Loan Default Prediction Using Machine Learning

Introduction

Predicting the likelihood of a borrower defaulting on a loan is a crucial task for financial institutions. Erroneously classifying a borrower as creditworthy can result in loans becoming non-performing assets. Conversely, if a deserving candidate is denied a loan, the lending institution may miss out on significant business opportunities. Machine learning can be effectively employed to train classification models capable of predicting the probability of loan default by a borrower, thereby aiding the decision-making process. The utilization of machine learning models becomes particularly imperative in cases of fast consumer loans, where lenders must promptly determine loan disbursal.

In this project, various machine learning models are built and evaluated using the loan default dataset. The project aims to address significant questions, such as:

1. Which classification model should be chosen?
2. What criteria and conditions should guide the selection of a specific model?
3. Is the selected model capable of generalization?
4. What are the crucial features in predicting defaults?

Methods and Materials

The projects is divided into 5 important steps Exploratory Data Analysis, Data Pre-processing, Model Traning and Evaluation and Model Selection.

In EDA step:

- Features from the dataset were divided into categorical and numerical features.
- Univariate and Bivariate analysis was performed.

In Data Pre-processing step:

- Information revealed in EDA was used to discard features with high number of unknown values.
- Categorical features were transformed into binary numerical features.

In the Model Training and Evaluation step:

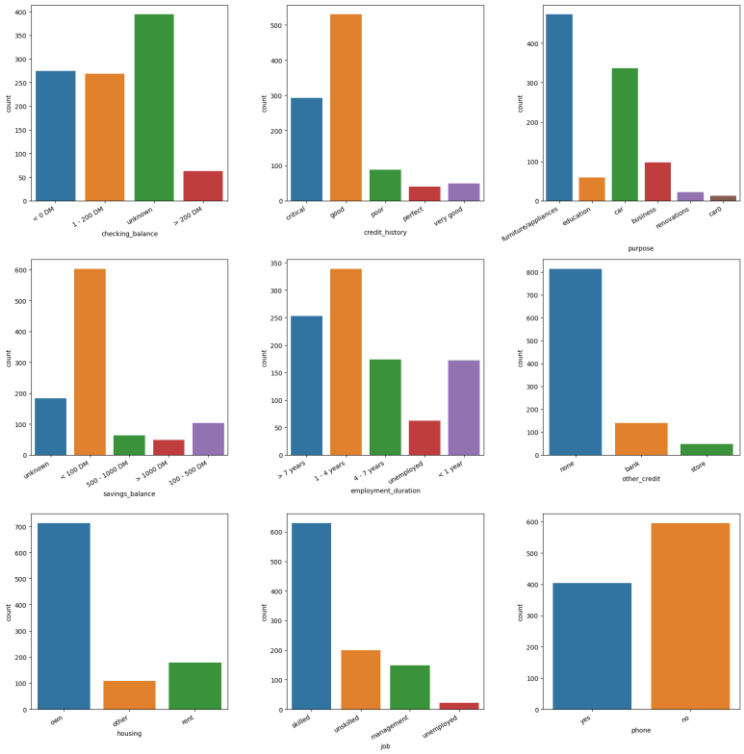
- Five models - Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Gradient Boost Classifier and Naïve Bayes were trained and evaluated.
- The models scored and evaluated using 10-Fold Cross Validation.
- The models were compared using four metrics – accuracy, precision, recall and F1 score.

In the Model Selection step:

- Random Forest and Gradient boost models were tuned to improve there performance metrics.
- Most important features for these models were identified.

