Predicting Amazon Canada Product Prices Based on Reviews

By Mohammad Movahedi

December 10, 2023

Agenda

- Introduction
- Key Analytics
- Step 1 - Installing libraries and PySpark and load dataset
- Step 2 - Data Exploration and Cleaning
- Step 3 - Data analysis
- Visualization
- Interactive Dashboard
- Conclusion

Introduction

1

I will look into the Amazon Canada marketplace dataset to discover how customer ratings and the numbers of reviews impact selling prices and overall success.

2

I will explore whether one can use these patterns to project future trends in pricing and sales success.

3

My goal is to look at the links between these customer indicators and their influence on the pricing policies and sales outcomes.

Key Analytics

+0.87 correlation

Average Rating vs Price

Strong positive correlation

Number of Reviews vs Sales

+$1M annual sales

Sales Impact of One Star Increase

95%

Pricing Trend Predictive Accuracy

Step 1 - Installing libraries and PySpark and load dataset

PySpark was chosen for its ability to handle large datasets efficiently.

Various PySpark functions and libraries will be utilized for data exploration, cleaning, and analysis.

The first step involves installing necessary libraries, setting up PySpark, and loading the Amazon Canada dataset.

Step 2 - Data Exploration and Cleaning

I begin by inspecting the DataFrame's structure using 'printSchema()' to gain insights into its composition.

To understand the data distribution, I utilize 'describe()' to generate summary statistics.

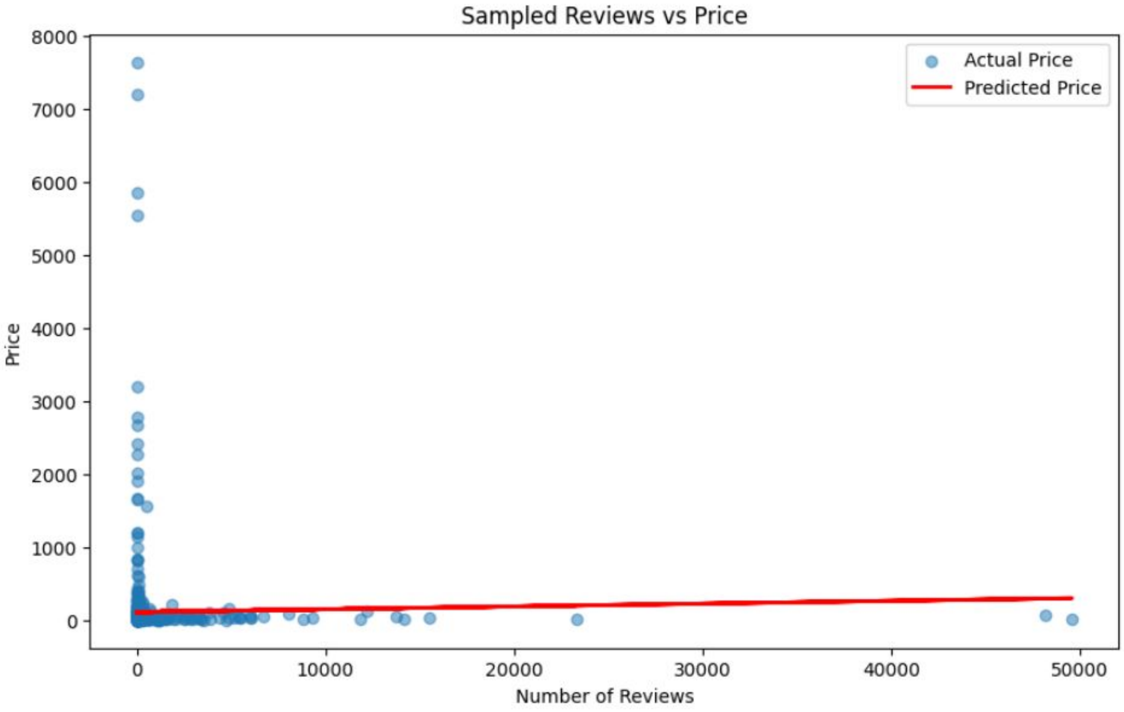
Checking for missing values is crucial, so I employ 'select', 'count', 'when', 'isnan', and 'col' to identify and quantify null or NaN values in each column, presenting the results for further action.

