THE UNIVERSITY
OF TEXAS AT DALLAS

BUAN 6341.002
Applied Machine Learning - S24

HOME CREDIT DEFAULT RISK

Group 1 Final Project submitted to:

Professor Ziyi Cao

Members:

Akshay Varma Penmatsa	AXP230091
Suman Bar	SXB220043
Rucheek Rajeev Kashyap	RXK230010
Varun Goals	VXG210035
Jagannadh Bora	JXB220028

I. Executive Summary

This project is centered around building a robust machine-learning pipeline for loan application risk assessment. Using comprehensive data from historical loan applications, we preprocess, analyze, and clean the data to develop models that identify critical factors for loan approval. The proposed solution will improve loan risk management and automate decision-making, significantly enhancing efficiency for financial institutions.

II. Introduction

i. Business Idea

The project's goal is to streamline the loan application process and reduce risks by leveraging machine learning for comprehensive data analysis. The system will use historical loan application data to train predictive models capable of detecting high-risk loans and assisting lenders in making more informed decisions.

Objectives

1. Assessing Credit Worthiness: To be used as a tool to assess Credit Worthiness of Prospective Borrower. Lenders can make informed decisions about whether to extend credit, ensuring responsible lending practices.
2. Risk Management: Data profiling of prospective borrowers enables lending institutions to effectively manage risk. By predicting whether a prospective borrower might default in the future, lenders can tailor their lending strategies to mitigate potential losses and maintain financial stability.
3. Access to Competitive Loans: This helps Borrowers with insufficient Credit History secure a competitive Loan. For individuals with limited or insufficient credit history, lenders can evaluate the creditworthiness of these borrowers more accurately, opening doors to better loan opportunities.

ii. Importance

Loan risk management is crucial for financial stability and profitability. Current manual and rule-based assessments are prone to errors and often overlook critical patterns. Our automated, data-driven solution improves the accuracy of these predictions, reducing the incidence of default loans and helping institutions save on losses.

iii. Data Source

Home Credit is an international consumer finance provider that focuses on responsible lending to people with little or no credit history. Home Credit finances both cash-less loan transactions and other commodities. They focus on responsible lending, primarily to people with little or no credit history. The company was founded in 1997 in the Czech Republic and expanded to Slovakia in 1999.

III. Data

i. Data Summary

The dataset for our project originates from Kaggle and is provided by Home Credit. This dataset, derived from mortgage loan applications approved by Home Credit, is structured into various categories. For our analysis, we've selected the main dataset, which contains the majority of the information necessary for our investigation.

Our primary focus within this dataset is on the target variable that indicates the status of the loan, distinguishing between active loans and those that have defaulted. To uphold privacy standards and safeguard sensitive information, we've taken careful measures to remove personally identifiable data, ensuring the confidentiality of individuals' personal details while conducting our analysis.

ii. Data Description

Overview

The dataset represents a collection of mortgage loan applications sourced from Home Credit. It is designed to support research and analysis related to loan default prediction and credit risk assessment. The credit default data consists of a high level of financial and personal information of clients; hence the dataset has been stripped of personally identifiable information to ensure privacy and compliance with data protection regulations.

Data Composition

The dataset comprises 307,511 rows, each corresponding to a unique loan application. The dataset contains 122 columns, with 120 features, one identifier, and one target variable. The identifier is used to distinguish individual records, while the target variable indicates whether a loan is active (0) or has defaulted (1). Notably, the dataset has 24,825 records in the positive class (defaulted loans), accounting for approximately 8.07% of the total data.

Feature Categories

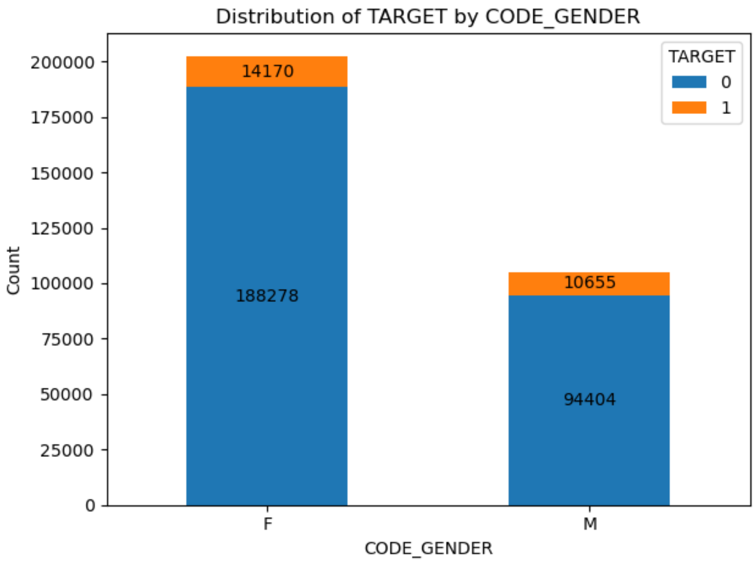
The dataset contains features that can be grouped into the following categories:

1. Demographic Information: This includes variables related to the applicant's gender, education, family status, and number of children.
2. Employment Information: This category contains details about the applicant's occupation, organization type, income, and employment history.

3. Geographic Information: These features provide insights into the region where the applicant lives, including normalized population metrics and housing type.
4. Incidental Information: They include information on contract types (e.g., cash or revolving), the credit amount, loan annuities, and the price of goods financed.
5. Existing Credit Scores: The dataset incorporates credit scores from three external rating agencies namely- Equifax, TransUnion, and Experian.

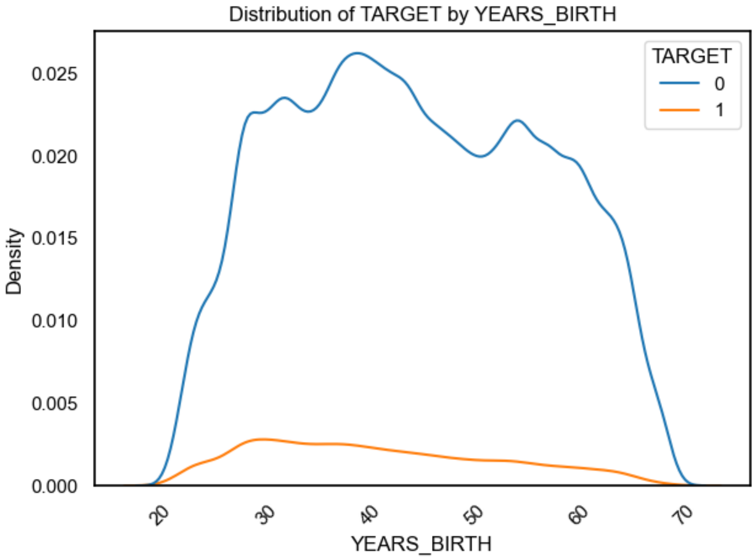
iii. Data Visualization

For the most commonly observed variables, we have performed Exploratory Data Analysis to examine the relationship with default.



