Best Stream

Movie Recommendation Engine

By Hendrik Schmidt

Our Agenda for Today

Business Objective

Data Insights

Model Selection

Evaluation

Consideration of
Ethical Issus

Business Objective

Recommendation Engine

Built a recommendation engine to select movies which the users will enjoy viewing

Assessment

Data Preparation

Team of three data scientists will undertake the project using statistical computing.

Prediction

Applying a number of machine learning methods to user data, in combination with external movie data, to predict the rating that a user will give a particular movie.

The **recommendation engine** is going to work based on the predicted movie rating for each user.

Target it to predict the ratings as precise as possible.

Data preparation

✓Data sets

- **Demographic variables** of the users and movies. **(age – gender – movie title – movie release date)**
- **Additional information** about each of the movies from IMDB. **(IMDB ratings – votes – average ratings by gender and age group)**

✓Variable removal and missing Values

- Removed Video Release Date (100% missing) and IMDB URL.
- All observations that had with missing values were changed to zero.

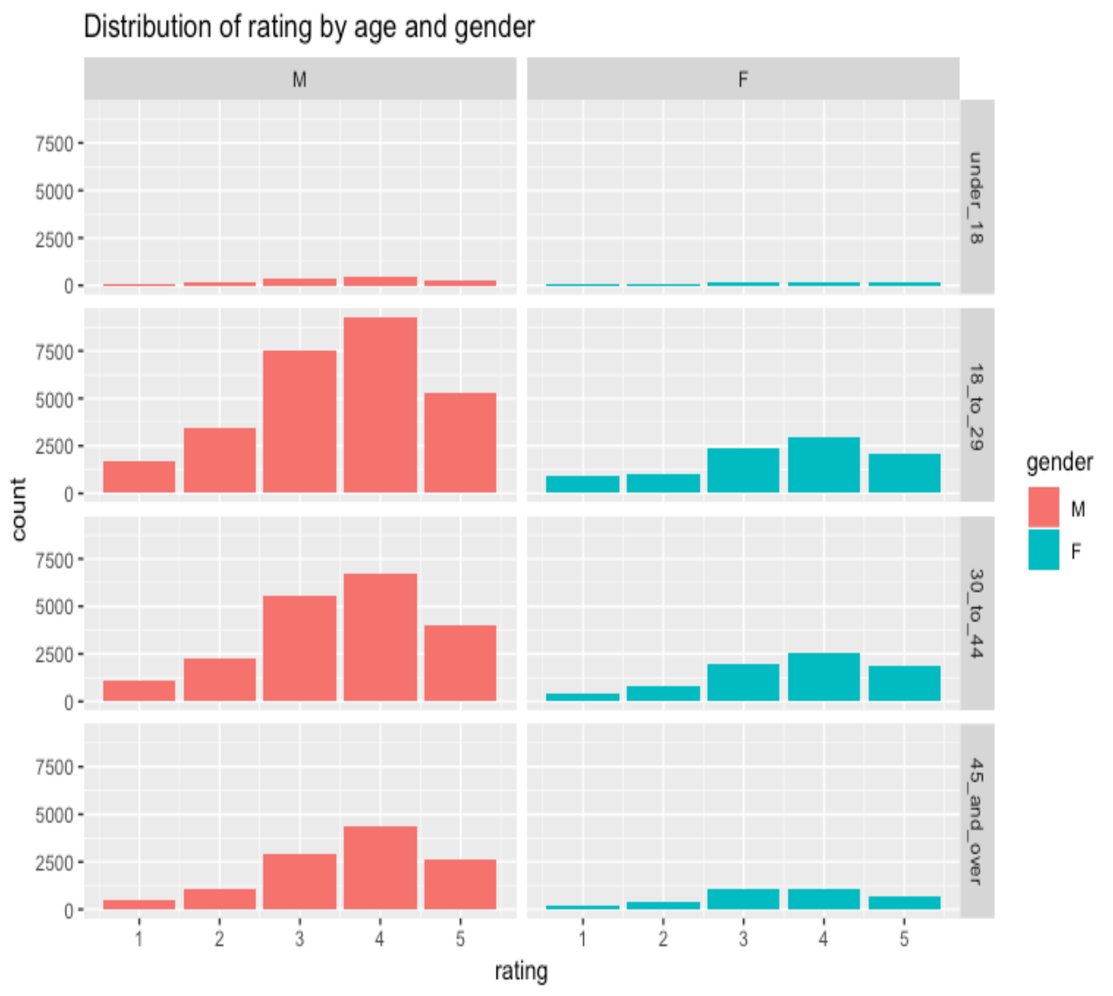
✓Creation of new features for better prediction

