imdb

June 23, 2018

Cap Stone Proposal Sai Kiran Rebba June 6th,2019.

Predicting the meta_score of a movie based on the information which is provided in the dataset.

Domain Background:

- Now a days everyone is watching movies and give rating or review for that whether it is good or bad.With this information I have an idea to predict the score of the particular movie based on the rating given by the audience and voting.
- I collected data from kaggle.com .Here it is the link:https://www.kaggle.com/nielspace/imdb-data.
- Using this data effectively we can be able to predict the metasore value.
- This problem must be solved because everybody is curious aboout the results of a movie rather than his/her examination.Some fans of that movie hero/heroine declared that the movie was blockbuster hit.So,I want to analyse the results based on the information given the audience not only some fans but all the audiences responses.
- The major motivation behind this project I want to know how Machine Learning works in the real world datasets and how much it can accurately predict the values.
- Here is the article where I get sufficient knowledge about my dataset.Link:https://www.linkedin.com/pulse/analyzing-imdb-movie-dataset-preetish-panda

Problem Statement:

- In this project I want to analyse the metascore of a movie based on the features like movienname,hero,Genre,Description,Director,Actors and ratings etc.
- By using all these features I can able to predict the metascore i.e how much score that movie can get out of 100.
- I decided to implement machine learning algorithms and certain techniques to optimize the performance of the applied models and eventually developing a model that could potentially predict the success rates of the respective movies through the evaluation of the 'MetaScore' value.
- This project consists of Data preprocessing,Data Exploration,Data Visualisation and various machine learning algorithms which are necessary for my project.

Datasets and Inputs:

- The data is selected from kaggle.com.The link to my dataset is :https://www.kaggle.com/nielspace/imdb-data.
- The dataset consists of 12 columns and 1000 rows.
- Features are:Rank,Title,Genre,Description,Director,Actors,Years,Runtime,Rating,Votes.Here the ouptut predicted from these features is 'Metascore'.
- The diverse features can be optimally filtered during the course of the project and the optimal features are then interpreted for evaluating the 'MetaScore' value that gives an indication of the success rate of a specific movie.

Solution Statement:

- Solution to this problem is mainly depends on the features which we are selected because here we want to predict the metascore based on the important features.
- It is supervised meachine learning algorithm in which we use regression for our problem.
- I will eventually works on algorithms like SVM's,decision trees etc and I use Gridsearch algorithm for model optimisation

Benchmark model:

- Since the output is continuous benchmark model I choose for this problem is linear regression.
- The benchmark model is a base model that has no intelligence about the data hence we start with findiing the r2_score which helps to select the goodness of fit for the predictions of a actual values.

Evalution Metrics:

- The current problem is purely a regression problem,since it takes features as inputs and tries to fogure out the metascore which is a continnuous output that will be helpful to detect whether the movie is good or bad based on the metascore value.
- Hence I decided to use 'coefficient of determination' as the performance metric that could be applied to check the performance of the scores obtained from the Bench mark model and the optimal model considered.

Project design:

- Following steps arre needed to design the project:
- Data Acquisition: First step of the project is acquiring the data which I choose from the Kaggle.
- Data preprocessing:It is the important step which decides the result of our entire model.In this we have clean and remove unnecessary data. Here we have to remove movie names and description about that movie because I think those columns are not necessary to predict the output.
- Data Exploration and visualisation:Noralising the data by using min_max_scaler which is neessary to avoid skewness we have to use log_transform. visualisations which tells us about the results very easily by graphs or something any visualisations.

- Model Evalution and validation: After normalising the data We have to evaluate the model by using the r2_score metric.If the score is above 0.7 that model is better otherwise we hae to try another model.First of all Model is tested for the performance of the bench mark model like linear regreassion and after we have to try on otherr models like SVM,Lasso,XGBoost,LightGBM etc .
- Optimisation: The naive model is optimised by the application of the GrisSearchCV which helps in tuning the parameters optimal for the model.