Data Science Capstone Project: Choose Your Own Project

Catherine Edis

February 2022

Contents

---

1 Introduction

1.1 Purpose

This project was undertaken to fulfill the requirements of the "Data Science: Capstone" course, which is part of the Professional Certificate in Data Science program offered by Harvard University through edX. The purpose of the project was to ... using the tools explained throughout the courses in the Professional Certificate program.

1.2 Dataset

Describe the dataset:

- what is it
- who published it
- when was it published and/or last updated

The dataset is available at: https://...

1.3 Goal of the Project

The aim of this project was to train a machine learning algorithm that used the data in ... to ...

The predictions generated by the algorithm were compared to the true ratings in the *validation* set using root mean squared error (RMSE). The goal was to achieve an RMSE of less than . . .

1.4 Key Steps Performed

The following key steps were performed.

1. Load the dataset, and create the training and validation sets?
2. Examine the structure of the datasets to understand the nature of the data and identify potential predictors that could be used in the modelling.
3. Where required, reformat the data to make it easier to work with.
4. Perform some basic checks on the quality of the data (for example, check for any missing values, view the range of ratings given to ensure they were in the range stated in the documentation).
5. Set up the objects required for modelling (including a training and test set generated from the original dataset, and a function for calculating the RMSE).
6. Use the training set to develop and train various prediction models, and calculate their RMSE against the test set.
7. Compare the outcome of the models developed and select the one with the lowest RMSE.
8. Run the selected model against the entire *edx* set and use it to predict ratings for the *validation* set.
9. Calculate the RMSE of the predicted ratings compared to the true ratings in the *validation* set, and determine if it met the target of less than . . .

These steps and the results that were produced are described in more detail in the following sections.

2 Analysis

2.1 Preparation

First, the code provided by *edx* was used to load the MovieLens dataset, and create the *edx* and *validation* sets. (For brevity, the code is not displayed in this report.)

Any additional libraries required for the data analysis were then installed and loaded.

```
# Install any packages required.
if(!require(scales))
  install.packages("scales", repos = "http://cran.us.r-project.org")
if(!require(lubridate))
  install.packages("lubridate", repos = "http://cran.us.r-project.org")

library(scales)
library(lubridate)
```

2.2 Examination of the Data

Both the datasets were then examined to understand their structure, including the:

- class of the set;
- number of observations (rows);
- number of variables (columns); and,
- name and class of each variable.

The first few rows in each set were examined to view some examples of content and the format of text fields.

The examination of the sets confirmed that:

1. Both sets had the same 6 columns.
2. The *edx* set contained 9,000,055 rows and the *validation* set contained 999,999 rows.
3. Each row appears to contain one rating given by one user for one movie.
4. For users, no further information is provided other than a unique identifying number ("userId"). This means that the set contains no user demographic data that could be incorporated into our analysis or modelling.
5. Each row includes the rating given, and the date and time the rating was given ("timestamp"). The timestamp is in a format that is not easily readable by a human.
6. Each row also includes:
    - a unique identifying number for the relevant movie ("movieId");
    - the title of the movie; and,
    - any genres the movie is associated with (eg, "Action", "Comedy").
7. The "title" variable includes the name of the movie followed by the year of release of the movie in parentheses (for example, "Boomerang (1992)").
8. Each movie has one or more "genres", with multiple genres separated by "|".

The sets were then checked for any missing values.

No missing values were found.

The *edx* set was explored further, and the number of unique movies and unique users calculated.

Another data quality check was performed, where the number of movie ids in the *edx* set was compared with the number of distinct combinations of movie id and title. If these two counts were different, this would have meant that one movie id had multiple titles recorded against it in the set, and some data cleaning would therefore be required.

The two counts were found to be the same, so no such data cleaning was required.

Next, the values and distribution of the ratings in the *edx* set was examined.

Finally, it was noted that, as well as the variables in the other columns, the timestamp and year of release could potentially be used in the modelling. To make this data easier to work with, both of the sets were reformatted to include these variables in individual columns, and in a more easily readable format.

2.3 Modelling Approach

The modelling approach adopted for this project was based on linear regression. Six models were trained using a training subset of *edx*. For each model, the accuracy of its predictions was measured using RMSE against a test subset of *edx* that contained known ("true") ratings.

The modelling started with three models used in the Machine Learning course:

- a "naive" model, which simply used the average rating of all movies to predict ratings;
- ...

In preparation for the modelling:

- a training and test set was created from the *edx* set. They comprised approximately 90% and 10% of the *edx* set respectively; and,
- a function was defined to calculate the RMSE.

The results produced for each model are included below.

3 Results

3.1 Naive Model - Average Rating of All Movies

The first model simply used the average rating of all movies to predict ratings.

As displayed above, when this model was run against the *edx* test set, its predictions were not very accurate - with an RMSE of 1.060054. The effect of other variables was then explored to determine if they improved the accuracy.

3.2 Predictions for Ratings in the Validation Set

The final model - the genres model - was run against the entire *edx* set, and then used to predict ratings for the *validation* set.

The RMSE calculated when the final model was used to predict ratings in the *validation* set was slightly less than 0.86490. Thus, the goal was achieved.

4 Conclusion

This project explored the use of linear regression to predict movie ratings in the MovieLens 10M dataset, with each variable available in the dataset incorporated into the modelling at some point.

Incorporation of each of the movie and user effects into the prediction model produced reasonable improvements on the RMSE against the *edx* test set. In contrast, incorporation of each of the time, year of release, and genres effects produced only slight improvements. It was sufficient however, to achieve the project goal of an RMSE of less than 0.86490 against the *validation* set.

There were some limitations on the analysis that could be performed, including:

- the inability to explore the effect of individual genres (as opposed to combinations of them) due to computing resource constraints; and,
- the inability to explore the effect of user characteristics on ratings, due to the lack of any user demographic data in the MovieLens 10M dataset.

While the goal of the project was achieved, avenues that could be explored in the future to further improve the accuracy of the predictions include the use of other techniques, such as:

- regularization;
- matrix factorization; and,
- other, non-linear, modelling techniques.