was performed to extract, aggregate, and create new features that better represent the underlying relationships within the data. This section provides a detailed explanation of the methodology, challenges faced, nomenclature of engineered features, and the significance of key features.

Methodology

The feature engineering process began with aggregating data from multiple tables into a unified structure. Each file, such as bureau.csv, credit_card_balance.csv, and installments_payments.csv, contained information at varying levels of granularity, making aggregation essential. For instance, the bureau.csv file records previous credits at the applicant level, requiring summarization to align with the primary table using the SK_ID_CURR key.

Aggregation techniques included calculating statistical metrics such as the mean, median, minimum, maximum, sum, and standard deviation for continuous variables. For categorical variables, counts and proportions were derived to represent their distributions. Temporal features, such as the days since the last credit application, were engineered to capture behavioral trends.

Ratios were also created to capture relationships between variables. Examples include:

- CREDIT_TO_INCOME_RATIO: The ratio of the total credit amount (AMT_CREDIT) to the applicant's total income (AMT_INCOME_TOTAL), representing affordability.
- ANNUITY_TO_INCOME_RATIO: The ratio of the loan annuity (AMT_ANNUITY) to total income, measuring the financial burden of repayments.

Finally, derived features like the percentage of life spent employed (DAYS_EMPLOYED_PERCENT) were calculated by normalizing employment duration (DAYS_EMPLOYED) by the applicant's age (DAYS_BIRTH). This feature adds context to employment stability in relation to the applicant's life stage.

Key Challenges and Solutions

1. Data Integration Across Files: The dataset was distributed across multiple tables with varying levels of granularity. Aggregating these datasets without losing meaningful information required careful handling.
   - ○ Solution: Relational joins using SK_ID_CURR and aggregation functions ensured that each applicant's profile was unified without duplication or data loss.

2. Handling Missing Values: Many features had significant missing values, which posed a challenge during aggregation and derivation of new features.
   - Solution: Missing values were addressed using imputation strategies (e.g., mean, median) or handled through domain-specific logic, such as filling missing external credit scores with averages.
3. Outlier Detection and Treatment: Unrealistic values, such as incomes in the millions for applicants seeking small loans, risked skewing the derived features.
   - Solution: Outliers were capped or removed using thresholds determined by domain expertise and statistical methods, such as z-scores or interquartile range (IQR).
4. Multicollinearity: Aggregating features often introduce high correlations between similar variables.
   - Solution: Correlation analysis was conducted to identify redundant features, which were then removed or combined into more meaningful aggregates.

Nomenclature of Engineered Features

The naming conventions for engineered features were designed to reflect their origins and the statistical operations applied. For example:

- PREFIX_VARIABLE_AGGREGATION: This format indicates the source file (PREFIX), the variable being aggregated (VARIABLE), and the type of aggregation (AGGREGATION).
  - Example: bureau_AMT_CREDIT_SUM_MEAN refers to the mean of credit amounts across all previous credits recorded in the bureau.csv file.
- RATIO_FEATURE1_FEATURE2: Indicates a ratio between two features.
  - Example: CREDIT_TO_INCOME_RATIO is the ratio of total credit amount to total income.

Key Features and Their Significance

1. bureau_DAYS_CREDIT_MEAN: This feature represents the average time between previous credit applications. A shorter interval indicates frequent borrowing, which could be a sign of financial stress.
2. bureau_AMT_CREDIT_SUM_DEBT_SUM: The total debt across all previous loans provides a direct measure of the applicant's financial burden. High values suggest increased repayment risk.
3. client_installments_DAYS_ENTRY_PAYMENT_MEAN: This feature captures the average delay between the due date and the actual payment date for installments. Larger values indicate inconsistent repayment behavior.
4. client_installments_AMT_PAYMENT_MIN_SUM: The sum of minimum payment amounts across all installments reflects the applicant's repayment discipline and ability to meet minimum obligations.
5. previous_NAME_CONTRACT_STATUS_Approved_COUNT: The count of previously approved loans gives insight into the applicant's creditworthiness and history of successful applications.
6. CREDIT_TO_INCOME_RATIO: This ratio measures the affordability of the credit amount in relation to the applicant's income. Higher values suggest financial overextension, which correlates with default risk.
7. DAYS_EMPLOYED_PERCENT: This derived feature normalizes employment duration by the applicant's age, offering a perspective on employment stability relative to life stage. Applicants with higher percentages are generally seen as more stable.

