Amazon Home & Kitchen Recommendation System

Lanxi Liu

URL: https://github.com/LanxyL/Capstone3---Recommendation-System

Why?

- Surge in Demand for Home Improvement in US:
  - Highest levels of home improvement spending in the history of US
  - Real estate market is booming because of low interest rate
- Recommendation System:
  - Prompts more sales
  - Improve customer shopping experience

Who might be the potential user?

Amazon User

- I am looking for some decorating ideas for my new house.
- How is the item I am browsing compared to others?
- What is the most popular / highest rated item in this category?

Marketing Team

- How to attract more user to buy products?
- How to increase the exposure of items to users?
- How to increase the amount of sales per order?

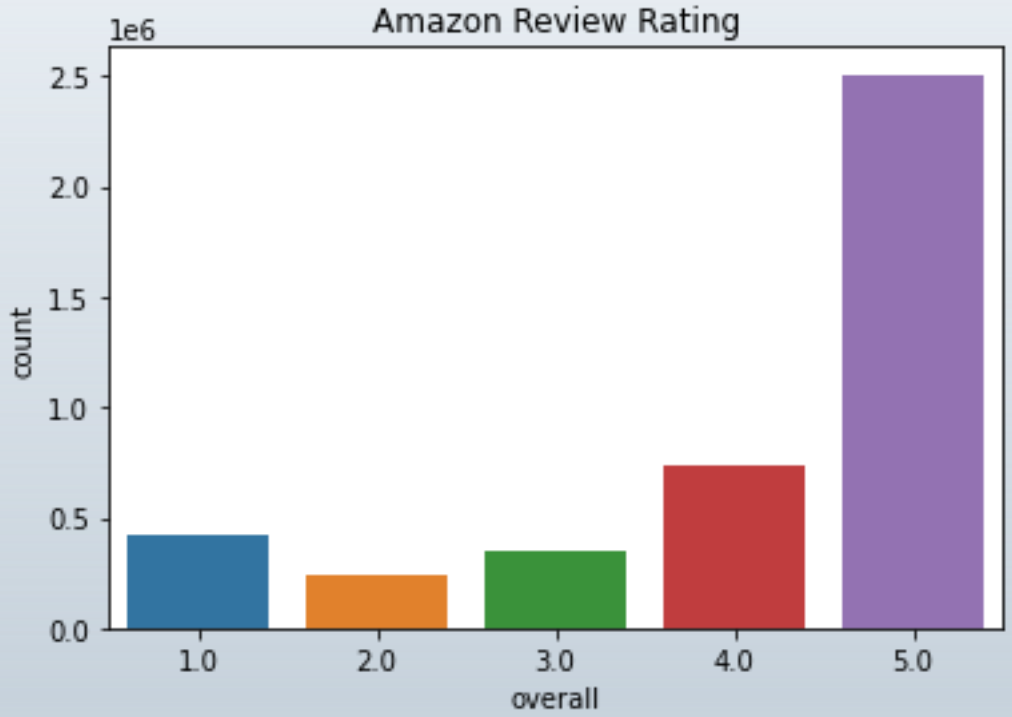
About the Data

- Data source: Amazon product data from lab lead by Professor Julian McAuley at UCSD.
- Two datasets: Reviews Data & Meta Data
- Reviews:
  - 4,253,926 entries
  - Info about review contents, summary, rating and review time, etc.
- Meta:
  - 436,988 entries
  - Info about sales rank, categories, item name, item description, product relationship, brand and price

Data Wrangling

- Source files in JSON format
- Many missing data on product information
- Key information: Reviewer ID, Item ID and ratings do not have missing data
- Dropping all redundant columns
- Convert time-dependent objects to datetime object
- Merge reviews and meta datasets

