Group Name: Tech Geeks
Name: Dhanashri Jagadale
Email: dhanashri.jagadale1998@gmail.com
College: Munster Technological University
Specialization: Data Science

FINAL REPORT

Problem Description:

The problem is to develop a predictive model that can assess the credit worthiness of potential future customers of a financial institution. The available data set consists of 807 past loan customer cases, each with 14 attributes including financial standing, reason for the loan, employment, demographic information, foreign national status, years of residence in the district, and the outcome/label variable Credit Standing, which classifies each case as either a good loan or bad loan. The objective is to build a model that accurately predicts the credit standing of new loan applications, using the available data as the training set. 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.

Business Understanding:

In the given problem statement, the business understanding stage involves understanding the financial institution's objective of assessing the creditworthiness of potential future customers. This stage requires a thorough understanding of the financial domain, including the lending process, loan evaluation criteria, and credit risk management. The data scientist needs to work closely with Maeve, the manager of the financial institution, to identify the specific problem to be solved, such as accurately predicting the credit standing of loan applications, which can reduce the risk of default and improve the profitability of the institution. It is also essential to define the key performance indicators, such as accuracy, precision, and recall, that the model should achieve to meet the business objectives. By gaining a deep understanding of the business context and goals, the data scientist can build a model that not only provides accurate predictions but also helps the financial institution make better lending decisions by identifying the key factors that determine creditworthiness.

PROPOSED SOLUTION:

The financial institution can develop a machine learning model that uses historical loan data to predict the likelihood of loan default. The model can be trained on a range of factors such as the applicant's credit history, income, employment status, loan amount, loan duration, etc. The

model can use different algorithms such as logistic regression, decision trees, or neural networks to learn the patterns and relationships between these variables and the likelihood of loan default. Once the model is trained and validated, it can be used to score new loan applications automatically. The model can generate a credit score for each applicant, indicating their creditworthiness and the risk of default. Based on these scores, the financial institution can decide whether to approve or reject the loan application, or adjust the loan terms such as interest rate or loan duration.

Benefits:

- Improved accuracy: Machine learning models can learn complex patterns and relationships between variables that are difficult to capture with traditional statistical methods. This can improve the accuracy of credit scoring and reduce the risk of default.
- Faster processing: Automated credit scoring can speed up the loan approval process and reduce the time and cost of manual assessment.
- Consistency: Machine learning models can apply consistent and objective criteria to all loan applications, eliminating subjective bias or errors that may occur in manual assessment.
- Better portfolio management: By accurately predicting credit worthiness, the financial institution can improve its loan portfolio management and reduce the risk of bad loans.
- Increased customer satisfaction: Faster and more accurate loan approval can improve customer satisfaction and attract more business.

Data understanding:

The dataset contains information about various individuals who have applied for loans. The data includes information such as the individual's checking account status, credit history, loan reason, savings account status, employment status, personal status, housing status, job type, foreign national status, months since checking account opened, residence time in the current district, age, and credit standing. There are 807 rows with 14 attributes.

Data cleansing and transformation done on the data:

1. Handling missing values

As the data contains missing values, we need to handle them as well. We can either remove the missing values or fill them with some values like the mean or median of the feature. The number of missing values is very small and there is no meaningful pattern to their occurrence therefore, Deleting missing values can be a reasonable approach. I have used dropna() function to remove any missing values before calculating the skewness. This is important because missing values can cause errors in the skewness calculation.

2. Handling skewness

Months.since.Checking.Acct.opened:Since the skewness value is greater than 1, this means the distribution is highly skewed to the right. In this case, we can apply a log transformation to reduce the skewness. After applying the log transformation, we can check the skewness value again to see if it has reduced.

Residence.Time.In.current.district: Since the skewness value is close to 0, this means the distribution is approximately symmetric. In this case, we don't need to apply any transformation.

Age: Since the skewness value is close to 1, this means the distribution is slightly skewed to the right. In this case, we can apply a log transformation to reduce the skewness. The log transformation can be applied.