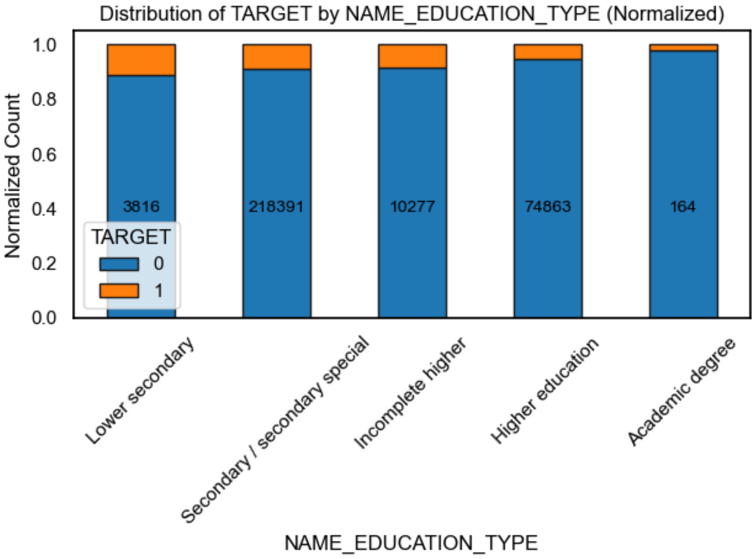
Distribution of Defaulters by Gender:

- Female borrowers are almost double that of Male ones. It could be that the dataset is from a region where Home Credit favors Female borrowers by providing Interest Rate Concessions or other Subsidies.
- Default rate of Female – 7%
- Default rate of Male – 10.14%



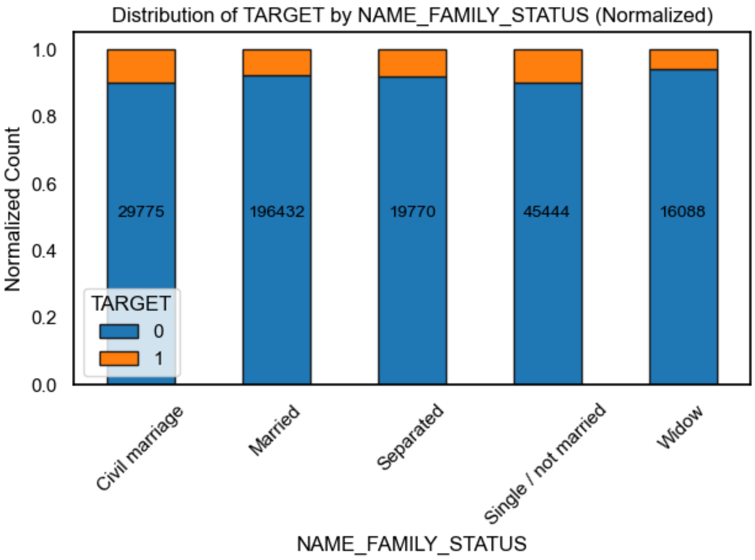
Distribution of Defaulters by age:

- Except for 20-30 and 60-70 year olds it can be said that loans are fairly uniformly granted across age groups by Home Credit, and negligible age discrepancy is observed.
- Distribution of Defaulters are, however, right skewed with younger people defaulting on more instances.



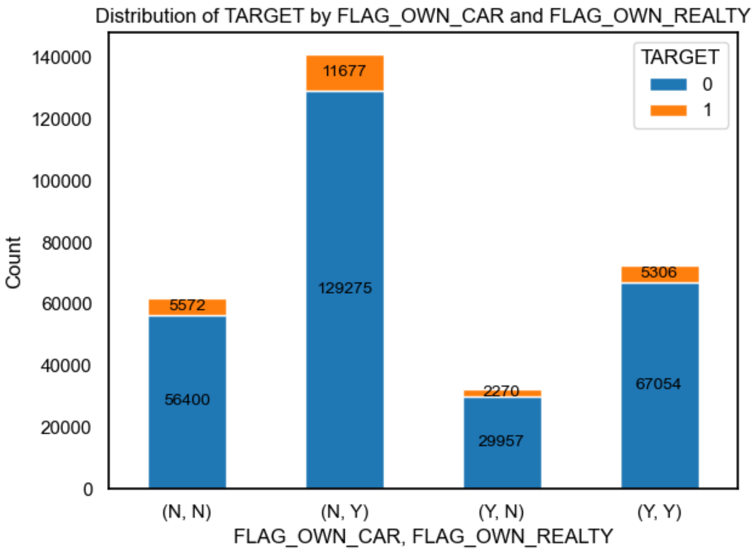
Distribution of Defaulters by Education level:

- We observe that most of the borrowers have their highest education at the Secondary level.
- Education level also seems to have a negative correlation with Default.



Distribution of Defaulters by Marital Status:

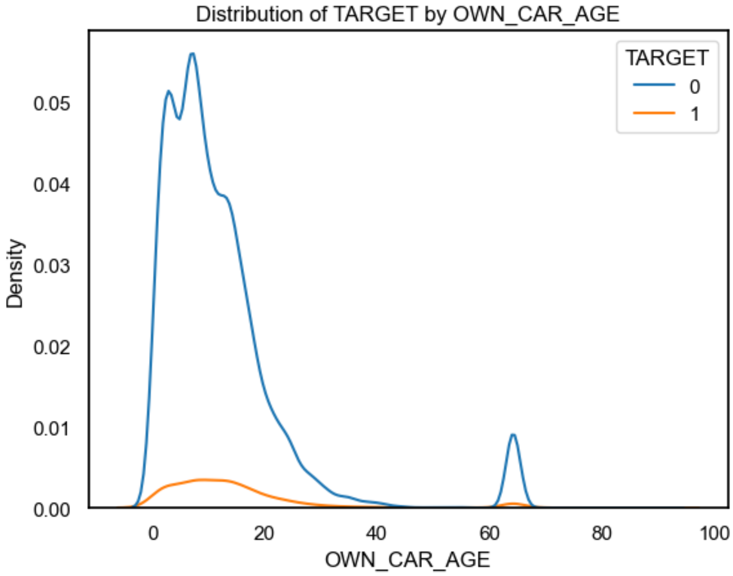
- As expected, more loans have been extended to married individuals.
- However, in percentage terms defaulters among each group does seem similar.