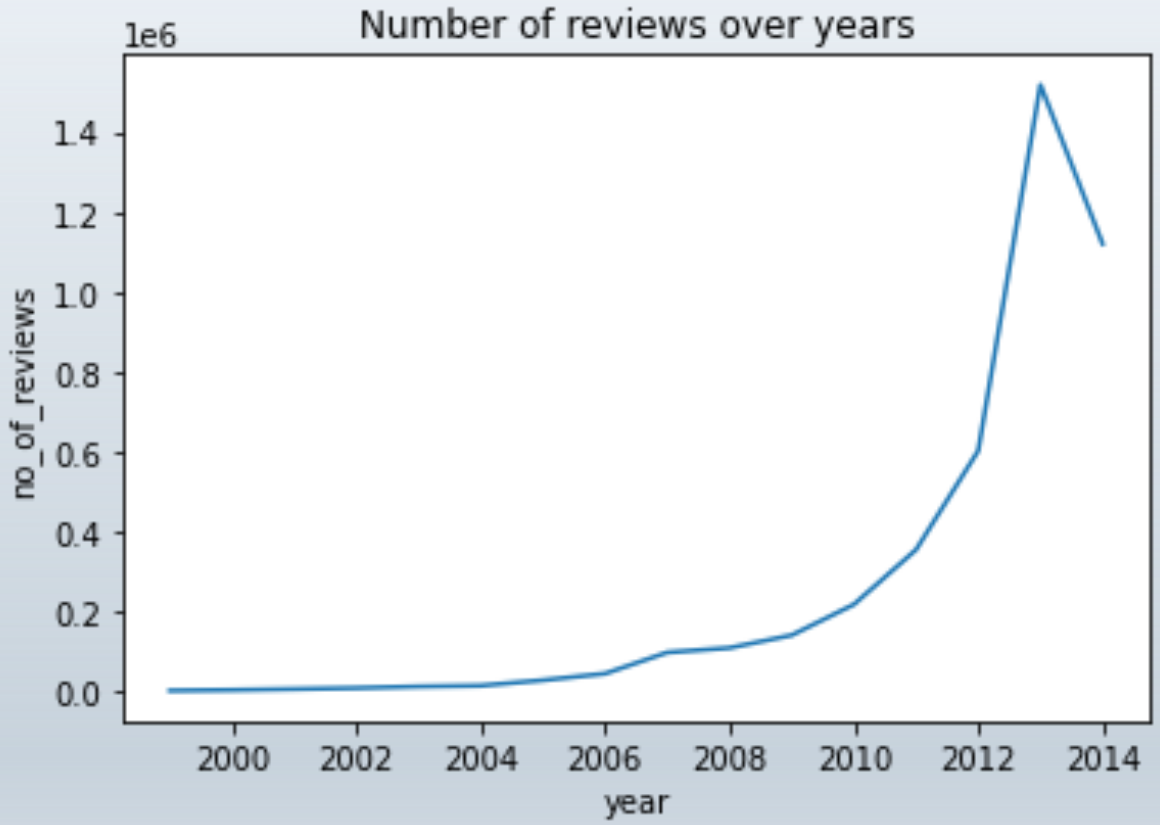
Exploratory Data Analysis



- Highly skewed
- Mean rating: 4.1
- Standard Deviation: 1.33
- Quantiles:
- 25%: 4 50%: 5 75%: 5