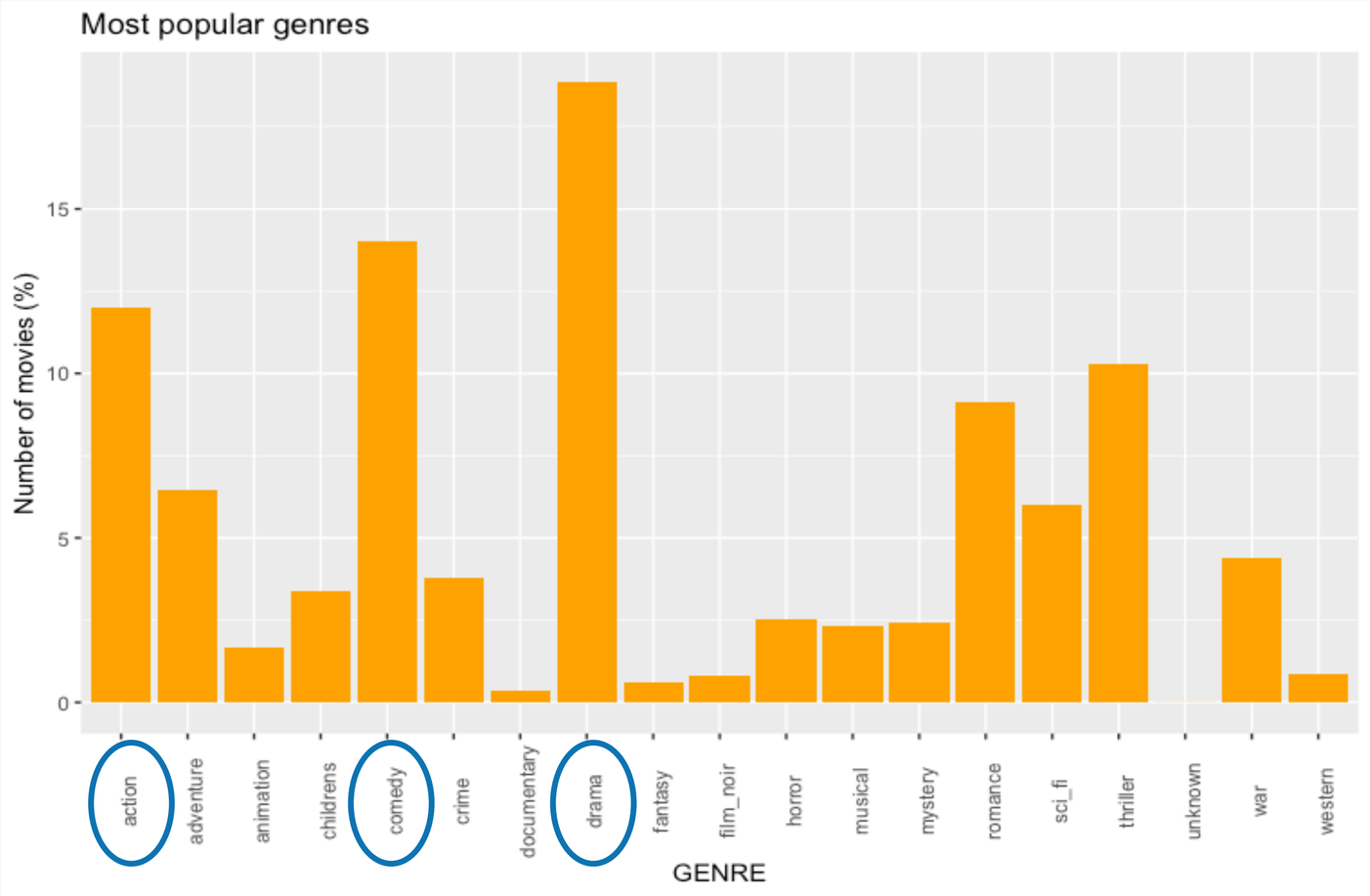
- Time between the movie rating and the release date
- Weight (IMDB rating out of 10 with respect to how many votes it received),
- Demographic Variable (item mean rating with respect to Age Band, Gender and Occupation of the user

Data Insights

1. **943 unique users** reviewed **1682 movies**
2. **Males** make up approximately **75%** of the **users**
3. **Drama**, **Comedy** and **Action** are **top 3** most **popular genres**. – **See plot**
4. Approximately **3% of the data** set observations were **missing**
5. **18-29** and **30-44 age bands** comprise the **majority of the data set**. – **See plot**
6. **Top three most watched movies**: **Star Wars** – **Contact** – **Fargo**

Data Insights – Plots





Model Selection and Assumptions

Complexity

Three different statistical models were used to predict the movie rating.