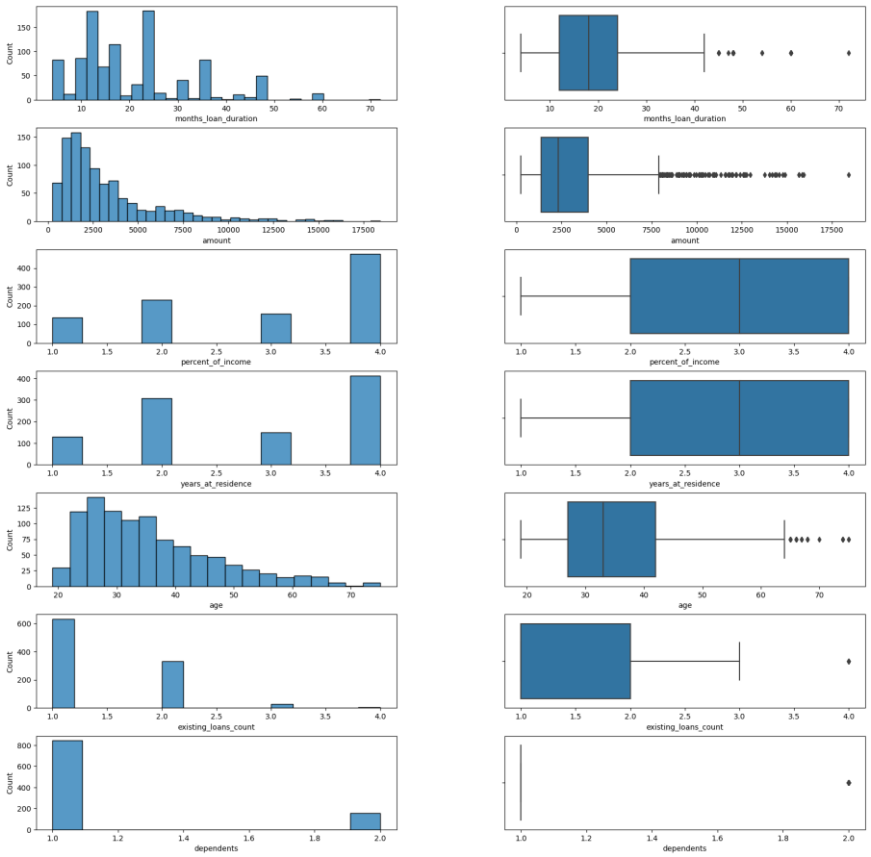
Important observations and analyses from these steps are summarised below.

From the below count-plots, we can see significant number of unknown/none values in *other_credit*, *checking_balance* and *savings_balance*. These features were later dropped in data preparation step.