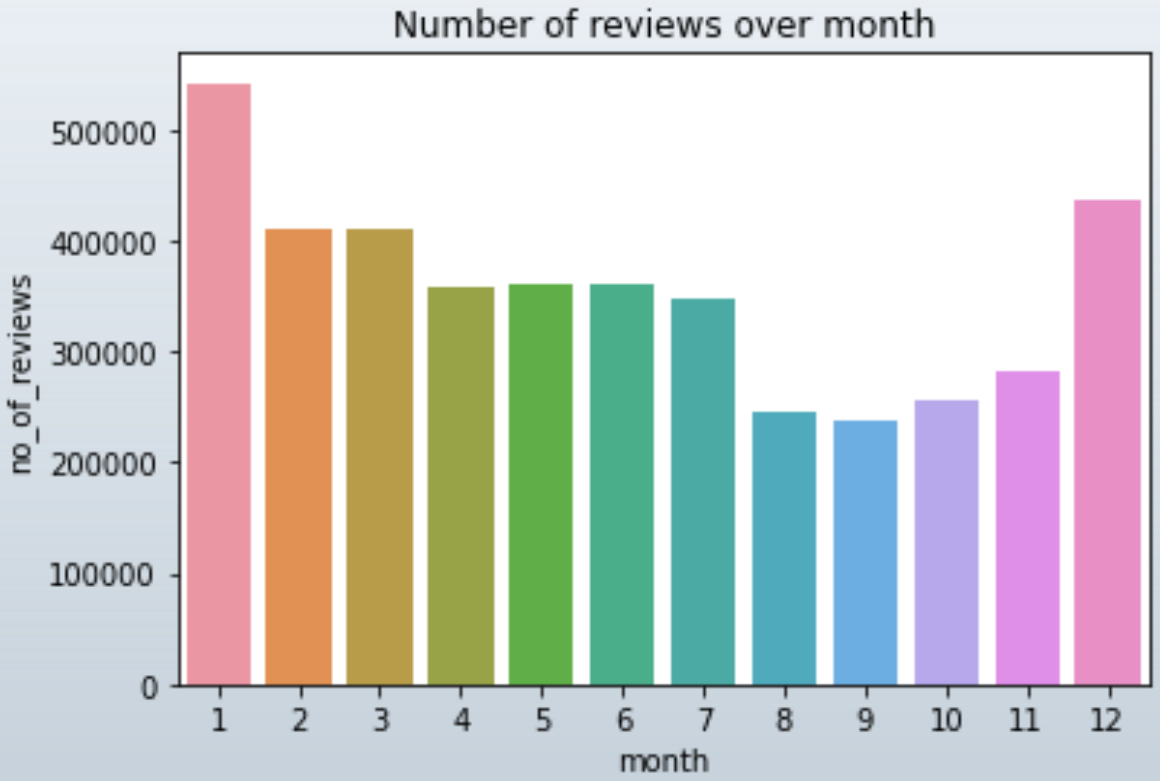
Number of Reviews over Years

- Exponentially increase over time
- Steep drop between 2013 to 2014
- Overall trend: More and more users using Amazon



Number of Reviews Over Month

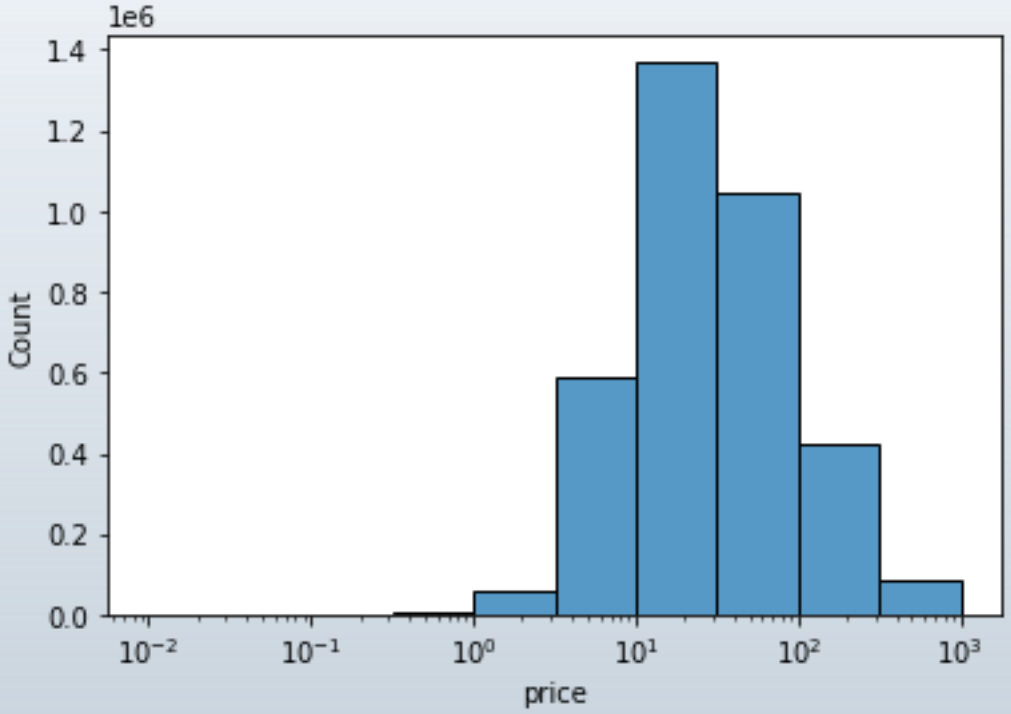
- Highest: January & December
- Lowest: August to November
- Possible reason: Holiday Season surges the sales