30%

Linear Regression

- Fast computation.
- Assumes linear relationship between variables and rating.
- Highly biased model towards linear relationships.

50%

Random Forest

- Slow computation time.
- Does not assume a linear relationship between variables (tree based).
- Ranks variables by their importance in the model.
- It can be biased towards the training data and then perform badly on different data set (validation).

70%

Gradient Boosting Method (GBM)

- Very slow computation time – requirement for AWS Server.
- Does not assume a linear relationship between the variables (tree based).
- Ranks variables by their importance in the model
- Has parameter to prevent bias towards the data on which the method is trained on.

Model Selection and Evaluation

- **Model's performance** is evaluated through a **Kaggle competition.**
- Used **Measure** for models **performance** is known as the **RMSE** (**R**oot-**M**ean-**S**quare **E**rror).
- The **lower** the RMSE – the **better** the **model** – the **smaller** the **error** for the predicted rating.

Model	Measure	Rank
Linear Regression	0.9978	2
Random Forest	1.1840	3
Gradient Boosting	0.9433	1

The predicted movie ratings for the recommendation engine, is going to be based on the Gradient Boosting Model – lowest Model Measure score.

Evaluation of the Methodology

- **The business objective** – build recommendation engine that recommends movies to users that they will enjoy.
- **GBM** is the **foundation** on which the recommendation system is build on
- Model to **predict** the **rating** a user would give to a movie is very **powerful**.
- Leads to potential **growth** in **revenue**.
- **GBM** – **Best performing model**. Produces **the lowest deviation from "true" rating**.
- **GBM** is tuned via **multiple parameters**.
- Further review into an optimal review of these parameters will be conducted.

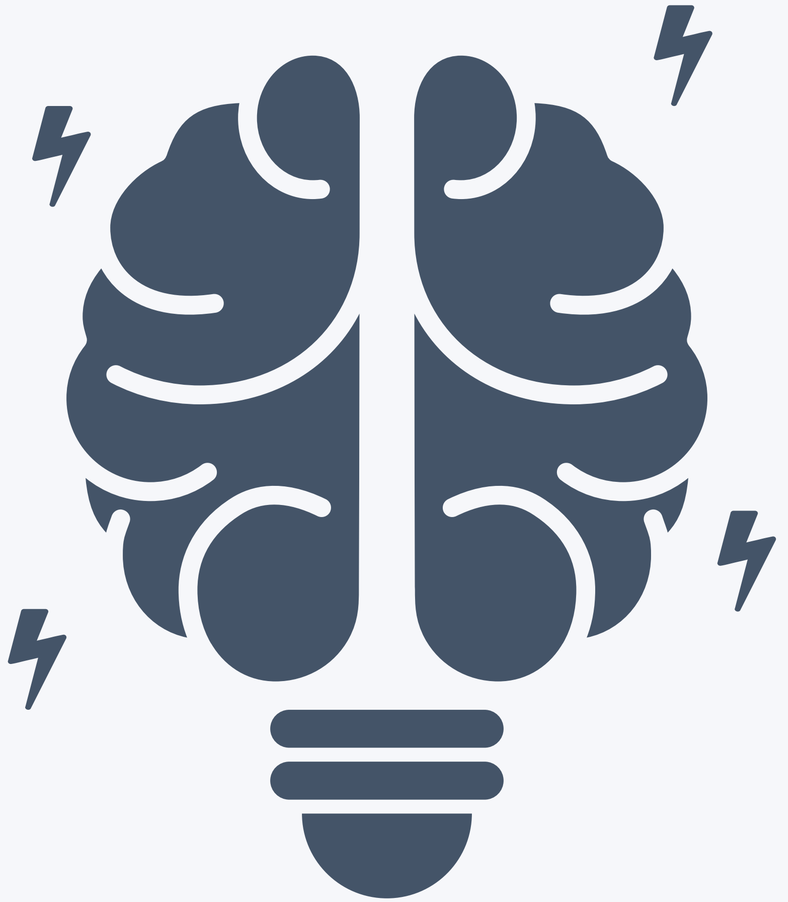
Consideration of Ethical Issus

The primary ethical issues:

- The **collection of demographic** data on each individual – Users may not have explicitly given consent.
- Aggregating data sets on the IMDB data, can potentially create **additional identifying features of users**.
- cultural, religious or political views can unintentionally be reaffirmed, influencing a user in a negative way.

Recommendations to mitigate identified ethical concerns

- making the **recommender service "opt-in"** – users have a freedom of choice
- "Reset" of the database to help ensure any user biases are not continually ongoing.



Thank you for your attention and have a nice day!