# TEAM 21

# PROJECT REPORT

# Housing Loan Repayment Risk Analysis

**Report by**:

Darshil Shah

Bhavya Parmar

**Appendix**

Introduction………………………………………………………………………………....3

Methodology…………………………………………………………...…………………...4

Description of Dataset………………………………………………….…………………...4

Data cleaning and Pre-processing……………………………………...…………………...5

Exploratory Data Analysis……………………………………………..…………………...8

Feature Engineering………………………………………………………………………..12

Model Selection and Evaluation…..……………………………………………………….13

Result and analysis…………………………………………………………………………16

Conclusion……………………………………………………………..…………………...17

Summary and Future Work…………………………………………………………………18

References……………………………………………………………..……………………19

**Introduction**

**Background:**

The “Housing Loan Repayment Risk Analysis” dataset is a very essential source that will go a long way in advancing the cause of credit risk assessment in the financial sector. It is a compilation of very many pieces of loan application data, starting from financial status and demographics to credit history used in understanding loan repayment behaviors. In this perspective, the dataset paves the way for data-driven decision-making that allows detailed research on the borrower's risk factors. It comes as the instrument of developing predictive models for loan management and, in a way, mirrors the industry's march towards analytics-based more precise and effective risk mitigation strategies.

### **Motivation:**

This initiative is propelled by the need to refine the precision of credit risk assessments within the finance industry, advancing beyond the confines of traditional assessment methods. By leveraging the “Housing Loan Repayment Risk Analysis” dataset, our goal is to harness sophisticated analytics and machine learning techniques to craft predictive models capable of accurately evaluating the risks associated with loan repayments. In essence, this project aspires to revolutionize risk management approaches, thus contributing to the stabilization of the financial sector and expanding the availability of credit, in harmony with the objectives of financial inclusion.

### **Goal:**

The core goal of this initiative is to craft and deploy sophisticated predictive models capable of accurately predicting challenges in loan repayment. Utilizing the extensive data housed within the “Housing Loan Repayment Risk Analysis” dataset, our objectives include:

- Elevating the precision and efficiency of credit risk evaluations in the finance industry.

- Uncovering crucial indicators of loan repayment challenges through advanced analytics and machine learning.

- Diminishing loan default rates by equipping financial entities with more accurate risk assessment tools.

- Refining the loan approval decision-making process to be more knowledgeable and fairer.

This project is set to transform the practices of credit risk management fundamentally, bolstering the lending system's stability and trustworthiness.

**Methodology**

* **Data Preprocessing and Cleaning**: To ensure the “Housing Loan Repayment Risk Analysis” dataset's accuracy and relevance, the initial step will involve cleaning and preprocessing the data. This process will include imputation of missing values, standardization of formats and data types, detection and correction of anomalies or outliers, and ensuring the data is in a state ready for analysis.
* **Exploratory Data Analysis**: We will engage in exploratory data analysis (EDA) using tools such as correlation matrices, scatterplots, and histograms to uncover patterns and trends within the loan data. This phase will involve visual representation of data through various graphs and charts and the application of statistical techniques to identify correlations between different factors, such as financial background, employment status, and loan specifics (e.g., amount, type, and repayment terms).
* **Feature Engineering**: To enhance the precision of our predictive models, we will create new variables or feature selection based on domain knowledge and statistical methods to identify significant predictors of loan repayment challenges. Key factors likely to influence our feature selection include applicants' financial backgrounds, employment status, loan amount, annuity, type of loan, repayment terms, and credit history. This step is pivotal in refining our models to accurately predict loan repayment risks. We will use methods like one-hot encoding on the categorical columns like employment status, type of loan which will help in accuracy of the model.
* **Cross-validation:** Is a technique to ensure robust model performance. The dataset is divided into equal parts (folds), and the model is trained and validated across these folds. This method provides a comprehensive assessment of the model's ability to generalize. It further includes:

1. Training Phase:In this phase, the model learns from the data by adjusting its parameters to predict outcomes based on input features. Typically, about 70-80% of the dataset is used for training, allowing the model to learn the relationship between inputs and outputs.
2. Testing Phase: In this phase assesses the model's performance on another set of unseen data—the test set. This final evaluation, using about 10-15% of the dataset, provides an unbiased measure of how the model will perform in real-world scenarios.

* **Model Selection and Evaluation:** We plan to utilize machine learning algorithms such as Logistic Regression, Decision Trees, Random Forests to predict loan repayment challenges. These models will be trained on historical data from the “Housing Loan Repayment Risk Analysis” dataset to evaluate their accuracy.

