Predicting German Credit Risk

SEP 767 Multivariate Statistical Methods for Big Data Analysis and Process Improvement

Prathamesh Joshi – joship14@mcmaster.ca

Master of Engineering – Systems and Technology

Student ID – 400485705

Under the guidance of Professor Dr Brandon Corbett – corbeb@mcmaster.ca

<u>This file contains all the images associated to the report.</u>

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 10 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Age               1000 non-null   int64
 1   Sex               1000 non-null   object
 2   Job               1000 non-null   int64
 3   Housing           1000 non-null   object
 4   Saving accounts   817 non-null    object
 5   Checking account  606 non-null    object
 6   Credit amount     1000 non-null   int64
 7   Duration          1000 non-null   int64
 8   Purpose           1000 non-null   object
 9   Risk              1000 non-null   object
dtypes: int64(4), object(6)
memory usage: 78.2+ KB
```

Fig 1: Dtype and non-null count

	Age	Job	Credit amount	Duration
count	1000.000000	1000.000000	1000.000000	1000.000000
mean	35.546000	1.904000	3271.258000	20.903000
std	11.375469	0.653614	2822.736876	12.058814
min	19.000000	0.000000	250.000000	4.000000
25%	27.000000	2.000000	1365.500000	12.000000
50%	33.000000	2.000000	2319.500000	18.000000
75%	42.000000	2.000000	3972.250000	24.000000
max	75.000000	3.000000	18424.000000	72.000000

Fig 2: Numerical features distribution

```
NUMBER OF UNIQUE VALUES PER FEATURE:

Age                  53
Sex                   2
Job                   4
Housing               3
Saving accounts       5
Checking account      4
Credit amount       921
Duration             33
Purpose               8
Risk                  2
dtype: int64
```

Fig 3: Displaying number of unique values per variable

```
male      690
female    310
Name: Sex, dtype: int64


own     713
rent    179
free    108
Name: Housing, dtype: int64


little        603
no-info       183
moderate      103
quite rich     63
rich           48
Name: Saving accounts, dtype: int64


no-info     394
little      274
moderate    269
rich         63
Name: Checking account, dtype: int64


car                    337
radio/TV               280
furniture/equipment    181
business                97
education               59
repairs                 22
domestic appliances     12
vacation/others         12
Name: Purpose, dtype: int64