Product Price Distribution

- Mean: $5.72
- STD: $8.74
- Quantiles:
- 25%: $12.99
- 50%: $26.20
- 75%: $59.99

If we ignore those very cheap items, the price distribution is almost a normal distribution. Most of items that got reviews range between 13 to 60 dollars.

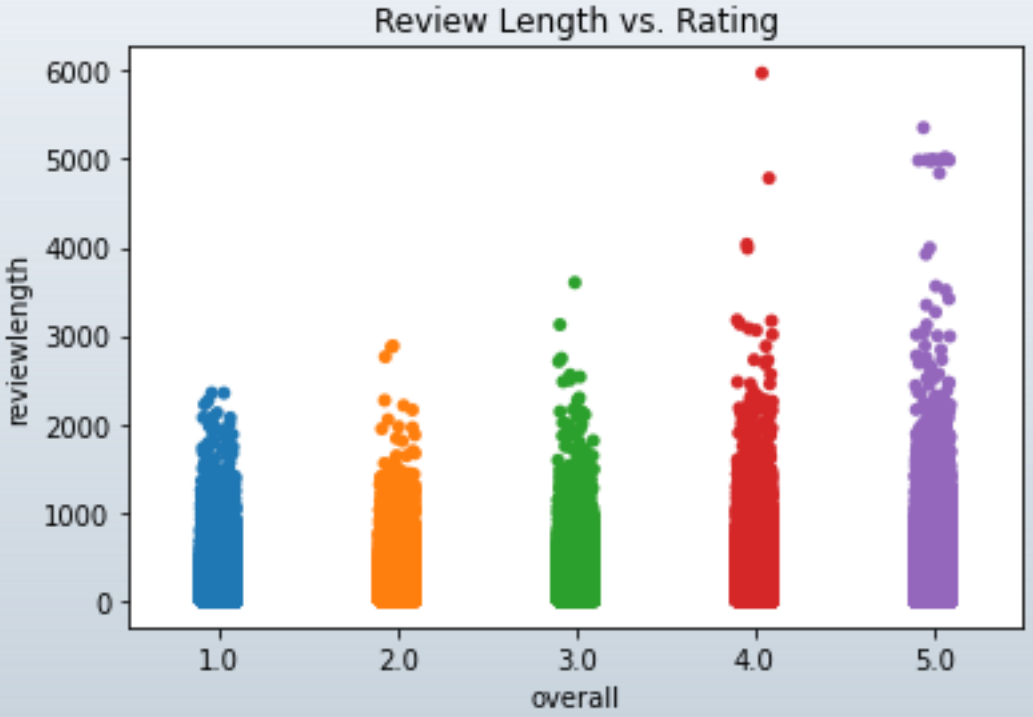


Price vs. Average Rating

- Items with higher price turn to get higher rating.
- Cheap items has wider range of ratings

