Milestone report Capstone Project

Project Title:

Movie Recommendation System

Project Proposal:

What is the problem to be solved?

Movie Recommendation System will mainly be used to guide the user what kind of movie they like watching depending upon the **other users which are similar to them** and all the **past movies** they have watched.

Who is this project beneficial for?

This project is for those company who is in the field of Recommending Movies to User such as **Amazon Prime, Netflix.** Also, this could be used by anyone who likes to watch movies. so, this project will recommend all those movies to the user which are similar to movies which the user has already watched.

What data are you using? How will you acquire the data?

The Data will be using is from the Kaggle website which is **Tmdb Dataset and The movie Dataset.**

How you'll solve this problem?

1. Data Wrangling: Remove all the Missing values and outliers and make sure that the data is clean.

2. Exploratory Data Analysis: I will plot the relation between different parameters in the dataset and based on that and based on that I can find the trends for storytelling.

3. Inferential Statistics: build a statistical relationship in the dataset.
4. Machine Learning Model: Content-based recommendation Algorithm.

What are your deliverables?

I will use the following tool to work on my project.

1. Jupyter Notebook
2. Python 3.7
3. Google doc for creating the document
4. Powerpoint for data Storytelling presentation

Data Wrangling Report:

1.Merging Two CSV file:

The first step which I did in this project was to merge the two datasets into one dataset on movie id. The first dataset was Movies.csv which were columns such as genres, release_date, revenue, vote count, vote average and the title of the movie.

The second CSV file is credits.csv which has columns such as cast which contains actors and supporting actors names, the crew(Name of director, editor, writer, and composer) and movie_id.

2. Dropping not necessary columns from the dataset.

The second step was dropping the unnecessary columns from the dataset which was not required in the dataset and even if they exist won't add any value to the dataset. So, it was better to remove such columns like,
Homepage, production_company, production_country, spoken language, tagline, some repeated columns such as title and movie_title.

3. Missing values:

After dropping the Columns which were not required in the dataset there were only 1 or 2 missing values in some column() so I drop that row from the dataset.

4. Outlier:

There was an outlier in budget and revenue but as this column will not be taking any part while recommending the movie to the user so it is better not to remove the outlier from the data frame.

5. Converting columns into proper format:

There are certain columns such as genres, cast, crew, keywords inside which information is stored in JSON format which is required for recommending the movie. So to extract hidden information from these columns we have to convert these columns into the proper format.
Like from the crew we wanted director, from cast we want the actor and actress etc..

Now the data is mostly clean and ready for the further process of Exploratory Data Analysis, Data Visualization, Machine Learning Model.

Exploratory Data Analysis And Storytelling:

variables that are significant in explaining the project question?

How the Recommendation will affect different variables into consideration?

There are many variables/ features which are taken into consideration for that is responsible for the Recommendation of a Movie.
These are as follows:

1. Popularity: One of the ways of recommending movie to the user is by popularity. If the movie is popular then we should recommend the movie to the user.
   Ex. Titanic is the most popular movie of all time then we should definitely recommend this movie to the user.
2. Vote Count: More the vote that means that majority of the people have liked the movie so in that case if a movie has some good amount of vote count then we should recommend that movie to the user.
3. Vote Average: Similar to vote count even vote average plays an important role. Good vote average value means good movie so we should recommend that movie to the user.
4. Cast, keyword, genres, director: if the user has seen a movie then depending upon the cast, director, genres of the movie we can recommend the movie to the user which are most similar to the movie user has seen.

Correlation between a pair of variables.

What does the pair of variables suggest when compared with each other?

1.Budget and revenue: Both are positively correlated with each other if one increases the other one will also increase.
2. Popularity and vote count: Both are positively correlated with each other if one increases the other one will also increase.
3. The vote count and revenue: Both are positively correlated with each other if one increases the other one will also increase.