**Description of Dataset**

This project focuses on utilizing the “Housing Loan Repayment Risk Analysis” dataset to develop predictive models aimed at improving the precision of credit risk assessments within the finance industry. The dataset includes comprehensive loan application data, such as financial background, employment status, family circumstances, and loan specifics of thousands of applicants.

The Key columns from this dataset are:

* **Applicant Financial Background**: This column contains detailed information about the loan applicant's financial history, including income level, existing debts, and overall financial health. It's crucial for assessing the applicant's ability to repay the loan.
* **Employment Status:** Indicates whether the applicant is employed, unemployed, self-employed, or retired, providing insights into their stable income sources.
* **Family Circumstances**: Includes data on marital status, number of dependents, and living situation.
* **Loan Amount:** The total amount of money requested by the applicant. This figure is key to determining the scale of the loan and its repayment structure.
* **Annuity:** The regular payment amount the applicant is obligated to pay towards the loan. It helps in understanding the loan's impact on the applicant's monthly budget
* **Type of Loan:** Specifies the loan's purpose, such as mortgage, personal loan and auto loan.
* **Repayment Terms:** Details of the length of the loan and any specific repayment conditions. It's critical for understanding the loan's duration and the borrower's long-term financial commitment.
* **Credit History:** A record of the applicant's past borrowing and repayments. This column is vital for assessing the borrower's reliability and predicting their future loan repayment behavior.
* NAME\_CONTRACT\_TYPE: Type of the loan (Cash vs Revolving)
* CODE\_GENDER: Gender of person taking loan.
* FLAG\_OWN\_CAR: To indicate if the person owns a car or not.
* AMT\_INCOME\_TOTAL: Total annual income the person earns
* AMT\_CREDIT: Loan amount the person applied loan for
* AMT\_ANNUITY:
* AMT\_GOODS\_PRICE: Collateral Amount present with the person
* NAME\_TYPE\_SUITE: To indicate loan type, if it is individual loan or joint loan
* NAME\_INCOME\_TYPE: Type of applicant whether employed, pensioner, etc
* NAME\_EDUCATION\_TYPE: Indicates level of education person has achieved
* NAME\_FAMILY\_STATUS: Whether single or married
* CNT\_CHILDREN : Indicates the count of children the person has
* NAME\_HOUSING\_TYPE:
* REGION\_POPULATION\_RELATIVE:
* DAYS\_BIRTH:
* DAYS\_EMPLOYED
* DAYS\_REGISTRATION
* TARGET:

**Data Source**: <https://www.kaggle.com/datasets/urstrulyvikas/house-loan-data-analysis>

**Dataset size**: 158 MB

A screenshot of a computer

Description automatically generated

When we execute df.info() on a DataFrame df, it prints out information about the DataFrame, including:

**Observation:**

1. **Range Index**: The range of row labels i.e. 307511
2. **Data columns**: The column names and data types of each column.
3. **Non-null counts**: The count of non-null values for each column.

This method is particularly useful for getting a quick overview of the DataFrame's structure, data types, and memory usage, especially when working with large datasets.

After getting an overview of the dataset we can now deep dive into understanding the data to get our target variable estimation which is our goal.

We will now proceed with the methodologies.

In order to preprocess and clean the raw data we do the following:

1. **Missing Values**:

The output we got shows that some columns have fewer non-null values than others. For example, the AMT\_ANNUITY column has 307499 non-null values, and the AMT\_GOODS\_PRICE column has 307233 non-null values out of 307511 total rows.

These columns have only a few rows of missing values so we decided to drop them.

1. **Exploratory Data Analysis:**

This step would involve visually and statistically exploring the dataset to gain insights into its structure, patterns, and relationships between variables. Techniques include summary statistics, data visualization (e.g., histograms, scatter plots, box plots), and correlation analysis.

We found the correlation between different features and plotted the graph between them.

For example: Below we find that the AMT\_ANNUITY is highly correlated to AMT\_CREDIT

A graph showing a graph of an annuity

Description automatically generated

Exploratory Data Analysis (EDA) is a crucial step to understand the characteristics of the data and identify potential patterns, relationships, and issues that may impact the model's performance. Based on the provided columns, here's a short paragraph on EDA for this dataset:

EDA should focus on examining the distribution and relationships among numerical features like: AMT\_CREDIT, DAYS\_BIRTH, AMT\_ANNUITY. Visualizations such as histograms, box plots, and scatter plots can reveal skewness, outliers, and correlations. For categorical features like NAME\_INCOME\_TYPE, NAME\_EDUCATION\_TYPE, bar plots and countplots can explore their distributions and potential impact on house loan default prediction.