good    700
bad     300
Name: Risk, dtype: int64
```

Fig 4: count of each unique value

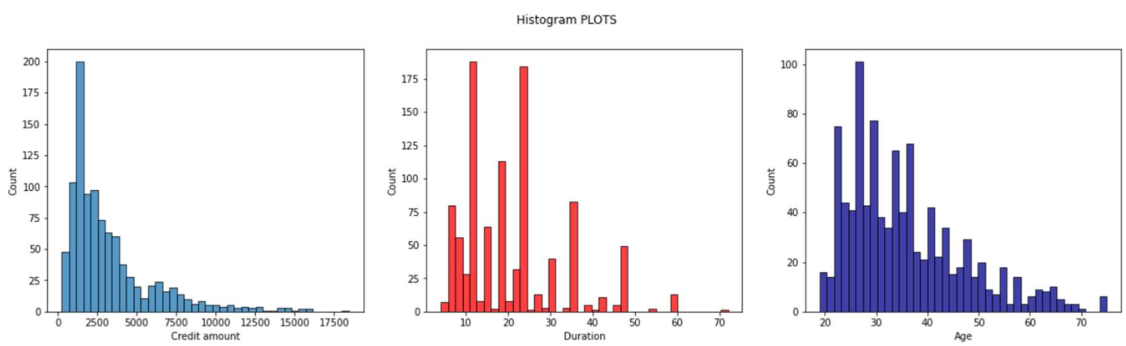


Fig 5: Univariate analysis of some of the features

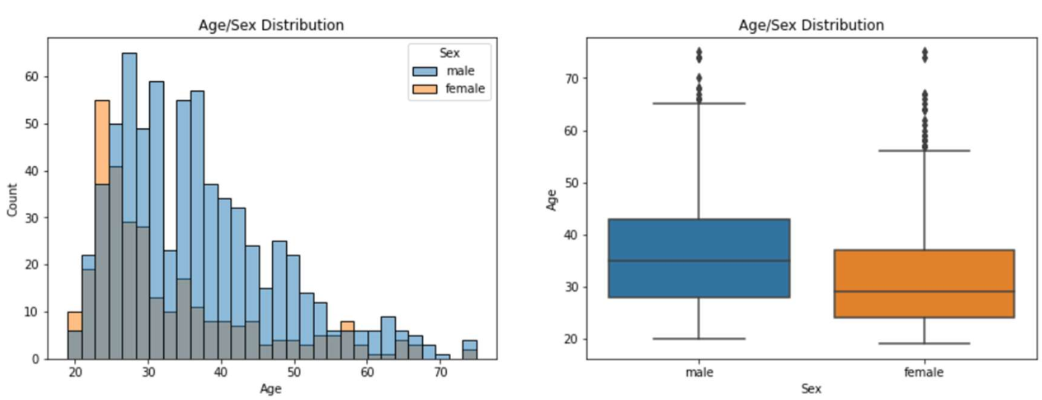


Fig 6: Bivariate analysis of Age and Sex variables

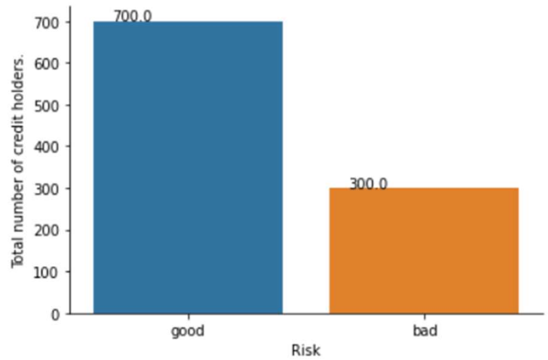


Fig 7: Countplot of risk – target variable

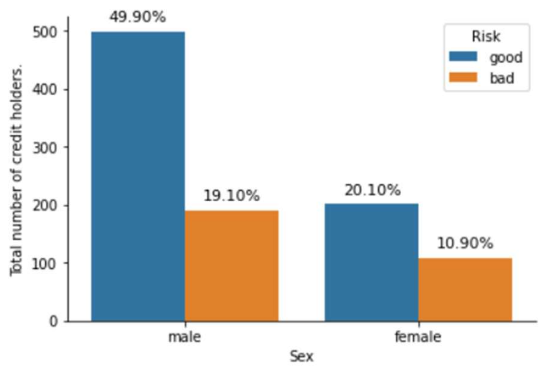


Fig 8: Distribution of Risk with respective to Sex

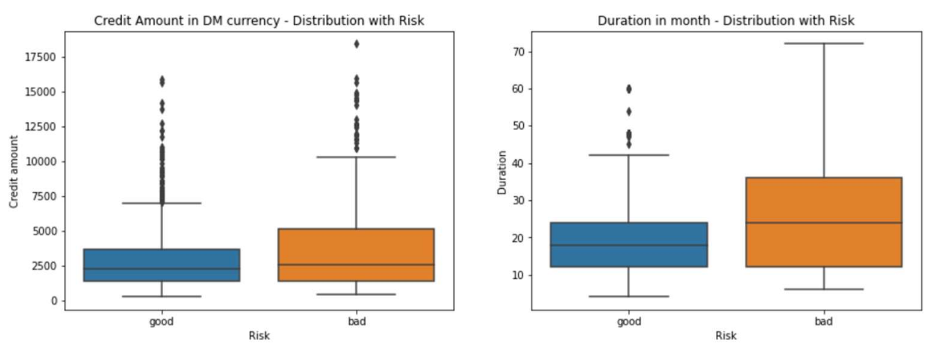


Fig 9: Risk vs Credit amount and Risk vs Duration

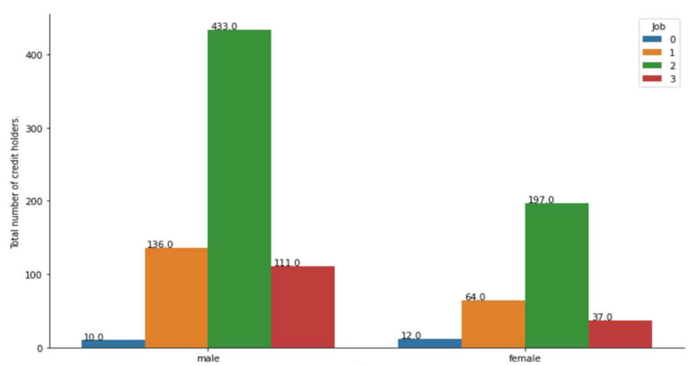


Fig 10: Countplot of job skill with sex

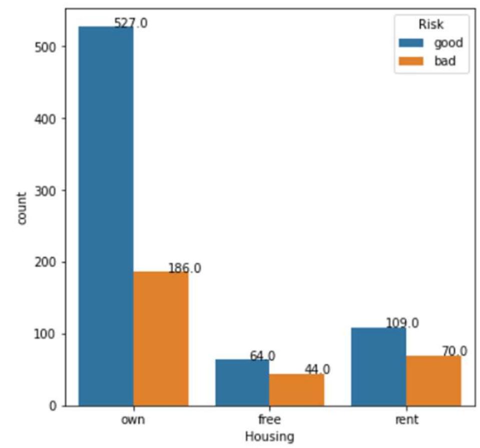


Fig 11: Housing and associate Risk count-plot

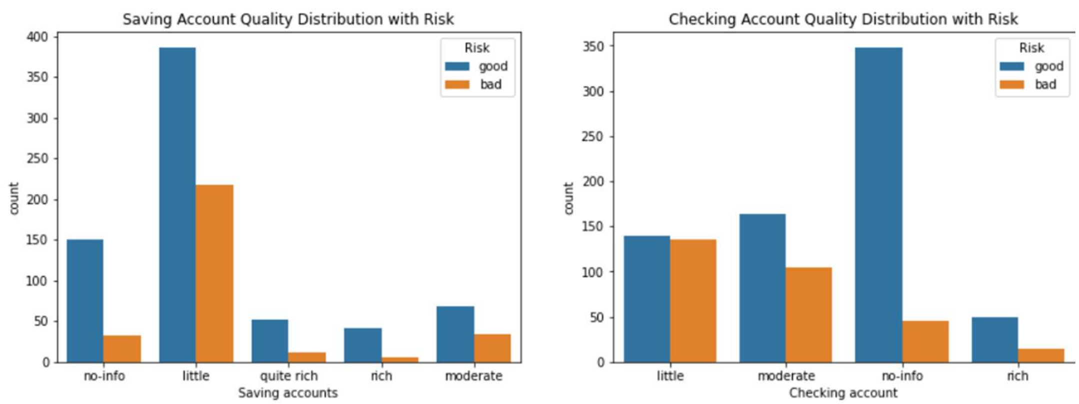


