

**Project Title: Analyze E-commerce Seller
Rating System Using Apache Spark**



Saima Sanjida Shila
sshila1@lsu.edu
10th October 2024

Table of Contents

1. Introduction

1.1 Motivation

1.2 Project Description

2. System Design

2.1 Chosen Big Data Frameworks

2.2 Chosen Datasets

3. Detailed Description of Components

3.1 Each Component Description

4. Evaluation & Test

4.1 Software Manual

4.2 Test Environment

4.3 Test Results

5. Conclusion

1. Introduction

1.1 Motivation

In this era of rapid growth of E-commerce businesses, customers satisfaction is the most priority for sellers. Seller ratings and reviews of a product shows how satisfied customers are and influences buying decisions. In this project, I am using big data tools to analyze and predict seller ratings on an e-commerce platform. From the results, we can understand and learn more insights about customer needs, product quality and seller's trends.

1.2 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.

2. System Design




2.1 Chosen Big Data Frameworks

I have used Apache Spark as my main tool for big data processing to build 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

2.2 Chosen 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

3. Detailed Description of Components

3.1 Each Component Description

- **Data Preprocessing:** I have preprocessed all the datasets. The initial size of the datasets was up to 300mb altogether. I have processed and cleaned some features from the datasets which is not necessary for the project criteria.
- **Data Loading:** I loaded each dataset into Spark Data Frames using PySpark read.csv() method. I selected only the columns needed for each of the part of the project.
- **Generate Seller Average Ratings:** In this component, the primary goal is to analyze customer ratings for various sellers and calculate an average rating for each. By using Apache Spark, I loaded the datasets in Data Frames and grouped it by seller ID to compute the average comparing with customer dataset. (test results in 4.3) This is very essential to identify trends in customer feedback and to predict the next best sellers. Additionally, I saved the average ratings of the seller in a CSV file to further use it for some more valuable insights of the e-commerce system.

- **Visualize the top 10 Best Sellers:** To gain some insights about who is the best seller in the e-commerce system. (test results in 4.3) I have used the average rating dataset which I stored in the first step of seller ratings. This useful dataset visualizes the top 10 best seller in the e-commerce system.
- **Predict Customer's Next Purchase:** This component focuses on predicting what products customers might purchase next. It generated from the customer's previous buy. (test results in 4.3) I have used ALS model filtering method to generate the next purchase for the customer. Here, customers are recommended some products for their next purchase.
- **Forecast the Next Best Seller:** I have used Linear Regression model to predict who will be the next best seller of the e-commerce system. I have also implemented linear regression model using Spark's MLlib to forecast the next potential seller. (test results in 4.3) This predictive functionality can be valuable for inventory management and marketing strategies. It also allows sellers to focus on promoting their products to increase their customer satisfaction.
- **Analyze and Visualize Monthly Sales Trends:** I performed time series analysis on the transactions dataset to observe sales trends over time. By sampling the data monthly and visualizing it, I identified patterns and seasonal sales in e-commerce. (test results in 4.3) This trend helps us gain insights that over time customers are preferring online shopping more than going into a mall.

4. Evaluation & Test

4.1 Software Manual

I installed and set up Apache Spark on my laptop locally. Also, I set up a distributed cluster system with one master node and one worker node. I connected Jupyter notebook and installed

PySpark in the environment with python. Moreover, I have used libraries like pandas, seaborn for data handling and visualizations and pyspark.ml library for machine learning model.

4.2 Test Environment

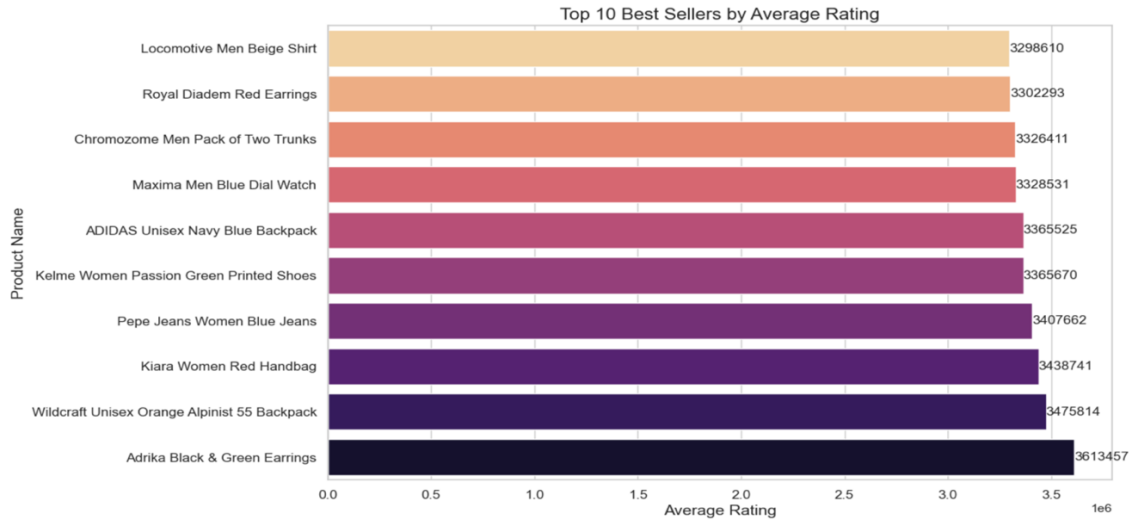
This projects test environment included:

- Apache Spark
- Python 3.11
- Jupyter Notebook

4.3 Test Results

```
+-----+-----+-----+
|    id| productDisplayName|    average_rating|
+-----+-----+-----+
| 2662|Murcia Women Broc...|1320385.4834749764|
| 4582|Wildcraft Unisex ...|1203044.7068965517|
| 7533|Nike Men's As T90...|1229085.6025316457|
|50738|Red Chief Men Mus...|1276089.2777777778|
| 8121|Fastrack Men Brow...|1207410.5935960591|
| 4873|Levis Kids Boy's ...| 1422504.750367107|
| 1868|Inkfruit Mens Fac...|1113600.7385740401|
| 4924|Gini and Jony Gir...| 1312095.405063291|
| 2888|Catwalk Women Wed...|1253116.6009345795|
| 2356|ADIDAS Black & Wh...|1153323.4444444445|
| 1939|Geonaute Women La...| 1150376.782186949|
|23765|Oakley Men White ...| 1236591.17721519|
| 5122|Wrangler Women's ...|1288775.6505050506|
|21148|s.Oliver Women Gr...| 892571.8933333333|
|46115|ADIDAS Men Olive ...| 1293387.829787234|
|30952|Fabindia Women Pr...| 898707.1724137932|
|43955|Scullers Men Pur...| 994589.4722222222|
|59961| Denim Men Sync Deo| 1402933.189189189|
|37719|John Players Men ...|1144914.1315789474|
|31724|Fabindia Women An...|1154472.6794871795|
+-----+-----+-----+
only showing top 20 rows
```

BIG DATA

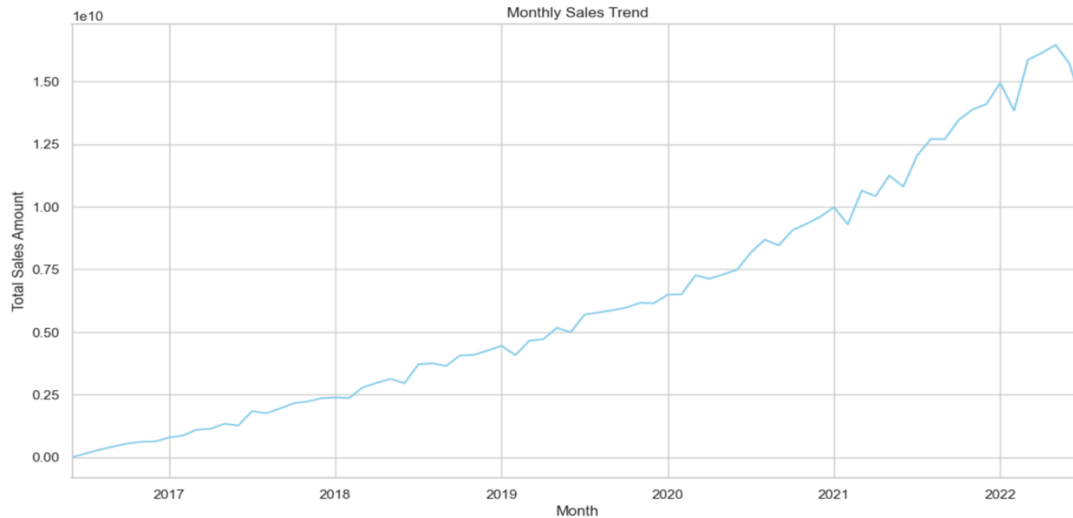


productDisplayName	average_rating	prediction
Nike Mean Team In...	1086552.3841131665	1190185.282933178
Ferrari Tee	1243243.359836901	1191962.1200702838
Puma Men Grey Sol...	1133484.1943573668	1189268.289879897
Puma Men Black Ne...	1085464.4828571428	1189356.492398966
Puma Men's Ballis...	1134777.8061728396	1189135.7251470948
Puma Men's Furore...	1226267.376101861	1189117.1973990065
Puma Power Cat Tr...	1058591.9331501832	1189072.052322678
Quechua Easy-to-C...	1054411.8915789474	1188889.6453379758
Kalenji Men's Sup...	1154166.5137111517	1191053.4775513525
Kalenji Mens Esse...	1130051.3569096844	1191053.216597154