3. Handling outliers

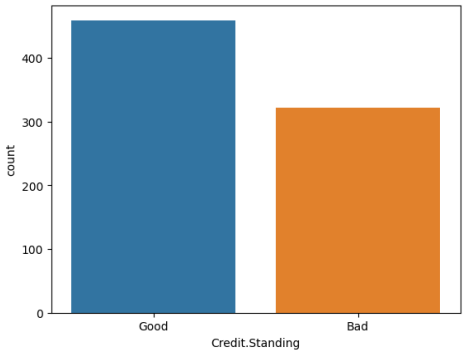
Outliers can have a significant impact on data analysis, so it's important to handle them appropriately. I have dropped the rows with outliers using the drop method with the index of the rows containing outliers

4. Next I have used the factorize function to factorize the categorical data to accurately analyze it further without any errors.
5. Changed the datatype of all numerical data to integer to avoid inconsistency in the data.
6. Renamed column name with appropriate naming conventions.
7. To build the best machine learning model for this dataset, it is important to first preprocess the data by encoding the categorical variables and scaling the numerical variables. I have done this using techniques such as one-hot encoding for categorical variables and standard scaling for numerical variables.

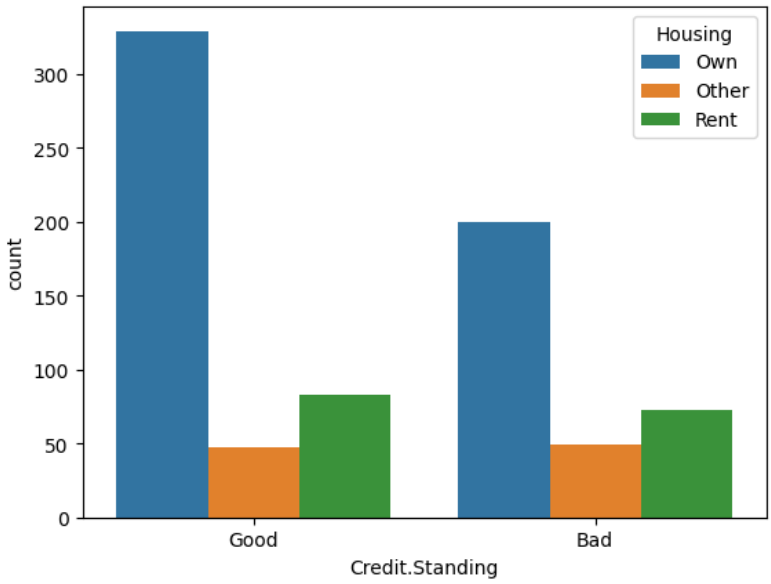
EDA Performed on the data:

1. Checked for the first five rows with the .head() function.
2. Checked the number of rows and columns with .shape.
3. Gained information about data types with .info() function.
4. Used .describe() function to gain statistical summary of the data.
5. Checked for missing values and dropped them.
6. Plotted histogram for numerical variables to check the distribution of data and found that there is a skewness in the data which I handled through log transformation.
7. Plotted boxplots to check the presence of outliers. There were very few outliers present in the data so dropped them to avoid errors in the result.

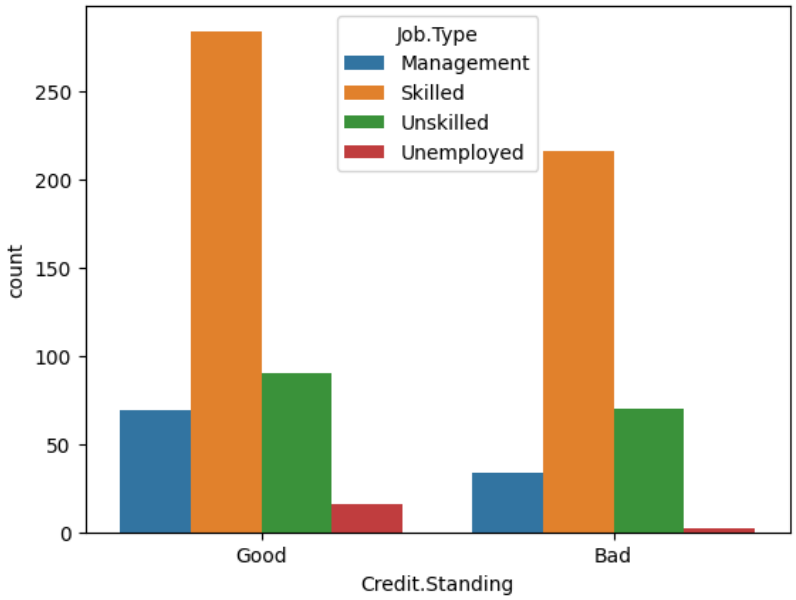
8. Used sns.countplot function to plot the count of good and bad credit standings and found that the dataset has more good credit standing records.



