IRE Second Deliverable - End-to-End System - MVP - Team blank

Team Members:

- Ainesh Sannidhi - 2019101067
- Anirudh Kaushik - 2020111015
- Gagan Agarwal - 2021201009
- Tanishq Chaudhary - 2019114007

Project Number 4: Classification/Regression and Recommendation

Project Description:
Given a set of movie metadata, along with plot summary, predict overall rating of the movie.
Given a set of users rating and a list of movies, predict a user specific rating for the movie.

1. Methodologies and Dataset

This project is broadly separated into 2 main tasks:

1. Predict rating of a movie given the movie's metadata and plot summary.
2. Predicting user-specific rating for a movie given that user's ratings for a limited set of movies.

This project is essentially a Recommender System, where Task 1 is a general recommender system where we do not have any information regarding the user who has provided the search query whereas Task 2 will use Recommender Systems that

allow us to get query results more specific to the user based on the information we already have on them from their previous movie reviews.

For this deliverable, we have created the MVP for the first task.

In order to make predictions based on strictly the user ratings, we initially use demographic filtering to calculate ratings for movies that we do not have ratings already for without taking into consideration the user-specific information. On these values, we trained regression models with respect to the movie-specific values such as runtime, actor, genre, etc. and find a correlation between these values and the ratings. The weights assigned to these parameters have been tweaked after testing as certain values will be much more important than others (such as director possibly being more important than runtime).

After trying multiple datasets, we used the Movies dataset, a subset of the MovieLens dataset, and merged and preprocessed it to work for our use case.

Movies Dataset - https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset

Our Preprocessed Dataset - https://drive.google.com/file/d/1zu97TwzyU2T8OVvQ2VMof-XqHDX_fxPf/view

For demographic filtering, we divided the various movie-specific parameters into numeric and non-numeric values.

1.1 Numeric Parameters

The numeric values we took are revenue, budget, runtime, popularity and release date. For numeric values, we directly trained regression models, specifically linear regression, logistic regression, stochastic gradient descent and linear support vector classification to find correlations between the numeric values and ratings.

Initially, we tried regression with Linear and SGD to find a specific value for rating, however, we later treated the task of rating finding as a classification task and rounded the ratings and predicted a whole number rating. This proved to be much more accurate and gave more realistic results.

1.2 Non-Numeric Parameters

After this, we tried the same task but with the non-numeric values that we had. The values we took here are directors, genres and actors. Based on the average ratings, along with a cutoff in case there weren't enough entries for that specific parameter, that each director, genre and actor had, we took the top 4-7 directors, genres and actors and assigned a weight to each of these. A linear combination of these values was taken to assign a score to each movie and arrive at a score for each movie with the highest score being proportional to the best movie in this recommendation system.

This was combined with the numeric parameters and all features were used and trained with the same regression and classification techniques as mentioned before.

1.3 Plot Summary

A major part of our project is to understand the movie plot summary. It is required for both the tasks of movie rating prediction and user-based movie recommendation.

Currently, the models simply use statistical measures such as TF-IDF, however, that only works on the surface level. It assumes there is no syntactic structure to words.

We go a level beyond and instead look at the semantic level of the plot, by adding on the BERT model. We consider things such as the depth and the content of the story. Thus, using BERT, we will get denser and much more meaningful vector representations for the same text.

2. Findings

R2 score was used to evaluate performance.

2.1 Numeric Parameters

Using regression to predict specific rating values,

Linear Regression - 0.19881318571349116

Stochastic Gradient Descent - 0.19652343917416404

Using classification to predict rating as a whole number,

Linear Regression - 0.4908256880733945

Linear SVC - 0.44954128440366975

2.2 Adding Non-Numeric Parameters

Using regression to predict specific rating values,

Linear Regression - 0.2799244941397596
Stochastic Gradient Descent - 0.27490078190351974

Using classification to predict rating as a whole number,

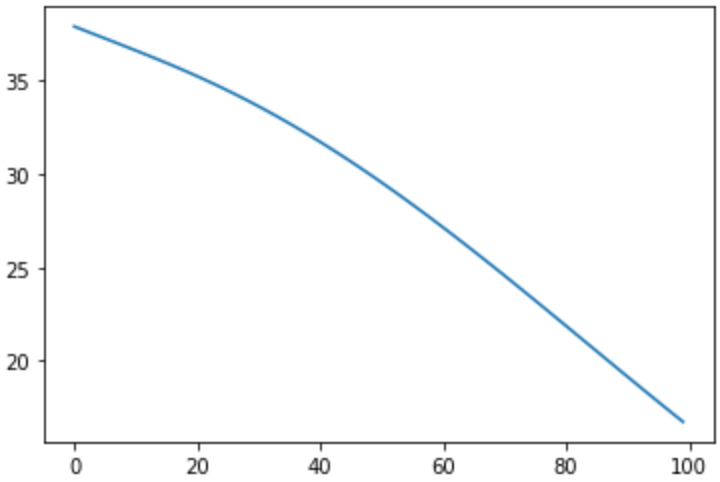
Linear Regression - 0.5367647058823529
Linear SVC - 0.5352941176470588

This accuracy of around 50% is ideal for our system and other tasks such as this one that predict human behaviour, R-squared values of around 50% are typical. We should not have an accuracy much higher than 50% for tasks that predict human behaviour as human behaviour is non-deterministic and a model that is able to accurately predict non-deterministic behaviour would make it deterministic.

2.3 Plot Summary

Using BERT to get a feature embedding did not give good results, even while using a deeper non-linear neural network. The scikit-learn LinearRegression was tried and it gave disappointing results.

Currently, using sBERT embeddings, 3 layers, our R2 score is -9.751597978418918



Losses plot shows that the model is training but is useless.

The plot summary does not give any insights into why a movie is good or not. The plot summary is purely just a description and does not provide much specific semantic information related to how good the movie actually was. It is instead just a representation of whichever writer or critic that wrote up that summary.

3. Code Links

Movie Rating Prediction Based on Purely Numerical Features - https://colab.research.google.com/drive/1lJAd2em5L0bZ_MyMU2EG3UnLcY-sC-Sn?usp=sharing

Movie Rating Prediction Based on Non-Numeric Values - https://colab.research.google.com/drive/1-KMAmDJcEo-dxtuOZHa-z7Y5gLDvdbsM?usp=sharing

BERT Model for Movie Rating Prediction Based on Movie Plot - https://colab.research.google.com/drive/1J6-FpBioeG7KgUzKkaRacDWFvm5DCMdi?usp=sharing

4. Updated Timeline and Methodology

We are currently ahead of schedule as we have completed the initial target for the second deliverable, i.e. task 1, and also tested the improvements (BERT and other NNs).

We will also discuss with the mentor about further improvements, specifically regarding the BERT model.

27th October: Implement Task 2 baseline: demographic filtering

2nd November: Implement Task 2: collaborative filtering

8th November: Implement Task 2: content-based filtering

15th November: Complete task 2, fine-tuning model hyperparameters.