MA-515 PROJECT REPORT

Foundations of Data Science

Basa Nikhilesh
2020mcb1234

Dr.Arun Kumar

➢The given dataset to analyse is 'Bankruptcy.csv'

→ Dataset contains 6819 observations and 96 characteristics.

→ Out of which one is dependent variable(Bankrupt?) and rest 95 are independent variables.

→ No variable column has null/missing values,

→ 93 columns are of float datatype and

3 columns are of int datatype(1 of them is target variable).

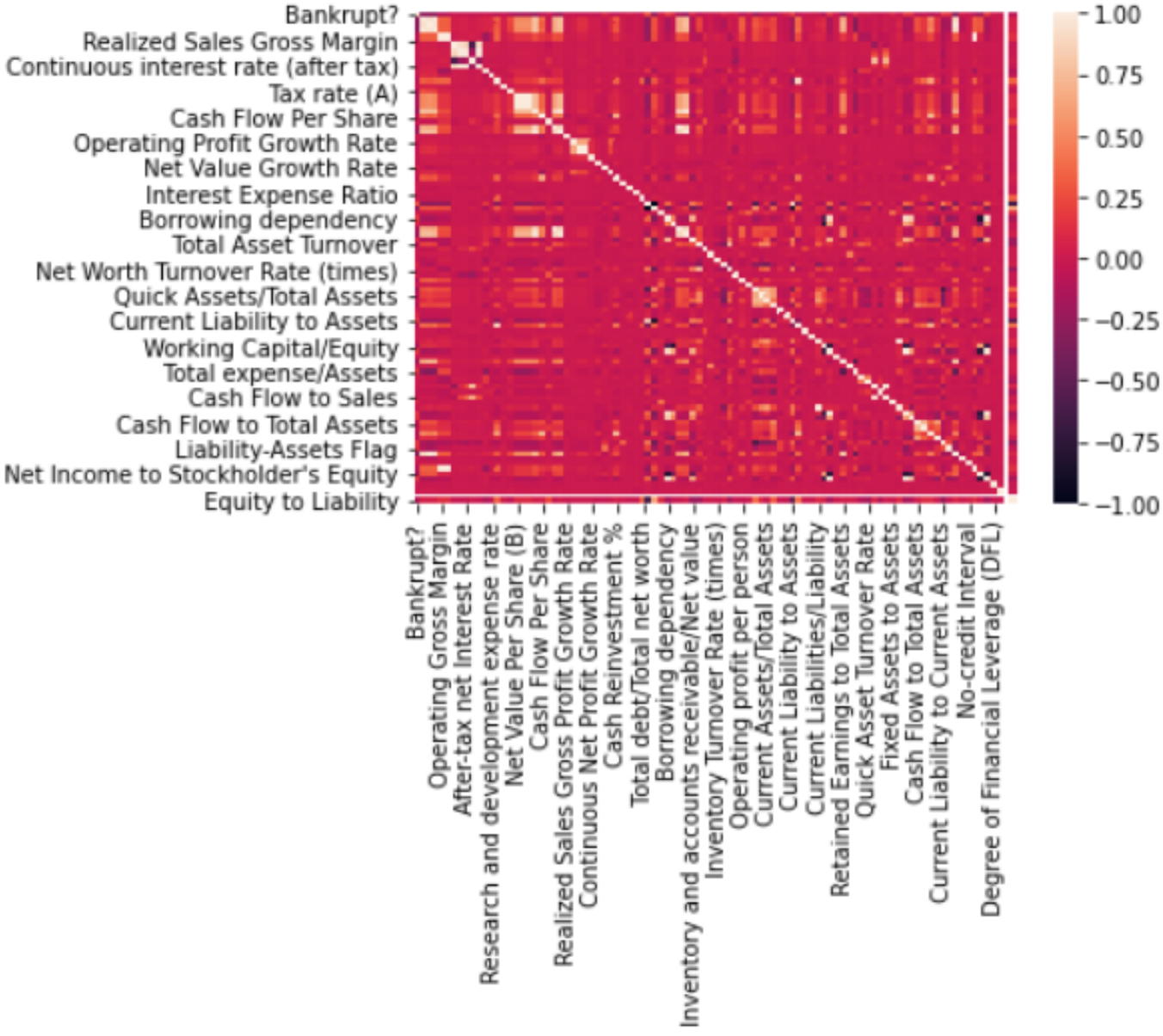
→ Target variable is discrete and categorical in nature.So,

→ It is a categorical problem,where we have to classify the new observations .