Results of Engineered Features

The feature engineering process yielded a range of meaningful and predictive features, each offering insights into applicant behavior, financial stability, and repayment tendencies. Below are the results, interpretations, and the meanings of key features, organized by their categorical segregation:

Behavioral Features

1. bureau_DAYS_CREDIT_MEAN: This feature represents the average time gap (in days) between an applicant's previous credit applications as recorded in the credit bureau. Shorter gaps indicate frequent borrowing, which is often a sign of financial stress. Applicants with lower values for this feature showed higher default rates, emphasizing the relevance of borrowing frequency as an indicator of repayment difficulty.
2. client_installments_DAYS_ENTRY_PAYMENT_MEAN: This feature calculates the average delay (in days) between the due date and the actual payment date for installments. It reflects the applicant's timeliness in meeting their financial obligations. Higher values for this feature were strongly associated with defaulters, indicating a lack of financial discipline and an increased likelihood of repayment challenges.
3. DAYS_EMPLOYED_PERCENT: This feature quantifies the percentage of an applicant's life spent in employment by dividing the number of days employed by their age in days. A higher percentage reflects greater employment stability. Lower percentages were found among defaulters, suggesting that employment continuity is a key determinant of financial reliability.

Financial Features

1. CREDIT_TO_INCOME_RATIO: This feature measures the ratio of the total credit amount (AMT_CREDIT) to the applicant's income (AMT_INCOME_TOTAL). It provides insight into how much debt an applicant is taking on relative to their income. Higher ratios indicate financial overextension and were found to correlate strongly with defaults, as borrowers with excessive debt relative to their earnings are more likely to struggle with repayments.
2. ANNUITY_TO_INCOME_RATIO: This feature represents the ratio of the loan's monthly annuity (AMT_ANNUITY) to the applicant's income (AMT_INCOME_TOTAL). It serves as a proxy for affordability, showing how much of an applicant's income is dedicated to repaying their loan. Higher ratios were linked to increased default rates, highlighting the financial strain of loans on applicants with limited disposable income.
3. bureau_AMT_CREDIT_SUM_DEBT_SUM: This feature aggregates the total outstanding debt from all previous loans recorded in the credit bureau. It captures the overall financial burden carried by an applicant. Higher values were observed among defaulters, illustrating that excessive debt levels contribute significantly to repayment risk.

Historical and Categorical Insights

1. previous_NAME_CONTRACT_STATUS_Approved_COUNT: This feature counts the number of previously approved loans for each applicant. It reflects the applicant's historical creditworthiness and reliability. Applicants with higher counts were more likely to repay their loans successfully, showcasing their ability to manage credit effectively.
2. EXT_SOURCE_MEAN: This feature is the average of three external credit scores (EXT_SOURCE_1, EXT_SOURCE_2, EXT_SOURCE_3) normalized to a standard scale. These scores provide an external benchmark of the applicant's credit risk. Lower average scores were strongly associated with defaults, making this feature a critical predictor of loan performance.
3. client_installments_AMT_PAYMENT_MIN_SUM: This feature sums the minimum payments made by the applicant across all installments. It highlights the regularity and sufficiency of installment payments. Non-defaulters showed higher and more consistent values for this feature, indicating that maintaining adequate minimum payments is a strong indicator of repayment ability.

Overall Insights from Results

Each engineered feature captures a distinct aspect of the applicant's financial and behavioral profile. Behavioral features like bureau_DAYS_CREDIT_MEAN and client_installments_DAYS_ENTRY_PAYMENT_MEAN offer insights into the applicant's borrowing frequency and repayment discipline. Financial features such as CREDIT_TO_INCOME_RATIO and ANNUITY_TO_INCOME_RATIO quantify the applicant's financial obligations relative to their income, directly reflecting affordability and risk. Historical and categorical insights like previous_NAME_CONTRACT_STATUS_Approved_COUNT and EXT_SOURCE_MEAN leverage past credit performance and external benchmarks to assess reliability.