Fig 12: Savings and Checking account distribution with Risk

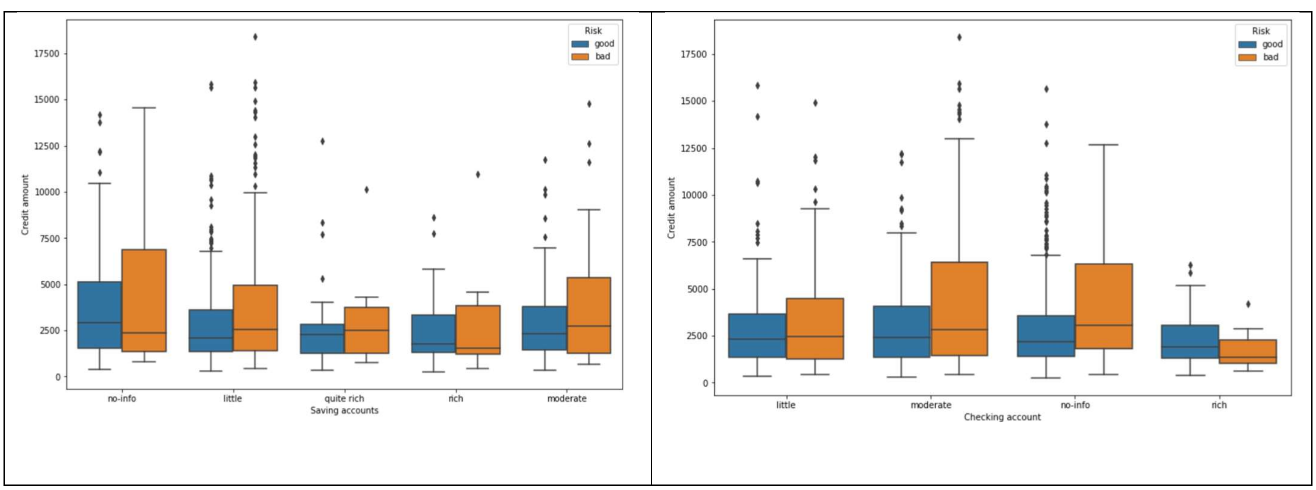


Fig 13: Savings and Checking account vs Credit amount with Risk

Purpose	Risk	Sex	
business	bad	male	27
		female	7
	good	male	51
		female	12
car	bad	male	66
		female	40
	good	male	177
		female	54
domestic appliances	bad	female	2
		male	2
	good	female	4
		male	4
education	bad	male	14
		female	9
	good	male	21
		female	15
furniture/equipment	bad	male	30
		female	28
	good	male	77
		female	46
radio/TV	bad	male	43
		female	19
	good	male	152
		female	66
repairs	bad	male	6
		female	2
	good	male	11
		female	3
vacation/others	bad	male	3
		female	2
	good	male	6
		female	1

Fig 14: Grouping purpose and risk variables with respect to sex

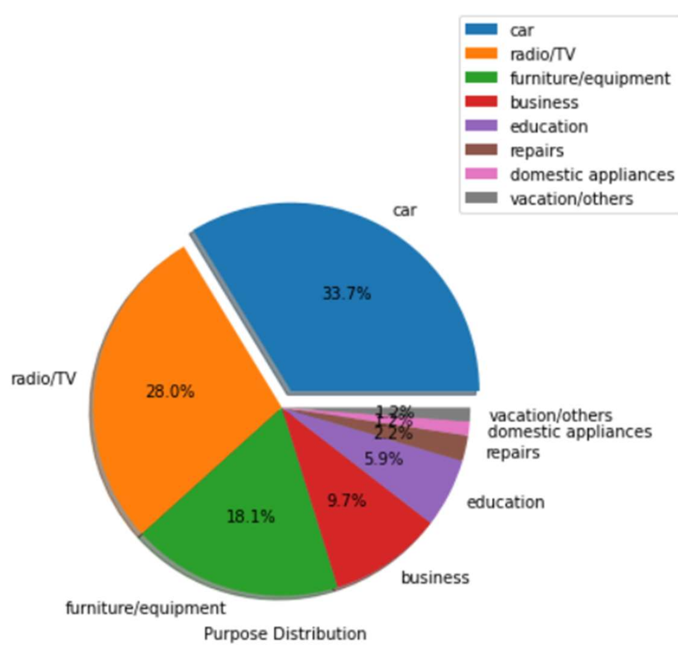