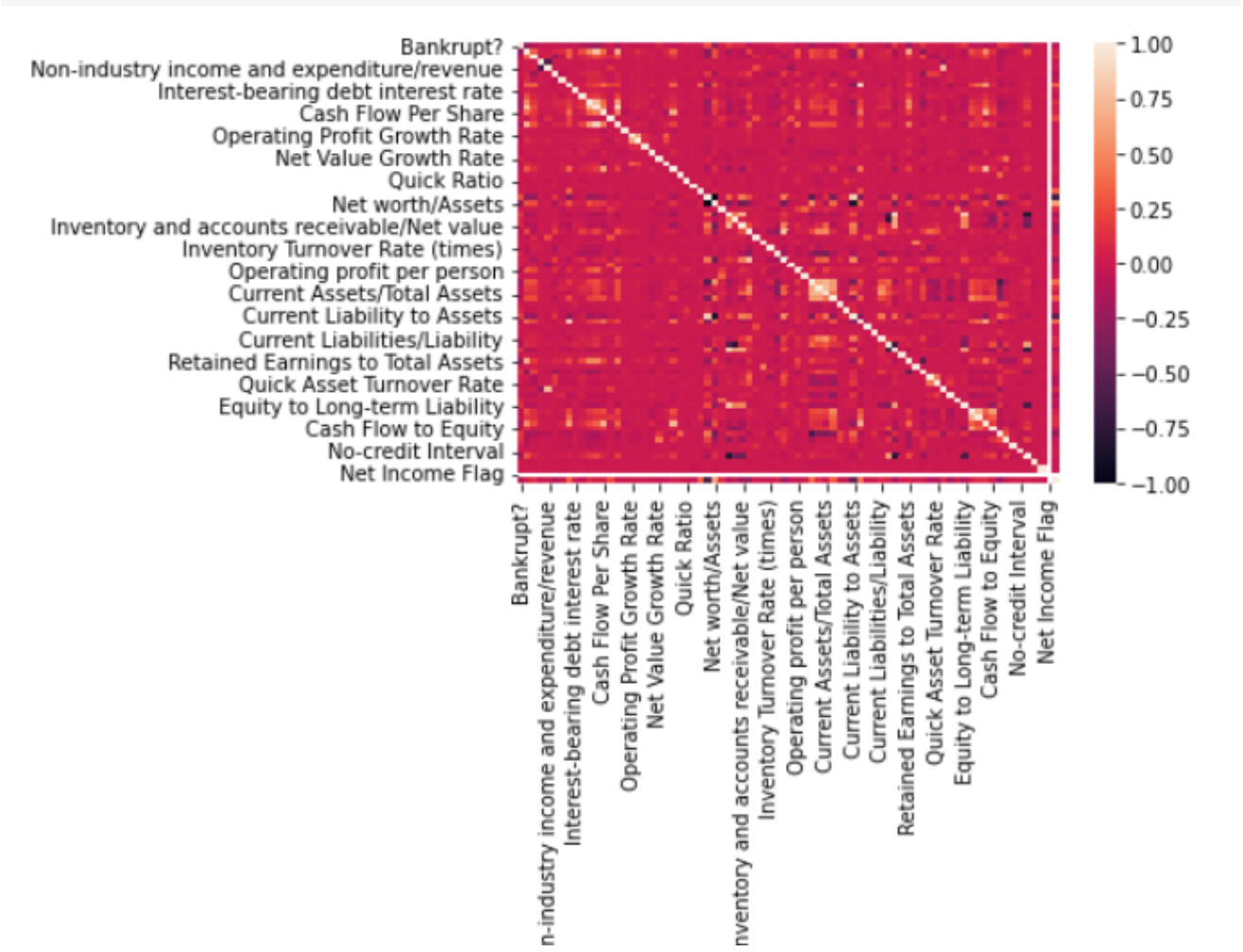
→ Only a few observations made for category 1(220 observations) and remaining fall into category 0.

→ This is basic glimpse of the data.

→ Let's explore the given data with graphs(known as heatmaps),Where we can find the correlation between various columns.