Distribution of defaulters in vehicle and property owners:

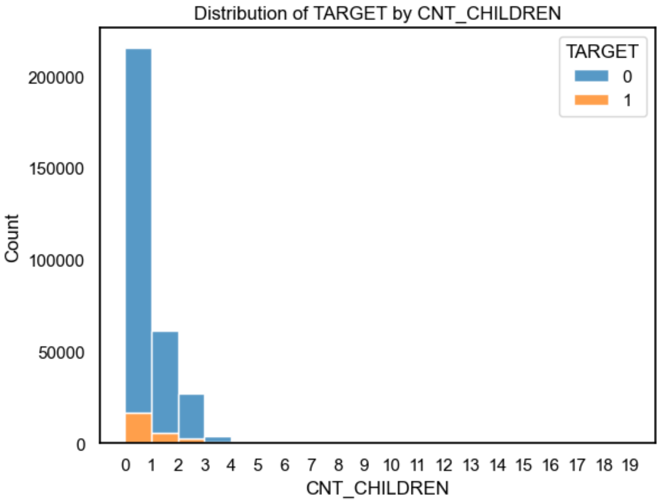
- It's well known that the materials one possesses say a lot about one's financial behaviour. Property and vehicles are two of the biggest financial assets for any common individual, so general wisdom suggests that such asset owners tend to be good investments as well.
- However, from the data we also observe a very interesting fact – the second bar corresponds to individuals who do not own a vehicle but own a property. We can see that the proportion of defaulters is highest in this category.

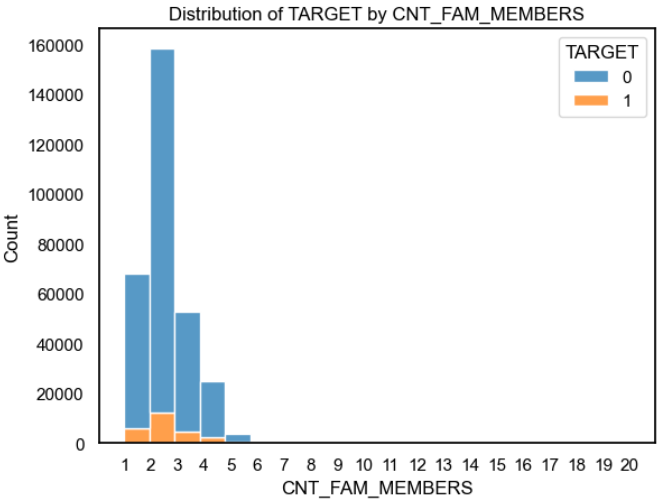
- Again, from the third bar, we see that people who do own a car but not a house are proportionately the lowest defaulters.
- This may suggest that people already having a house may risk the foreclosure on an additional property due to default more so than people who depend on their single house for sustenance.



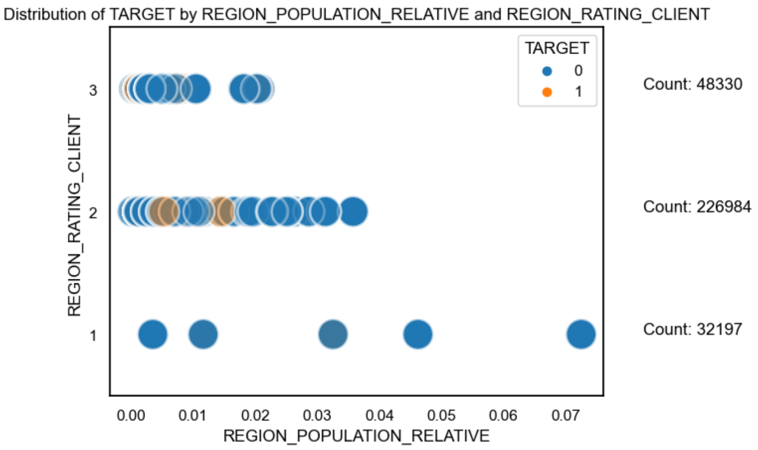
Distribution of defaulters by owned vehicles age:

- Nothing in the data suggests any particular relationship to defaulting on loan payments.
- One interesting thing to observe is that vintage car owners have very low proportion of defaulters, which makes sense because they're generally rich hobbyists.

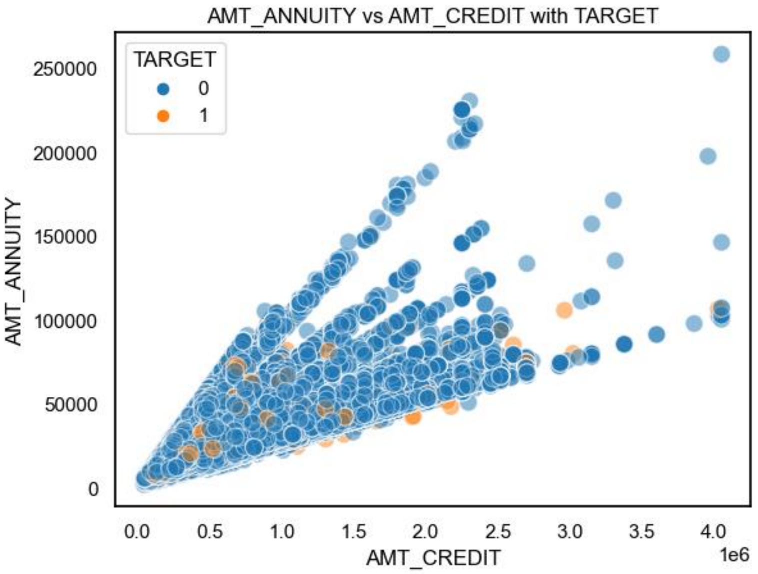




The above two graphs show that number of children and family size of the household cannot serve as good predictors of our target.

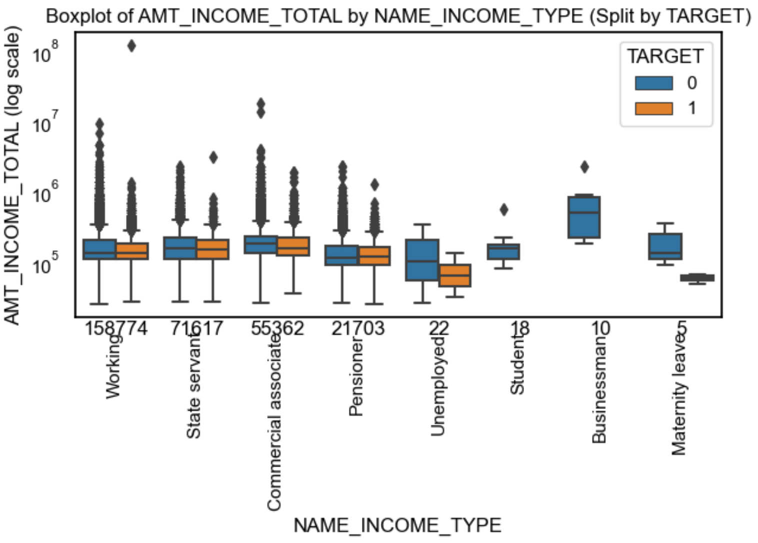


- Home Credit has internal ratings for the region of the client. From the scatterplot, we see that Region 1 includes 5 specific regions ranging all the way from low to high populations.
- Region 3 has mostly has clients from low population centers. Most of the clients are from Region 2, which are low to medium population centers.
- Similar trend in the proportion of Defaulters in also seen, where we have most Defaulters from Region 2 and also none from Region 1.