The engineered features revealed strong patterns that distinguished defaulters from non-defaulters. Notably, features like EXT_SOURCE_MEAN, CREDIT_TO_INCOME_RATIO, and bureau_AMT_CREDIT_SUM_DEBT_SUM emerged as the most influential predictors, providing clarity and interpretability essential for effective credit risk assessment. The enriched dataset enabled a granular understanding of applicant behavior and repayment potential, laying the foundation for robust predictive modeling.

4.3 Feature Selection

Feature selection is a crucial step in machine learning workflows, especially when working with high-dimensional datasets like the one used in this analysis. The aim is to retain the most informative features, eliminate redundancy, and ensure computational efficiency while maintaining model accuracy. The methodology, challenges faced, and key findings for feature selection are discussed in detail below.

Methodology

1. The dataset contained a total of 1638 features after the feature engineering process, which included attributes from both Bureau and Previous Loans datasets. Such a high-dimensional dataset posed challenges in terms of computational cost, model overfitting, and interpretability, necessitating feature selection.
2. The dataset included categorical variables that were transformed using one-hot encoding. This process converted categorical attributes into binary features, increasing the feature count to 1758 for both the training and testing datasets. Ensuring consistent alignment between the datasets was a critical step before proceeding further.
3. Collinearity Analysis:
    - Correlation Matrix: A correlation matrix was computed to identify highly collinear features. High correlation (above 0.9) between features indicates redundancy, as such features carry overlapping information.
    - Feature Removal: The upper triangle of the correlation matrix was scanned, and one feature from each highly correlated pair was removed. This process resulted in a significant reduction, with 893 features eliminated, leaving 865 remaining.
4. Handling Missing Data
    - Thresholding: Features with over 75% missing values were identified and removed. This threshold was chosen to balance the retention of meaningful information with the need to avoid bias introduced by excessive imputation.
    - Feature Removal: 18 columns were dropped, further reducing the feature count to 847.
5. Feature Importance Evaluation

    - A Random Forest Classifier was employed to compute the importance of each feature. Random Forest models are well-suited for this task because they inherently measure the contribution of each feature to prediction accuracy through reductions in impurity.
    - Cumulative Importance Threshold: Features were ranked based on their importance scores. A threshold of 95% cumulative importance was applied, meaning only the top features contributing to 95% of the variance in the target variable were retained.
    - Final Selection: This process reduced the dataset to 498 features, which collectively explain the majority of the variance while eliminating noise.
6. Dataset Finalization The resulting datasets for modeling consisted of:
    - Training Data: 500 features, including SK_ID_CURR and TARGET.
    - Testing Data: 499 features, excluding TARGET.

Key Challenges and Solutions

1. The large number of features (1638) posed significant computational challenges and increased the risk of overfitting. Systematic reduction through collinearity analysis and feature importance evaluation reduced the dataset to a manageable size without sacrificing predictive power.
2. Redundant features with high correlation could lead to instability in model predictions. A correlation threshold of 0.9 was used to identify and remove highly correlated features, ensuring a diverse and independent feature set.
3. Many features had substantial proportions of missing values, which could skew model performance. A strict threshold of 75% missing values was applied, and imputation strategies were avoided for highly missing columns to maintain data integrity.
4. While reducing features, it was essential to retain those with high predictive importance to maintain model accuracy. Feature importance scores from the Random Forest model provided a robust criterion for selecting relevant features while discarding noise.

Final Results

1. The final dataset retained 498 features that explained 95% of the cumulative variance in the target variable. This reduction significantly improved computational efficiency while preserving the dataset's predictive capabilities.
2. The Random Forest model identified several key features as highly important for predicting loan defaults:
    - EXT_SOURCE Variables: These normalized external credit scores (e.g., EXT_SOURCE_1, EXT_SOURCE_2, EXT_SOURCE_3) consistently ranked among the top predictors, highlighting their relevance in assessing applicant creditworthiness.
    - Bureau Data: Aggregated features from the Bureau dataset, such as bureau_DAYS_CREDIT_mean (average time since credit applications) and bureau_CREDIT_DAY_OVERDUE_max (maximum days overdue), provided insights into applicant borrowing behavior and financial stress.
    - Loan Amount and Income: Features like AMT_CREDIT (loan amount), AMT_ANNUITY (loan repayment amount), and AMT_INCOME_TOTAL (total income) were critical for evaluating the applicant's financial capacity.
    - Behavioral Features: Variables such as DAYS_BIRTH (age in days) and DAYS_EMPLOYED (employment duration) reflected life stage and job stability, both of which are closely tied to repayment behavior.
3. Segmentation by Dataset

