Analyze E-commerce Seller Rating System Using Apache Spark

Saima Sanjida Shila

Louisiana State University, Baton Rouge, LA

Your Project Description

I have analyzed an E-commerce platform's seller rating system by processing large-scale transaction, product and customer datasets using Apache Spark. The analysis focuses on identifying trends and insights in seller ratings based on customer information, such as ratings, purchases and reviews about products. This analysis aims to help e-commerce platforms enhance customer experience and improve seller performance.

- Generate Seller Average Ratings based on the datasets.
- Visualize the top 10 Best Seller based on customer ratings.
- Predict Customers Next Purchase.
- Forecast The Next Best seller using Linear Regression.
- Analyze and Visualize time series of sale trends.

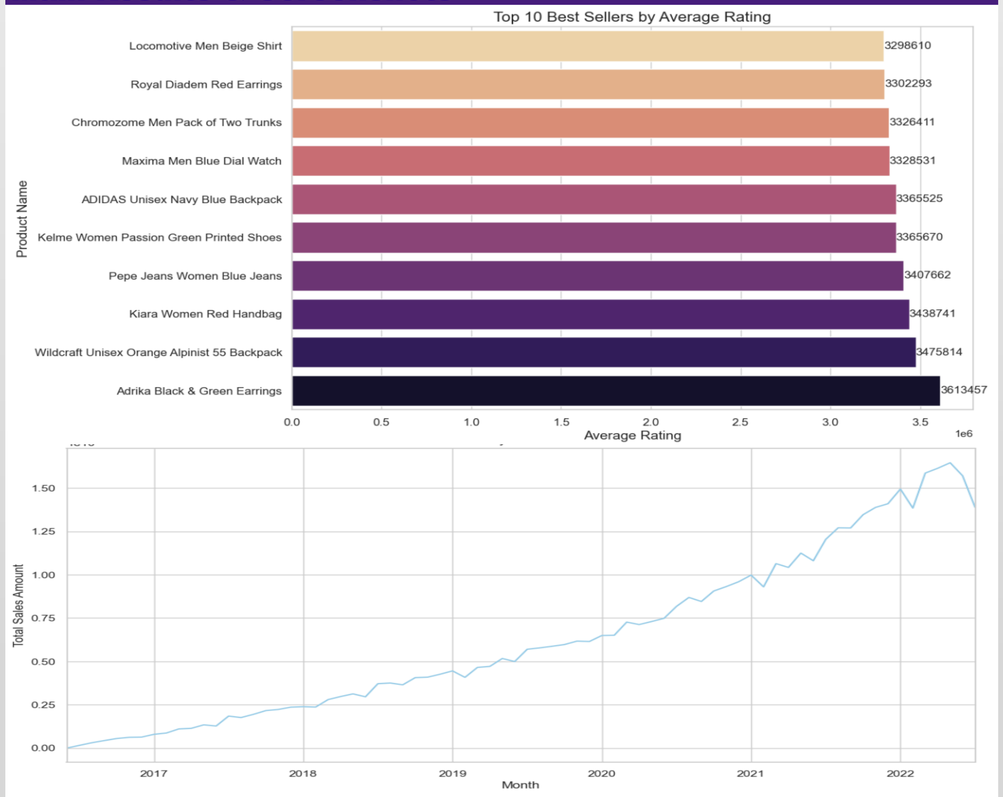
Your Development Details

- I have used Apache Spark locally in my laptop to create the distributed system.
- Also, I have used Framework Jupyter Notebook to implement my codes and to visualize the outcomes of the system.
- Initially, I faced technical issues setting up the distributed system.

Big Data Frameworks

I have used Apache Spark framework for the E-commerce seller rating system. Additionally, I have used Jupyter Notebook, Matplotlib, Seaborn and ML models to generate some insights about the seller rating system.

Main Results & Screenshots




```
+------+--------------------------------------------------
|userId|recommendations
+------+--------------------------------------------------
|26    |[{59058, 2134923.5}, {46074, 1965412.1},
|27    |[{6639, 2535520.0}, {58482, 2168415.0}, {
|28    |[{22082, 1.1263976E7}, {10294, 1.0905442E
|65    |[{28991, 696154.06}, {13240, 415007.62},
|76    |[{39551, 9929264.0}, {13157, 9315780.0},
|81    |[{12900, 2114426.0}, {46074, 2041500.8},
|108   |[{32803, 5848038.0}, {9262, 5416562.5}, {
|115   |[{18958, 1635810.1}, {26171, 1263241.1},
|126   |[{46726, 1.5918573E7}, {51702, 1.5169332E
```

Datasets

I am using multiple datasets from Kaggle. Customer data, product data and transactions data of an e-commerce platform. Moreover, I have made a CSV of average ratings from these datasets and used it to further generate some useful insights.
URL: https://www.kaggle.com/datasets/bytadit/transactional-ecommerce?

customer.csv	Today, 1:32 PM	4.4 MB
product.csv	Today, 1:33 PM	3.2 MB
transactions.csv	Today, 1:37 PM	158.5 MB
avg_rating.csv		

Conclusion & Future Work

- I have learned the functionality of Distributed system on machines locally.
- Using cluster system on a real-life dataset and using analyzing and visualization techniques.

❑ I will try to extend the project further by using some more Machine Learning Models like SVM, Decision Tree to predict more future insights on the seller rating system.

File Name	#lines
Project/seller_rating_system.py	30
Project/visualization_best_sellers.py	32
Project/visualize_insights.py	20
Project/recommender_system_for_customers.py	70
Project/prediction_best_seller.py	60