Review Length vs. Average Rating

- Longer the review, higher chance for a high rating



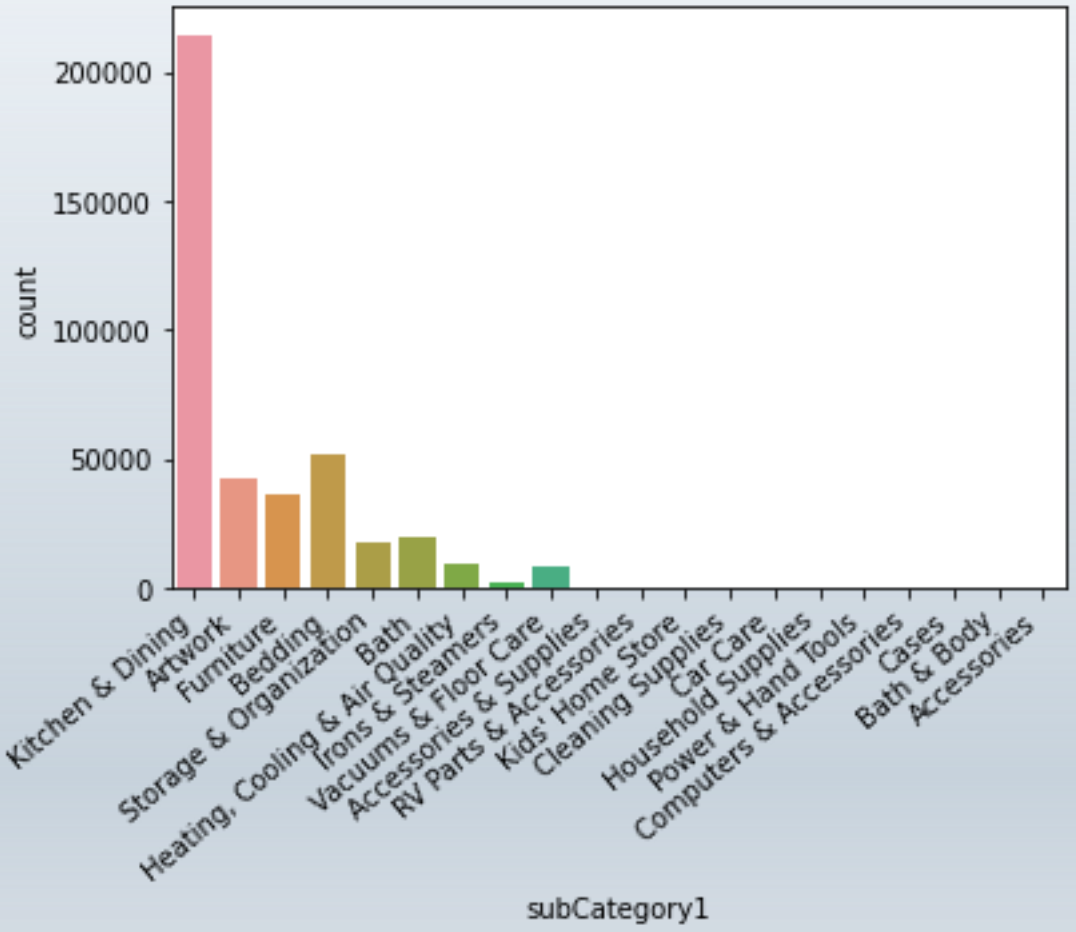
Rating over Years

- Highest of average rating on 2000
- 1999, 2004 and 2005 have lower median rating

Category Reviews Count

- Kitchen & Dining category has most reviews
- Total 2,575,376 entries
- Use Kitchen & Dining subcategory for Modeling purpose
- (Only include items with >10 reviews & users gave>10 reviews)
- Final dataset size:

  81,948



Modeling

- Type: Supervised Learning
- Data Separation: 80% - Train Set; 20% - Test Set
- Feature Engineering:
  - Average rating of user
  - Average rating of item
  - Number of users rate the item
  - Number of items user rated
  - Etc.
- Metrics: RMSE & MAE
- RMSE weights more for large error, so it is more preferable