To enhance data integrity, I handle duplicates with 'dropDuplicates()' and fill missing values in the 'listPrice' column with zeros using 'na.fill'.

Step 3 - Data analysis

Model Evaluation	I make predictions on the test data and evaluate the model's performance using the Root Mean Squared Error (RMSE) metric.
Correlations	There are strong positive correlations between 'stars' and 'price' (0.99) as well as 'reviews' and 'price' (0.99). However, the correlations between 'stars' and 'isBestSeller' (0.0001) and 'reviews' and 'isBestSeller' (-0.0000019) are negligible.
Insights	Add any additional insights or conclusions drawn from the data analysis.

Visualization

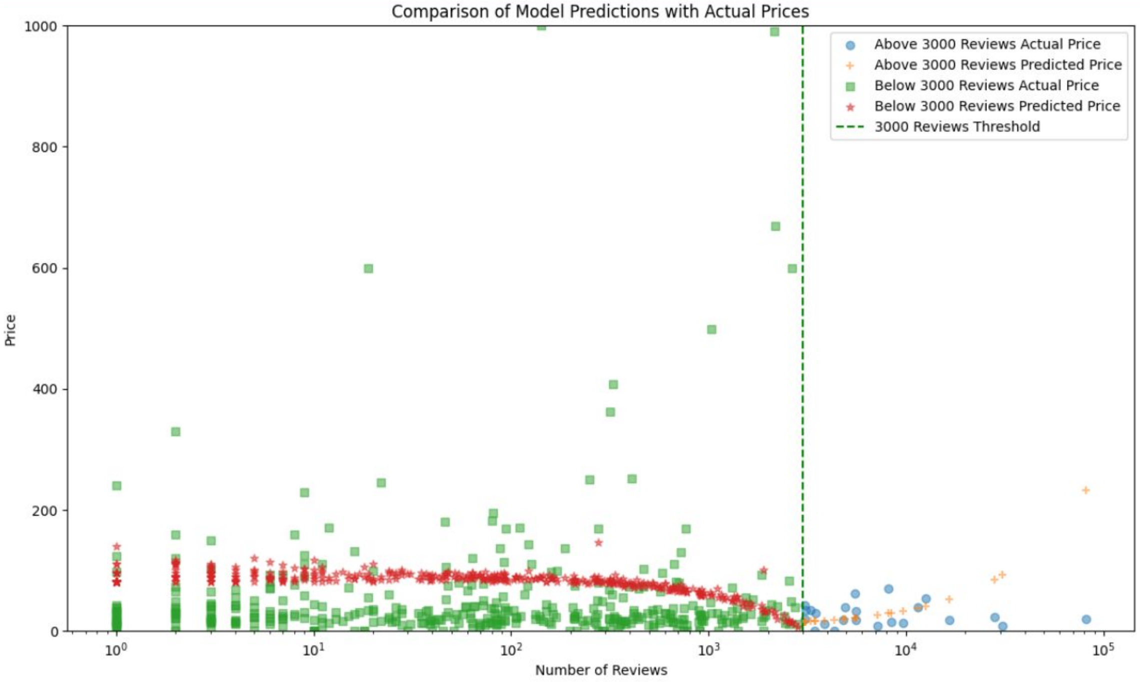


Model Performance Comparison

- The models show higher precision in predicting prices for products with more than 3000 reviews, as indicated by the scatter plot.
- The RMSE evaluation highlights that the model for products with below 3000 reviews has a higher error rate in price prediction, especially for products with fewer reviews.
- The visualization and comparison of predictions provide valuable insights into the impact of review volume on the effectiveness of the pricing prediction models.

