The dataset has 1000 records with no missing values. 700 of these are labelled as good while the remaining 300 are labelled as bad. There are 21 columns with one of them being the label. Of the 20 features, 7 are numeric and 13 are non-numeric. Two features, duration and credit amounts show a positive collinearity of 0.62. No missing value makes the analysis simpler. A small discrepancy in the feature credit history is noticeable as the category ‘all paid’ and ‘no credits/all paid’ essentially represent the same thing.

After doing the EDA, I ran logistic regression. The plot of confusion matrix for logistic regression indicates that the models are good at identifying good credits as good credits with an accuracy of 83% whereas it identifies a bad credit correctly only 52% of the time.

I then implemented KNN algorithm and chose 5 neighbors after plotting the error rate vs number of neighbors. The plot of normalized confusion matrix for KNN indicated that this model is also better at identifying good credits and underperforms in identifying bad credits correctly. This could be possibly because 70% of data is labelled as good credit.

Following are the key insights found on analyzing the relationships of the attributes with the label: -

1. Checking Status: 80% of bad credits have a checking status less than 200’. 88% of people with ‘no checking’ as their checking status are good credits.

2. Credit History: 83% of people who have critical/other existing credit are good credits. 86% of good credits have their credit history of either ‘existing paid’ or ‘critical/other existing credit’.

3. Purpose: 83% of the times when the purpose is ‘used car’, credits are good whereas with new car good credits fall to 61%. 8 out of 9 retraining purposes yields to a good credit. Also 78% of times radio/tv results in good credit.

4. Credit Amount: Average credit amount of a bad credit is $3938 whereas for a good credit is $2985.

5. Saving Status: 72% of bad credits have a saving status of less than 100’. 82% of no known savings have good credit.

6. Employment: 77% of people with 4 to 7 years of present employment are good credits with the highest number of good credits falling in the 1 to 4 years of employment.

7. Installment Commitment: 53% of bad credits have an installment rate of 4% of disposable income.

8. Personal Status: 57% of Good credits are single male.

9. Other Parties: 80% of people with another debtor/guarantor are good credits. 70% of those with no other parties are good credits.

10. Property: 79% of clients who own real estate are good credits and 69% of people with a life insurance are good credits.

11. Age: Average age of a person who is a good credit is 36.2 years while those of bad credit is 34 years.

12. Other payment plans: 72% of people with no other payment plan make a good credit.

13. Housing: 75% of people with good credits own their residence.

14. Number of existing credits at this bank: 68% of people with bad credits have 1 existing credit whereas 72% of those with 2 existing credits are good credits.

15. Jobs: 70% of skilled employees are good credits. Surprisingly, 72% of unskilled employees are also good credits.

16. Number of dependents: 70% of people with 1 dependent are good credits.

17. Foreign worker: 89% of non-foreign workers are good credits.