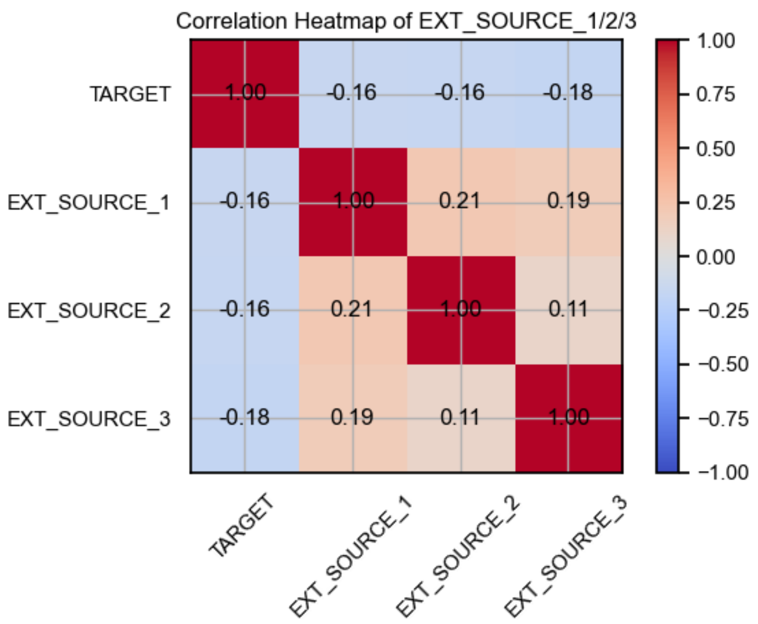


AMT_ANNUITY vs AMT_CREDIT

- AMT_CREDIT is the original loan amount and AMT_ANNUITY is the repayment a borrower has to make in a year.
- The resulting scatterplot is obvious. For similar loan amounts, different borrowers get different interest rates, and chose to pay in different tenures. So, annuity amount would vary based on other factors in conjunction with loan amount.
- Default seems to have no relation with the loan amount as we see occurrence of defaults in all loan amounts.



The boxplot shows us that income levels of the defaulters are not much different from the non-defaulters.



Correlation Heatmap of the Target along with credit scores from three external credit rating agencies.

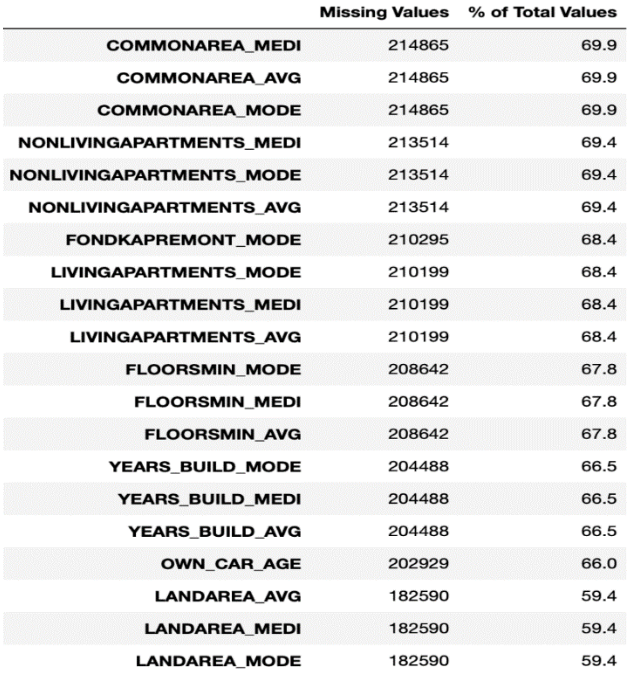
- Scores from these credit Bureaus rely on loan and payment information sent by banks and financial institutions. As such, even though they are independent and have their own method of scoring an individual borrower, they generally reflect the same information and can be used interchangeably in most cases.
- A good repayment discipline results in a higher credit score, and as such is visibly correlated negatively with our Target, although it's a weak correlation.

IV. Analysis

i. Data Cleaning

Handling the missing values:

In our data preprocessing phase, we encountered a dataset with dimensions of (307511, 122), indicating 307,511 instances with 122 features. However, upon closer inspection, we identified several columns that contained many missing values.



	Missing Values	% of Total Values
COMMONAREA_MEDI	214865	69.9
COMMONAREA_AVG	214865	69.9
COMMONAREA_MODE	214865	69.9
NONLIVINGAPARTMENTS_MEDI	213514	69.4
NONLIVINGAPARTMENTS_MODE	213514	69.4
NONLIVINGAPARTMENTS_AVG	213514	69.4
FONDKAPREMONT_MODE	210295	68.4
LIVINGAPARTMENTS_MODE	210199	68.4
LIVINGAPARTMENTS_MEDI	210199	68.4
LIVINGAPARTMENTS_AVG	210199	68.4
FLOORSMIN_MODE	208642	67.8
FLOORSMIN_MEDI	208642	67.8
FLOORSMIN_AVG	208642	67.8
YEARS_BUILD_MODE	204488	66.5
YEARS_BUILD_MEDI	204488	66.5
YEARS_BUILD_AVG	204488	66.5
OWN_CAR_AGE	202929	66.0
LANDAREA_AVG	182590	59.4
LANDAREA_MEDI	182590	59.4
LANDAREA_MODE	182590	59.4

- To address this issue, we implemented a strategy to drop columns with more than 60% missing values, aiming to maintain data integrity while reducing dimensionality and computational complexity. Following this approach, the shape of our training data was transformed to (307511, 105), resulting in a new dataframe comprising 105 columns.
- By selectively removing features with a significant proportion of missing values, we ensured that our dataset remained robust and conducive to subsequent analysis and modelling efforts.
- To handle the missing values, we utilized the SimpleImputer class from the scikit-learn library to handle missing values in our dataset. The strategy employed for imputation was based on replacing missing values with the median of each respective numerical feature.

ii. Data Transformation

In order to include our categorical variables in the machine learning models, we applied encoding techniques tailored to the nature of each variable.

1. Label encoding:

For binary categorical variables, which encompass only two unique values, we employed label encoding. This involved transforming the categorical values into numerical representations, assigning 0 and 1 to each unique category using the LabelEncoder from the scikit-learn library. By doing so, we ensured that these binary categories were appropriately encoded for subsequent analysis and modelling.

