Machine Learning Model Outcomes: Loan Prediction Project

Executive Summary Report

Prepared by Robes Fokoueng – February 2024

Project Overview

We want to develop a robust and reliable machine learning model to predict whether to authorize or not a loan to a customer.

Key Insights

We built 4 classification models: Decision Tree, KNN, Random Forest and Logistic Regression.

Random forest (RF) and Logistic regression (LR) performed exceptionally well.

Logistic regression model with a better recall score (0.976), precision score (0.838) and an accurracy of 85.37 % was selected as champion.

Credit_History has the highest correlation with Loan_Status (a positive correlation of 0.52). Therefore, our target value is highly dependent on this column.

Despite the good performance of the selected model, we do not recommend using it because our target value is highly dependent on only one features. And in practice, the loan is not granted only based on the credit history, but also on certain features such as, income, coapplicant income if applicable and loan amount.

Next Steps

We recommend to:

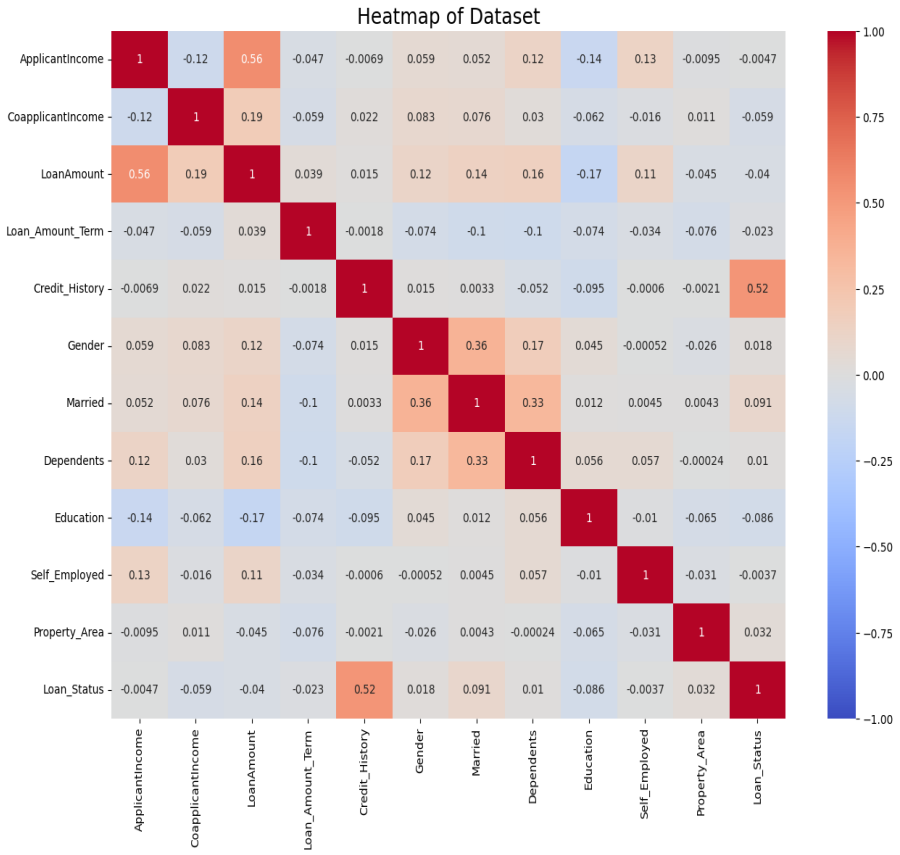
Collect/add more the dataset.

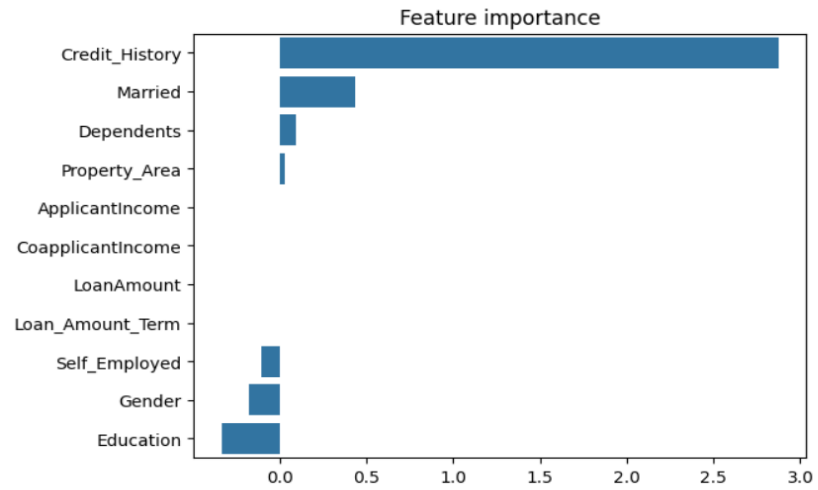
Build Random Forest model and XGBOOST model, using hyperparameters tune and evaluate them.

Apply Cross validation: Data will split into training, validation and test sets. It is particularly useful when working with smaller dataset. The goal is to evaluate the model's robustness and avoid overfitting.

Details

- Correlation heatmap.





- In the logistic regression model above, `Credit_History`, `Married`, `Dependent and `Property_Area` have the positive importance. These variables are most helpful in predicting the outcome variable.