Prediction and Recommendation System for Airbnb

Team Perceptrons

Manisha Shivshette
Arshiya Pathan
Tarun Arora

Dataset

- Review Details (802k records)
- Calendar Details(16.2M records)
- Size of Dataset: 910.7MB

Problems Addressed

- How to interpret text review ?
- Which listings should be recommended to user?
- What should be optimum price of listing to increase profit ?

System Design

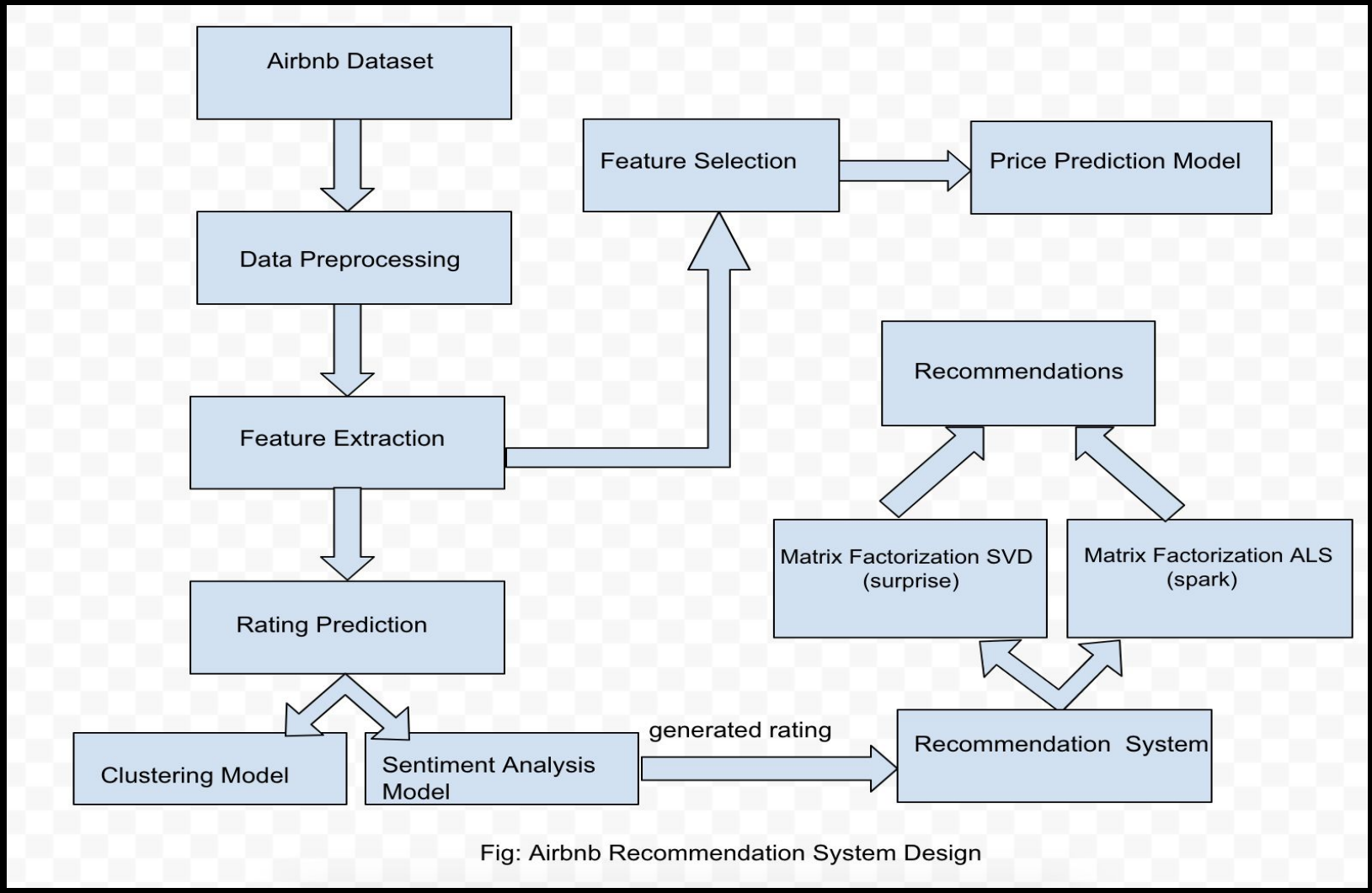


Fig: Airbnb Recommendation System Design

Solutions

Rating Prediction

- Rating Prediction using K-Means Clustering
  - Create TF-IDF Matrix from input review comments
  - Generate Clusters of similar review comments
  - Identify the cluster containing maximum positive tokens
  - Replace the cluster label with the rating

- Rating Prediction using Sentiment Analysis
  - Clean Review Comment text data
  - Using NLTK Sentiment analyzer generate polarity scores
  - Using positive and negative polarity scores, compute the rating

Solutions (Cont.)

