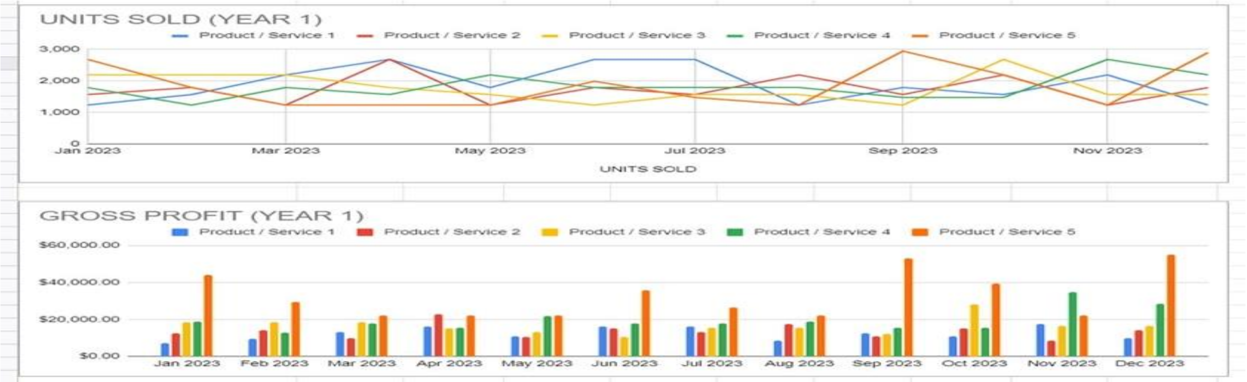
Predictive Modelling for Revenue Forecasting in Retail, Manufacturing, and Distribution Industries

Executive Summary of Sales Revenue Prediction Project.

Problem Statement:

This project aims to predict the sales revenue for specific products or product categories in different industries, primarily focusing on manufacturing, retail, and distribution sectors. The critical question this project seeks to answer is how businesses can not only forecast the total sales revenue but also accurately predict the detailed sales revenue of each specific product or product category based on various product features or attributes. This granular level prediction can greatly assist businesses in making informed decisions related to analysis, investment, budgeting, and better management of their operations.

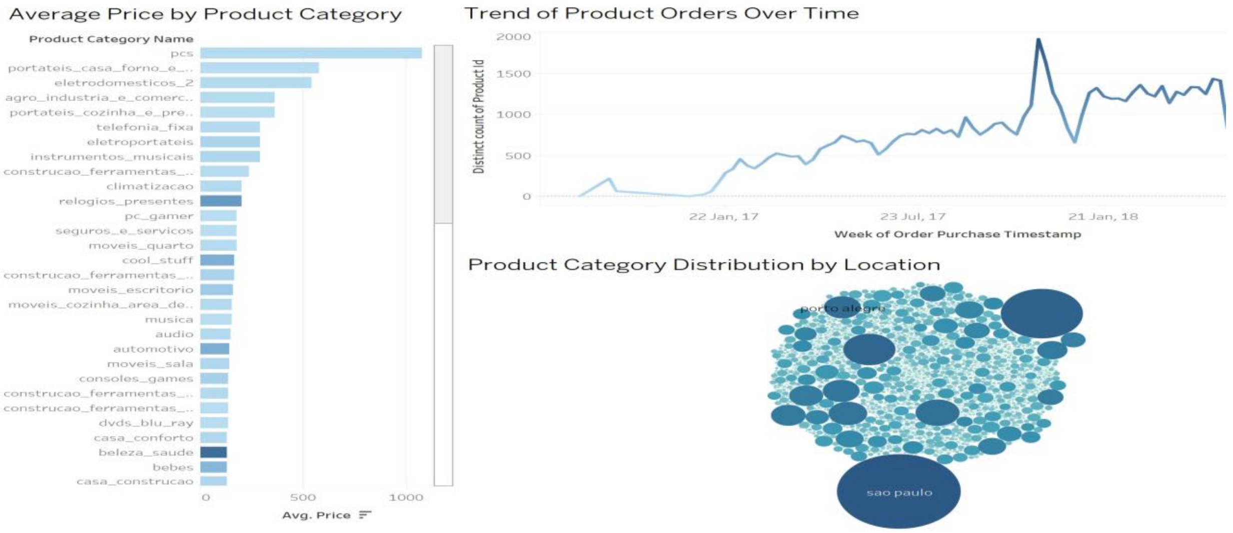


Background and Importance:

With the advancements in data science, businesses can now manage their data, analyze past activities, identify gaps and issues in their operations, and predict future trends. Data science can help find the best pattern of sales revenue for each product based on its features, thereby predicting its future sales revenue.

