Exploratory Data Analysis

Agenda

- Introduction
- Problem Description
- EDA
- Proposed Modeling Technique
- Conclusion

PROBLEM STATEMENT

develop a predictive model that can assess the credit worthiness of potential future customers of a financial institution.

Introduction

- The objective is to build a model that accurately predicts the credit standing of new loan applications
- The model should be able to identify the key factors that determine creditworthiness and provide insights to help the financial institution make better lending decisions.

About Data

The available data set consists of 807 past loan customer cases

---

14 attributes like

financial standing, reason for the loan, employment, demographic information, foreign national status, years of residence in the district

---

outcome/label variable

Credit Standing

Back to Agenda Page

Data Cleaning

Handling Missing Values

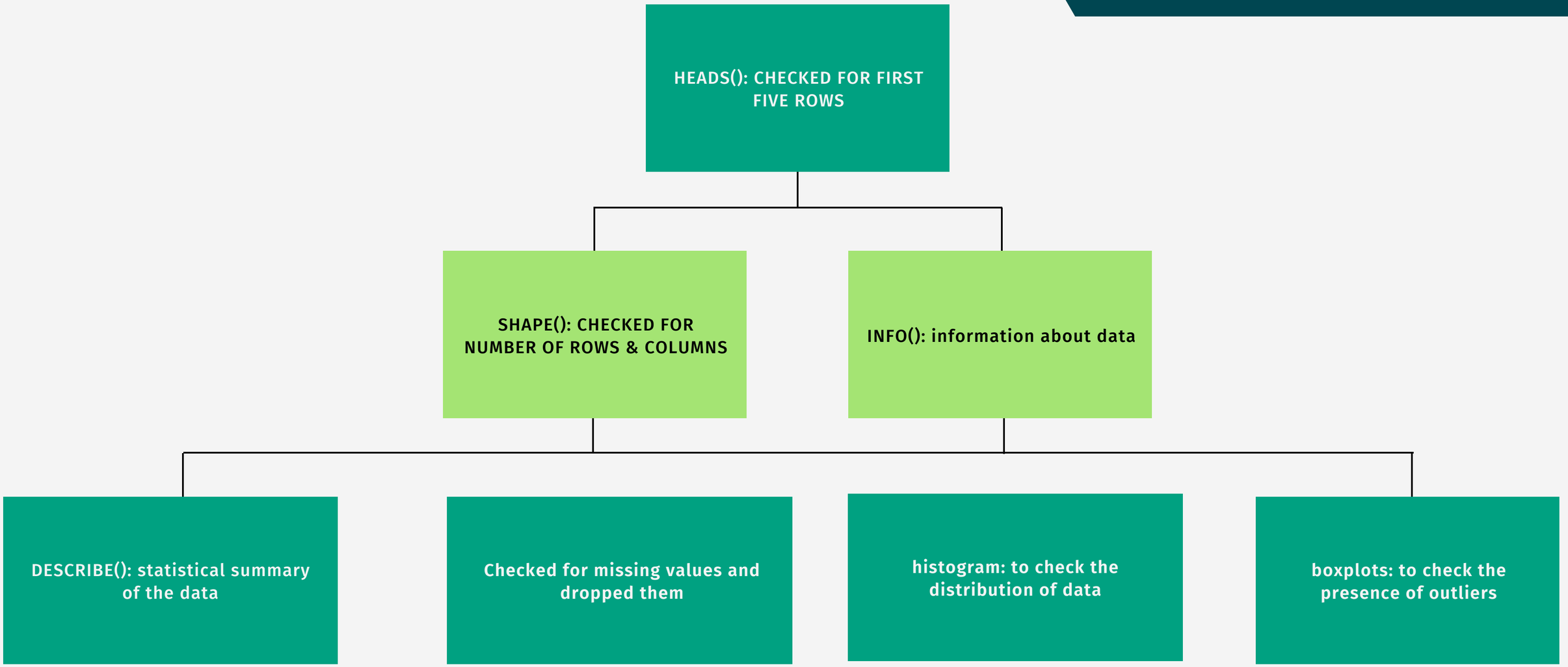
- number of missing values is very small and there is no meaningful pattern
- used dropna() function to remove any missing values