2. One-hot encoding:

On the other hand, for categorical variables with more than two unique values, we utilized one-hot encoding to create binary dummy variables. This process involved expanding each categorical variable into multiple binary columns, each representing a unique category. We leveraged the get_dummies() function from the pandas library, setting drop_first=True to avoid multicollinearity issues by dropping the first dummy variable for each category.

By combining label encoding and one-hot encoding techniques, we effectively prepared our categorical data for robust analysis and modelling, enhancing the predictive power of our machine learning algorithms.

iii. Data Modelling

Splitting The Dataset

In the dataset splitting phase, we divided our data into features and target labels. The features were stored in variable X, while the target label, representing the 'TARGET' column, was stored in variable y. We then split the data into training and validation sets using the train_test_split function from scikit-learn, with a test size of 20% and a random state of 42 for reproducibility.

Subsampling for Model Training

To optimize the training process and efficiently tune hyperparameters, a subsampling approach was utilized. A 10% subsample of the original training data (X_train) was randomly selected without replacement to form X_train_subsample and y_train_subsample. This subsampling strategy helps in reducing the computational cost and time, especially useful when dealing with large datasets or when performing computationally expensive model training and validation techniques.

iv. Model Training

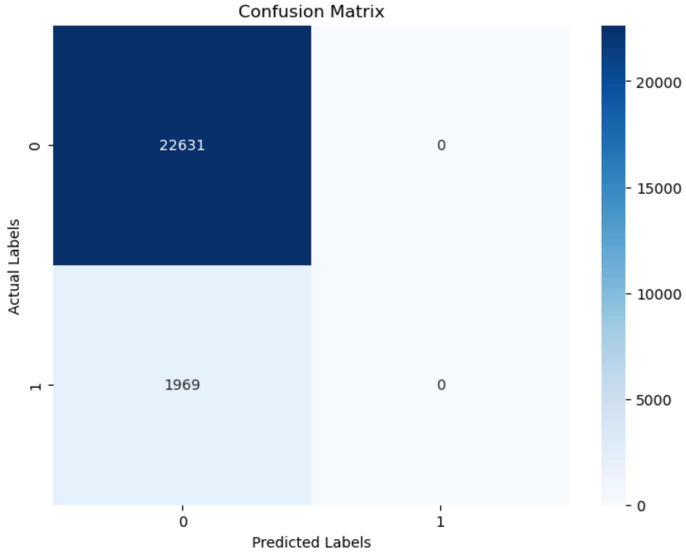
1. Logistic Regression

The Logistic Regression model was initialized with a maximum iteration parameter of 1000 to ensure convergence. Training was performed on the subsample dataset, which contains 2,6000 observations with 1,9690 (approximately 75.7%) belonging to the negative class (0.0) and 6310 (approximately 24.3%) to the positive class (1.0). This indicates a class imbalance within the subsample.

Model Performance

The model achieved an accuracy of 92.00% on the training subset. This high accuracy rate suggests that the model is capable of distinguishing between the two classes effectively within the subsampled dataset.

Confusion Matrix Analysis



The provided confusion matrix shows the results from the Logistic Regression model when evaluated using a subsample of the training dataset for hyperparameter tuning.

The absence of True Positives and the presence of a high number of False Negatives show that the model is heavily biased towards predicting negatives, which might be a result of class imbalance, lack of representative features for the positive class in the training data, or inadequate model complexity to capture the nuances distinguishing the positive class instances.

Implementation of SMOTE for Class Balancing

To address the significant class imbalance observed in the initial model evaluation, SMOTE was applied to the training data. This technique generates synthetic samples from the minority class (class 1 in this case) to achieve a balanced distribution between the two classes. After resampling, both classes had an equal count of 22,631 instances, effectively eliminating the class imbalance issue.

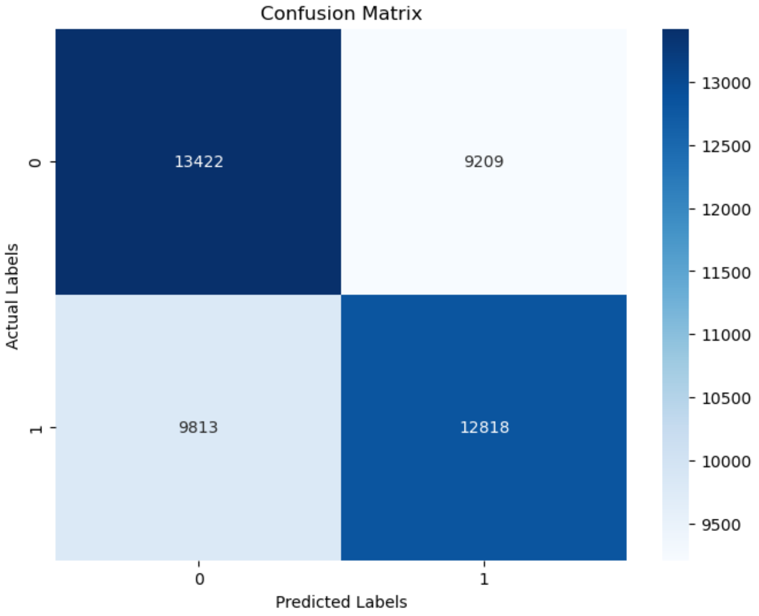
2. Logistic Regression Retraining

The Logistic Regression model was retrained using the newly balanced dataset. The model was set up with a maximum of 1,000 iterations to ensure convergence.

Performance Metrics

- The model achieved an accuracy of 57.97% on the resampled training data. This accuracy is notably lower than the initial accuracy observed on the unbalanced data, which can be attributed to the increased challenge of distinguishing between more equally represented classes.

Confusion Matrix Analysis

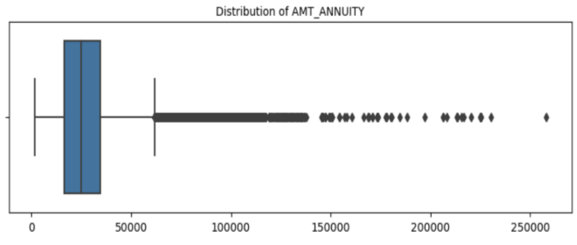


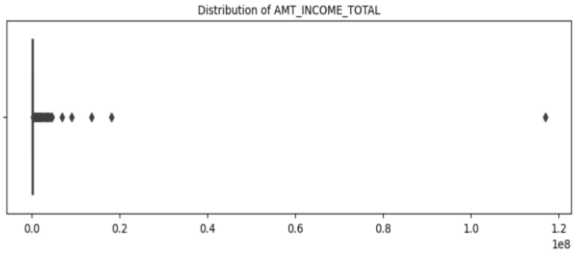
- Sensitivity: The model has significantly improved in identifying the positive class, as evidenced by 12,818 true positives, compared to zero in the initial unbalanced model.
- Specificity: There is a decrease in the model’s ability to correctly identify all negative cases, with 9,209 false positives indicating a trade-off between sensitivity and specificity.
- Overall Evaluation: The balance between sensitivity and specificity is a common challenge in classification tasks, particularly in scenarios involving previously imbalanced classes. The increase in false positives and negatives is indicative of this trade-off.

