**Home Credit Default Risk (HCDR)**

**(Applied Machine Learning)**

**FP\_Group\_16\_Risk Wranglers**

1. Penugonda, Viswa Suhaas
2. Muvva, Sai Sumanth
3. Chintha, Teja Naidu
4. Kasi Nathan Deepak

**Project Description:**

The Home Credit Default Risk (HCDR) project is focused on designing a predictive model that assesses the probability of a client repaying a loan, utilizing a range of alternative data sources. These sources consist of telecommunications and transactional information, which provide valuable insights into the creditworthiness of clients who have limited or no credit history. By accurately forecasting clients' repayment capabilities, financial institutions can make well-informed lending decisions, which ultimately leads to a favorable borrowing experience for clients and a reduction in default risks.

The model aims to address the challenges faced by clients with inadequate credit histories when trying to secure a loan. Traditional credit scoring methods often fail to accurately capture the creditworthiness of such individuals, as they rely heavily on existing credit histories. The HCDR project seeks to bridge this gap by incorporating alternative data sources that can provide a more comprehensive view of a client's financial behavior.

This innovative approach to credit risk assessment will not only empower financial institutions to make smarter lending decisions, but also create opportunities for clients who would have otherwise been overlooked due to their lack of credit history. As a result, the HCDR project has the potential to significantly impact the financial services industry, promoting financial inclusion and expanding access to credit for underserved populations.

**Data:**

The dataset for this project is provided by the Home Credit Group as part of the Kaggle competition. The data is split into several files, including:

1. **application\_{train|test}.csv**
   * This is the main table, broken into two files for Train (with TARGET) and Test (without TARGET).
   * Static data for all applications. One row represents one loan in our data sample.
2. **bureau.csv**
   * All client's previous credits provided by other financial institutions that were reported to Credit Bureau (for clients who have a loan in our sample).
   * For every loan in our sample, there are as many rows as number of credits the client had in Credit Bureau before the application date.
3. **bureau\_balance.csv**
   * Monthly balances of previous credits in Credit Bureau.
   * This table has one row for each month of history of every previous credit reported to Credit Bureau – i.e the table has (#loans in sample \* # of relative previous credits \* # of months where we have some history observable for the previous credits) rows.
4. **POS\_CASH\_balance.csv**
   * Monthly balance snapshots of previous POS (point of sales) and cash loans that the applicant had with Home Credit.
   * This table has one row for each month of history of every previous credit in Home Credit (consumer credit and cash loans) related to loans in our sample – i.e. the table has (#loans in sample \* # of relative previous credits \* # of months in which we have some history observable for the previous credits) rows.
5. **credit\_card\_balance.csv**
   * Monthly balance snapshots of previous credit cards that the applicant has with Home Credit.
   * This table has one row for each month of history of every previous credit in Home Credit (consumer credit and cash loans) related to loans in our sample – i.e. the table has (#loans in sample \* # of relative previous credit cards \* # of months where we have some history observable for the previous credit card) rows.
6. **previous\_application.csv**
   * All previous applications for Home Credit loans of clients who have loans in our sample.
   * There is one row for each previous application related to loans in our data sample.
7. **installments\_payments.csv**
   * Repayment history for the previously disbursed credits in Home Credit related to the loans in our sample.
   * There is a) one row for every payment that was made plus b) one row each for missed payment.
   * One row is equivalent to one payment of one installment OR one installment corresponding to one payment of one previous Home Credit credit related to loans in our sample.
8. **HomeCredit\_columns\_description.csv**
   * This file contains descriptions for the columns in the various data files.

**Metrics:**

1. **ROC AUC:** The primary evaluation metric used for this project is the Area Under the Receiver Operating Characteristic curve (ROC AUC). This metric captures the trade-off between true positive rate (sensitivity) and false positive rate (1-specificity), making it suitable for imbalanced classification problems like credit default prediction. While the F1 score can also be a useful metric, the ROC AUC provides a more comprehensive understanding of the classifier's performance across different decision thresholds, which is important in credit risk assessment.
2. **F1 Score:** The F1 score is another metric that can be used in this project, especially when dealing with imbalanced datasets. It is the harmonic mean of precision and recall, and ranges from 0 to 1, where 1 represents the best possible score. The F1 score balances the trade-off between precision (the proportion of true positive predictions among all positive predictions) and recall (the proportion of true positive predictions among all actual positives). This metric is useful when both false positives and false negatives are important to consider.

A picture containing text

Description automatically generated

1. **Balanced Accuracy:** Balanced accuracy is an alternative to the traditional accuracy metric, which may not be suitable for imbalanced datasets. Balanced accuracy is defined as the average of the true positive rate (sensitivity) and true negative rate (specificity). It is more robust to class imbalance, as it considers both the performance on the minority class (defaulters) and the majority class (non-defaulters).
2. **Precision-Recall AUC (PR AUC):** The Precision-Recall Area Under the Curve (PR AUC) is another option for evaluating the performance of the model. Unlike the ROC AUC, PR AUC focuses on the positive class (defaulters) and is less sensitive to the presence of a large number of true negatives. This metric can be more informative when dealing with imbalanced datasets, as it measures the relationship between precision and recall across different classification thresholds.
3. **Gini Coefficient:** A financial industry metric measuring inequality and credit risk. Ranging from 0 (perfect equality) to 1 (complete inequality), it identifies the concentration of defaults across borrower segments. Including the Gini Coefficient in our project helps optimize the model by addressing credit risk distribution and catering to domain-specific needs.

**Baseline Models:**

The project will begin with the development of baseline models using standard machine learning algorithms such as:

1. Logistic Regression: A simple linear model for binary classification problems.
2. Random Forest: An ensemble method that constructs multiple decision trees to improve prediction accuracy and reduce overfitting.
3. Gradient Boosting Machines (GBM): A powerful ensemble method that builds decision trees sequentially, minimizing a loss function iteratively.

**Baseline Pipeline:**

1. **Data Preprocessing:** Clean and pre-process the data, including handling missing values, outliers, and categorical features encoding.
2. **Feature Engineering:** Create new features based on domain knowledge and by combining existing features to improve model performance.
3. **Feature Selection:** Identify the most relevant features using techniques such as Recursive Feature Elimination (RFE) or LASSO regularization.
4. **Model Training:** Train the baseline models using cross-validation and tune hyperparameters using grid search or random search.
5. **Model Evaluation:** Assess the models' performance using the ROC AUC metric and choose the best performing model.

**Other Planned Pipelines:**

1. **Advanced Feature Engineering:** Explore additional feature extraction techniques, such as aggregation, interaction, and polynomial features.
2. **Dimensionality Reduction:** Apply techniques like PCA (Principal Component Analysis) or t-SNE (t-Distributed Stochastic Neighbour Embedding) to reduce the number of features while retaining most of the information.
3. **Hyperparameter Optimization:** Use more sophisticated methods like Bayesian optimization or genetic algorithms for tuning model hyperparameters.
4. **Ensemble Methods:** Combine different models, such as stacking or bagging, to improve overall performance.
5. **Deep Learning:** Experiment with deep learning models like feed-forward neural networks or Long Short-Term Memory (LSTM) networks to capture complex patterns in the data.