only showing top 10 rows

userId	recommendations
26	[{59058, 2134923.5}, {46074, 1965412.1}, {3574, 1946569.8}, {52749, 1885802.2}, {29407, 1827122.9}]
27	[{6639, 2535520.0}, {58482, 2168415.0}, {20344, 2160296.8}, {22769, 2152804.2}, {21181, 2126297.5}]
28	[{22082, 1.1263976E7}, {10294, 1.0905442E7}, {1578, 1.0846498E7}, {57589, 1.0844808E7}, {47416, 1.0676829E7}]
65	[{28991, 696154.06}, {13240, 415007.62}, {32831, 395000.1}, {29548, 364476.47}, {57072, 359390.25}]
76	[{39551, 9929264.0}, {13157, 9315780.0}, {8237, 9290119.0}, {50567, 9016579.0}, {51702, 8921341.0}]
81	[{12900, 2114426.0}, {46074, 2041500.8}, {26171, 2029840.5}, {4905, 2025671.9}, {43827, 1963250.1}]
108	[{32803, 5848038.0}, {9262, 5416562.5}, {48448, 5209621.5}, {3574, 5041358.5}, {52166, 4670107.5}]
115	[{18958, 1635810.1}, {26171, 1263241.1}, {20245, 1236024.1}, {18405, 1205743.5}, {55500, 1109960.5}]
126	[{46726, 1.5918573E7}, {51702, 1.5169332E7}, {43235, 1.5156189E7}, {52675, 1.2137075E7}, {11629, 1.1907307E7}]
133	[{16804, 2103105.5}, {3645, 1464491.4}, {21691, 1322993.0}, {46985, 1184222.4}, {39930, 1138807.8}]
155	[{32803, 7575830.5}, {42822, 7170596.0}, {4380, 6301963.5}, {50045, 6004348.5}, {10235, 6000915.0}]
183	[{23313, 5141278.0}, {12079, 4128551.0}, {24574, 3951468.8}, {40648, 3757326.2}, {39291, 3658509.0}]
210	[{39551, 453264.9}, {57589, 432433.12}, {28835, 412960.94}, {54609, 399536.5}, {8221, 395436.3}]
211	[{28991, 1.632464E7}, {54087, 1.2360478E7}, {57072, 1.1042577E7}, {39011, 1.0835896E7}, {8221, 1.0796051E7}]
236	[{39551, 9771168.0}, {8221, 9003150.0}, {24429, 7984011.0}, {33624, 7609701.5}, {54609, 7462904.5}]
243	[{39551, 1.6462296E7}, {45837, 1.2648311E7}, {8221, 1.1169256E7}, {4096, 1.1083628E7}, {29548, 1.0956591E7}]
255	[{38130, 1.5120839E7}, {16804, 1.3796891E7}, {57072, 1.3514204E7}, {23313, 1.232482E7}, {15194, 1.1762591E7}]
300	[{28991, 4832460.0}, {57072, 4323388.0}, {16804, 4289302.5}, {27815, 4175791.5}, {11629, 4165499.0}]
321	[{8237, 2536947.0}, {24574, 2485924.5}, {8760, 2259387.5}, {28166, 2157408.2}, {4696, 2081745.8}]
322	[{28991, 7523332.0}, {39551, 7459119.0}, {25451, 7336832.0}, {50567, 6816840.0}, {58509, 6490848.0}]

only showing top 20 rows



5. Conclusion

In this project, I have used Apache Spark to analyze and predict seller rating in e-commerce. With big data tools, I efficiently processed large amounts of data and created a model to predict seller ratings. Future work could involve using advanced machine learning model like decision tree and SVM to learn more insights about the e-commerce system to improve business decisions.

6. Appendix

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

File Name	#Spark Operations
Project/seller_rating_system.py	30
Project/recommender_system_for_customers.py	30
Project/prediction_best_seller.py	20