CCE Proficiency course, IISC. 2022

Course name: AI-ML 2022 Project Report

Title: OTT REVENUE PREDICTION MODEL[NETFLIX]

Submitted by

VAGDEVI P

VIBHUTI DEMBI

Introduction:

Over **T**he **T**op platform is a direct-to-consumer video(media) content platform. There are several OTT platforms available around us including Netflix, Hulu, YouTube, Disney Hotstar, Zee5, Amazon Prime Video, Jio Cinemas etc.,

There are different OTT revenue models: SVOD (Subscriber Video On Demand), AVOD (Advertising Video On Demand) and Hybrid. Here in this project, an attempt is made to predict Netflix revenue based on number of subscribers ie SVOD model.

The project predicts revenue of Netflix OTT platform using different models: Linear Regression, Decision Tree, Random Forest and KNN models. At the end, a comparison is made across models based on Mean Squared Error (MSE) parameter.

Dataset:

The dataset required for the project can be downloaded from link: Kaggle-Netflix Revenue and Users. This dataset contains Revenue and User Stats of many OTT platforms from 2011-21 and the current project uses Netflix related information.

For the project, revenue, expenditure for content, profit, number of subscribers data are used.

Preprocessing:

1. **Import required libraries**: The project required libraries like
    a. Pandas to define the data objects
    b. NumPy to perform array computations
    c. Matplotlib for mathematical and graph plotting operations
    d. Seaborn for graph plots
    e. Sci-kit learn for importing regression models
2. **Import dataset and explore data:**

    The dataset required has been imported to a data object using pandas and read using read_csv() function.

All the data available are numerical data. To be more specific, Revenue, Profit, Content Spend features are continuous data while Year and Number of subscribers features are discrete numerical data

3. **Identifying and handling missing values:**

   This is one of the steps of data cleaning. Missing values are observed in the dataset and are replaced/imputed with the median of that feature values.

4. **Extract dependent and independent variable:**

   The relationship between the features and features with revenue(target) is understood using pairplot () and lmplot () functions.

5. **Split the dataset:**

   The datapoints available is spilt into two separate sets: test and training sets. In the project split ratio of 75:25 (training: testing) is used.

Regression Models:

Linear Regression, Decision Tree, Random Forest, and K-Nearest Neighbor regression models are used to predict Netflix revenue.

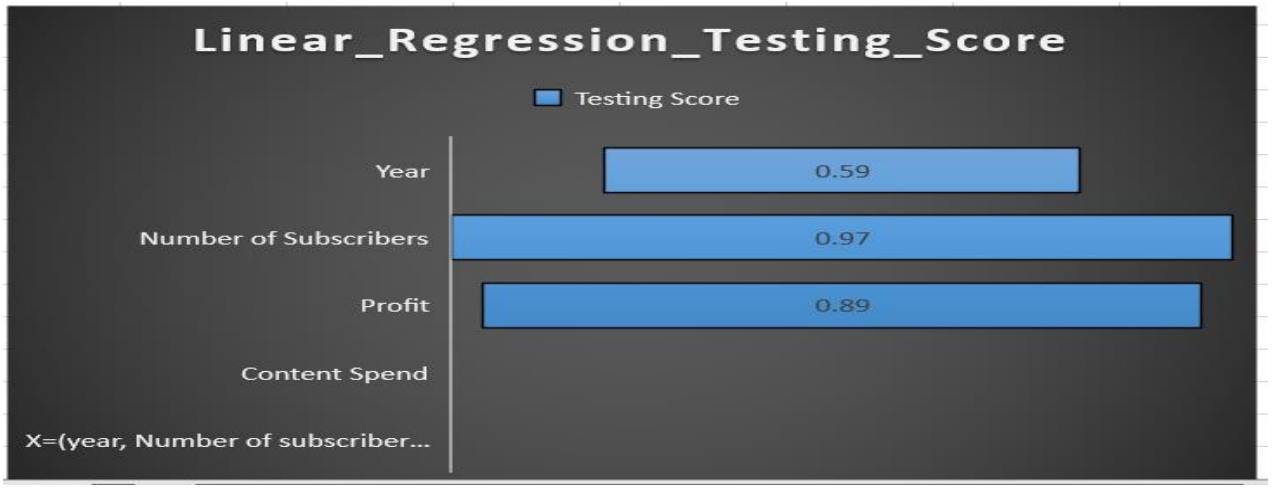
Independent variable X: Subscribers/Year/Content Spend/ in million dollars/Profit in million dollars

