ID/X PARTNERS DATA SCIENTIST
PROJECT BASED INTERNSHIP

FINAL TASK

STEPHANIE YOLANDA

RAKAMIN ACADEMY

BUSINESS UNDERSTANDING

Prediction of credit risk involves predicting risky and non-risky borrowers in order for financial companies/institutions to reduce the risks of lending money to borrowers that would fail to fulfil their obligations of paying the loan.

ANALYTIC APPROACH

To predict credit risk of borrowers, the approach to be taken is by training a machine learning model and gaining predictive insights from the model.

DATA

REQUIREMENTS

A credit risk prediction model requires data of previous/existing borrowers of the company, including both the information of the borrower and information of the loan.

COLLECTION

ID/X Partners has collected the required data to be used for the credit risk prediction model.

DATA

UNDERSTANDING

The raw dataset includes 466285 rows of data and 74 columns. However, 34 columns contain null values. That problem is tackled by either removing whole columns or imputing the data with new values.

Additionally, the dataset contains some information that would not be necessary for the prediction model, such as address and zip code, and therefore those unnecessary data are removed.

DATA

PREPARATION

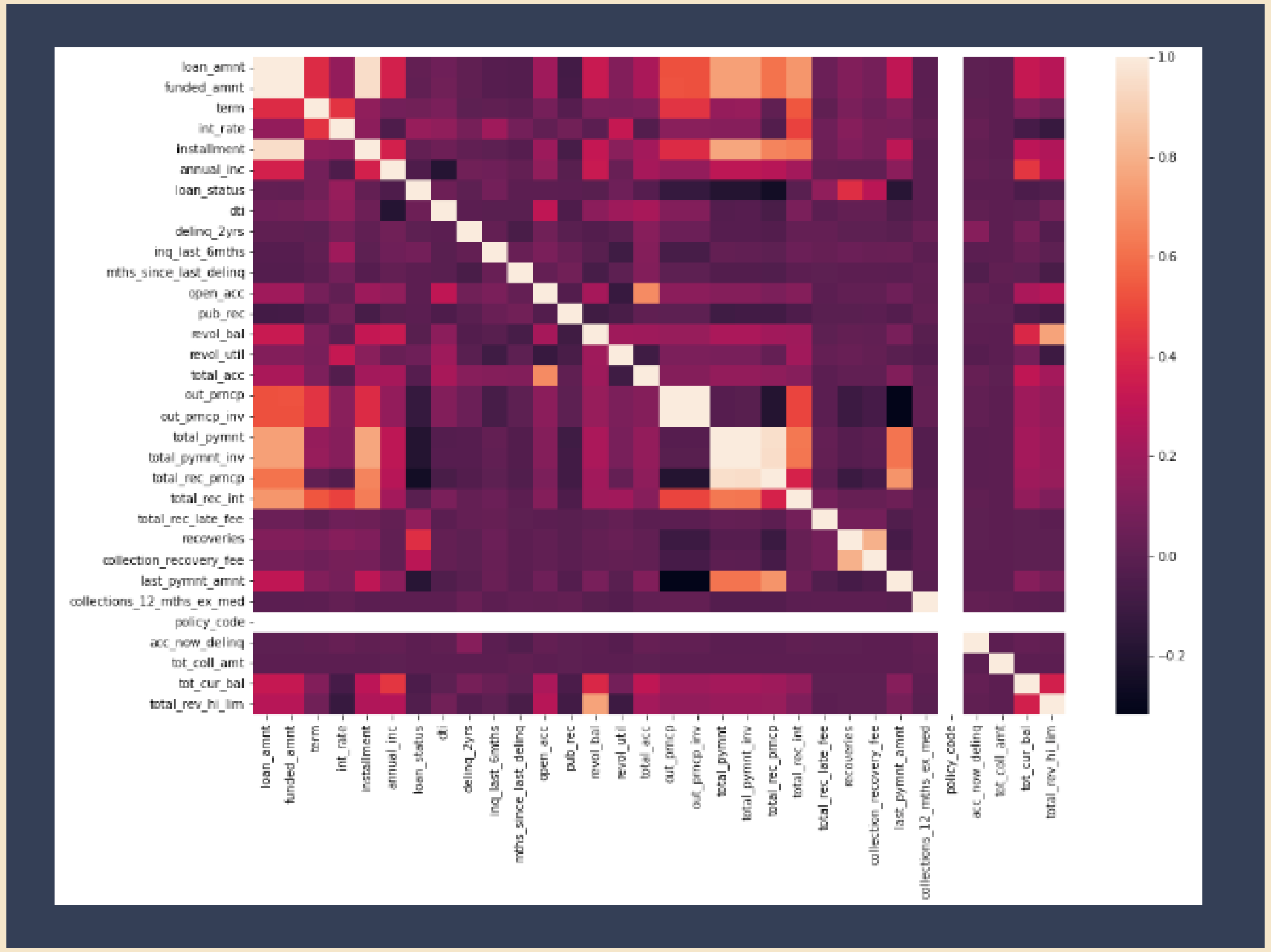
The dataset is cleaned in terms of data representation and type. It is then encoded and normalized for a suitable input of the model. The dataset is then split into train and test data in a 80:20 ratio.

The target variable chosen from the dataset, the loan status, is created as the labels and removed from the train and test data.

EXPLORATORY DATA ANALYSIS (EDA)

CORRELATION

The correlation of the features to the target variable is important for understanding which features are unnecessary and not beneficial for the model performance.



Low correlating features are removed from the dataset entirely.

MODEL

BUILDING

The classifier chosen for this project is the Random Forest classifier. The model is trained with the train data and tested on the test data.

Model parameters:
n_estimators: 100
max_depth: None
max_features: sqrt

EVALUATION

Accuracy score: 0.967
Precision score: 0.982
Recall score: 0.863
ROC-AUC score: 0.863

The model performed well using the given parameters. However, with better resources, implementing hyperparameter tuning may improve the model.