**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| Shajad [shehzadglocal786@gmail.com](mailto:shehzadglocal786@gmail.com) |
| **Please paste the GitHub Repo link.** |
| Github Link:- <https://github.com/Shajad121/Credit-card-fefault-prediction.git> |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)** |
| This project is about the Credit card default prediction analysis. When we start this project, it is very difficult to understand and look very complex. In first step we upload the data and then we start to find the duplicate values and nulls values and missing values etc. in this dataset we found the some values are missing in Cast and Directors columns so we decide to remove that colums also in release year there are sum values also missing we fill that values with mean and we also fill the country columns null values with mode in. After completion of 1st step in 2nd step we plot the data in form of graph, pie plot and bar plot and line plot and also box plot, and in box plot we find some outlier so we remove the outliers. And we also find some hidden information by plot the graph. In this dataset there are two types of contents 30.86% is Tv shows and 69.14% is movies includes. We have reached a conclusion from our analysis from the content added over years that Netflix is focusing movies and TV shows (From 2016 data we get to know that Movies is increased by 80% and TV shows is increased by 73% compare). From the dataset insights we can conclude that, most number of TV Shows released in 2017 and for Movies it is 2020. On Netflix USA has the largest number of contents. And most of the countries preferred to produce movies more than TV shows. And then India comers on number 2 hos produce the largest movies and the tv show. Most of the movies are belonging to 3 categories. TOP 3 content categories are, International movies, dramas, comedies. there are 35.1% movies and tv shows are from the United State. 12.7% movies and tv shows are from the India. 5.5% tv shows and movies from the United Kingdom.3.1% tv shows and movies are from the Japan. 2.5% movies and shows are from the South Korea. We also devide the data into month year and the day of month and add that columns into the data frame and we also create some Dataframe in our data set. In text analysis (NLP) I used stop words, removed punctuations, stemming & TF-IDF vectorizer and other functions of NLP. Applied different clustering models like K-means, hierarchical, Agglomerative clustering, DBSCAN on data we got the best cluster arrangements. By applying different clustering algorithms to our dataset . we get the optimal number of cluster is equal to 3 .So we decide to take the as no of cluster in this data because it is best no in our conclusion. |