APRIL 29, 2023

DATA STORM 4.0 TEAM REPORT

Team Name: Team MilkMan

Team Members:

- Minidu Thiranjaya
- Themira Chathumina
- Navinda Perera

Kaggle Username and Display Name:
DataStorm120

GitHub repository link:
https://github.com/ThemiraChathumina/data-storm-team-120

Highest Total F1 score achieved: 0.66

APPROACH IN BRIEF

In this project, we aimed to explore the business performance of 124 shops using data visualization and machine learning techniques. We used Python as our programming language and Jupyter notebooks in Google Colab environment to conduct our analysis.

We first examined the data for any patterns or trends that could inform our subsequent steps. Then, we computed various average values for each shop, such as total income, monthly income, average customer spending, and so on.

These values served as general features that characterized the business situation of each shop. We prepared a dataset of 124 records, one for each shop, with these features as variables.

FEATURE ENGINEERING IDEAS WORKED AND REASONS FOR DESIGNING THEM.

We then experimented with different models to classify the shops based on their business performance. We initially used general average values for each shop as features, such as total sales, average transaction value, number of unique customers, most popular item, total quantity sold, and average item price. However, these features did not yield satisfactory predictions.

Upon further investigation of the historical transaction data, we noticed that the prices of the same item varied across shops and over time. We also observed that the customer demand for an item was inversely related to its price.

Based on these insights, we decided to add a new feature for each shop-item pair: the total revenue generated by that item for that shop. This was calculated by summing up the product of the price and the quantity sold of

that item for each transaction. We also kept some general features such as shop area in square feet, sales per square foot, and average transactions per day. After adding these details, the accuracy of the model improved significantly.

The following features were included in the final dataset.

shop_id, shop_area_sq_ft, shop_profile, num_unique_customers, sales_per_sq_ft, avg_transaction_value_per_sq_ft, avg_item_price_per_sq_ft, avg_transactions_per_day, avg_customer_spending, avg_customers_per_sq_ft and a feature for all unique item in historical transactions dataset.

FINAL MODEL AND HOW IT WAS REACHED.

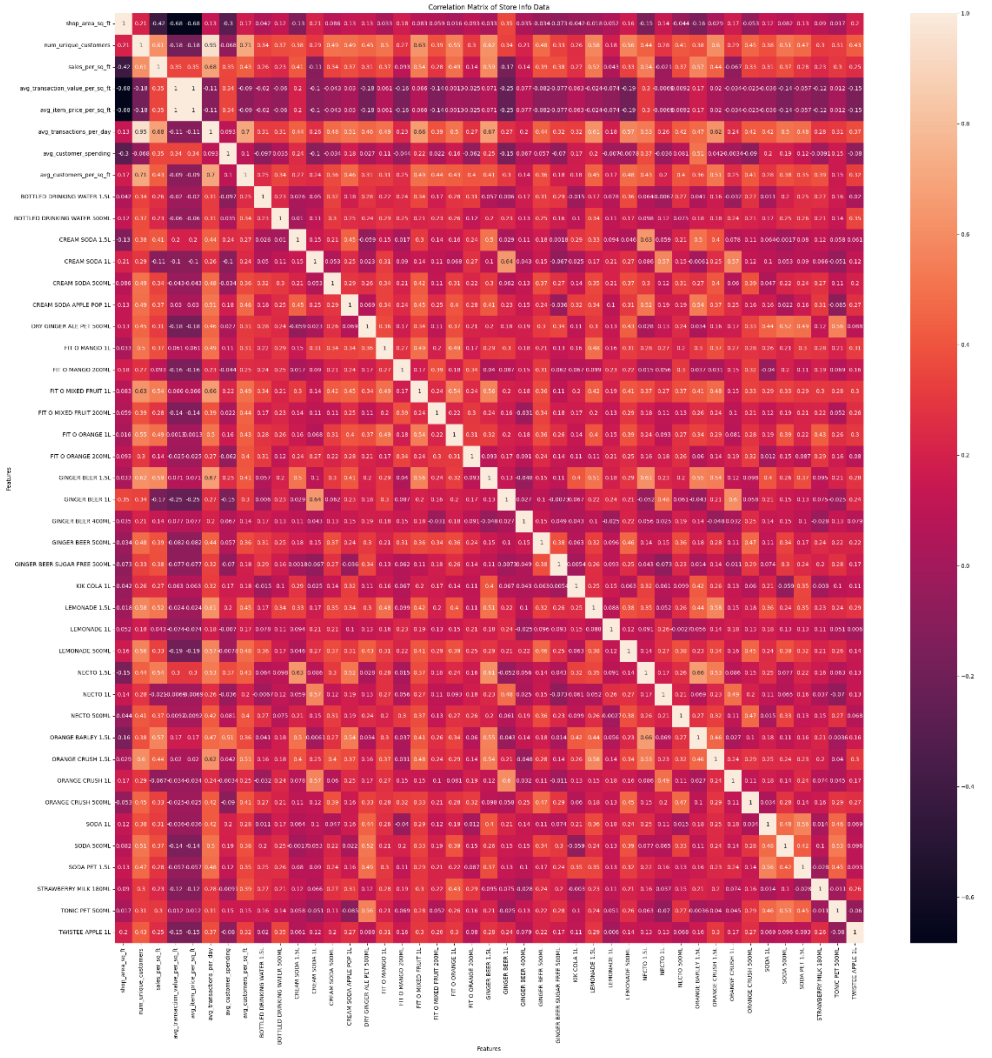
One of the main challenges of this project was to select the most relevant features for our machine learning model that could accurately classify the shops based on their business performance. To do this, we used two statistical methods: mutual information and chi square scores. Mutual information measures how much information one variable provides about another variable, while chi square scores test the independence of two categorical variables.