Algorithms

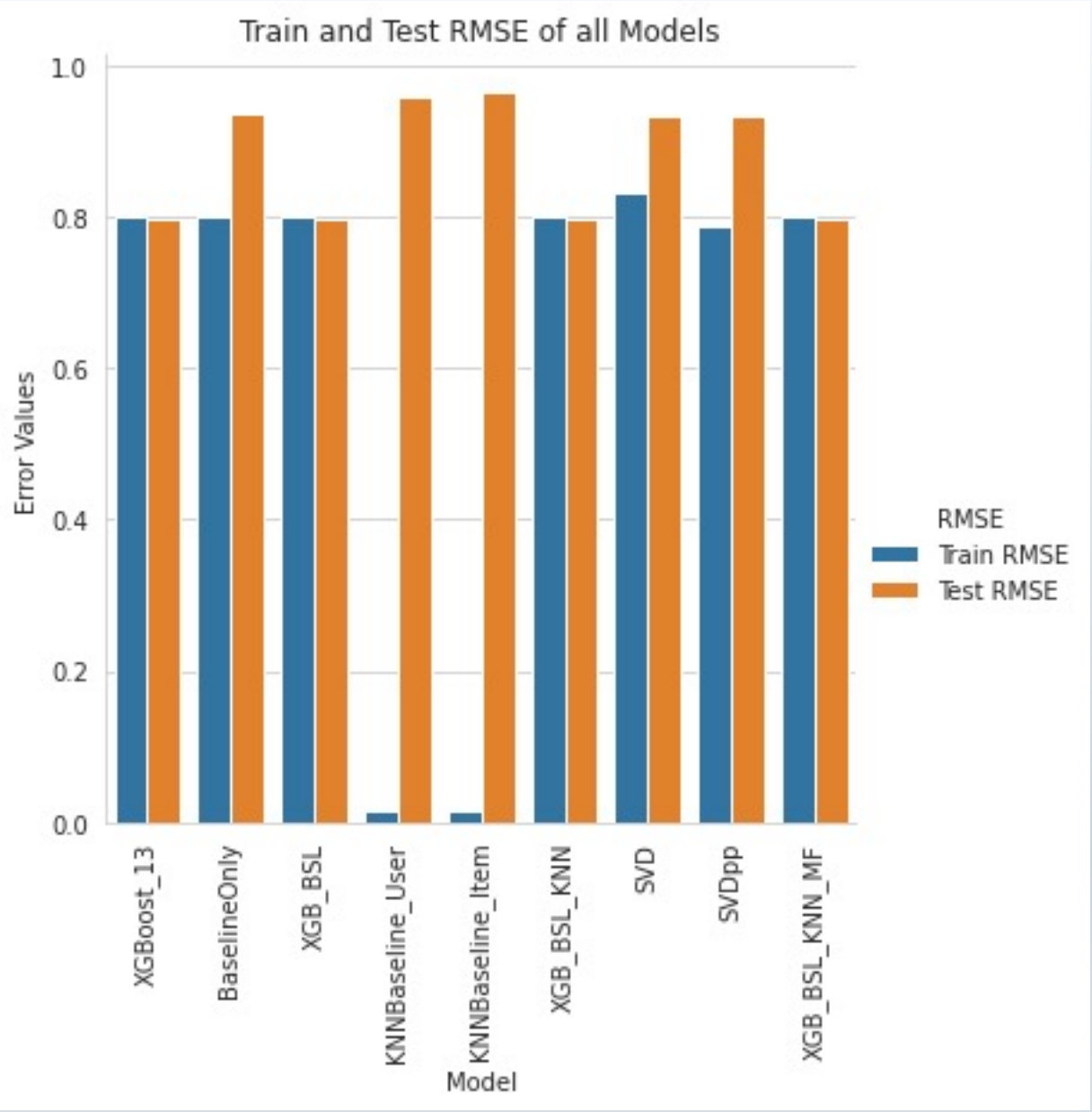
- 4 Base Model:
  - 5-star model (RMSE 1.168)
  - Content-filtering based model with mean rating of users (RMSE 1.016)
  - Collaborative-filtering based with mean rating of items (RMSE 0.98)
  - Collaborative-filtering based with mean rating of items bought together (RMSE 0.972)

Surprise Library & XGBoost: (Individual & combinations of ML algorithms)