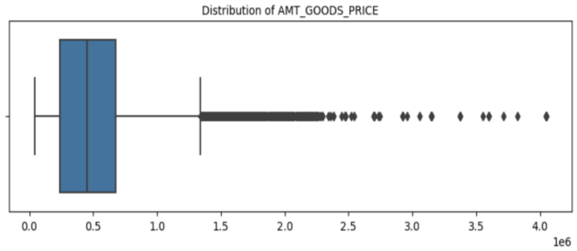
3. Further Pre-processing of Data

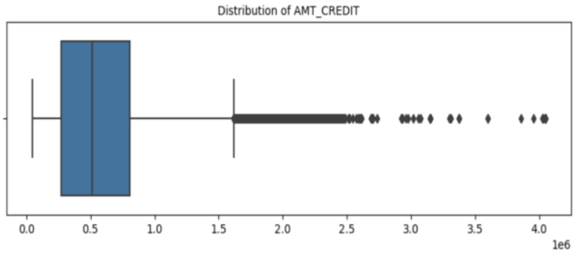
Handling the outliers: In our data exploration phase, we identified several important numerical features that may contain outliers, including 'AMT_INCOME_TOTAL', 'AMT_CREDIT', 'AMT_ANNUITY', and 'AMT_GOODS_PRICE'. To visually inspect the distribution of these features and identify potential outliers, we created box plots for each feature using Seaborn.

Subsequently, we implemented a robust outlier detection and removal method utilizing the Interquartile Range (IQR) technique.

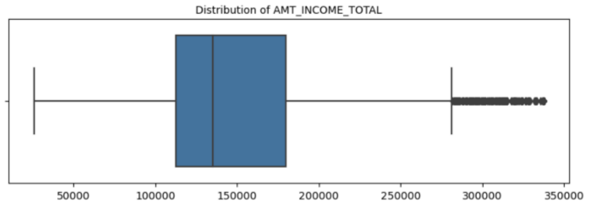


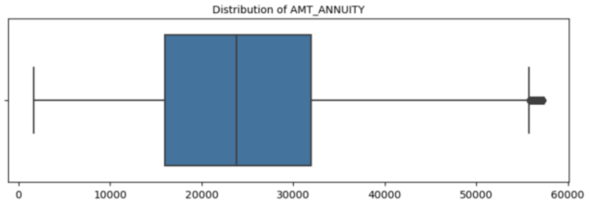


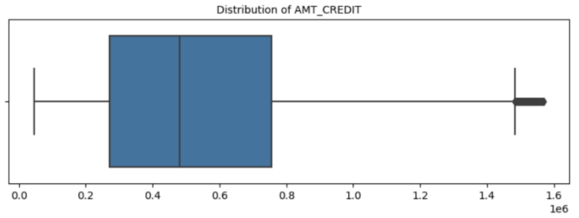


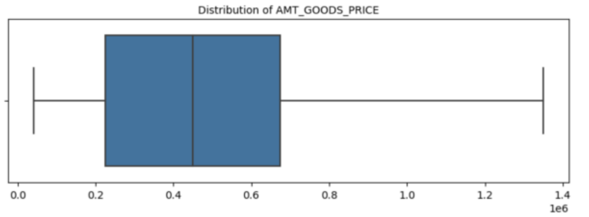


This iterative process was applied to each numerical feature, resulting in the removal of outliers and ensuring that our dataset remained free from potentially influential data points that could skew our analysis and modelling efforts.







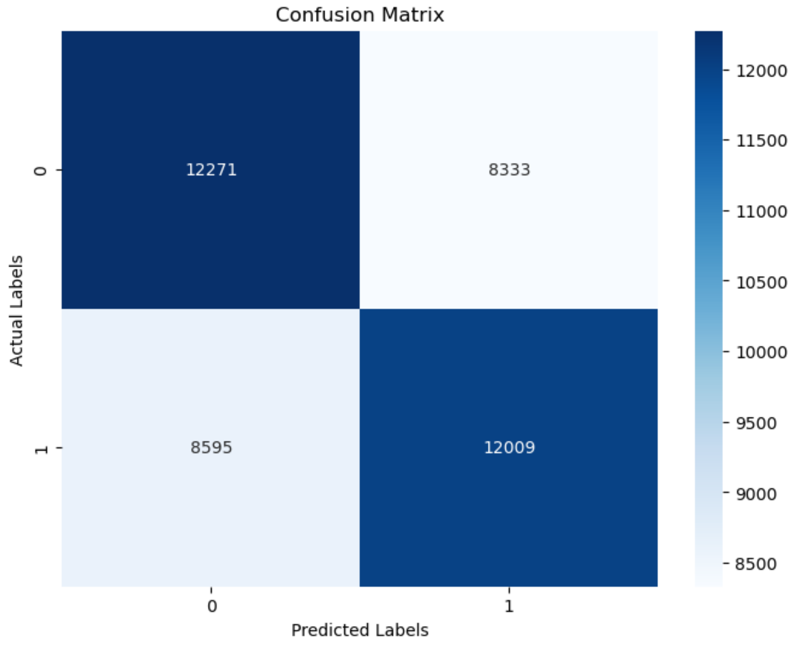


Finally, we quantified the number of outliers filtered for each feature, providing insight into the impact of our outlier removal procedure on the dataset.

AMT_INCOME_TOTAL - Filtered 14035 outliers
AMT_CREDIT - Filtered 5752 outliers
AMT_ANNUITY - Filtered 4998 outliers
AMT_GOODS_PRICE - Filtered 1868 outliers

4. Updated Model Performance After Handling Outliers

The Logistic Regression model was further refined by addressing outliers in the dataset. Outliers can significantly skew the results of logistic regression, which assumes a linear relationship between the log odds and the predictors.



Confusion Matrix

Model Accuracy: The accuracy of the model, post-outlier handling, is approximately 58.92%, which shows a slight decrease compared to the accuracy reported before addressing outliers (57.97%). This marginal improvement suggests that while outlier management has somewhat enhanced model performance, the balance between sensitivity and specificity continues to pose challenges.