- ○ Bureau Features: Primarily focused on external credit history, these features provided aggregated metrics such as mean and maximum values, which effectively summarized credit behavior over time.
- ○ Previous Loans Features: Captured trends in past loan applications, including approval rates, credit amounts, and repayment statuses, offering a historical perspective on borrower reliability.

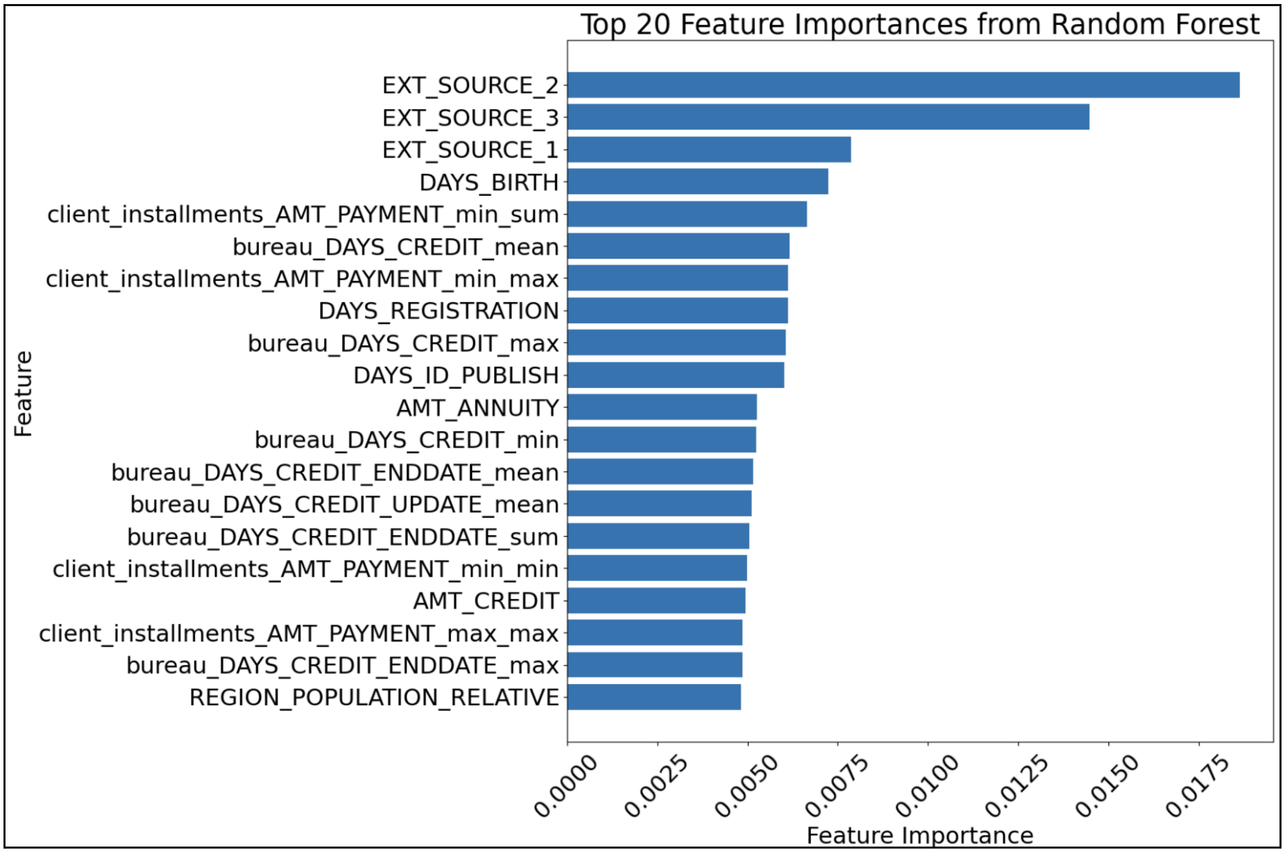


Figure 4.3.1: Top 20 features based on feature importance from a standard Random Forest model

The feature selection process was meticulous, balancing the need to reduce dimensionality with the goal of retaining predictive power. The resulting dataset is well-suited for robust modeling, with a focus on features that directly influence loan repayment behavior and default risk. This refined feature set not only enhances model performance but also improves interpretability, enabling actionable insights for credit risk management.