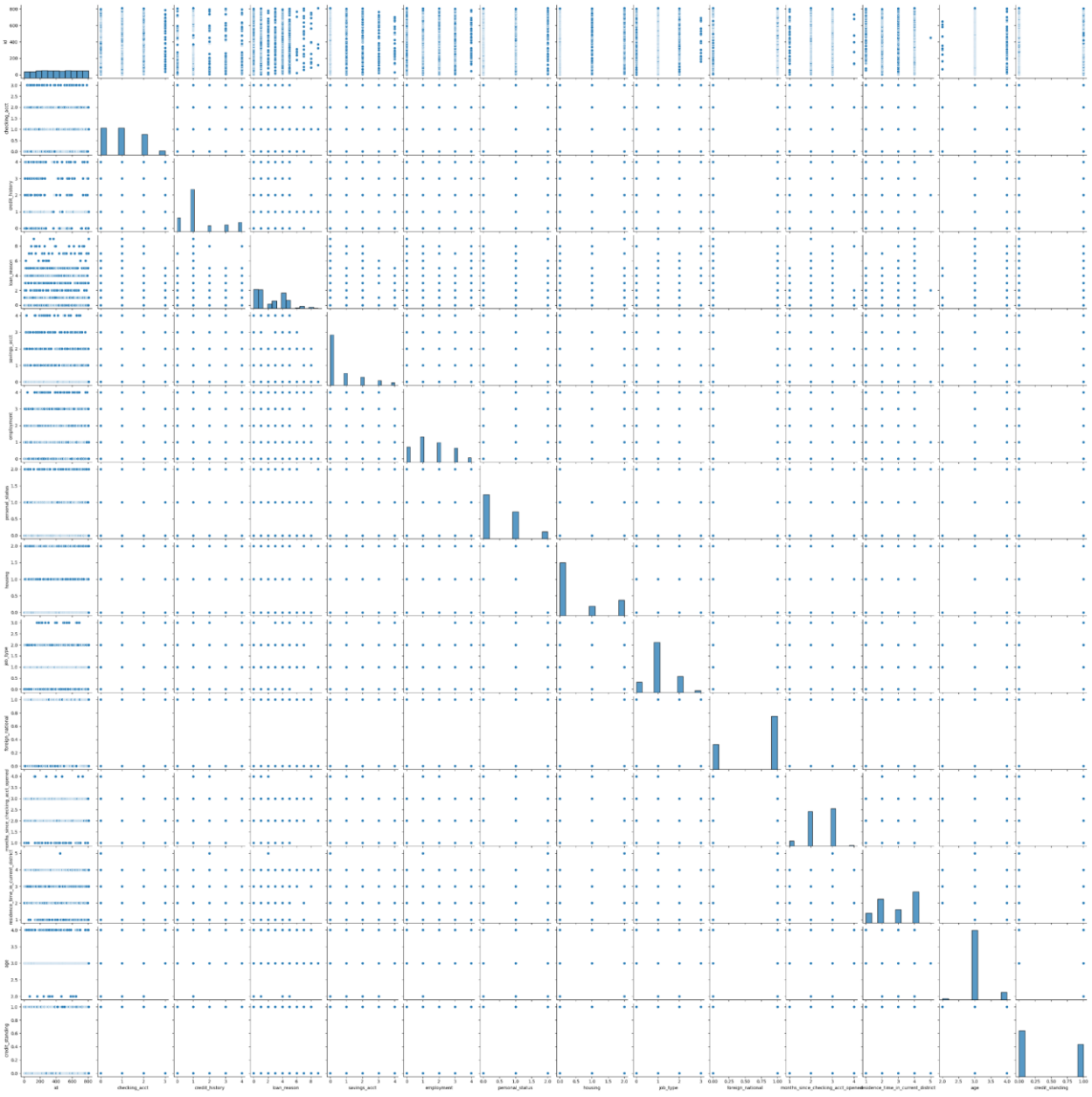
9. Next plotted the credit standing by using housing type as the differentiating category. The result shows that people with their own house are more likely to have good credit standing.