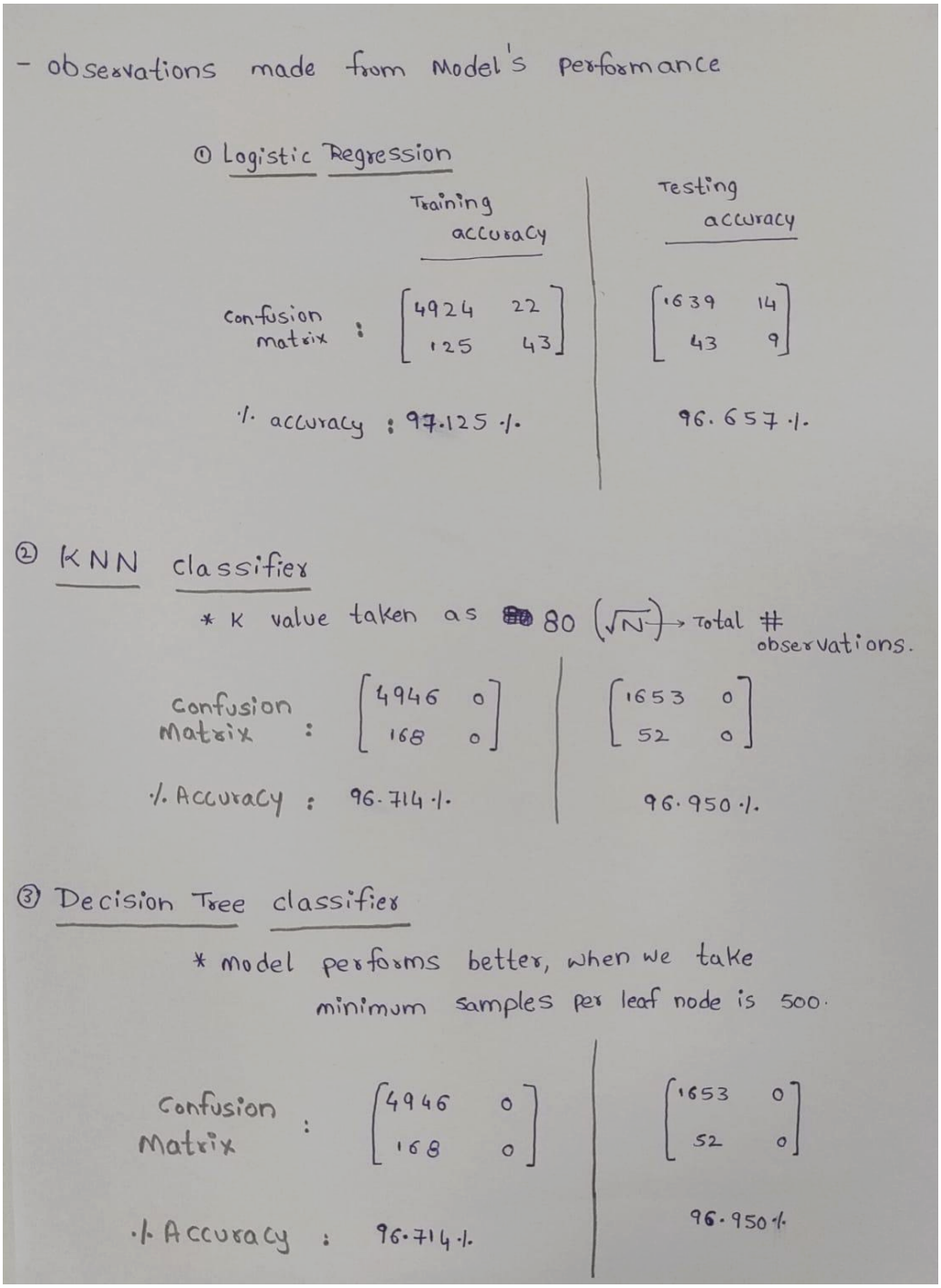
→ Remove the columns those have very high correlation(>=0.9).



→ By putting threshold as 0.9, we were able to remove 18 columns and still we can able to analyse the data most accurately and can reduce the computational cost.

→ Split the Dataset into training(75%) and testing data, and did Feature scaling(if some columns are much larger than others,then we will scale those values).

→ We have completed Basic Exploratory analysis and Data preprocessing .Now,fit the training data into Models and we evaluate model's performance .(by seeing confusion matrix,training accuracy and testing accuracy).



- Here ,in KNN classifier ,I took K-value as 80[sqrt(N)],
- Minimum 500 samples per leaf node is optimal choice for stopping tree to grow.
- Based on training and testing accuracy,
    1. Logistic regression model performs well on training data than testing data.
    2. KNN and Decision tree classifiers perform well on testing data than training data.

AUC and ROC curve :

The area under ROC curve (AUC) results are considered excellent for AUC values between 09-1, good for AUC values between 0.8-0.9, fair for AUC values between 0.7-0.8, poor for AUC values between 0.6-0.7 and failed for AUC values between 0.5-0.6,

AUC Scores:

Logistic regression	0.8916
KNN classifier	0.91642
Decision tree(500/leaf)	0.91390
Decision tree(250/leaf)	0.87721
Decision tree(375/leaf)	0.90936

➢Based on AUC scores:

1. Before it seems like KNN and decision tree have same training and testing accuracy,But from AUC scores we can say that KNN performs better than Decision tree classifier for given data set.

➢Final conclusion is , KNN classifier performs better than other models for given data set.