5 Modeling

5.1 Evaluation Metric

We chose AUC (Area Under the Curve) as the primary evaluation metric. AUC measures the relationship between true positive and false positive rates, making it particularly useful for our business problem, where distinguishing between defaults and successful repayments is critical.

5.2 Models Tested

We evaluated the following models, each with its unique strengths and trade-offs:

1. **Logistic Regression** (with and without regularization):
    - A simple yet effective baseline model.
    - Lasso (L1 regularization) provided slight improvements in AUC.
2. **Random Forest**:
    - Captures non-linear feature interactions.
    - Performance improved after hyperparameter tuning but remained suboptimal.

3. **Support Vector Machine**:
    - Did not perform well due to being computationally expensive and inefficient on large datasets.

4. **Boosting Models**:
    - XGBoost and AdaBoost effectively handled class imbalances and captured complex relationships in the data.

To compare models, we trained them on a subset of **10,000 rows** from the training dataset using an **80-20 train-validation split**. Boosting models and logistic regression models emerged as the top performers.

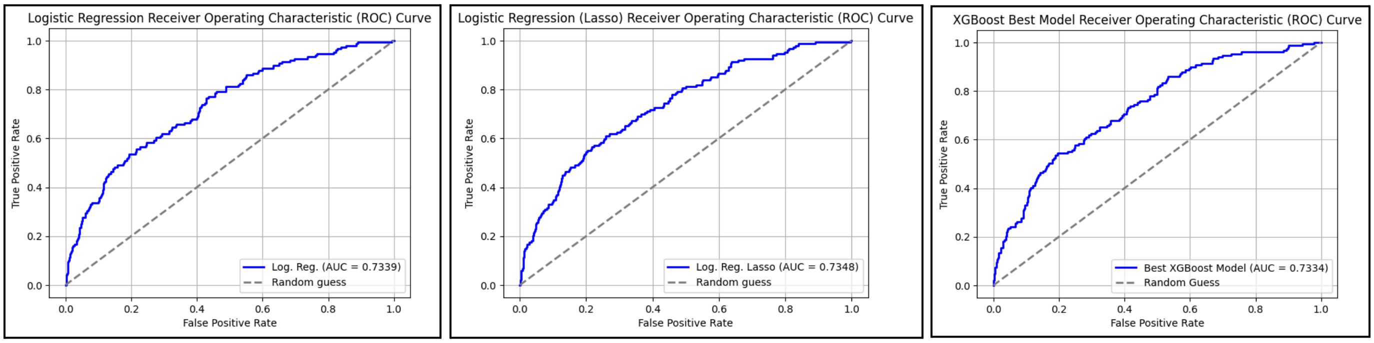


Fig 5.2: AUC Curves of Top 3 Best Performing Base Models

S.No.	Model	AUC
1.	Logistic Regression	0.7339
2.	Logistic Regression (Lasso)	0.7348
3.	Logistic Regression (Ridge)	0.7336
4.	Random Forest	0.6770
5.	Random Forest (best params) (Cross Validation)	0.7239
6.	Support Vector Machine	0.6677
7.	Support Vector Machine (best params) (Cross Validation)	0.6735
8.	XGBoost	0.6740
9.	XGBoost (best params) (Cross Validation)	0.7334
10.	AdaBoost	0.7004
11.	AdaBoost (best params) (Cross Validation)	0.7310

Table 5.2: AUC Scores of Base Models Tested

5.3 Model Selection

We Selected XGBoost as our final model for fine tuning and evaluation due to the following reasons:

1. It achieved a Validation AUC of 0.7334 which is one of the highest scores we achieved from our comparative analysis of the base models.
2. It handles NaN values and large datasets efficiently.
3. It effectively addresses class imbalance.
4. It captures complex, non-linear relationships in the data.
5. Its performance significantly improved after cross validation.