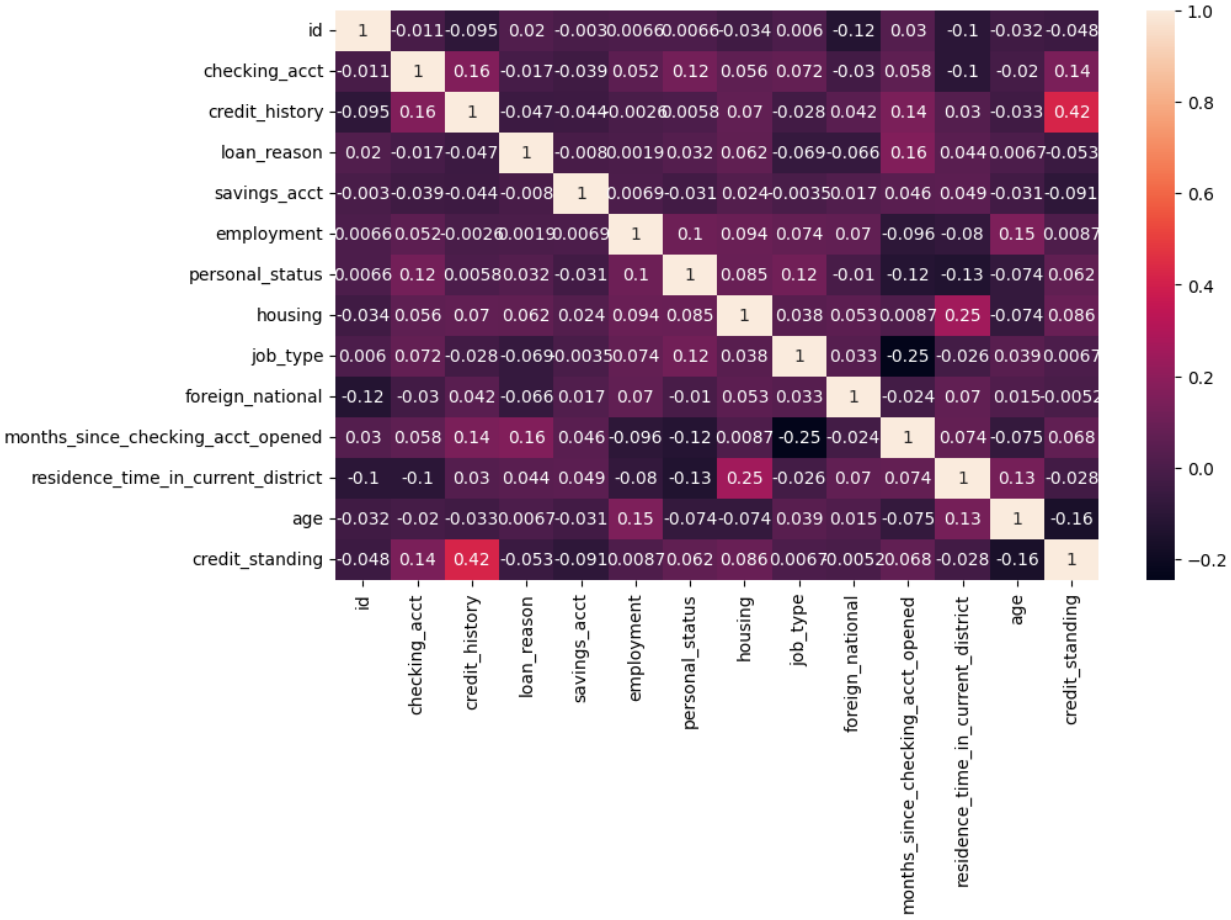
10. Next, plotted the credit standing count with the job type as a differentiating variable. Skilled people are more likely to have good credit standing. To my surprise even if the people are unemployed they are most likely to have good credit standing.



11. Next plotted the pairwise relationship among the numerical variables using sns.pairplot function and there is mostly no linear relationship between any of the two variables



12. Plotted heatmap to verify the correlation and it is evident that there is no strong negative or positive correlation between these variables and hence no linear relationship.



It seems there is a non relationship between feature variables from the Exploratory data analysis. Decision tree or Random forest machine learning algorithms can be assumed to give best results in such scenarios. Overall, the best machine learning model for this dataset will depend on the specific data and It is important to carefully evaluate and compare the performance of different algorithms before selecting the best one for this dataset. So I will be building different models to compare the performance and select the best model at the end.

Modeling techniques used:

Trained 6 different models to choose the best one.

Logistic Regression Accuracy: 66.35 %
K-Nearest Neighbors Accuracy: 65.40 %
Decision Tree Accuracy: 75.50 %
Random Forest Accuracy: 76.79 %
Support Vector Machine Accuracy: 70.68 %
Naive Bayes Accuracy: 64.42 %

PROPOSED MODELING TECHNIQUE

DECISION TREE AND RANDOM FOREST ARE PROVIDING ALMOST SIMILAR ACCURACY BUT THE RANDOM FOREST WOULD BE THE BETTER CHOICE FOR FOLLOWING REASONS:

- 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

In conclusion, implementing a machine learning model for credit scoring can provide significant benefits to financial institutions by improving the accuracy and speed of loan processing, reducing the risk of default, and increasing customer satisfaction.