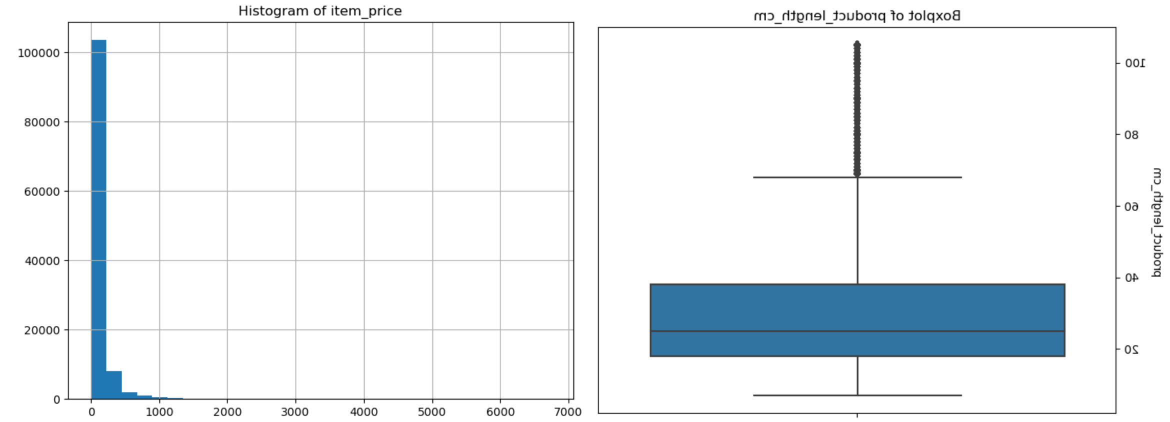
Dataset Overview:

The dataset used in this project is based on the Olist E-commerce public dataset of orders made at Olist Store. The dataset contains information from 100k orders from 2016 to 2018 made at multiple marketplaces in Brazil.



Data Preprocessing and Feature Engineering:

Data preprocessing began with aggregating all datasets into one CSV file. The irrelevant columns were removed, and duplicate rows were dropped. Null values were also dropped due to their low percentage. Feature engineering involved creating a 'sales revenue' feature by multiplying the 'item count' with the 'item price.' Some categorical variables were also converted into numerical variables for analysis.



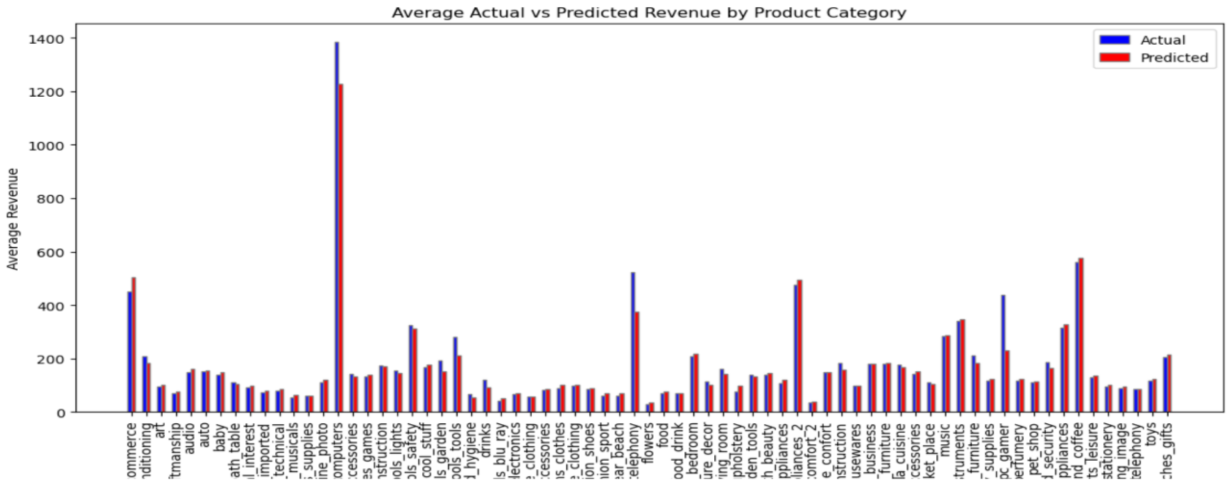
Insights, Modeling, and Results:

Three models were tested: Linear Regression, Decision Tree, and Random Forest.

1. Linear Regression Model: The model had an R-squared value of 85% for the training set, 66% for the validation set, and 82% for the test set, indicating a good predictive capability. However, it had a high level of error, suggesting the need for further optimization.
2. Decision Tree Model: This model exhibited an R-squared score of 78%, indicating a reasonable degree of prediction accuracy. Further optimization improved the model's accuracy, particularly when predicting unseen data in the test set.
3. Random Forest Model: The model showed variable performance with accuracy ranging from 61.7% to 86.2%. After optimization, it demonstrated a strong fit with an R^2 score of 89% on the training set, and a good generalization capability with an R^2 score of 85% on the test set.

Findings and Conclusions:

The Linear Regression model provided decent predictive accuracy but struggled to maintain this accuracy across the entire dataset. The Decision Tree model proved useful in predicting sales revenue for different products, correctly estimating about 79% of the sales changes. The Random Forest model emerged as a robust and reliable model for sales revenue prediction, underscoring its potential applicability in strategic business decisions.



END