Fig 15: Pie chart of purpose for loan

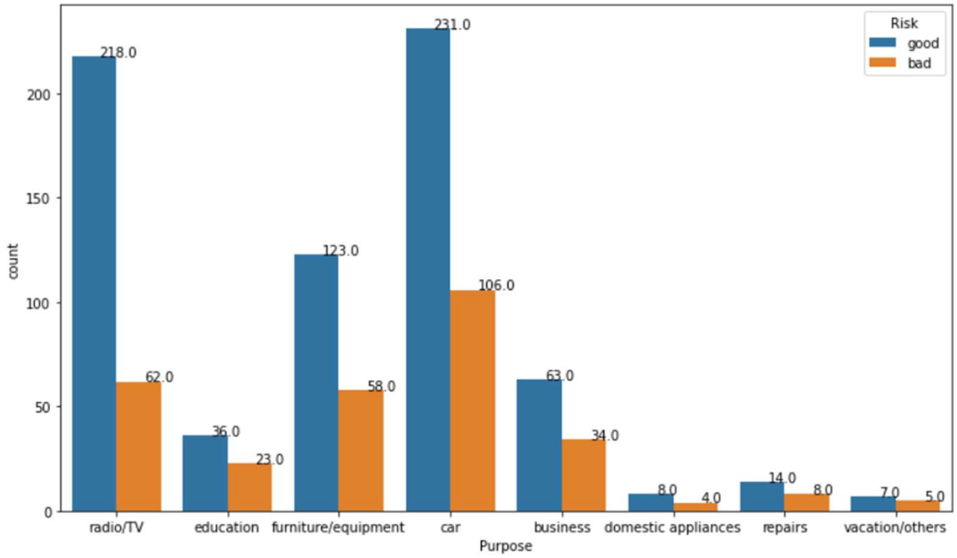


Fig 16:Distribution of Purpose with Risk

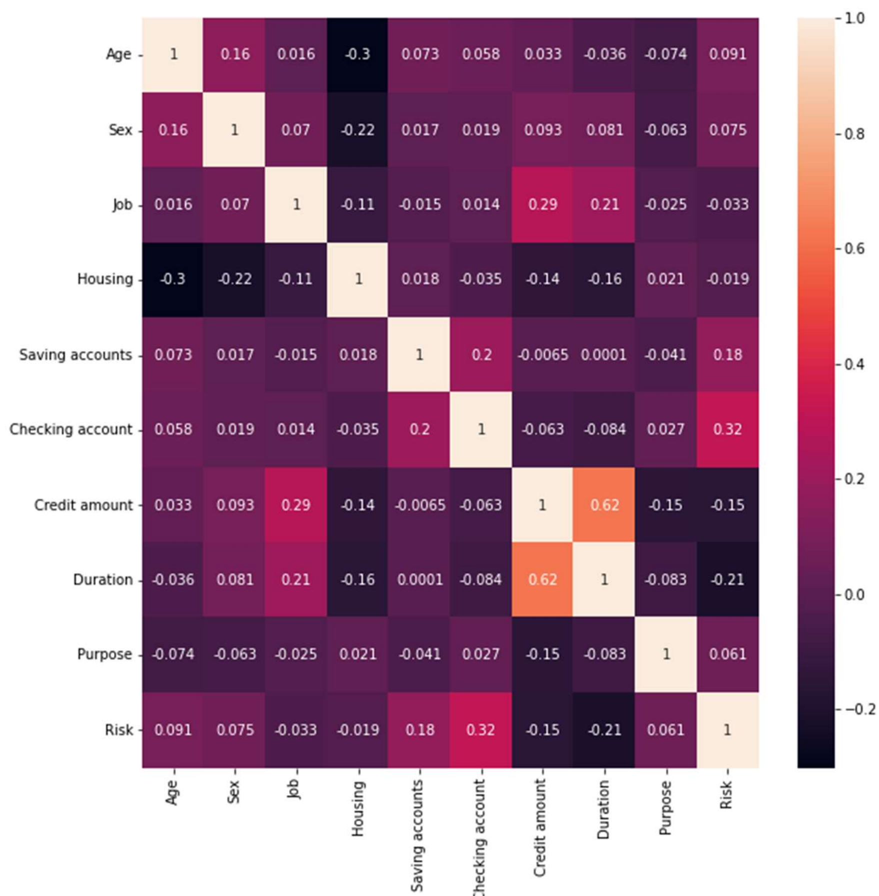
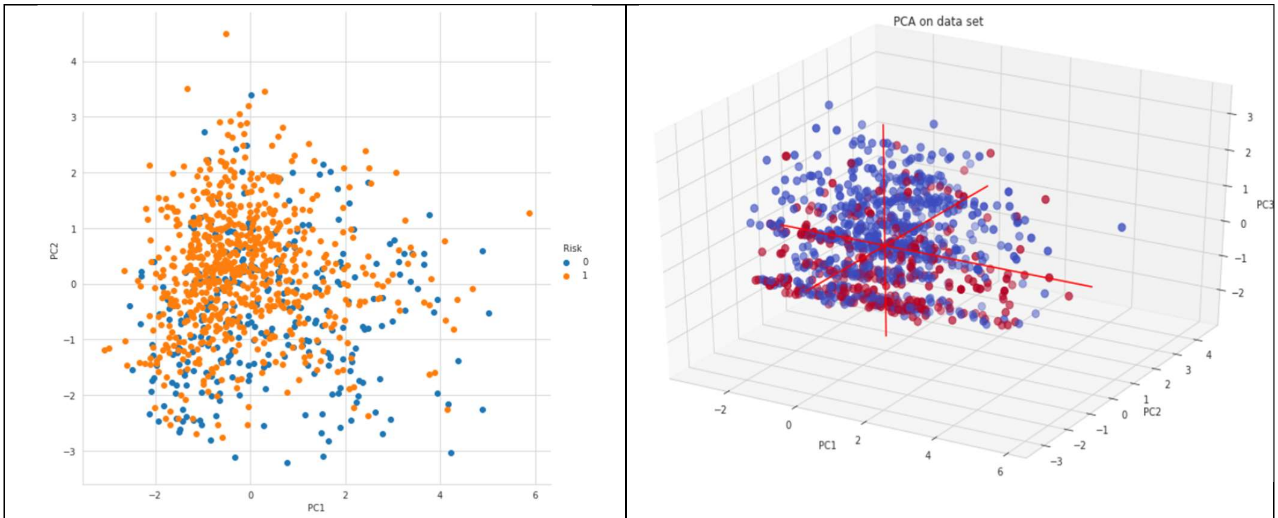


Fig 17: Heatmap of all features with Risk

	Age	Sex	Job	Housing	Saving accounts	Checking account	Credit amount	Duration	Purpose
0	2.766456	0.670280	0.146949	-0.133710	0.955847	-1.344000	-0.745131	-1.236478	1.073263
1	-1.191404	-1.491914	0.146949	-0.133710	-0.706496	-0.265348	0.949817	2.248194	1.073263
2	1.183312	0.670280	-1.383771	-0.133710	-0.706496	0.813303	-0.416562	-0.738668	0.061705
3	0.831502	0.670280	0.146949	-2.016956	-0.706496	-1.344000	1.634247	1.750384	0.567484
4	1.535122	0.670280	0.146949	-2.016956	-0.706496	-1.344000	0.566664	0.256953	-0.949853

Fig 18: Standardize Data