Handling Skewness and Outliers

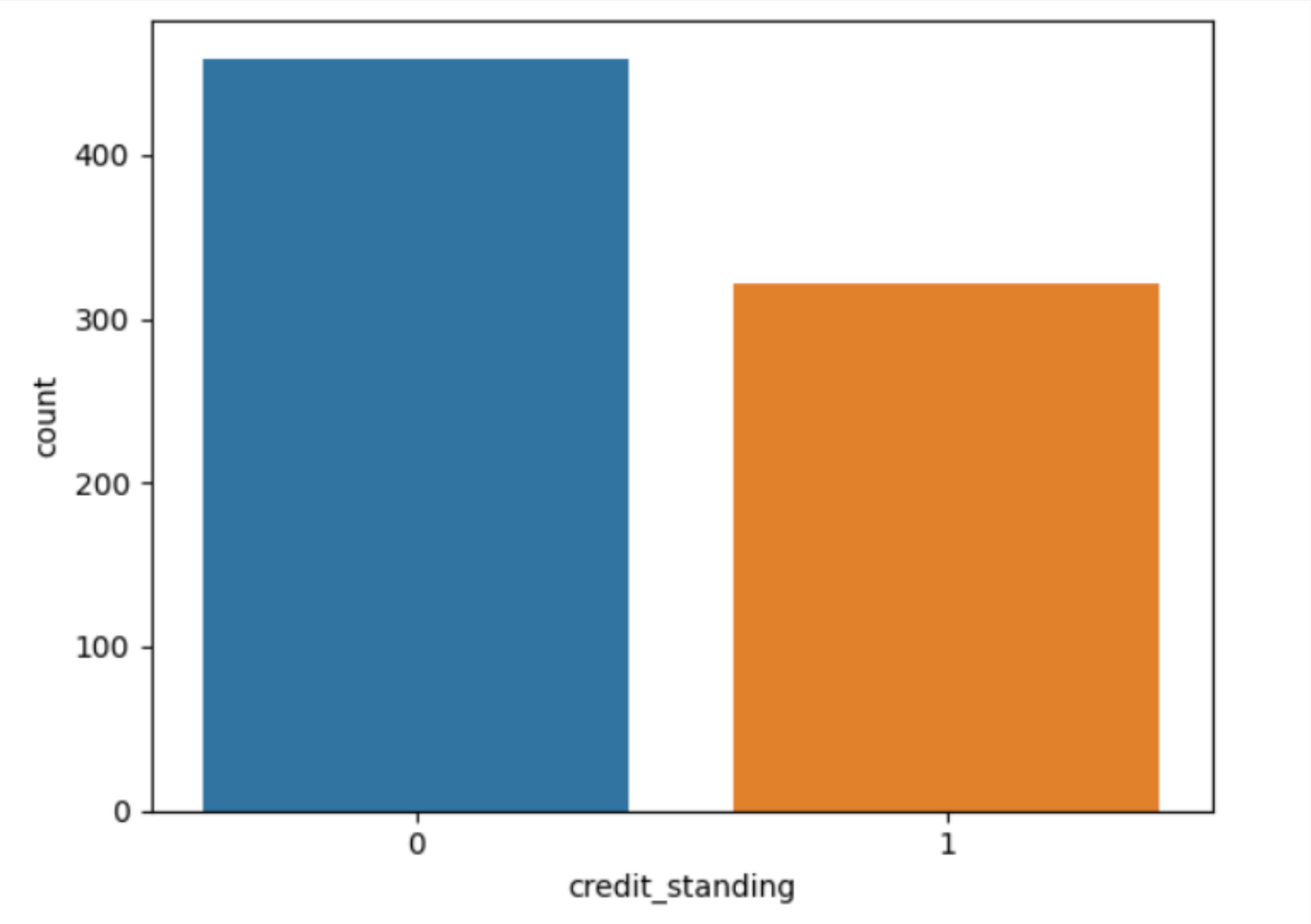
- Applied logarithmic transformations to the columns with skewness.
- dropped the rows with outliers using the drop method