Recommendation System

- Approach 1 - Matrix factorization - SVD
  - Tuning the parameters using Grid search
  - Evaluate the model. Compare biased and unbiased accuracy.
  - Using K fold cross validation train the model.
  - Build the test set by considering users and all the listings which user has not given reviews for.

- Approach 2 - Matrix factorization - Alternating Least Squares (ALS)
  - Train and tune the hyperparameters such as maximum iterations, rank and regularization parameter
  - Improve the performance score using k-fold cross validation
  - Test the model on test set and calculate RMSE.
  - Recommend top 3 listings to all users

Analysis of Rating Prediction

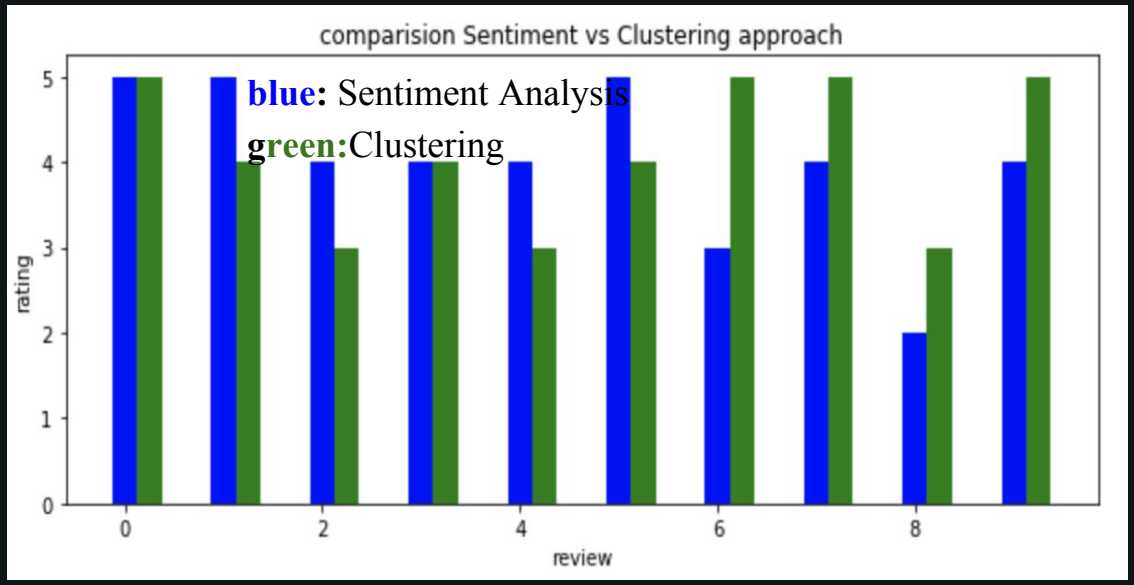


Fig:Comparative Evaluation

Advantages of Sentiment Analysis over Clustering approach

- Scalable
- Evaluates negative component i.e differentiates “good” from not “good”
- Independent of review length