**Categorical Variables:**

In categorical variable EDA, we analyze distribution and relationships. Using visualizations like bar plots, we explore frequency distribution, identify outliers, and understand patterns. Cross-tabulation and heatmaps reveal interactions between categories. Bar plots and box plots help assess the relationship with the target variable. We address missing values, outliers, and consider variable transformations. Statistical tests determine significance, guiding subsequent analysis. These insights inform feature engineering and model building.

Top of Form

A pie chart with numbers and a number of people

Description automatically generated with medium confidence

**Analysis:**  Here we find that the people in the Working category have more applications than any other type.

**A blue pie chart with a blue circle and a blue circle with a blue circle with a blue circle with a blue circle with a blue circle with a blue circle with a blue circle with a blue

Description automatically generated**

**Analysis:**  Majority of the loan applicants have higher/ Secondary education with them.

**Correlation between different columns**

A chart with red and blue squares

Description automatically generated

**Feature Engineering**

Now we would be transforming existing features to improve the performance of machine learning models. We will compare many different columns to see which when combined will help us predict the home loan defaulter more accurately. The goal is to extract relevant information from the raw data and create input features that enhance the predictive power of the models.

After cleaning the dataset and understanding the data we will now prepare the dataset in a way so that when we run the dataset under a model, we would be able to get the best results.

* **Label encoding**

Label encoding is a pivotal aspect of feature engineering, particularly for preparing categorical variables for machine learning models. By converting categorical data into numerical format, label encoding enables seamless integration of these features into the model's training process. It assigns unique numerical labels to each category, preserving their ordinality and facilitating interpretation. This transformation enhances model performance by allowing it to effectively learn from categorical information and capture underlying patterns in the data. Overall, label encoding plays a crucial role in enriching the feature space and optimizing the model's predictive capabilities.

* **Combining columns**

We will combine the input numerical columns and the encoded categorical columns to train and find the relationship between the columns.

We find there are 35 categorical columns that we can use to find the relationship and train our model for optimal results.

* **Balance the dataset**

We find that the output target of the dataset is highly imbalanced. The loan defaulter are less than 10% from total applications

A blue and orange pie chart

Description automatically generated

To solve this problem, we made use of 3 parts of the sample from the target 0 values and target 1 value and balanced the dataset with total 148824 samples.

A blue and orange pie chart

Description automatically generated

* **Scaling the numerical columns**

We have scaled the numerical columns such as AMT\_CREDIT, AMT\_INCOME as they have higher value which will affect the model by assigning higher weights to these columns and will lead to incorrect results.

* **Splitting the dataset into Train and Test set**

To test the model accuracy and precession we have divided our dataset into train and using train\_test\_split module from sklearn library

* **Model Selection and Evaluation**

Finally, we would be choosing appropriate machine learning algorithms for the predictive task based on the dataset characteristics and problem requirements. Techniques include evaluating various algorithms and selecting the best-performing ones based on performance metrics and finding which algorithm would give the best scores that would help us predict the home loan defaulters that are listed in the dataset.

Each of these methodologies plays a crucial role in the predictive modeling process, from data preparation and exploration to feature engineering and model selection. By systematically applying these techniques, we can develop accurate and reliable predictive models.

1. **Logistic Regression**

Leveraging the linear relationship between predictor variables and the target variable, linear regression aims to estimate the coefficients of these variables to form a linear equation that best fits the data and converting it into the probably using sigmoid function to classify as defaulter.

In the context of home loan defaulter risk prediction, predictor variables typically include features like annuity, income earned, etc. By fitting a linear regression model to historical loan data, the algorithm learns the relationships between these features, enabling it to make predictions for new instances. Additionally, techniques such as feature scaling, regularization, and feature selection are often applied to enhance model performance and mitigate overfitting.