5.4 Model Evaluation

The final XGBoost model was trained and validated on the complete dataset, which contains over **300,000 data points**. We used an **80-20 train-validation split** for this evaluation, with cross-validation to further fine-tune the hyperparameters.

Performance Metrics:

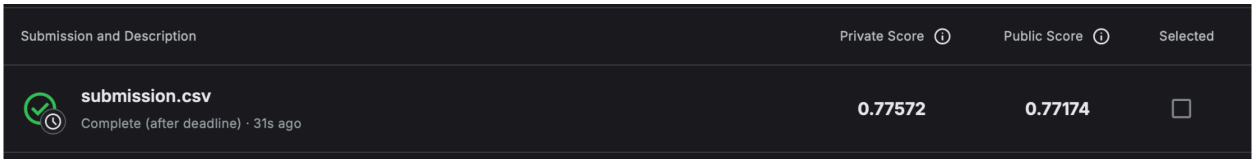
- The final XGBoost model achieved a **validation AUC score of 0.7783**, reflecting its ability to distinguish between loan defaults and repayments effectively.
- This AUC score indicates a significant improvement over the baseline XGBoost model and other competing models.

Cross-Validation:

- Cross-validation was employed to optimize the model's hyperparameters and ensure its performance generalized well across different subsets of data.
- The selected hyperparameters ('subsample': 0.6, 'reg_lambda': 2, 'reg_alpha': 0.1, 'n_estimators': 200, 'max_depth': 3, 'learning_rate': 0.2, 'gamma': 0.1, 'colsample_bytree': 0.8) were found to provide the best balance between accuracy, complexity, and generalization.

Kaggle Submission:

- To validate the generalizability of the model, we submitted predictions to the Kaggle competition from which the dataset originated.
- The Kaggle submission score closely matched the validation AUC score (0.7783), demonstrating that the model effectively learned patterns in the data without overfitting.
- This alignment with Kaggle scores provides external validation of the model's robustness and reliability.



Submission and Description	Private Score ⓘ	Public Score ⓘ	Selected
submission.csv Complete (after deadline) · 31s ago	**0.77572**	**0.77174**	□

Fig 5.4: Kaggle Submission Scores

Handling Challenges:

- **Class Imbalance**: The dataset's significant imbalance between loan repayment (TARGET=0) and loan default (TARGET=1) was addressed through XGBoost's built-in support for weighted loss functions.
- **Missing Data**: XGBoost's ability to handle missing values directly in the training process ensured efficient and accurate modeling without requiring complex imputation techniques.
- **Large Dataset**: The scalability of XGBoost allowed it to process over 300,000 rows efficiently, which was a limitation for some other models, such as Support Vector Machines.

6 Recommendations and Impact

Based on the insights derived from our predictive model, we have formulated targeted recommendations for different stakeholders within the real estate market. These recommendations are aimed at leveraging the model's findings to optimize investment strategies, improve property values, and assist buyers and sellers in making informed decisions.

Real Estate Agents:

- Evaluate House Prices with Given Features: Agents can use the predictive model to accurately assess house prices based on specific property features, enhancing their ability to advise clients on fair pricing and investment returns.
- Guide Premium and Budget Buyers: Premium buyers should be directed towards the Crawford neighborhood due to its higher property values, while budget buyers might find more affordable options in the Stone Brook neighborhood, aligning purchase decisions with financial strategies and lifestyle preferences.

Real Estate Developers:

- Design Two-Story Dwellings: Developers should focus on constructing two-story dwellings, which are shown to be highly valued in the market, especially in premium neighborhoods.
- Prioritize Living Space Quality and Garage Size: Emphasize the quality of living spaces and the size of garages in new developments, as these features significantly influence house valuations.

Home Owners:

- Assess House Value: Owners should periodically assess their property's market value using the predictive model, considering current market trends and property conditions.
- Evaluate Targeted Renovations to Increase Value: Repaint and remodel interior and exterior finish as enhancements to the interior and exterior finish can significantly boost property appeal and value. Ensuring the basement is finished as completing or renovating the basement can add substantial usable space and increase the property's overall value. Remodel kitchens as modern and well-equipped kitchens are crucial for increasing home valuation, reflecting current buyer preferences for high-quality amenities.