```
sum(pca.explained_variance_ratio_)
```

0.3738162048504894

```
sum(pca_1.explained_variance_ratio_)
```

0.5023911780108355

Fig 19: Visualisation of PCA with 2 and 3 components respectively

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
0	-0.536581	2.121703	-1.203952	0.250195	0.743649	0.303601	2.403770
1	1.151461	-2.346665	0.253069	1.232846	1.326862	0.161107	-0.760142
2	-0.808558	1.486696	-0.758179	-0.269251	0.250938	-0.267851	-1.070221
3	2.897207	0.057036	-1.695903	0.623460	1.313375	-0.317518	0.040774
4	1.850461	1.165030	-2.052197	-0.793020	0.423994	0.593500	0.269726

Fig 20: PCA with 7 components

```
array([0.21666756, 0.15714864, 0.12857497, 0.11200901, 0.09432997,
       0.09354757, 0.08568509])
```

```
sum(pca_2.explained_variance_ratio_)
```

0.8879628224022866

Fig 21: Cumulative Explained Variance ratio by 7 components

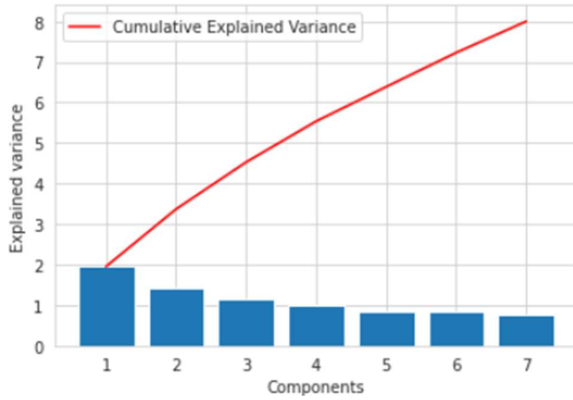


Fig 22: Barplot with Explained variance for each PC

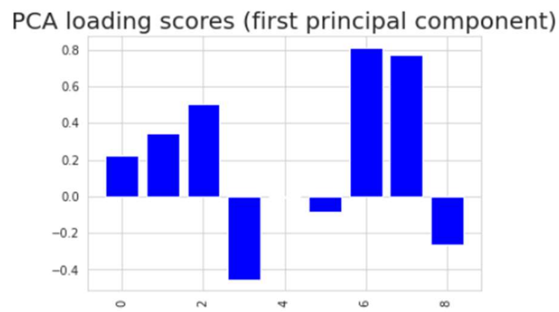


Fig 23: PC1 loadings visualization where 0 to 8 represent variable id

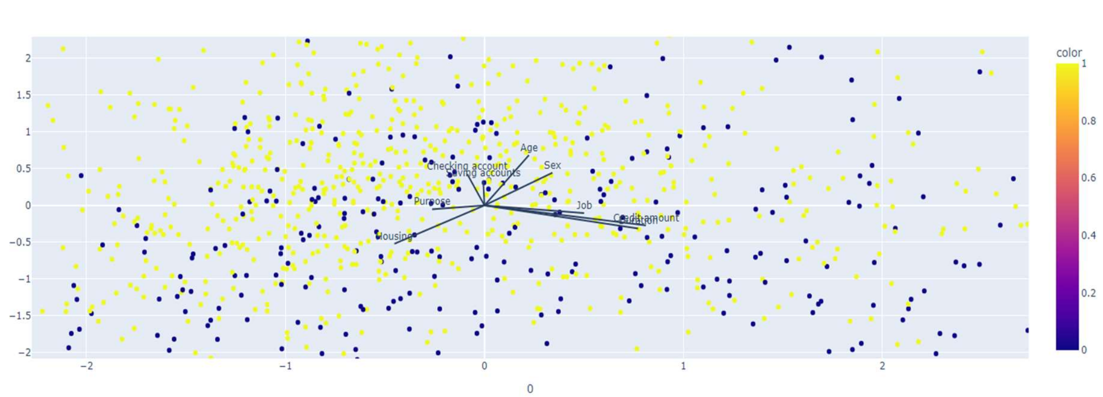


Fig 24: PCA Biplot – combination of loadings and score plot

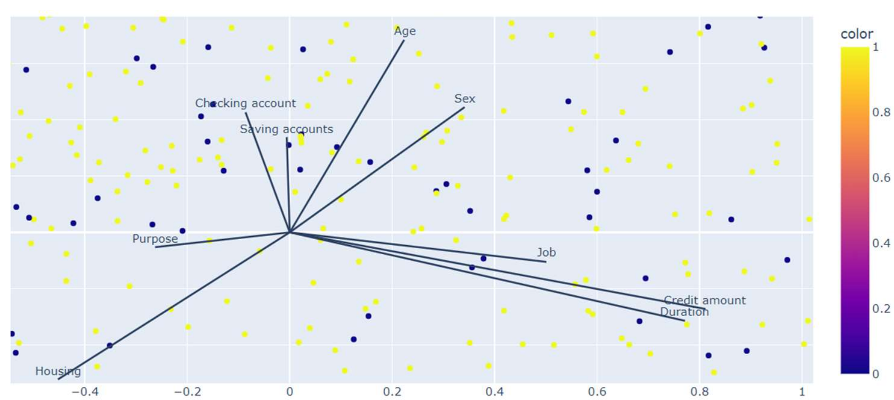


Fig 25: zoomed image of the above

Logistic Regression:

```
Training Time  LG  : 0.018
Testing Time  LG  : 0.022
```

Fig 26: Training and Testing time

```
Classification report :
              precision    recall  f1-score   support

           0       0.64      0.31      0.41        75
           1       0.76      0.93      0.83       175

    accuracy                           0.74       250
   macro avg       0.70      0.62      0.62       250
weighted avg       0.72      0.74      0.71       250

Train Accuracy Score   :  0.732
Test Accuracy Score   :  0.74
Area under curve :  0.6161904761904762
```

Fig 27: Classification Report Logistic Regression

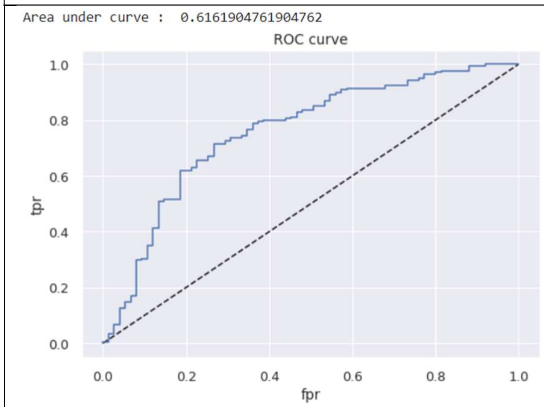


Fig 28: Area under the curve

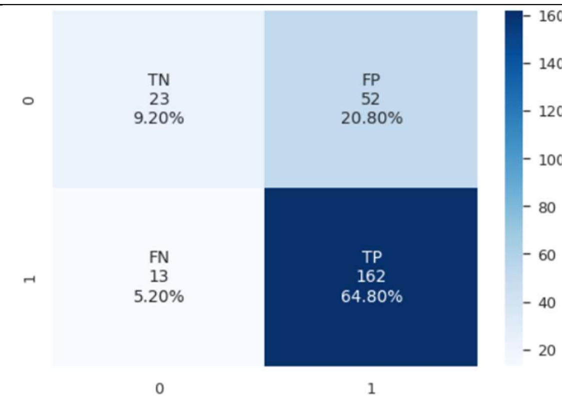


Fig 29: Confusion matrix

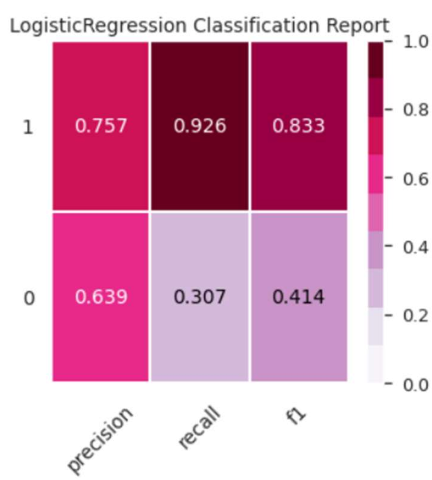


Fig 30: Precision, recall and f1 score

KNN:

```
Training time KNN   --  0.018277883529663086
Testing time KNN   --- 0.0382997989654541
```

Fig 31: Training and Testing time

