**PREDICTIVE ANALYSIS**

**-** Coffee Shop Using Sales Data **-**

**Project Overview**

In this project, we aimed to analyze and predict profit and loss for a coffee shop based on its sales data. The goal was to clean and prepare the data, dive into performance metrics, and use machine learning to forecast future profits. With the help of PySpark and some visualizations through Matplotlib, we were able to uncover insights and trends that could help optimize the coffee shop’s operations.

**Objectives**

1. **Clean the Data:** Prepare the dataset by fixing any inconsistencies, filling in missing values, and removing duplicates.
2. **Analyze Profit and Loss:** Calculate revenue, cost, and profit for each product to understand financial performance.
3. **Trend Analysis:** Look at how sales and revenue change over time to spot any patterns.
4. **Store Performance:** Evaluate how different store locations are performing financially.
5. **Predict Future Profits:** Use machine learning to forecast potential future profits.
6. **Understand Feature Impact:** Determine which factors are most important in predicting profit or loss.
7. **Scenario Analysis:** Simulate different scenarios to see how changes might affect profits.
8. **Loss Mitigation:** Identify any products that are losing money and suggest ways to address these issues.

**Methodology**

1. **Data Cleaning:**

We started by loading the dataset into Spark and cleaning it up. This involved checking for missing values and duplicates, and making sure all text data was consistent by converting it to lowercase.

Cleaning Processes:

* Inconsistencies Corrected:

Standardized text fields (e.g., store\_location, product\_category, product\_type, product\_detail) by converting all text to lowercase.

* Missing Values**:**

Identified columns with missing values. In this dataset, no missing values were found, but processes for handling them are described.

* Duplicates Removed**:**

Checked for and removed duplicate rows to ensure the dataset’s accuracy and integrity.

1. **Profit and Loss Analysis:**

We calculated the total revenue and cost for each product, assuming a cost price of 70% of the unit price. By aggregating this data, we were able to determine the profit or loss for each product.

1. **Trend Analysis:**

We broke down the sales data by month and year to observe trends over time. This included visualizing monthly sales quantities and revenue to identify any significant patterns.

1. **Store Performance:**

We assessed how each store location was performing by calculating total sales and revenue. This helped us see which locations were the most successful.

1. **Profit Prediction:**

We used a Random Forest Regressor model to predict future profits. By preparing our data with StringIndexer and VectorAssembler, we could train the model and evaluate its performance using metrics like RMSE and R-squared.

1. **Feature Importance:**

We analyzed which features were most influential in predicting profit or loss. This involved creating visualizations to highlight the importance of each feature in the model.

1. **Scenario Analysis:**

We created some hypothetical scenarios to see how changes in transaction quantity and unit price would impact profit. This helped us understand potential outcomes under different conditions.

1. **Loss Mitigation:**

We checked if there were any products with negative profit margins. Fortunately, the analysis showed that all products were making a profit, so this section was more about confirming that there were no losses to address.

**Challenges**

1. **Data Cleaning:**

Ensuring the dataset was clean and consistent was a bit tricky. It took careful checking to handle missing values and remove duplicates.

1. **Feature Engineering:**

Converting categorical features into numerical values for the model required some extra steps, including using StringIndexer and VectorAssembler.

1. **Model Training:**

Choosing and tuning the Random Forest model involved some trial and error. We had to experiment to find the best settings for accurate predictions.

**Conclusion**

Overall, the project provided valuable insights into the coffee shop’s financial performance. We were able to see how different products and store locations were performing, and predict future profits with reasonable accuracy. The analysis highlighted key areas for improvement and provided a solid foundation for making data-driven decisions. Going forward, refining the model with additional features or exploring other algorithms could further enhance our predictions.