Dalhousie University

CSCI5408 DATA WAREHOUSING AND MANAGEMENT

Assignment Report

PROJECT PLAN DOCUMENT

Submitted By :
Deeksha Behara :B00784704
Tushar Gupta :B00782699

Supervisor:
Suhaib Qaiser

July 2, 2018

Contents

1 Role of each team member

The work for the project has been divided among the team mates as the following.

1.1 Team Role : Data Engineer

1. Understanding the dataset[1]

   Exploring the attributes,data types, data size, columns, range and usefulness of the individual attributes to solve the problem statement.

2. Cleaning the dataset

   (a) *The cleaning of the dataset includes several steps such as filling missing values, converting the texts to lowercase etc.

   (b) Imputing Null Values : Assigning the null values to valid data in the data frames. In some cases , removing null values to make the data useful for the next processes.

   (c) Removing irrelevant columns: Identifying the columns which don't contribute to the problem statement and removing the irrelevant columns from the dataset.

   (d) Normalizing the dataset if required : After identifying the range of the attributes defined in the dataset, the values will be normalized to a specific range if required.

   (e) Binning: Combining two or more columns of the datasets to individual columns.

   (f) Cluterting : Grouping values in clusters and then detect and remove outliers (automatic or manual)

   (g) Aggregations: Performing regression by fitting the data into the regression functions.

1.2 Team Role: Data Scientist

1. Objective : Stating the objective of the problem, why are we solving it and how we are solving it. Having a clear understanding of how we are gonna be tackling the problem and what results we should get from solving the problems.

2. Importing Data : Reading the data into the IDE, having full accessibility to modify it and play with it.

3. Data Exploration : Learning about the columns, what all information is given in the dataset. What kind of analysis can be done on it , based on the columns it could be either a inference analysis gaining insights into the database from what has happened Or it could be a predictive analysis by learning what has happened what could happen in future given a set of attributes. If there is no target column then unsupervised learning algorithms like Kmeans, Dbscan algorithms would be used to find out clusters of dataset with similar properties/attributes , if there is a target column then depending on its values if its continuous or discrete supervised learning algorithms would be divided into regression and classification respectively. To gain more knowledge on the dataset EDA(Exploratory data analysis) techniques could be performed.

4. Baseline Modelling : Performing some basic machine learning techniques like linear regression , kmeans to get a baseline on which we could improve upon.

5. Evaluation of models : using evaluation metrics like confusion matrix or using ROC-AUC or MSE scores to test our analysis. Based on these results we would tweak our model burther.

6. Secondary Modelling : Performing complex algorithms like ANN, LSTM or Random Forest to increase accuracy and compare with our base models , and see if after performing the same tweaks on the the new model how do they perform against the base model.

7. Conclusion and Reporting : After the results are obtained we would be explaining the whole process and document it in the report.

2 Objective of the project

The motivation of the project is identify the probability of the movie being successful based on attributes given in the dataset given in the dataset.

- **Identifying movie genre, time of release are correlated to the success of the movie and how the trend is changing over time.**
- **Predicting weather a movie would be a hit or not based on give attributes.**
    1. Attributes would be decided with EDA techniques finding which attributes correlate with the target variable most.
- **Visualizing the results using suitable graphs/charts with matplot library/Tableau.**

2.1 Dataset Description:

The dataset contains metadata for all 45,000 movies listed in the Full MovieLens Dataset. The dataset consists of movies released on or before July 2017. Data points include cast, crew, plot keywords, budget, revenue, posters, release dates, languages, production companies, countries, TMDB vote counts and vote averages. This dataset also has files containing 26 million ratings from 270,000 users for all 45,000 movies. Ratings are on a scale of 1-5 and have been obtained from the official GroupLens website. [Kaggle]

3 Programming language, tools, and resources required for the project :

3.1 Validations:

1. Scripting Language : Python
2. Packages involved : pandas, keras , numpy , matplotlib, re, sklearn.
3. IDE : Pycharm,Spyder, Jupyter Notebooks
4. Visualization Tool : Tableau
5. Project Management Tool: Github

4 Work breakdown Structure

The tasks for the project is divided among the team mates in the following way.

1. Setting up the environment for the project
2. Includes installation of tools and configuring them
3. Using Python, java for preprocessing and preparing the the data for the next steps
4. Performing analytics on the data and deriving useful information from the processed data.
5. Visualizing the data using tableau or Matplotlib.

5 Value proposition

The recent works include the prediction of ratings and the box office numbers. In these works, there were several drawbacks for eg., the tools built were very subjective in nature and did not perform any trend analysis. Some projects didn't use many parameters for predictions. It is evident that there are several influencing parameters that defines the success of the movies. The goal of project also includes identifying ,determining and incorporating the other potential parameters that defines the success rate of the movies.

6 Milestones/Sprints:

Sprint	Story	Short Task Description	Timeline	Ownership
1	Setting up the environment	Installing required packages for python.	Week 1	Deeksha
2	Data Preprocessing	Checking for null values, empty strings,	Week 1	Tushar
3	Data Analytics	Performing analytics on the data	Week 2	Tushar
4	Visualization	Using suitable visualization techniques to display the results.	Week 3	Deeksha
5	Testing	Performing Validations and testing the individual modules	Week 4	Deeksha/Tushar

Figure 1: The sprint of the project until the project delivery.

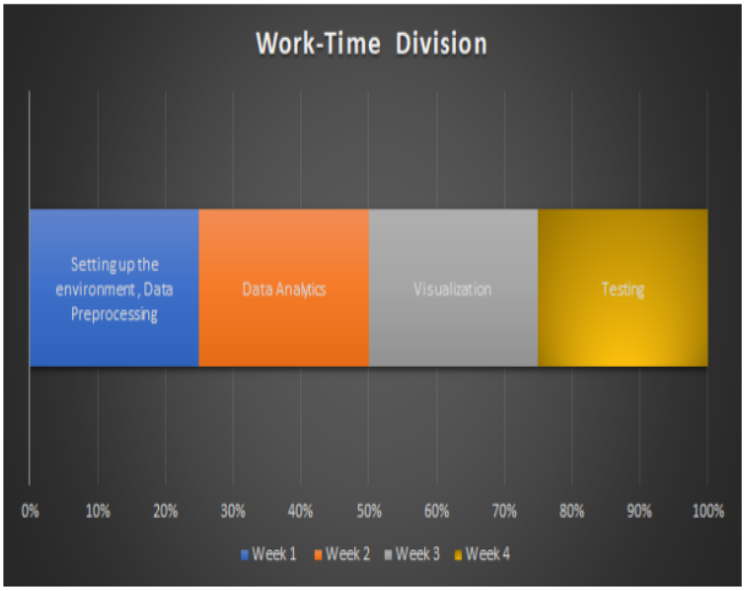

Figure 2: An over view of the time line for the sprints displayed in the table above.

References

[1] François Chollet et al. Kaggle. https://www.kaggle.com/rounakbanik/the-movies-dataset#movies_metadata.csv, 2015.