Analysis of Recommendation System

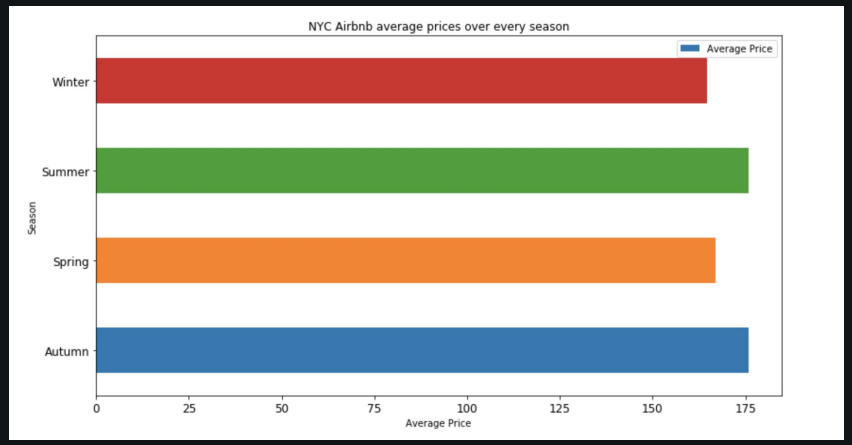
SVD

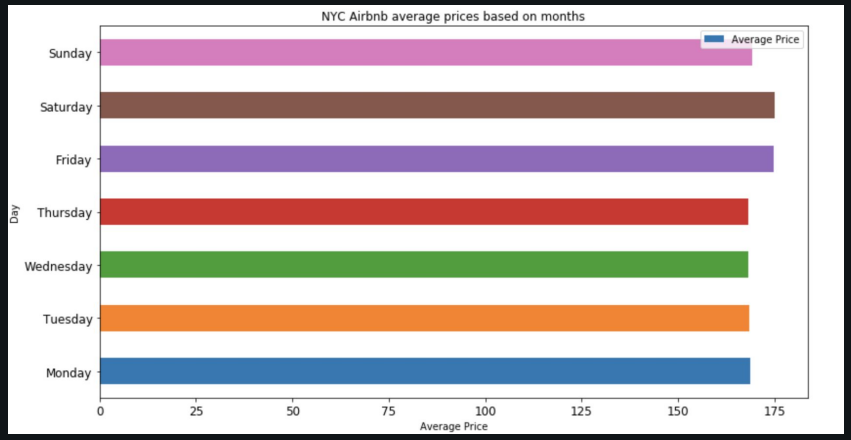
- SVD gives better accuracy over ALS
- RMSE for model using SVD is 0.85
- Good for medium sized dataset.
- Not suitable for very large and sparse dataset

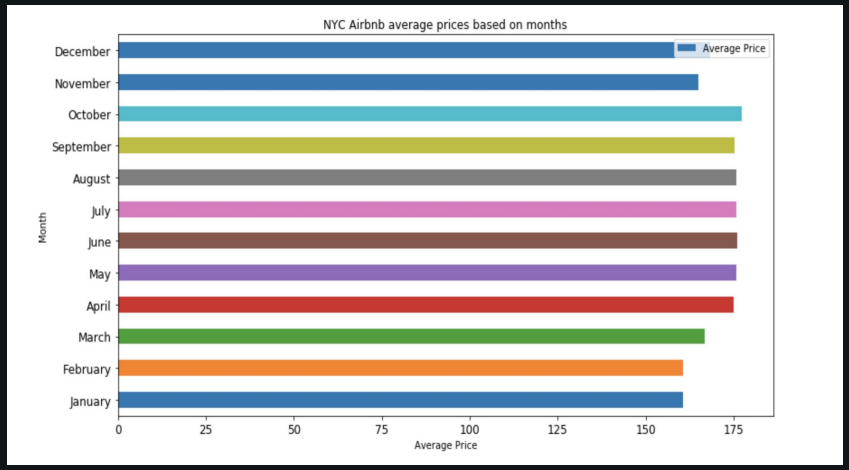
ALS (Spark)

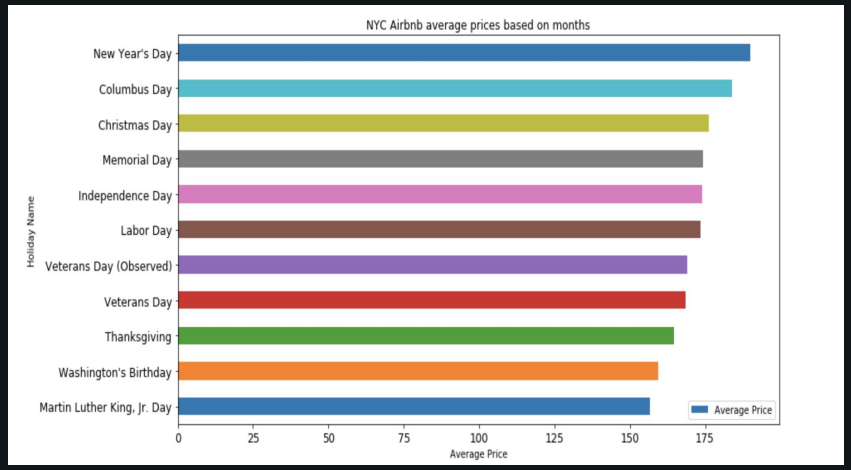
- Good for recommendations using large and sparse dataset.
- Accuracy is not as good as SVD.
- RMSE is 2.1

Price Prediction Analysis









Difficulties

- Very large dataset
- Lot of effort in preprocessing the data
- Evaluation of predicted rating

Things that worked well

- Effective use of data with various pre processing techniques
- Models creation using clustering and sentiment analysis
- Matrix factorization using SVD and ALS supported large and sparse dataset to predict ratings

Things that did not work well

- As the AirBnb dataset is very large and users have given only one or two reviews, user based and item based algorithm could not be used to build and evaluate model.
- Matrix built is very big. Difficult to predict ratings.

Conclusion

After evaluating clustering and sentiment analysis based approaches, we decided to use rating generated by the sentiment analysis model to train the recommendation algorithm.

After evaluating different collaborative filtering algorithms ,we concluded that matrix factorization is the best algorithm to provide recommendations for the large and sparse dataset like Airbnb dataset.

Also in order to provide recommendations over the large dataset, Spark's ALS implementation is more suitable.

Building this module helped to build skills required for 1) data preprocessing 2) building ML models for large-scale data 3) evaluating methods. 4) visualize data and use it for predictions.