Dependent variable Y: Overall revenue generated in million dollars

a. Linear Regression Model:
   - ➔ LinearRegression () is imported from sci-kit library
   - ➔ Feature selection is done by checking the MSE of each feature against revenue. The one with lower MSE is chosen for prediction (Graph 1).
   - ➔ Training and testing scores are computed along with R2 score, MSE(Mean Square Error) and RMSE (Root Mean Square Error) which help evaluate the linear regression model (Table1, Graph 2 and 3)

Graph1: Number of subscribers has lower MSE and thus helps in better predicting Revenue

Graph2: Number of subscribers when chosen as an independent variable shows better training score



Graph3: Number of subscribers when chosen as an independent variable shows better training score

Linear_Regression_Model	MSE	Training Score	Testing Score
Year	20.17	0.35	0.59
Number of Subscribers	2.33	0.83	0.97
Profit	4.83	0.92	0.89
Content Spend	78.15	0.02	-127.81
X=(year, Number of subscribers, Profit, Content Spend)	105.93	0.95	-0.37

Table1: Training and Testing score comparison of features

b. Decision Tree Regression Model

➔ DecisionTreeRegressor () function is imported from sci-kit library

c. Random Forest Regression Model

➔ RandomForestRegressor () function is imported from sci-kit library

d. K- Nearest Neighbors Regression Model

➔ KNeighborsRegressor () function is imported from sci-kit library and chose nearest neighbors' parameter to be 2

For all the above models,

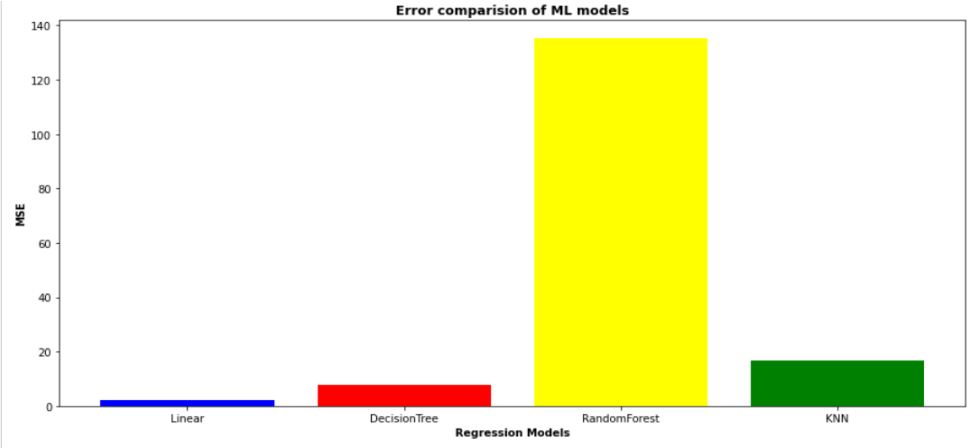
➔ Number of subscribers is taken as an independent variable to predict Revenue (target/dependent variable)

➔ Model is evaluated by computing training and testing scores, R2 score, MSE and RMSE

Table below shows MSE, Training and Testing score comparisons of regression models used in the project. From the data, it is evident that Linear regression model has least MSE and proves to be best regression model for the current dataset used. However, one should also notice Decision Tree regression model has better training and testing scores.

#of_Subscribers	Linear_Regression	Decision_Tree	Random_Forest	KNN_Regressor
MSE	2.33	7.89	135.17	16.71
Training Score	0.83	1.00	0.96	0.87
Testing Score	0.97	0.90	0.83	0.79

Graph below shows that MSE for Linear Regression model is very low when compared to other regressor models and is a better predictor model.



Conclusion:

For a given set of data used in the project, Linear Regression model suits best. However, with further tuning the hyper parameters in other regression models, we might see better predictions and lower errors. Also, using other machine learning techniques like Deep Neural Networks could help make result more accurate.

From the OTT point of view, these prediction models help in making business critical decisions and answer questions like how much and how quickly a company can intend to grow? how likely to plan for investments? how quickly and what are the actions to be taken to improve business if there is a fall in revenue?

With the help of these insights, it becomes easier to identify trends in market and make quicker business decisions.

References:

1. https://www.kaggle.com/datasets/azminetoushikwasi/ott-video-streaming-platforms-revenue-and-users
2. https://selectra.in/blog/how-ott-earn
3. https://scikit-learn.org/
4. https://pandas.pydata.org/
5. https://matplotlib.org/
6. https://numpy.org/
7. https://seaborn.pydata.org/