- XGBoost regression
- BaselineOnly
- KNN-Baseline
- SVD
- SVD++

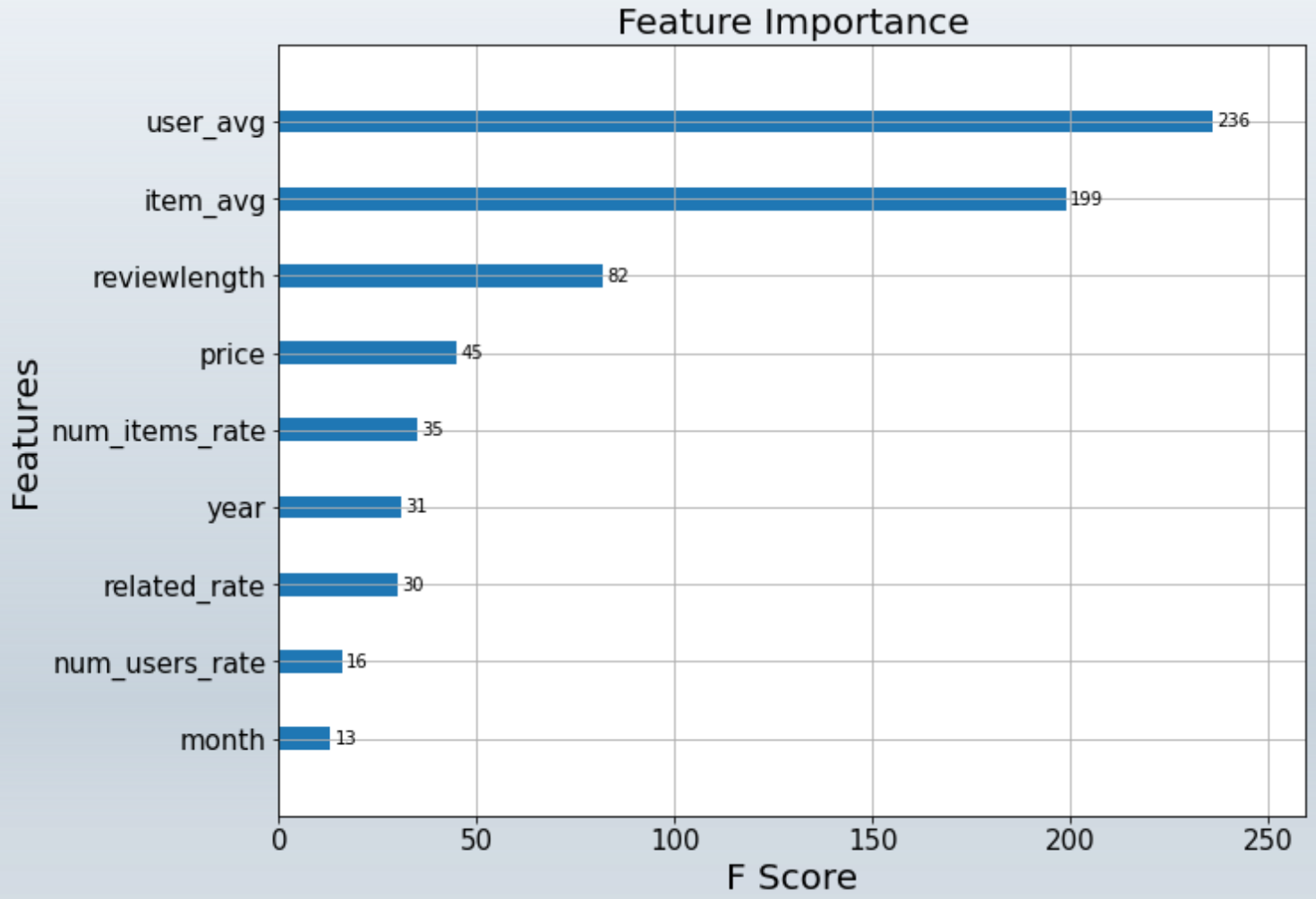
Model Performance

- XGBoost regression gave the lowest RMSE: 0.795
- The combination of different algorisms didn’t improve the XGBoost model and the difference was tiny.



