Movie Recommendation System - Capstone Project Report

Rashmy Patwari

23 February 2021

Contents

1 Project Summary

The goal of this project is to create a movie recommendation system similar to the ones used by NETFLIX. A smaller version of the MovieLens dataset is used with 10 million ratings. The dataset is divided into 2 sets, **EDX**, for training and **Validation** for evaluation.

Both the datasets have the following features.

- **userId** `<integer>` that contains the unique identification number for each user.
- **movieId** `<numeric>` that contains the unique identification number for each movie.
- **rating** `<numeric>` that contains the rating of one movie by one user. Ratings are made on a 5-Star scale with half-star increments.
- **timestamp** `<integer>` that contains the timestamp for one specific rating provided by one user.
- **title** `<character>` that contains the title of each movie including the year of the release.
- **genres** `<character>` that contains a list of pipe-separated of genre of each movie. There are about 20 Genres.

The objective of the project is to choose a recommendation model based on RMSE lower than **0.87750**

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$$

2 Initial Data Exploration

edx dataset

The `edx` dataset contains approximately 9 Millions of rows with 70.000 different users and 11.000 movies with rating score between 0.5 and 5. There is no missing values (0 or NA).

```
##   Users Movies
## 1 69878  10677
```

Missing Values

```
##    userId   movieId    rating timestamp     title    genres
##         0         0         0         0         0         0
```

First 6 Rows of edx dataset

```
##    userId movieId rating timestamp                          title
## 1:      1     122      5 838985046                Boomerang (1992)
## 2:      1     185      5 838983525                 Net, The (1995)
## 3:      1     292      5 838983421                 Outbreak (1995)
## 4:      1     316      5 838983392                 Stargate (1994)
## 5:      1     329      5 838983392 Star Trek: Generations (1994)
## 6:      1     355      5 838984474       Flintstones, The (1994)
##                            genres
## 1:                Comedy|Romance
## 2:          Action|Crime|Thriller
## 3:  Action|Drama|Sci-Fi|Thriller
## 4:       Action|Adventure|Sci-Fi
## 5: Action|Adventure|Drama|Sci-Fi
## 6:       Children|Comedy|Fantasy
```

3 Dataset Pre-Processing and Feature Engineering

Initial data exploration reveals that the Genres are Pipe (|) separated values. For estimation precision, it is required to separate them.

4 Build, Evaluate and Analyze Models

4.1 Model 1 - Naive Mean-Baseline model

The formula for computing this is

$$Y_{u,i} = \hat{\mu} + \varepsilon_{u,i}$$

With $\hat{\mu}$ is the mean and $\varepsilon_{i,u}$ is the independent errors sampled from the same distribution centered at 0.

The RMSE on the **Validation** data is `1.0525579`. This is way off from the target RMSE of < 0.87. This clearly states that the model is not optimal.

4.2 Model 2 - Considering the bias that, some movies are rated higher than others.

The formula used is:

$$Y_{u,i} = \hat{\mu} + b_i + \epsilon_{u,i}$$

With $\hat{\mu}$ is the mean and $\varepsilon_{i,u}$ is the independent errors sampled from the same distribution centered at 0. The b_i is a measure for the popularity of movie i, i.e. the bias of movie i.

The RMSE on the `validation` dataset is **0.9410700**. It better than the Naive Mean-Baseline Model, but it is also very far from the target RMSE (below 0.87) and that indicates poor performance for the model.

4.3 Model 3 - Considering the User Effects.

The second Non-Naive Model consider that the users have different tastes and rate differently.

The formula used is:

$$Y_{u,i} = \hat{\mu} + b_i + b_u + \epsilon_{u,i}$$

With $\hat{\mu}$ is the mean and $\varepsilon_{i,u}$ is the independent errors sampled from the same distribution centered at 0. The b_i is a measure for the popularity of movie i, i.e. the bias of movie i. The b_u is a measure for the mildness of user u, i.e. the bias of user u.

The RMSE on the `validation` dataset is **0.8633660** and this is very good. We need to explore further with the Genres effect.

4.4 Model 4 - Check the Genre effects

The formula used is:

$$Y_{u,i} = \hat{\mu} + b_i + b_u + b_{u,g} + \epsilon_{u,i}$$

With $\hat{\mu}$ is the mean and $\varepsilon_{i,u}$ is the independent errors sampled from the same distribution centered at 0. The b_i is a measure for the popularity of movie i, i.e. the bias of movie i. The b_u is a measure for the mildness of user u, i.e. the bias of user u. The $b_{u,g}$ is a measure for how much a user u likes the genre g.

The RMSE on the `validation` dataset is **0.8632723** and this meets our target. Adding Genre did not significantly change much from the Movie+User model. Regularization can improve the performance just a little.

The regularization method allows us to add a penalty λ (lambda) to penalizes movies with large estimates from a small sample size. In order to optimize b_i, it necessary to use this equation:

$$\frac{1}{N}\sum_{u,i}(y_{u,i} - \mu - b_i)^2 + \lambda \sum_i b_i^2$$

reduced to this equation:

$$\hat{b_i}(\lambda) = \frac{1}{\lambda + n_i}\sum_{u=1}^{n_i}(Y_{u,i} - \hat{\mu})$$

4.5 Model 5 - Regularized Movie based model. Regularized to eliminate noisy estimates

The RMSE on the `validation` dataset is **0.9410381** and it looks a definite improvement over just the Movie Based Model.

4.6 Model 6 - Regularized movie + user based model

The RMSE on the `validation` dataset is **0.8627554** and it looks a definite improvement over just the Movie+User Based Model. We will try to improve it by including the Genre.

4.7 Model 7 - Regularized Movie,User and Genre.

The RMSE on the `validation` dataset is **0.8627554** and this is the best of the models. The Regularized Movie+User+Genre Based Model improves just a little the result over the Non-Regularized Model, not a significant improvement.

5 Results

```
##                                    model      RMSE
## 1                Naive Mean-Baseline Model 1.0525579
## 2                       Movie-Based Model 0.9410700
## 3                        User-Based Model 0.8633660
## 4             Movie+User+Genre Based Model 0.8632723
```

```
## 5            Regularized Movie-Based Model 0.9410381
## 6       Regularized Movie+User Based Model 0.8628015
## 7 Regularized Movie+User+Genre Based Model 0.8627121
```

6 Conclusion

Analyzing the RMSEs of the above models, it is evident that **Movie Id** and **User Id** are better contributors than **Genre**. Having said that, the models indicate over training. Regularization helps in reducing the effect of variables and get the best meeting our target goal of $<$ than **0.87**.