5. Data Scaling

Min-Max Scaling

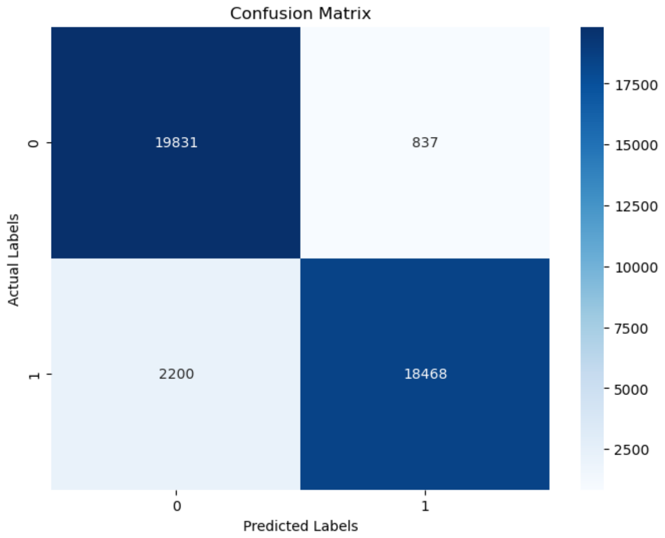
We employed the Min-Max scaling method to standardize the numerical features in our dataset. To initiate the scaling process, we instantiated the MinMaxScaler object from the scikit-learn library, setting the feature_range parameter to (0, 1). This configuration instructs the scaler to transform the features to a range between 0 and 1, facilitating uniformity in the scaling transformation.

To identify the numerical features requiring scaling, we iterated through the columns of the DataFrame and selected those with maximum values exceeding 1. This criterion effectively captures features with a wide range of values, indicating the need for normalization to bring them within a consistent range.

Subsequently, we applied the scaling transformation to the selected features using the fit_transform() method of the MinMaxScaler object. This process rescaled the numerical values of each feature to fall within the specified range of (0, 1), thereby normalizing the data distribution and improving the numerical stability of our dataset.

Enhanced Model Performance After Feature Scaling

Following the application of feature scaling to the dataset, a significant improvement in the Logistic Regression model's performance was observed. Scaling is a critical preprocessing step, especially for algorithms like Logistic Regression, which can be sensitive to the range of input features. It ensures that each feature contributes proportionately to the prediction, preventing features with larger scales from dominating the model's decision process.



Model Accuracy

The accuracy of the model has dramatically increased to approximately 92.65%. This notable increase compared to previous iterations (where accuracy was around 58.59%) underscores the importance of appropriate feature scaling in improving model predictions.

Analysis of Results

- Reduction in False Positives and False Negatives: The substantial decrease in both false positives and false negatives highlights the model's enhanced capability to correctly classify both classes without being biased towards one. This balance indicates a well-tuned model that effectively recognizes the patterns for both the positive and negative classes.

- Improvement in True Positives and True Negatives: The significant increase in true positives, alongside a robust count of true negatives, suggests that the model is now much better at handling the classification task, making reliable predictions across the board.

6. Model Selection

After data pre-processing, in our model selection process, we considered several classifiers to evaluate their performance in predicting loan default risk. The classifiers considered include Logistic Regression, K-Nearest Neighbors, Decision Trees and Support Vector Machines (SVM).

By utilizing pipelines, we ensured consistency in preprocessing steps across different classifiers, facilitating fair comparison of their performance. Through grid search with cross-validation using GridSearchCV, we systematically explore hyperparameter combinations to optimize each classifier's performance.

Hyperparameter Tuning

1. Logistic Regression: We utilized a regularization parameter (C) to control model complexity, with values ranging from 0.001 to 10, aiming to find the optimal balance between fitting the training data and avoiding overfitting.
2. K-Nearest Neighbors (KNN): We adjust the number of neighbors considered for classification, exploring options of 5, 10, and 15 neighbors to determine the optimal neighborhood size for accurate predictions.
3. Decision Tree: We tune tree depth, minimum samples for splitting, and minimum samples per leaf to control tree complexity and prevent overfitting, with options including unlimited depth or maximum depths of 10 or 20 and varying minimum sample thresholds.
4. Support Vector Machine (SVM): We adjust the penalty parameter (C) to balance the classification error and margin width, exploring a range of values from 0.001 to 10 to optimize the SVM's ability to generalize to unseen data.

V. Results

1. Logistic Regression:

Best parameters for Logistic Regression: {'C': 1}

Best score for Logistic Regression: 0.877

2. K-Nearest Neighbors (KNN):

Best parameters for KNN: {'n_neighbors': 5}

Best score for KNN: 0.8257

3. Decision Tree:

Best parameters for Decision Tree: {'max_depth': 20, 'min_samples_leaf': 10, 'min_samples_split': 2}

Best score for Decision Tree: 0.8959

4. Support Vector Machine (SVM):

Best parameters for SVM: {'C': 0.001}

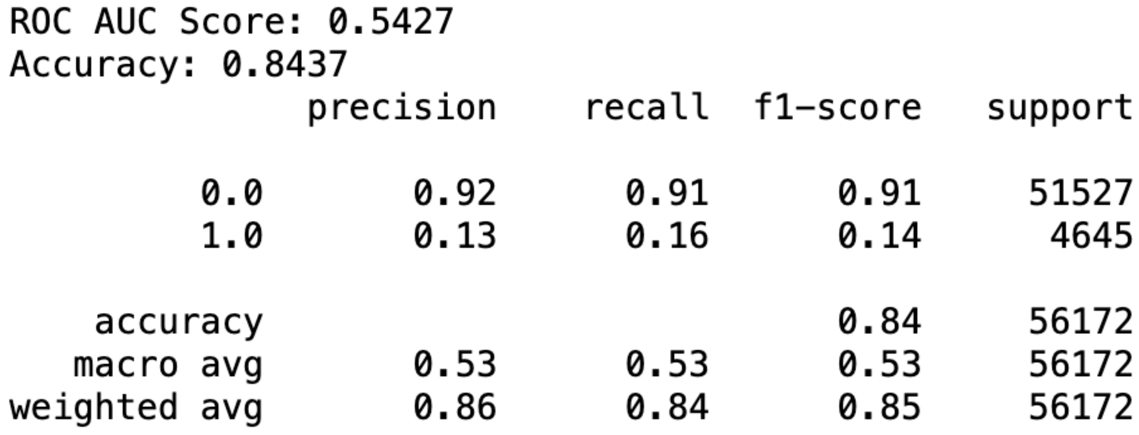
Best score for SVM: 0.5005

Model Evaluation

The Decision Tree model provided the best results among the models evaluated and was selected for final predictions. The model achieved a best score of approximately 0.896 during cross-validation on the training set, indicating robust performance.

- ROC AUC Score: The model achieved an ROC AUC Score of 0.5427, which suggests moderate discriminatory ability.
- Accuracy: The accuracy on the test set was 84.37%, which is relatively high, although this metric alone can be misleading due to the class imbalance present in the dataset.