We calculated these scores for each feature with respect to the target variable and ranked them accordingly. We then tried various combinations of features with different machine learning models, such as Logistic Regression, K-Nearest Neighbors, Support Vector Machines, Decision Trees, and Random Forests. We evaluated each model using accuracy and f1 scores after cross validation.

We found that Random Forest Classifier performed the best among all the models, with an average accuracy and f1 scores falling between 0.6 and 0.65. We also tuned the parameters of the Random Forest model, such as

the number of trees, the maximum depth, and the minimum samples split, to further improve the average f1 scores.

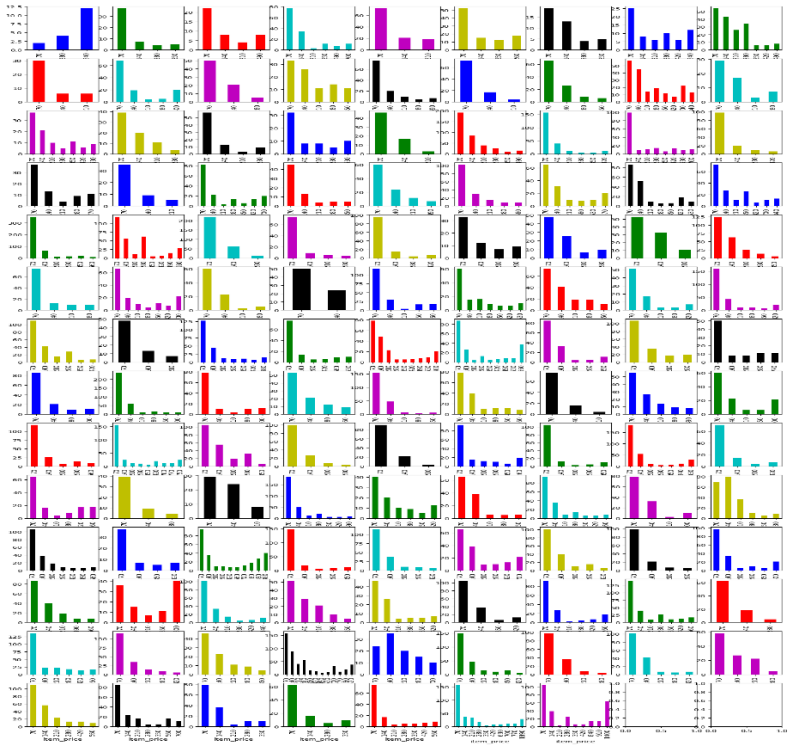
Before training the model, we performed some data preprocessing steps to ensure the quality and consistency of the data. These steps included replacing outliers with median values, filling missing values with zero in shop-item features, dropping rows with missing values in the historical transactions data, scaling the data using Min Max scaler to bring all the features to a common range, and removing some highly correlated and unusefull features.



BUSINESS INSIGHTS.

Upon conducting an observational study of the items sold in the shops, it was found that there were instances where the same item was being sold at different prices within the same shop. Furthermore, it was observed that when these items were sold at a lower price point, they generated a higher revenue for the shop.

Based on the observation that selling items at a lower price generated higher revenue for the shop, one possible business insight or recommendation could be to consider implementing a pricing strategy that involves lowering the prices of certain items to increase sales volume and overall revenue



.

An example, Variation of sales vs selling price for the sales of bottled drinking water 1.5l for each shop. This clearly shows that the sales for lower prices is very much higher than large price hence increasing revenue than larger priced bottle.

The higher sales of orange crush 1.5l compared to orange crush 1l suggest that customers prefer to buy larger quantities of the product at a higher price point, rather than smaller quantities at a lower price point. A possible recommendation based on this insight is to leverage the strong brand equity and customer loyalty of orange crush 1.5l and to explore opportunities to expand its distribution, promotion, and innovation strategies to increase its market share and profitability.

The very low sales of chocolate milk 180ml and lime crush juice 1l suggest that these products are not appealing to the target customers, or that they face strong competition from other brands or categories. This could indicate that these products have a low awareness, a poor positioning, a weak differentiation, or a high price relative to the alternatives.

The high footfall of customers on Sundays indicates that this is the most convenient or attractive day for shopping for most people. A possible recommendation based on this insight is to optimize the store operations, staff allocation, inventory management, and marketing communication on Sundays to ensure a smooth and pleasant shopping experience for the customers and to maximize the sales and profitability of the stores.