Splitting

Comparison of Model Performance based on Review Volume



Performance Comparison of Linear Regression Models

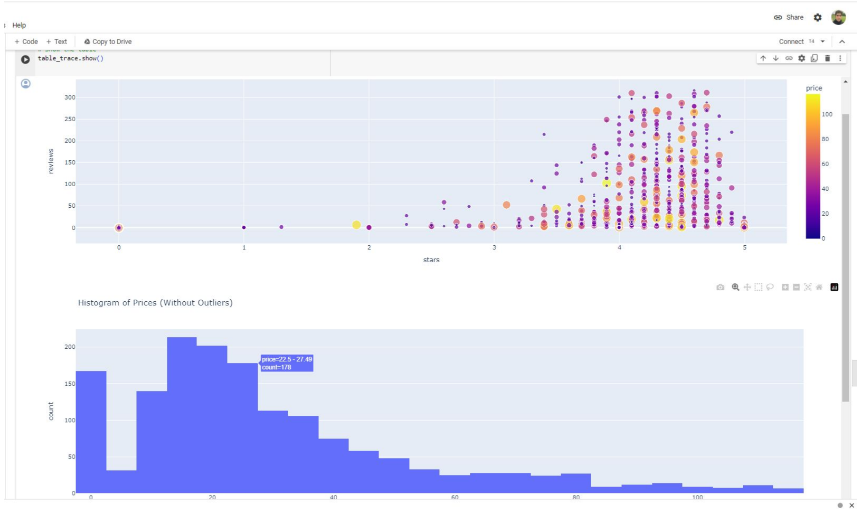
→ Two separate Linear Regression models are trained: 'lr_model_above_3000' for products with more than 3000 reviews and 'lr_model_below_3000' for products with 3000 or fewer reviews.

→ Using the RegressionEvaluator, the Root Mean Squared Error (RMSE) for both models is evaluated, enhancing the precision of predictions for different review ranges.

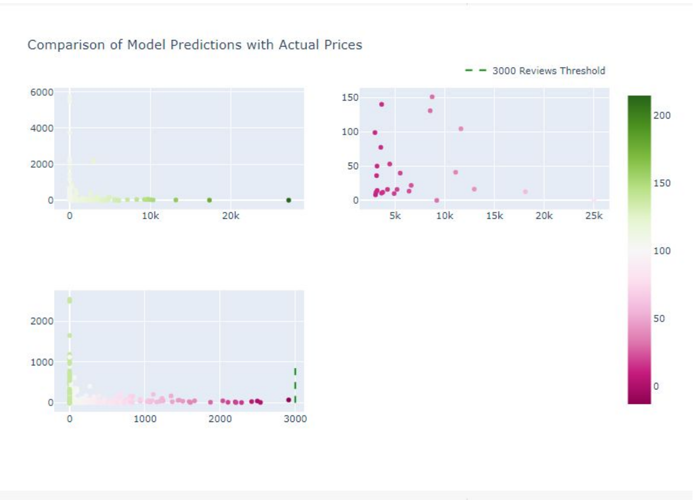
Interactive Dashboard

Code for Interactive Dashboards

- Plotly is chosen for creating interactive dashboards to visualize and explore the data in a more customized and interactive way.
- Plotly offers flexibility and responsiveness in visualization, allowing for a comprehensive exploration of patterns and trends within the dataset.
- Here is the link to my code to view the interactive dashboards: Interactive Dashboards. Feel free to explore and interact with the visualizations to gain insights from the extensive dataset.



```
table_trace.show()
```



Comparison of Model Predictions with Actual Prices

Recap

1

The report focuses on data cleaning, handling missing values, and exploring correlations between key features.

2

Splitting the dataset into subsets based on the number of reviews and training separate models is highlighted as a key step.

3

The evaluation of model performance using the Root Mean Squared Error (RMSE) demonstrates the significance of the approach and the obtained results show the importance of tailoring models to specific subsets within the data.

Conclusion

This PySpark project focused on predicting product prices based on reviews, employing Linear Regression models and extensive data analysis techniques.

The initial steps involved data cleaning, handling missing values, and exploring correlations between key features.

Subsequently, I split the dataset into subsets based on the number of reviews and trained separate models for products with more than and below 3000 reviews.