Classification Metrics

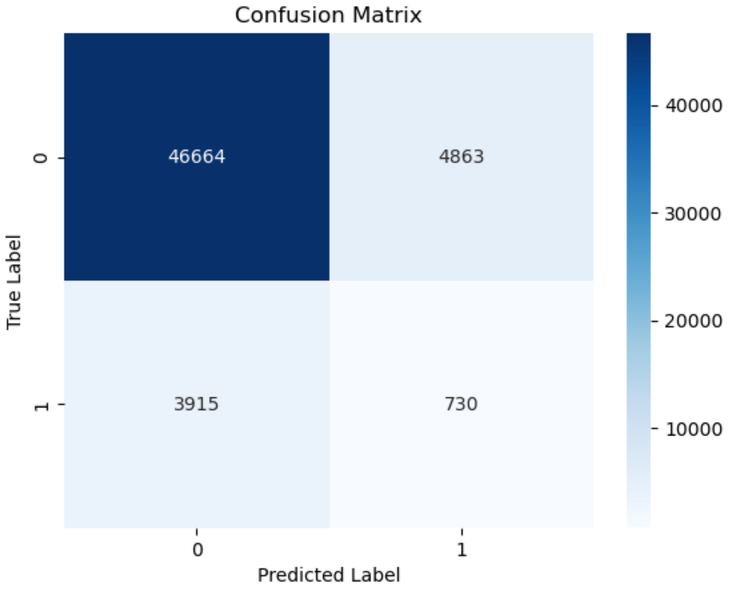


```
ROC AUC Score: 0.5427
Accuracy: 0.8437
              precision    recall  f1-score   support

         0.0       0.92      0.91      0.91     51527
         1.0       0.13      0.16      0.14      4645

    accuracy                           0.84     56172
   macro avg       0.53      0.53      0.53     56172
weighted avg       0.86      0.84      0.85     56172
```

The Classification report provides the following insights:

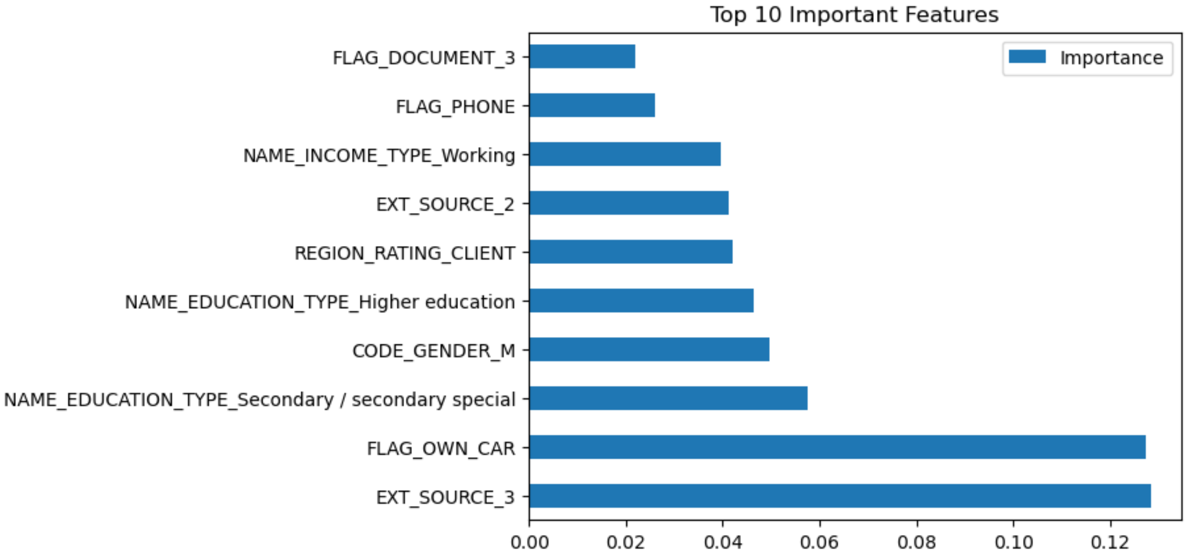
- Precision: Low for the positive class (0.13), suggesting a high number of false positives.
- Recall: Moderately low for the positive class (0.16), indicating many positives were missed.
- F1-Score: The harmonic mean of precision and recall for the positive class was also low (0.14), reflecting the difficulty of predicting the positive class correctly.

Confusion Matrix Analysis



The confusion matrix displayed significant true positives and true negatives but also a substantial number of false positives and false negatives, reflecting the challenges in accurately classifying the positive class.

Feature Importance



The feature importance plot reveals key attributes that most significantly impact the predictions of the model. Understanding these features can provide insights into the factors driving the outcomes and can inform strategies for both model refinement and operational decisions. Here's a breakdown of the top 3 features listed by their importance:

1. EXT_SOURCE_3: The most influential feature, likely representing an external data source or score. Its high importance indicates that it carries substantial predictive power, possibly summarizing applicant creditworthiness based on external evaluations.
2. FLAG_OWN_CAR: Indicates whether the applicant owns a car. Surprisingly, this feature ranks high, suggesting a potential correlation between car ownership and the target variable, such as credit risk or loan repayment behavior.
3. NAME_EDUCATION_TYPE_Secondary / secondary special: Reflects the level of education. This feature's prominence in the model highlights the role of education level in influencing the applicant's profile and associated risks.

VI. Discussion

i. Takeaways from Data Mining Analysis

1. Effective Data Cleaning: Ensuring data integrity through meticulous cleaning processes is crucial. Removing duplicates and handling missing values are essential steps in preparing data for analysis.
2. Outlier Detection and Handling: Identifying and managing outliers is vital for maintaining the reliability of predictive models. Outlier management helps in achieving more generalized and robust models.
3. Feature Encoding and Data Transformation: Proper encoding of categorical variables and normalization of numerical features are critical for the smooth performance of machine learning algorithms.
4. Class Imbalance Management: Addressing class imbalance with techniques like SMOTE is fundamental in training models that perform well across all categories of data.

ii. Interpretation and Analysis for Business Managers

The data analysis offers insights into risk patterns associated with loan defaults, which can be instrumental for:

- Enhancing Credit Decision-Making: Leveraging predictive analytics to refine credit scoring models, thereby facilitating more accurate loan approval decisions.
- Risk Management Strategies: Developing targeted strategies to mitigate risks associated with potential loan defaults.
- Customer Segmentation: Identifying customer segments that exhibit different risk profiles, enabling tailored marketing and risk management strategies.

iii. Other Recommendations

1. Continuous Monitoring and Updating of Models: Regularly update models to incorporate new data and trends, which helps in maintaining the accuracy and relevancy of the predictions.
2. Exploration of Advanced Analytical Techniques: Investigate the use of advanced machine learning and statistical techniques to enhance predictive accuracy.
3. Integration of Additional Data Sources: Augment existing data with additional external data sources to enrich the analysis and provide deeper insights into customer behavior.