German Bank Loan Default Prediction Report

Introduction:

In this project, we go deeper into the analysis of the factors that lead to loan defaults among applicants within the German lending landscape. We analyze German Credit Risk dataset that captures various details of loan applicants and their credit histories to establish contribution of these factors to likelihood of loan default and credit risk assessment. The objective of this analysis is to construct predictive models that help in identifying potential defaults.

In the lending space, understanding and mitigating risks associated with lending are crucial. Financial institutions, such as Non-Banking financial companies (NBFCs) and banks rely on credit risk assessment models to minimize risk of loan defaults and make informed decision regarding loan approvals. These models use historical data to predict possibility of loan default after taking into consideration the attributes and the contribution of these attributes on loan defaults. This enables financial institutions to better manage their exposure to risk.

In this analysis, we identify a set of interesting questions that we wish to explore the relationships between various applicant features such as loan amount, credit history, checking balance, loan duration and likelihood of loan default. We will also evaluate different machine learning models to determine best model to predict loan defaults and the ways in which predictive models contribute to informed and data-driven lending decisions.

Methods and Materials:

In this analysis, we perform Exploratory Data Analysis(EDA) on the dataset which includes distribution key features and visualizing relationship between different predictors and target variable default. As part of data pre-processing, We perform missing data imputation and merging of categories.

We pre-processed the data by encoding categorical variables. We used Ordinal Encoding for sequential categorical predictors. Then, we split the dataset into training and testing sets and performed model training. We used the trained model to predict the loan default. We also used oversampling technique to improve recall for loan defaults.

To predict loan default, we used several machine learning algorithms such as Gradient Boosting, Bagging , Random Forest and XGBoost, SVM, Voting Classifier Hard, Voting Classifier Soft. We also performed hyper tuning of parameters for algorithms to improve loan default prediction through recall.

Results:

The key findings from this project are Voting Classifier Soft voting shows highest recall of 94% which is important metric for identification of loan defaults. Other models like Gradient Boosting, Bagging, Random Forest, SVM, Voting Classifier Hard showed recall values of 80%, 89%, 89%, 89%, 89%, 91% respectively.

After hyperparameter tuning of models, we also get high recall of 94% for XGBoost and it performs better as compared to other models in prediction of recalls after fine tuning.

Our analysis also highlighted importance of certain features such as amount, duration of loan, age and credit history in influencing loan default outcomes.

In addition to this, EDA phase gives more insights such as loan duration, amount and loan default suggested that more loan duration or amount results into more defaults.

Discussion:

Voting Classifier with soft voting and XGBoost after hyper parameter tuning position as the right choice for identifying potential defaults. The performance of Gradient Boosting, Bagging, Random Forest, SVC in this analysis was moderate and needs further fine tuning of hyper parameters. After hyperparameter tuning, XGBoost shows best recall indicating high accuracy in identification of loan defaults.

These results are interpreted while taking into consideration feature importance. Applicants with more loan duration and more amount are likely to default on loans which aligns with the observation in lending industry. These insights can aid financial institutions in making informed decisions regarding loan approvals.

However, there are certain limitation in this analysis. The dataset is small and might lack certain important feature for prediction of loan defaults. Additionally, model performance could vary with different datasets and some of the aspects not captured in this dataset might have influence on loan defaults.

Conclusions:

In conclusion, this analysis has provided insights about relationships between applicant attributes and loan default outcomes. The exploration of various models highlighted efficacy of Voting Classifier with soft voting and XGBoost in prediction of recall for loan default. EDA also highlighted patterns in the dataset. By understanding these insights, financial institutions can refine the lending practices and develop strategies to mitigate risk of loan defaults. Further research could leverage more factors and data to provide real-world loan default predictions.