```
Classification report :
              precision    recall  f1-score   support

           0       0.71      0.29      0.42        75
           1       0.76      0.95      0.84       175

    accuracy                           0.75       250
   macro avg       0.73      0.62      0.63       250
weighted avg       0.74      0.75      0.71       250

Train Accuracy Score   :  0.7466666666666667
Test Accuracy Score   :  0.752
Area under curve :  0.6209523809523809
```

Fig 32: Classification Report KNN

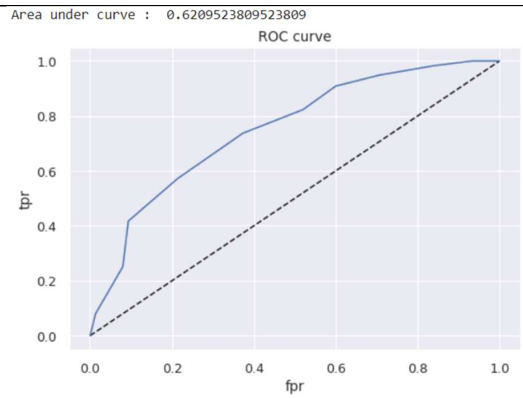


Fig 33: Area under the curve

Fig 34: Confusion matrix

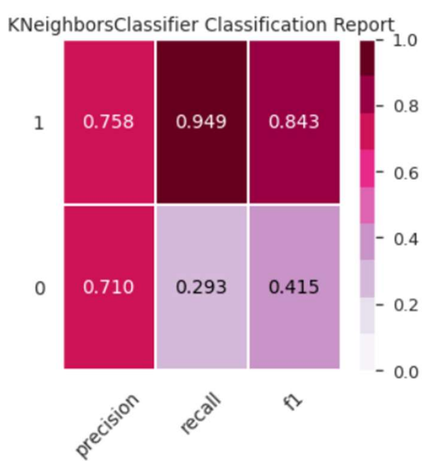


Fig 35: Precision, recall and f1 score

Decision Tree Classifier:

```
Training time DTC ----  0.015047788619995117
Testing time DTC ----  0.0021305084228515625
```

Fig 36: Training and Testing time

```
Classification report :
              precision    recall  f1-score   support

           0       0.48      0.20      0.28        75
           1       0.73      0.91      0.81       175

    accuracy                           0.70       250
   macro avg       0.60      0.55      0.55       250
weighted avg       0.65      0.70      0.65       250

Train Accuracy Score   :  0.796
Test Accuracy Score   :  0.696
Area under curve :  0.5542857142857143
```

Fig 37: Classification Report DTC

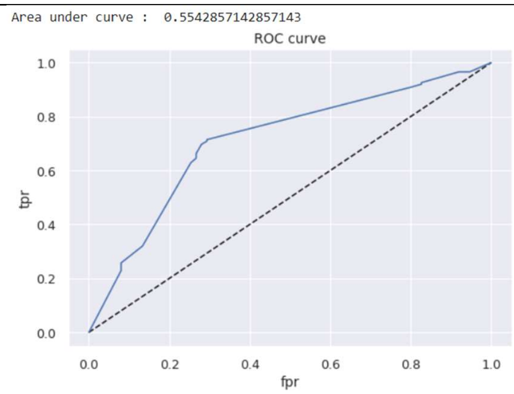


Fig 38: Area under the curve

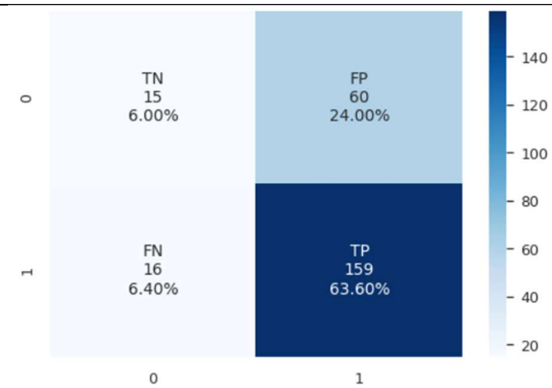


Fig 39: Confusion matrix

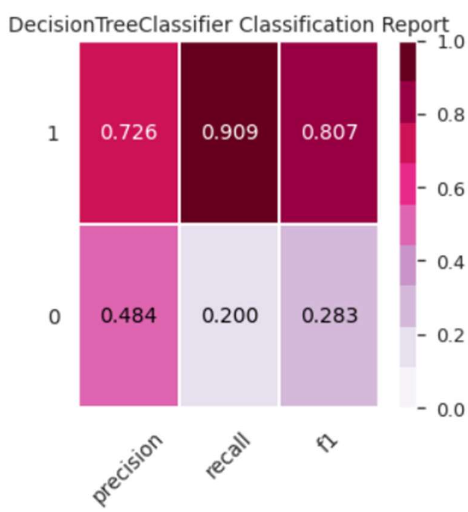


Fig 40: Precision, recall and f1 score

Random Forest Classifier:

```
Training time RFC ----  0.28087306022644043
Testing time RFC ----  0.030716896057128906
```

Fig 41: Training and Testing time

```
Classification report :
              precision    recall  f1-score   support