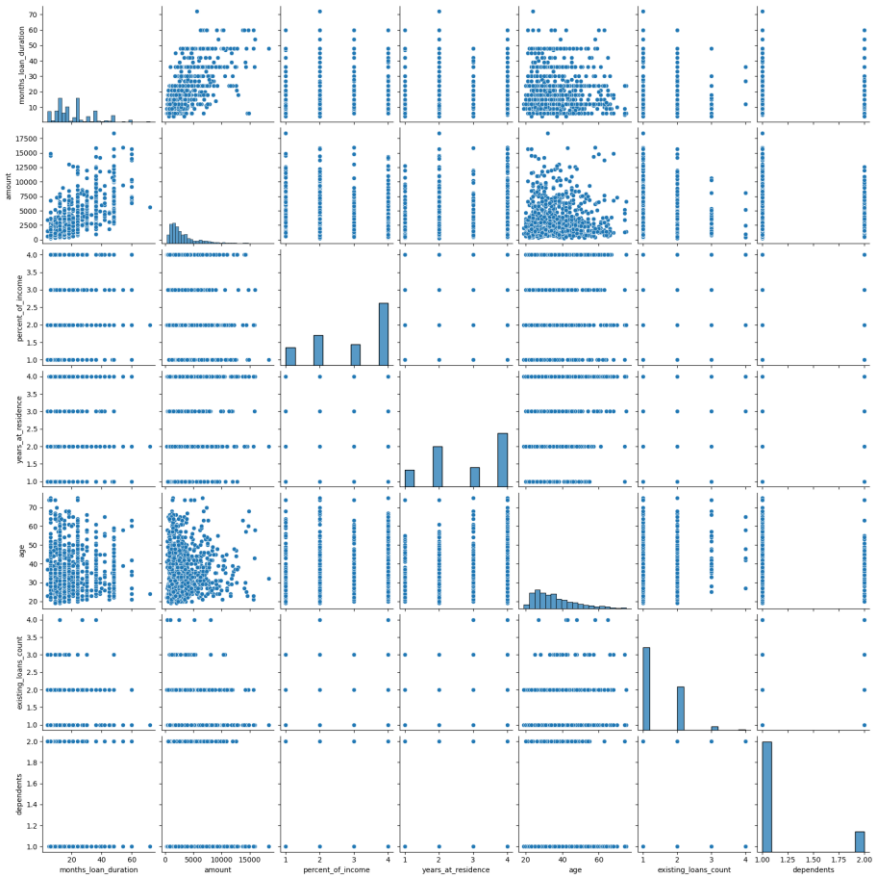


From the Hist-plots and Box-plots we can see:

1. Age and Amount plot are positively skewed.
2. There are very few loans with duration more than 40 months.
3. Loans of amount greater than 7500 are also not common.
4. It is not common for people with age greater than 65 to have loans

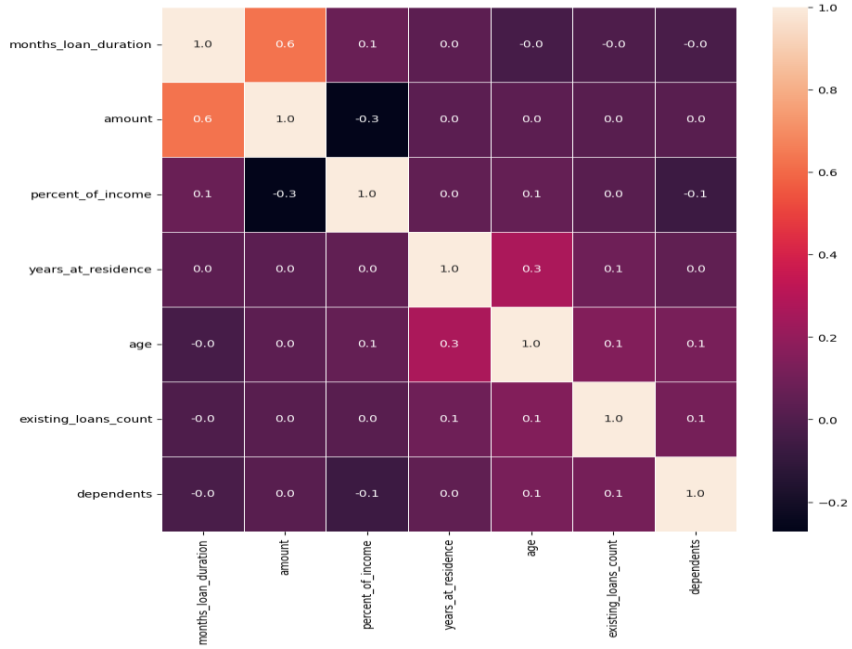


There is evidence of positive relationship between *month_loan_duration* and amount as seen in one of the pair-plot in the below image.



Correlation matrix reveals positive correlation between amount and *month_loan_duration*, which was also observed in their respective pair-plot. The positive correlation also makes sense as higher loan amount will require more time to repay.

Also, there is negative correlation between amount and percent of income.



Results

Model Training and Evaluation Results of the models is summarised in the below table.

model_name	mean_test_accuracy	mean_test_precision	mean_test_recall	mean_test_f1
Logistic Regression	0.722	0.578323	0.266667	0.361069
Decision Tree	0.66	0.434257	0.42	0.419354
Random Forest	0.708	0.535503	0.236667	0.317132
Gradient Boost	0.717	0.559676	0.283333	0.371434
Naive Bayes	0.703	0.511001	0.34	0.406391

Amongst all the models, Logistic regression performs best in terms of accuracy and precision. The decision Tree has the lowest accuracy and precision score but the best recall score. Naïve Bayes has the best f1 score along with the Decision tree model. Random Forest and Gradient Boost are similar in terms of all the metrics.