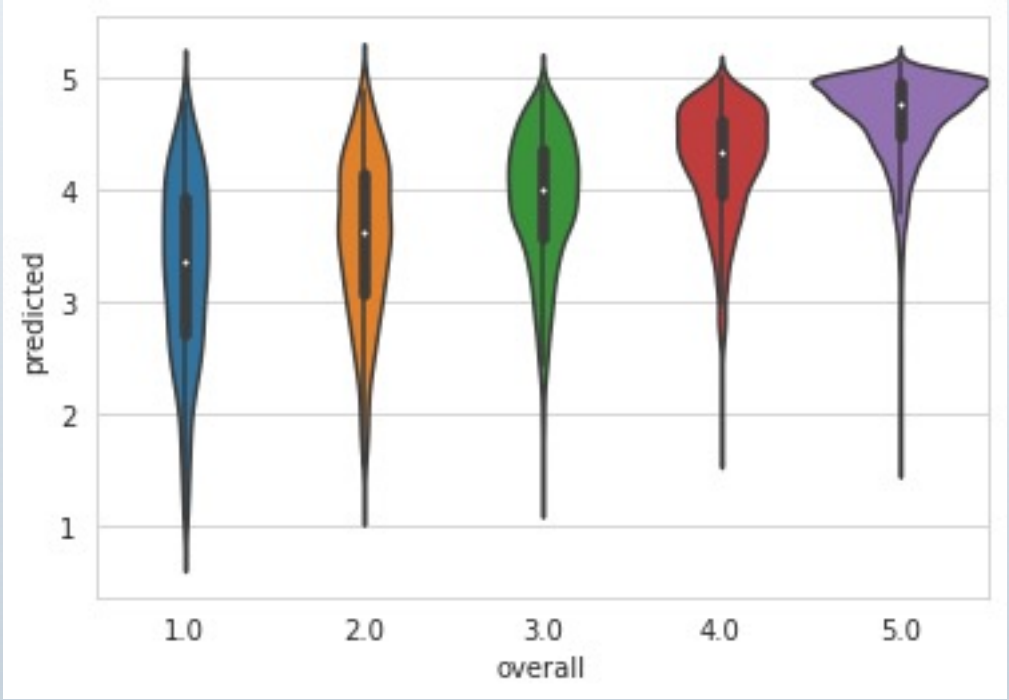
Important Features

- Top 3 features which have biggest influence on rating prediction:
  - Average rating of user
  - Average rating of item
  - Review length



Prediction

- Predicted ratings skew to the higher values
- Distribution range gets wider when original rating gets lower
- Mean Absolute Error = 0.54
- The error may lead to one rating lower or higher



Prediction with/without brands and models

- User can get the top ten items recommend to them based on the prediction model

	asin	predicted	title
0	B005HJH2NM	5.000862	DeLonghi Red Lattissima Plus Nespresso Capsule...
12	B004YWEY8E	4.996239	Anolon Advanced Bronze Hard Anodized Nonstick ...
25	B0000Y73UQ	4.996239	Kuhn Rikon Duromatic Top Model Energy Efficien...
37	B004VMAC8I	4.985146	Vitamix 1782 TurboBlend, 2-Speed
50	B001CEPYVS	4.982747	Anolon Advanced Bronze Collection Hard Anodize...
63	B00AYCUNVU	4.982747	Circulon Symmetry Chocolate Hard Anodized Nons...
78	B00851TPAM	4.982747	Oster VERSA 1400-watt Professional Performance...
93	B000MAKVLQ	4.982747	Zojirushi NP-HBC10 5-1/2-Cup (Uncooked) Rice C...
106	B003V8A4KY	4.980407	Wusthof Classic 3-Piece Essentials Set with Ch...
117	B004Z915M4	4.979287	Excalibur 3900B 9 Tray Deluxe Dehydrator, Black

Conclusions

- XGB Regression Model provided the best results
- Future improvements:
  - Use more recent data and whole data

    (there are no data between after 2014; trimmed data was used for this project)
  - More features

    (Collect more information about users and items; NLP about the review content)
  - Edge Cases

    (Recommend to new users with no review history)
  - More Algorithms

    (There are more algorithms to try)