EDA Perfomed

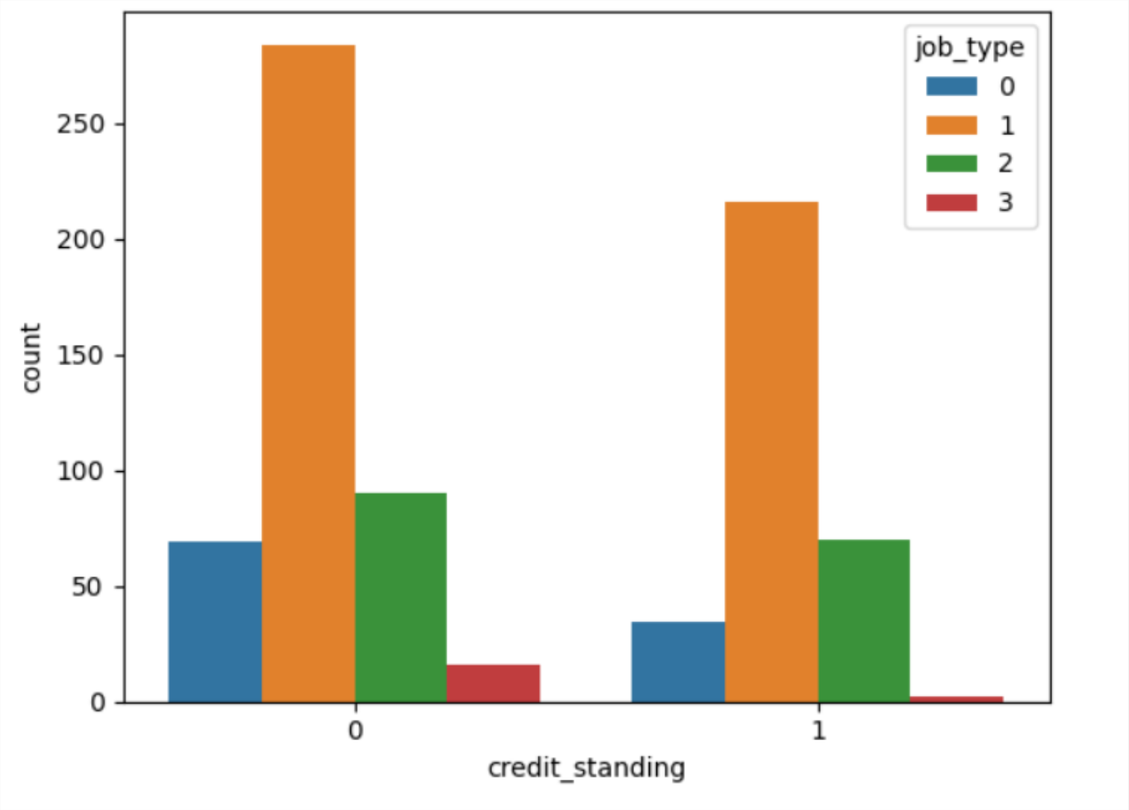
Back to Agenda Page

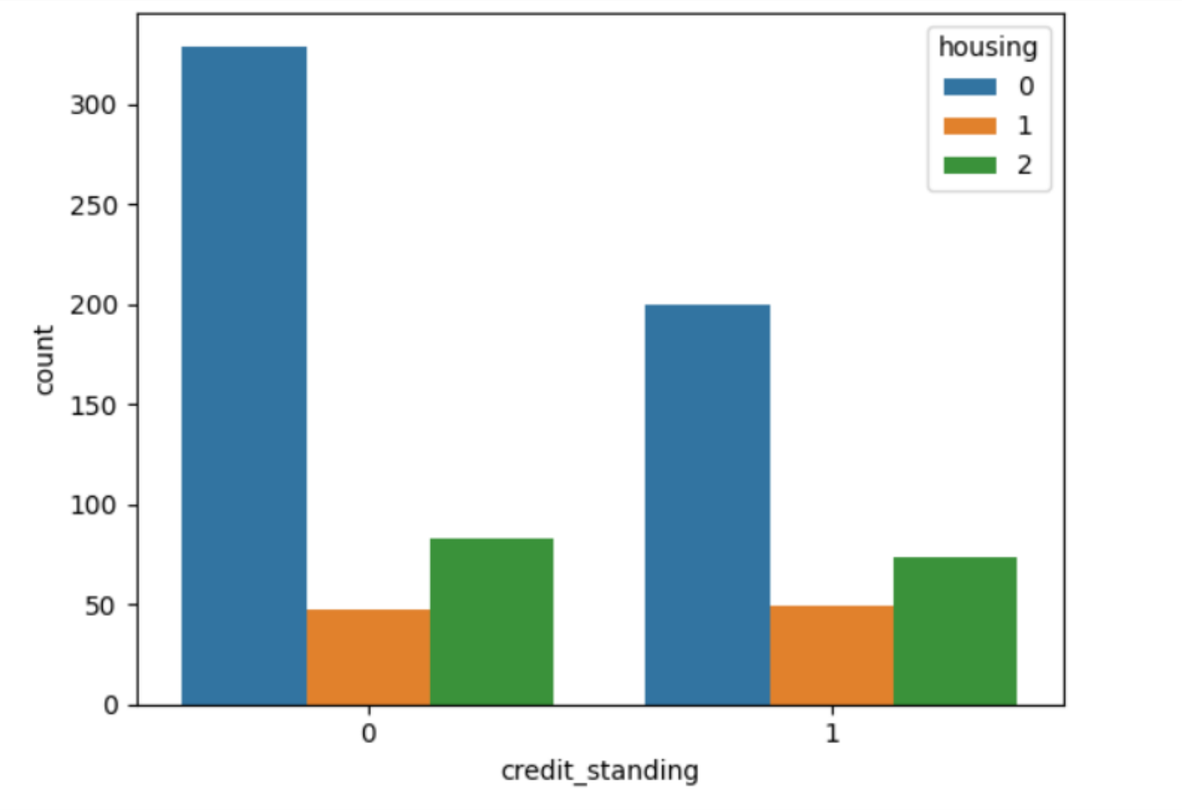


count of good and bad credit standings

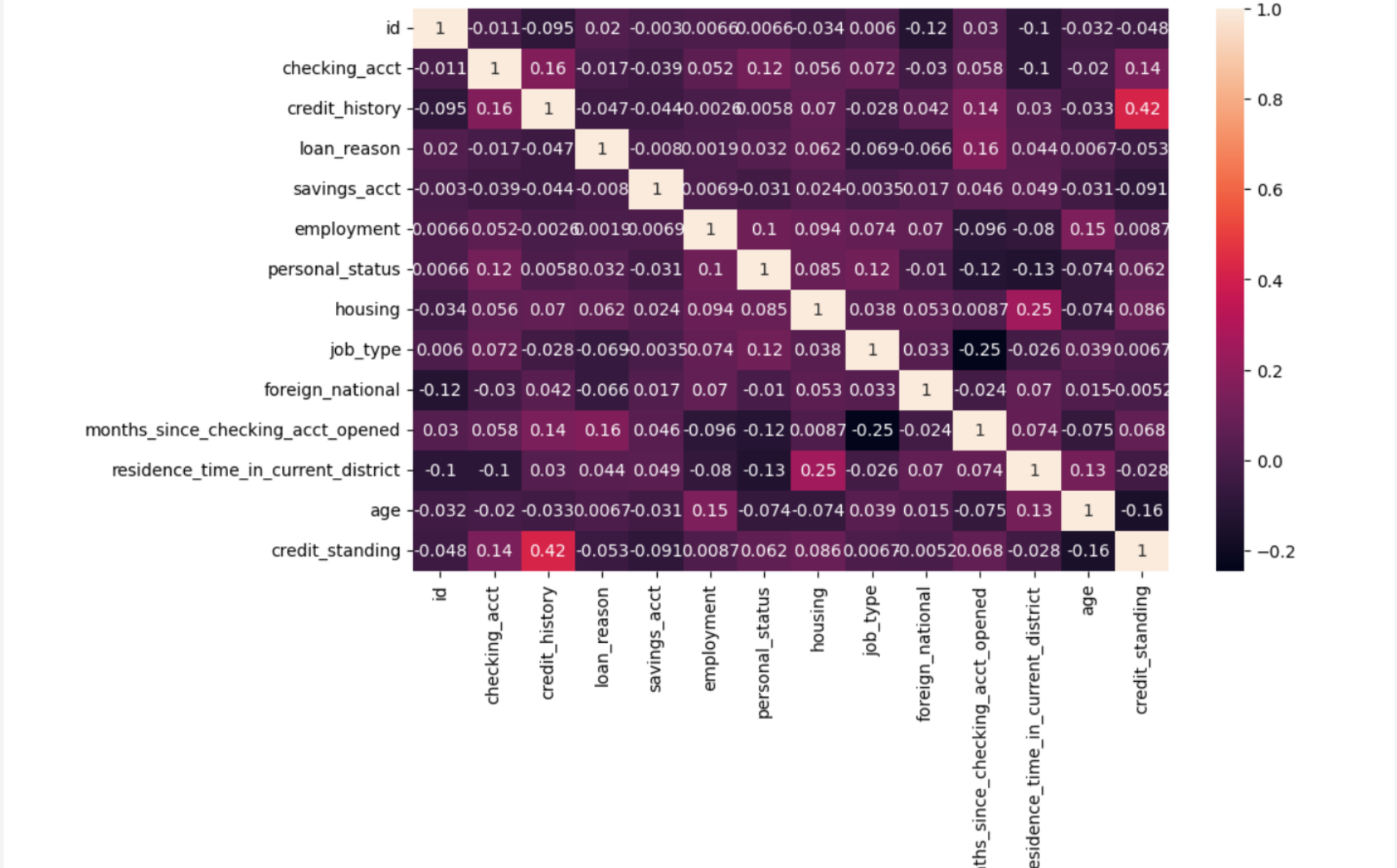


Factors influencing credit standing count





CORRELATION BETWEEN VARIABLES



Proposed Modeling Technique

Random Forest Algorithm

- Random Forest is a versatile algorithm that can handle both classification and regression tasks.
- It can also work well with both numerical and categorical data.
- Random Forest is an ensemble learning method that combines multiple decision trees to make predictions.
- By combining multiple trees, it reduces the risk of overfitting, which can occur when a model learns the training data too well and performs poorly on new data.
- Random Forest can handle missing data well. It can make use of available data to predict missing values and does not require imputation of missing data.
- Random Forest is less sensitive to outliers compared to other models like linear regression
- Random Forest provides a measure of feature importance, which can be useful in understanding the most important features
- Random Forest can handle large datasets with many features efficiently.

CONCLUSION

Based on detailed dive into data and exploratory data analysis, Random forest could be a best suitable model for this business problem.