**Results:**

A blue squares with white text

Description automatically generated

A screenshot of a computer

Description automatically generatedTop of Form

Bottom of Form

Results: We achieved max f1 score of 0.62 using Logistic RegressionBottom of Form

1. **Decision Tree:**

Decision trees are powerful tools used in machine learning for predictive modeling, and they can play a crucial role in predicting house loan defaulters. By analyzing historical data on loan applicants and their repayment behaviors, decision trees can identify patterns and factors that are predictive of whether a borrower is likely to default on their house loan. These factors may include variables such as credit score, income level, employment status, debt-to-income ratio, and past credit history. By constructing a decision tree model based on these factors, lenders can assess the risk of potential borrowers and make more informed decisions on whether to approve a loan application. Additionally, decision trees provide transparency in the decision-making process, as they can illustrate the sequence of factors that lead to the prediction of loan default, helping lenders understand the underlying reasons behind the predictions and enabling them to take appropriate risk management measures. Overall, decision trees serve as valuable tools for lenders in accurately predicting house loan defaulters and mitigating financial risks associated with lending.

**Hyperparameter tuning:**

Hyperparameter tuning is a crucial step in optimizing the performance of decision tree models for predicting house loan defaulters. By adjusting hyperparameters such as the maximum depth of the tree, minimum samples required to split a node, and the criterion for splitting, the model's predictive accuracy can be enhanced. Techniques like grid search or random search can be employed to systematically explore different combinations of hyperparameters and identify the optimal configuration that minimizes prediction errors. For example, increasing the maximum depth of the tree may capture more complex relationships in the data but also risks overfitting, while setting a higher minimum samples per leaf node may prevent overfitting by enforcing a more generalizable model. Through hyperparameter tuning, decision tree models can be fine-tuned to achieve the best possible performance in predicting house loan defaulters, thus aiding lenders in making more accurate risk assessments and informed lending decisions.



Decision Tree with max\_depth of 2

A diagram of a diagram

Description automatically generated

Results:

A blue squares with white text

Description automatically generated

A screenshot of a computer

Description automatically generated

Results: Using Decision Tree we achieved F1-score of 0.87

1. **Random Forest**

Random Forest, a powerful ensemble learning technique, significantly contributes to predicting house loan defaulters by leveraging the collective intelligence of multiple decision trees. By constructing a multitude of decision trees on random subsets of the training data and aggregating their predictions, Random Forest effectively mitigates the risk of overfitting and enhances the model's generalization ability. Each decision tree in the Random Forest learns from different features and instances, capturing diverse patterns and nuances in the data. This diversity among the trees helps in identifying complex relationships and critical factors influencing loan default, such as credit history, income stability, and debt-to-income ratio. Furthermore, Random Forest provides robustness against outliers and noisy data, improving the reliability of predictions. By combining the strengths of individual decision trees, Random Forest delivers accurate and stable predictions of house loan defaulters, empowering lenders to make informed decisions and manage lending risks effectively.

A blue squares with white text

Description automatically generated

A screenshot of a computer

Description automatically generated

Result: Random Forest achieved the best F1 score of 93%

**ROC Curve:**

The Receiver Operating Characteristic (ROC) curve is a graphical representation used to evaluate the performance of a classification model, such as Random Forest, across different thresholds. It plots the true positive rate (sensitivity) against the false positive rate (1 - specificity) at various threshold settings.

AUC stands for Area Under the ROC Curve, and it quantifies the overall performance of the model. An AUC value of 0.65 suggests that the Random Forest model has moderate discriminatory power in distinguishing between the positive and negative classes.

In the context of predicting house loan defaulters, an AUC of 0.65 indicates that the Random Forest model is better than random guessing but may still have room for improvement. It means that there is a 65% chance that the model will rank a randomly chosen positive instance higher than a randomly chosen negative instance.

Interpreting the ROC curve with an AUC of 0.65, you would typically see the curve bending towards the top-left corner of the plot, indicating that the model is making meaningful predictions. However, it might not be achieving optimal performance, and further optimization or feature engineering may be necessary to improve the AUC score.

A graph with a line

Description automatically generated

**Conclusion**

In conclusion, Random Forest emerges as a superior model compared to decision trees and logistic regression for predicting house loan defaulters due to its robustness, accuracy, and generalization ability. Unlike decision trees, Random Forest mitigates overfitting by aggregating predictions from multiple trees trained on random subsets of the data, thus capturing diverse patterns and enhancing model stability. Additionally, Random Forest outperforms logistic regression by handling nonlinear relationships and interactions among features more effectively. Its ensemble approach not only provides better predictive performance but also offers resilience against outliers and noisy data. Overall, Random Forest stands out as the optimal choice for lenders seeking accurate and reliable predictions of house loan defaulters, empowering them to make informed decisions and effectively manage lending risks.

**References:**

https://www.kaggle.com/datasets/urstrulyvikas/house-loan-data-analysis

<https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html>

https://medium.com/@kapiljain1989/loan-default-prediction-using-machine-learning-e39f2f968f94