PROJECT PROPOSAL

Home Credit Default Risk

Introduction

Many people struggle to get loans due to insufficient or non-existent credit histories. And, unfortunately, this population is often taken advantage of by untrustworthy lenders.

Home Credit strives to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience. In order to make sure this underserved population has a positive loan experience, Home Credit makes use of a variety of alternative data including telco and transactional information--to predict their clients' repayment abilities.

While Home Credit is currently using various statistical and machine learning methods to make these predictions, they're challenging Kagglers to help them unlock the full potential of their data. Doing so will ensure that clients capable of repayment are not rejected and that loans are given with a principal, maturity, and repayment calendar that will empower their clients to be successful.
Source: https://www.kaggle.com/competitions/home-credit-default-risk/overview

Data Overview

According to the data there are 307511 rows and 122 columns

Data Pre-Processing

1. **Cleaning the data:** In order to perform skewness analysis, I am converting the data with negative values to abs because there are predictors in the data that have entirely negative values and cannot be utilised to calculate skewness.

2. **Centering and Scaling:** Predictor value centre scaling is the simplest and most popular data transformation. Subtracting the average predictor value from all of the values centres the predictor variable. Every value of the predictor variable is split by its standard deviation, same like when scaling data.

3. **Skewness of the data:** Additionally, due to the skewness (either left or right), the data must be modified. The skewness may be eliminated by substituting log, sqrt, or inverse data. Given that skewness only accepts numeric variables, only the integer and numeric variables should be used for the df.

4. **Transformation to resolve Outliers:** An outlier is a sample that deviates significantly from the data's average. It is crucial to confirm that the values are valid from a scientific standpoint when one or more samples are recognised as outliers. Outlier resistance is a feature of several prediction models, including SVMs and tree-based models. It is crucial to scale the data before performing a spatial data transformation if a model is thought to be sensitive to outliers. Since I have not read about the spatial data transformation in great detail and the book does not go into great detail about it, I will not be using it.

5. **Handling Missing Values:** Understanding the reason for the missing values is crucial. The absence of values may be due to their structural absence. Determining if the absent value is connected to the result is crucial. Do not mix up missing values with censored data, which involves knowing something about the value even if the exact value is missing. If there are only a few predictors with missing values, exploratory data analysis (EDA) is the optimal choice. Creating a model to impute missing values is possible when a variable with many missing values is strongly correlated with another variable that has few missing values.

    a. A widely used method is the K-nearest neighbor model. The benefit is that the filled-in values will always fall within the range of the values in the training set. The drawback is that you must utilize the entire training set to fill in missing values for the variable.

    b. Another method employed is carrying out linear regression between the predictor that shows strong correlation and one with incomplete data. In certain instances, the amount of missing data is significant and warrants exclusion of this predictor in future modeling tasks.

6. **Data Reduction and Feature Extraction:** Principal component analysis (PCA) is frequently utilized as a method for reducing data. This technique aims to discover the linear combination of predictors referred to as (PCs). The PC is described as the linear combination of predictors that captures the highest amount of variability among all potential linear combinations. The coefficients assist in comprehending the component weights and identifying the most crucial predictors for each PCs. Prior to conducting a PCA analysis, it is recommended to first adjust the skewed predictors and then standardize and center them. PCA analysis, as an unsupervised learning method, does not take into account the modeling goal or outcome variable when summarizing variation.

7. **Adding Predictors:** It is typical to break down a categorical predictor into a group of more detailed variables. Dummy variables are consistently n-1, where n represents the levels of the variable. Whether we include all variables depends on the model type we plan to use. When using a model like linear regression that is sensitive to that data, it is crucial to use n-1. On the other hand, utilizing n (entire array of irrelevant variables) could enhance the understanding of the model. One Hot Encoding is another name for dummy variables. Different methods of handling dummy variables should be considered based on the nature of the categorical feature, such as whether it is ordinal or not. One illustration of an ordinal yet categorical attribute is the size options for a shirt (XS, S, M, L, XL). In that situation, the characteristic needs to be converted to a numeric value (XS = 1, S = 2, etc.). An illustration of a categorical feature, not an ordinal one, is gender (M, F). In that situation, the characteristic must be "One Hot Encoded."