           0       0.67      0.45      0.54        75
           1       0.79      0.90      0.84       175

    accuracy                           0.77       250
   macro avg       0.73      0.68      0.69       250
weighted avg       0.76      0.77      0.75       250

Train Accuracy Score   :  1.0
Test Accuracy Score   :  0.768
Area under curve :  0.6780952380952381
```

Fig 42: Classification report RFC

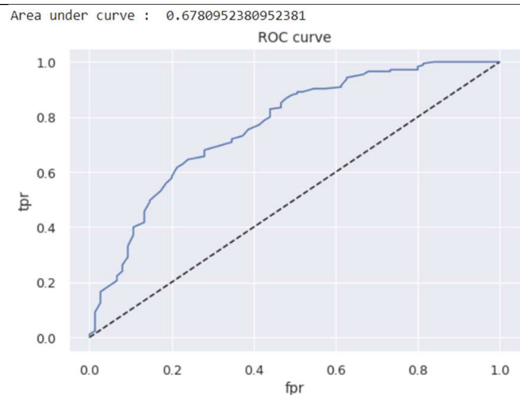


Fig 43: Area under the curve

Fig 44: Confusion matrix

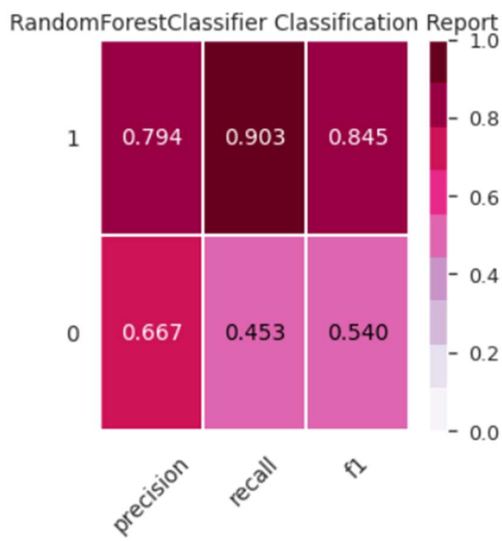


Fig 45: Precision, recall and f1 score

XGBoost Classifier:

```
Training time XGBC ----  0.17719149589538574
Testing time XGBC ---- 0.001890420913696289
```

Fig 46: Training and Testing time

```
Classification report :
              precision    recall  f1-score   support

           0       0.74      0.43      0.54        75
           1       0.79      0.94      0.86       175

    accuracy                           0.78       250
   macro avg       0.77      0.68      0.70       250
weighted avg       0.78      0.78      0.76       250

Train Accuracy Score   :  0.864
Test Accuracy Score   :  0.784
Area under curve :  0.681904761904762
```

Fig 47: Classification Report XGBoost Classifier

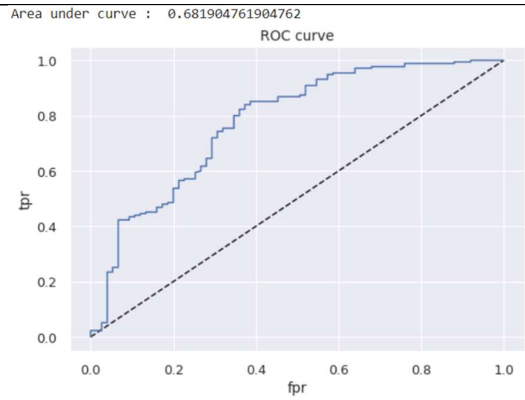


Fig 48: Area under the curve

Fig 49: Confusion matrix

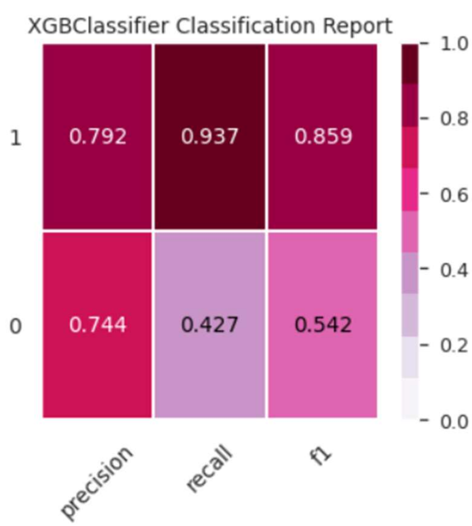


Fig 50: Precision, recall and f1 score

Random Forest Classifier			XGBoost Classifier		
	feature	importance		feature	importance
6	Credit amount	0.249119	5	Checking account	0.412734
0	Age	0.178325	7	Duration	0.105873
7	Duration	0.158723	6	Credit amount	0.085121
5	Checking account	0.129879	4	Saving accounts	0.077524
8	Purpose	0.095747	1	Sex	0.072554
4	Saving accounts	0.065215	0	Age	0.066395
2	Job	0.050208	3	Housing	0.065942
3	Housing	0.042522	8	Purpose	0.057584
1	Sex	0.030263	2	Job	0.056273

Table 1 : Feature importance Random Forest Classifier and XGBoost Classifier

Classifiers	Test Accuracy
Decison Tree Classifier	0.696
Logistic Regression	0.740
KNN	0.752
Random Forest Classifier	0.768
XGBoost Classifier	0.784

Table 2: Model Test Accuracy Comparison