PROPOSAL FOR CAPSTONE PROJECT
PREDICTING BOX OFFICE REVENUE

DOMAIN BACKGROUND :-

The films industry has always seen rapid growth and some movies are making so much of buisness whereas opposite to this some movie did not make much buisness. There are multiple factor on which the success of movie depends. On the basis of some important features we can predict that how much buisness movie will do.Earlier, the films industry used to utilize the knowledge of specific industry trends, the basic rule of thumb approaches and traditional wisdom and intuition to predict the buisness of particular films. This method was never very accurate or reliable.With the emerging technology, Machine learning can predict movie buisness more accurately.

Here are some link of research paper in which some models were proposed on predicitng box-office result

- https://www.researchgate.net/publication/326497327_Improving_Box_Office_Result_Predictions_for_Movies_Using_Consumer-Centric_Models
- https://www.researchgate.net/publication/313455341_Predicting_Movie_Box_Office_Profitability_A_Neural_Network_Approach
- https://pdfs.semanticscholar.org/a799/c0196d6a08c3420528fbd6906392114ec752.pdf

And further more details are given in DATASETS AND INPUTS section of this article

PROBLEM STATEMENT:- Kaggle is currently hosting a competition for predicting the movie's worldwide box office revenue.

DATASETS AND INPUTS:- The dataset is already available on kaggle. They have presented with metadata on over 7,000 past films from The Movie Database to try and predict their overall worldwide box office revenue. Data points provided include cast, crew, plot keywords, budget, posters, release dates, languages, production companies, and countries.And also they are providing with two different dataset in csv format, one for training i.e. train.csv and other for testing i.e. test.csv. More details from kaggle website:-

Details For Training File:- In this dataset, you are provided with 7398 movies and a variety of metadata obtained from The Movie Database (TMDB). Movies are labeled with id. Data points include cast, crew, plot keywords, budget, posters, release dates, languages, production companies, and countries. DATA SHAPE:- `(3000, 23)`

Details For Testing File:- You are predicting the worldwide revenue for 4398 movies in the test file.

DATA SHAPE:- `(4398, 22)`

These are the names of columns given by Kaggle:-

(*id,belongs_to_collection,budget,genres,homepage,imdb_id, original_language, original_title, overview, popularity, poster_path, production_companies, production_countries, release_date, runtime, spoken_languages, status, tagline, title, Keywords, cast, crew, revenue*)

Link to download dataset and to get more details :-

https://www.kaggle.com/c/tmdb-box-office-prediction/data

Above are given datasets format by kaggle but while working on these dataset we need to divide train dataset to train and validaton dataset, to check the performance of the model

SOLUTION STATEMENT:- As we have done in the projects of nanodegree. First we have to check the dependency of different factors with revenue. If some normalisation is required for the values of parameter that will be done.And next Step will be to build a model by choosing appropriate regression model.Along with this random seed also need to set so that anytime we run result does not change. Cross-Validation can also be used for better results.

BENCHMARK MODEL:- For the benchmark model, I have decided to implement the AdaBoostRegressor on the splitted train dataset(the given train dataset will be splitted into two validation and train) and then will test on the splitted validation dataset.Whatever score will be given by this model, that will be considered as benchmark score. After the final solution model, trained and tested on the same splitted dataset as mentioned above, it will be checked that wheather final solution gives better result or not.

EVALUATION METRICS:- On the kaggle,Submissions are evaluated on Root-Mean-Squared-Logarithmic-Error (RMSLE) between the predicted value and the actual revenue. Root-Mean-Squared-Logarithmic-Error (RMSLE) is the one of evaluation metric used in multiple regression problems. It can be used when we don’t want to penalize huge differences when both the values are huge numbers.Also, this can be used when we want to penalize under estimates more than over estimates.

OUTLINE OF THE PROJECT DESIGN :-

1. Visualization of the data(null values,mean,std,etc) and relationship among different attributes of data
2. Dealing with null values(drop rows or fill by 0) depending on relationship among attributes
3. Handling text columns(using LabelEncoder)

4. Selecting a good model from some supervised regression model (LinearRegression, DecisionTreeRegressor, SVM XGBRegressor),the model that will give better score from all will be considered as final solution and further will be compared with BenchMark model.
4. Fine tune the model:- Selecting best parameters value (Cross-Validation,grid search can be used)
5. Evaluating the final score