Algorithms for Predicting Repayment Abilities

1. **Logistic Regression**

    Logistic regression is a statistical technique employed for tasks involving binary classification, such as predicting if borrowers will return borrowed funds. It uses the logistic function to transform linear combinations of predictors into probabilities,

depicting the connection between predictor variables and a binary outcome. This enables the understanding of model coefficients, showing the impact of each predictor on the probability of repayment. The model trains using maximum likelihood estimation to determine the best coefficients, and predictions are based on probabilities above a threshold (typically 0.5). Assessment of performance is based on metrics such as accuracy, precision, and AUC-ROC. Logistic regression relies on the assumption of linear associations between predictors and the log-odds of the outcome, and also requires observations to be independent. Within Home Credit Group, this method assists in pinpointing borrowers who are anticipated to settle their loans, aiding in the advancement of financial inclusion.

2. **Support Vector Machine (SVM)**
    - SVM aims to identify the hyperplane that maximizes the margin between two classes, ensuring that the closest data points, known as support vectors, are as far away as possible from the hyperplane. The algorithm can handle both linear and non-linear data through the use of kernel functions, which transform the input space into higher dimensions. Common kernels include linear, polynomial, and radial basis function (RBF), allowing SVM to adapt to complex data distributions.
    - During training, SVM minimizes a cost function that balances maximizing the margin and minimizing classification error. Once trained, the model can classify new data points based on which side of the hyperplane they fall on. Performance is evaluated using metrics such as accuracy, precision, recall, and F1-score.
    - In the context of predicting clients' repayment abilities for Home Credit Group, SVM can effectively separate clients who are likely to repay from those who are not, even when the relationship between features and the target variable is complex. This capability makes SVM a valuable tool in enhancing financial inclusion initiatives.

3. **K-Nearest Neighbors (KNN)**
    - K-Nearest Neighbors (KNN) is a straightforward yet powerful supervised learning algorithm used for classification tasks, making it particularly useful for predicting clients' repayment abilities in the context of Home Credit Group. KNN operates by determining the 'k' nearest data points in the feature space to a given client and predicting the outcome based on the majority class of those neighbors. To classify a new client, the algorithm calculates the distance—often using metrics like Euclidean distance—between the client and all other points in the training dataset.
    - One of the main advantages of KNN is its simplicity and interpretability, which can be beneficial when assessing client characteristics to predict repayment likelihood. However, the algorithm can be computationally intensive, especially with large datasets, since it requires distance calculations for all training samples during prediction. Additionally, KNN is sensitive to irrelevant features and the scale of the data, making feature scaling an important preprocessing step.
    - In the context of Home Credit Group, KNN can effectively identify clients who are similar to those with known repayment behaviors, enabling the classification of new clients as likely to repay or default. This capability supports the organization's goal of enhancing financial inclusion by ensuring that loans are provided to clients who demonstrate the ability to repay.

4. **Random Forest**
    - Random Forest is an ensemble learning technique that combines multiple decision trees to enhance prediction accuracy and reduce the risk of overfitting. In the context of predicting clients' repayment abilities for Home Credit Group, Random Forest is particularly valuable due to its capability to handle complex datasets and a wide range of predictor variables.
    - The algorithm works by training a large number of decision trees on random subsets of the data and features, creating a diverse set of models. When making predictions, it aggregates the outputs of these trees—typically through majority voting for classification tasks—to produce a final result. This approach not only improves accuracy but also provides robustness against noise and irrelevant features.
    - One of the key advantages of Random Forest is its ability to assess feature importance, helping to identify which factors most significantly influence clients' likelihood to repay. This insight can guide strategic decision-making and enhance the design of financial products tailored to clients' needs.
    - In the context of Home Credit Group, employing Random Forest can significantly improve the accuracy of predicting repayment behavior, ultimately supporting efforts to promote financial inclusion and responsible lending. By accurately identifying clients who are likely to repay their loans, Home Credit can better allocate resources and offer suitable financial solutions to underserved populations.

Conclusion

We will assess the performance of the algorithms—Logistic Regression, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Random Forest—using critical metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. This thorough evaluation will enable us to identify the most effective model for predicting clients' repayment abilities and understanding the key factors influencing these decisions. By utilizing these insights, we can enhance our strategies, align with Home Credit Group's mission of promoting financial inclusion, and ensure that loans are responsibly offered to clients who are capable of repayment, ultimately empowering underserved communities.

TEAM DETAILS

1. Srinath Bulusu - 1002197525
2. Manas Pavan Sai Kallaganti - 1002209866
3. Pranay Mohan Kanakabandi - 1002198613
4. Sai Prabhakar Gattu -1002209363