Home Buyers:

- Assess House Value Based on Features and Location: Buyers should evaluate potential homes based on essential features and location to ensure investment in properties that meet their needs and financial considerations.
- Assess Possible Features and Locations Given a Budget: Understand what features and neighborhoods are feasible within a set budget to maximize the value and satisfaction of their investment. Buyers should explore neighborhoods like Stone Brook for affordability and consider features typically available within their budget, such as smaller lot sizes or older homes needing renovation. Premium neighborhoods like Crawford offer higher-end features like modern kitchens and spacious garages but at a higher cost.

These recommendations are designed to provide actionable advice tailored to the specific needs and roles of each stakeholder group, enabling them to make strategic decisions that align with market dynamics and personal or business goals. By applying the insights gained from the predictive model, stakeholders can enhance their competitive edge, achieve better financial outcomes, and realize optimal property values in the real estate market

7 Future Work

As the predictive model evolves, expanding its capabilities and enhancing its accuracy are pivotal for capturing real-world market dynamics more effectively. The application of the model across different cities could benefit significantly from the inclusion of more comprehensive datasets. By analyzing diverse markets with a variety of data inputs, the adaptability and predictive performance of the model could be refined to better suit the nuances of each location. Integrating local economic indicators such as employment rates and income levels could also enhance the model's ability to respond to economic shifts. These metrics can provide deeper insights into how market fluctuations influence house prices, offering a more robust tool for stakeholders in the real estate sector.

Consumer behavior is another area where deeper insights could improve model accuracy. Utilizing surveys and behavioral studies to capture shifts in buyer preferences and trends could help in anticipating future market demands and adjusting housing price predictions accordingly. Finally, incorporating real-time market data could significantly increase the model's responsiveness and accuracy. Access to up-to-date information on listings, sales, and market fluctuations could transform the model into a more dynamic and immediately applicable tool in real estate transactions. These enhancements could ensure that the predictive model remains a vital asset for real estate agents, developers, and consumers, keeping pace with market changes and consumer trends.

8 Conclusion

In this project, we meticulously examined the Home Credit Default Risk dataset to address the critical challenge of predicting loan defaults. The methodology encompassed a series of well-structured steps, including data cleaning, feature engineering, feature selection, and model development, each contributing significantly to the predictive model's robustness and interpretability. Data cleaning ensured that inconsistencies, missing values, and outliers were addressed to enhance data reliability. Feature engineering and selection extracted meaningful insights from raw data and refined the feature set, balancing dimensionality reduction with predictive power.

The final model, based on XGBoost, demonstrated impressive performance in distinguishing defaulters from non-defaulters, validated through both internal cross-validation and external Kaggle submission. The results underscored the importance of features such as external credit scores, credit-to-income ratios, and repayment behaviors, which provided actionable insights into credit risk assessment.

This project highlights the transformative potential of data science in advancing financial inclusion. By leveraging diverse data sources and applying sophisticated analytical techniques, the model offers a nuanced approach to credit evaluation that balances institutional risk management with broader societal benefits. While challenges such as data imbalance and feature redundancy were carefully managed, the work sets the stage for future enhancements, including the integration of real-time data and localized economic indicators.

The findings not only serve as a robust tool for financial institutions but also underline the importance of continuous innovation in credit risk modeling to foster a more equitable and efficient financial ecosystem.

9 Appendix

9.1 Group Contribution

We would like to emphasize that the success of this project is attributed to the equal and significant contributions of all group members. Each phase, including modelling, data analysis, result generation, and the drafting of the presentation and final report, was reviewed by every member of the team.

Kaartikeya Panjwani (kp3291) - EDA, Comparative Analysis of Models and Final model evaluation including kaggle submission
Aditya Mittal (am13294) - Data Pre-processing, Feature engineering, EDA and Feature Selection
Saurabh Paul (sp6296) - Data collection, Business problem and inference of results
Teja Reddy (sa8238) - Data collection Project presentation and data analysis

9.2 Data Source and Code Repository

DataSet : https://www.kaggle.com/competitions/home-credit-default-risk/data
Notebooks : https://drive.google.com/drive/folders/1C1G1w4fqGmpnPlhfX7e6oO2BKvS5NmW0?usp=sharing