Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

Team Member's Name	**Email id**
Syed Aquib	syedaquib153@gmail.com

Contribution:

Syed Aquib :

- Data Wrangling
  - Loading and Pre-processing
  - Structuring data
  - Enriching data
- Data Mining
- Data Analysis
- Visualizations
  - Bar graphs and Distribution Graph
- Machine Learning -- Modelling and Predicting using Algorithms
  - Applied different clustering models like **Kmeans, hierarchical, Agglomerative clustering, DBSCAN** on data we got the best cluster arrangements
- Observation
- Summarization
- Conclusions
- Technical Document
- Power Point Presentation

Please paste the GitHub Repo link.

Github Link:-

https://github.com/syedaquib153/Netflix-Movies-and-Tv-Shows

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

Project Name: Credit Card Default Prediction

This project is aimed at predicting the case of customers default payments in Taiwan. From the perspective of risk management, the result of predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification - credible or not credible clients.

Approach:

As a individual I read the data present in the file and gone through the details in each and every column. The data set was huge in which some of the data was not required for the analysis so the data was cleaned by dropping some unwanted columns and created new data frame, with the columns we required for the analysis. The first problem we faced was the name of the columns which was not proper and the nan values present in the data. We renamed the columns by using dictionary format and replaced all the nan values to zero in int dtype and unknown in object dtype by using replace syntax. Each and every column were compared to gain the insights about the data. I worked individually to gain some insights by doing the exploratory data analysis using python. Cleaning the dataset, analysing the data and visualizing the data by plotting the data into different graph and charts so that the trend and relationship between the various indicators can be understand easily, Modelling and Predicting the model using Machine learning algorithms which model is best to predictor .

Conclusion:

Logistic Regression model has the highest recall but the lowest precision, if the business cares recall the most, then this model is the best candidate. If the balance of recall and precision is the most important metric, then Random Forest is the ideal model. Since Random Forest has slightly lower recall but much higher precision than Logistic Regression, I would recommend **Random Forest**.