Upon performing parameter tuning for the random forest and gradient boost models, the newly trained models had the following performance metrics:

Random Forest	Gradient Boost
Accuracy = 0.7366666666666667 Precision = 0.7727272727272727 Recall = 0.18681318681318682 F1_score = 0.3008849557522124	Accuracy = 0.71 Precision = 0.5370370370370371 Recall = 0.31868131868131866 F1_score = 0.4

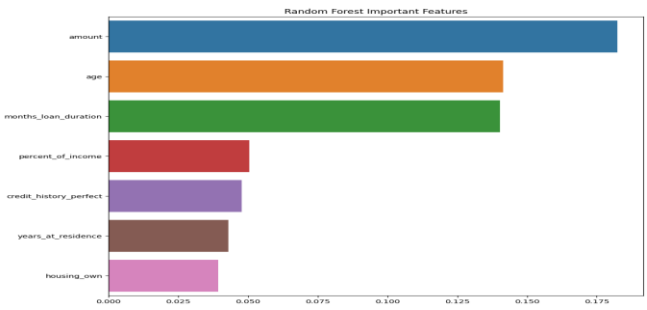
After parameter tuning, Both Random Forest and Gradient boost had significant improvement in accuracy. Random forest also improved on precision score.

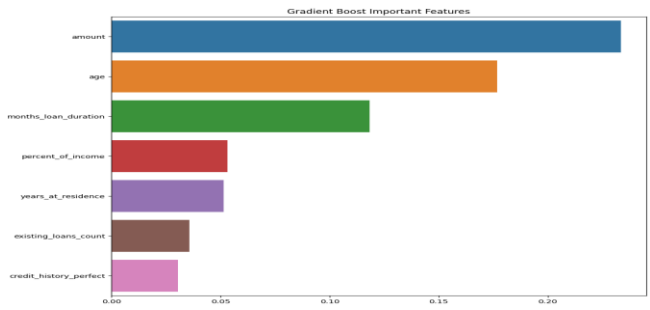
Discussion

The selection of the model depends on the prioritized performance metric. Depending on the cost of false positives and false negatives, we can prioritize one metric over the other. Accuracy emerges as a superior metric when the cost of false positives and negatives is similar, and the classes maintain balance. Precision proves its utility when the goal is to minimize the cost of false positives, effectively reducing type-1 errors. Recall finds its application when the objective involves reducing both false positives and type-2 errors. The F1 score becomes particularly valuable when aiming to strike a balance between precision and recall, especially in scenarios with imbalanced classes.

If prioritizing accuracy, we can consider any models except for the Decision Tree. When focusing on precision, the choice naturally falls on the Random Forest. Decision Trees and Naïve Bayes stand as favorable options based on recall and F1 scores.

To streamline models, one effective approach involves eliminating the least important features. The image below presents the most significant features obtained through Random Forest and Gradient Boosting.





Both model consider amount, age and loan duration in month, as the most important parameters. However, Gradient boost prefers existing loan count and random forest prefers own housing.

Conclusions

In conclusion, the choice of model depends on the specific business area that the bank aims to prioritize. If the bank's goal is to enhance its business by increasing the number of loans granted, it becomes imperative to mitigate false positives – instances where loans are inaccurately classified as prospective defaults. In this context, the Random Forest model becomes the obvious choice due to its high precision score achieved through parameter tuning. On the other hand, if the focus is on reducing exposure to risky loans, the preferred model would be one that minimizes false negatives while maintaining a high recall score, namely Decision Trees and Naïve Bayes.