Data Science Capstone: DATS 6501_10

Final Report

Revenue Prediction & Customer Analytics for Supermarket Data

Instructor: Dr. Edwin Lo

Presented By

Purvi Jain

Sowmya Maddali

Date: 6th May 2024

ABSTRACT

Our project uses a variety of machine learning and deep learning models to forecast supermarkets' income for the following day based on a multitude of product categories. The main goal of our project is to use feature engineering techniques to improve forecasting accuracy. We were also able to learn more about customer preferences when it came to the supermarket's own brand over their competitors by conducting exploratory data analysis.

During the modeling phase, we evaluate our models based on their Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and coefficient of determination (R^2 score) to gauge their effectiveness. This approach not only streamlines operational efficiencies but also serves as a scalable model for other retail entities within the competitive United Kingdom (UK) supermarket landscape.

In the end, our project contributes to the field of retail analytics by highlighting how complex analytical methods can be used in real-world scenarios to forecast supermarket sales, enabling retailers to make better business decisions.

Table of Contents

1. Introduction

Machine Learning is being used in a variety of industries as a result of advancements in the area and an increase in processing capacity. The retail sector is not an anomaly. Using sophisticated forecasting algorithms to more accurately estimate future sales and enhance product allocation and ordering procedures is one way that machine learning is being used in the retail industry.

Enhancing forecasting models for the retail sector can yield numerous benefits. Products are more readily available to consumers, and retailers are seen as a trustworthy source of commodities. For the stores, better forecasting performance can mean less waste from overstocking, which can have detrimental economic effects; more sales because understocking can reduce sales because of low product availability; and better staffing distribution. Consequently, the stores may benefit from higher prediction accuracy in a variety of ways, including:

1 **Pricing Strategies:** Forecasts that are precise enable the successful implementation of dynamic pricing. Supermarkets have the ability to modify their prices in response to the expected variations in demand, which can be caused by various circumstances.
2 **Strategic Decision Making:** By offering insights into future trends, forecasting aids in long-term strategic planning. This can affect choices about product ranges, workforce levels, and even store layout.
3 **Enhancing Competitiveness:** Being able to anticipate and react quickly to market trends can provide supermarkets with a competitive advantage where a small margin can have a huge impact.

2. Literature Review

In time series analysis, ARIMA modeling is widely regarded as one of the most sophisticated techniques for time series forecasting within the realm of statistical learning tools. This methodology comprises three fundamental components: the Autoregressive (AR) part, the Moving Average (MA) part, and the pivotal requirement of ensuring the stationarity of the underlying time series. Assessing market efficiency is a crucial metric in determining the maturity of a given market. Although a direct correlation between market volatility and inflation may not always be apparent, extended periods of heightened market volatility have been observed to potentially trigger inflationary pressures. Notably, this volatility is not restricted to traditional financial markets but is also observable within the commodities market. This sector serves as a derivatives market for commodities, facilitating effective hedging strategies, which, in turn, can contribute to inflationary dynamics. **[1]**

Forecasting grocery items can not only avoid excessive stockpiling but also meet customer demand. This can reduce the grocery store's losses and increase the grocery store's turnover. On the other hand, considering features that affect sales is also the core of feature engineering. This paper makes a forecast about the sales of merchandise in a large chain store. To solve this problem, the author uses Machine Learning to solve this regression problem. The Light Gradient-Boosting Machine (LGBM) model was used, and the time series is divided into different data periods. The final accuracy of the model was 35%. The disadvantage of the model is that training required a very large dataset, but not all conditions were met for this project. **[2]**

Improved sales forecasting for individual products in retail stores can have a positive impact both on the environment and the economy. With the increase in computational power, there has been an interest in applying ML to such problems as well. In this project, the authors have used two years worth of sales data and weather data to investigate models such as XGBoost, ARIMAX, LSTM, and Facebook Prophet. In the end, XGBoost and LSTM outperformed all the other models, as they had the best RMSE value when compared with the other models. **[3]**

3. Dataset Description

The dataset used in this project comprises sales data from various supermarkets across the UK. The primary source of this data is available on Kaggle, which aggregates sales figures from five major UK supermarkets—Aldi, ASDA, Morrisons, Sainsbury's, and Tesco.

3.1 Original Dataset Composition

The original dataset consists of several key variables, such as:

- **Supermarket:** The name of the supermarket (Eg: Aldi, ASDA, etc)
- **Prices:** The Price of a good
- **Per unit price:** The price per unit of a good
- **Unit:** The unit the good is measured in (Eg: kg, l, etc)
- **Name:** The name of the good
- **Date:** The date the information
- **Category:** The category of the good
- **Own brand:** Whether the good is own brand or not
- **Revenue:** Overall revenue generated by the product
- **Quantity:** Number of times the product has been sold

3.2 Data Augmentation and Feature Engineering

To improve the dataset's utility, the following steps were undertaken:

I. **Handling Missing Values:**

Initial data cleaning involved addressing missing and inconsistent data entries, particularly in the 'Prices', 'Unit' and ‘Name’ fields. Specifically, missing values in price were imputed by leveraging the price data from the days immediately before and after the missing entry. However, for products lacking product names, we opted to exclude these entries from our analysis as they lacked sufficient identifiers for accurate categorization and subsequent analysis.

II. Dealing with date datatype:

To better understand underlying patterns, the date variable was decomposed into several time-related features: the date initially recorded as a string was converted into a date-time format. Subsequently, this date-time was split into discrete components—day, month, and year—along with the calculation of the week of the year. To capture and leverage cyclical patterns in the data—important for understanding daily and weekly seasonality—we incorporated trigonometric transformations, specifically the sine and cosine of the day and the week. These transformations helped our models recognize and adjust to the inherent cyclical nature of time-related data, enhancing predictive accuracy.

III. Lagged Features:

To capture temporal dependencies crucial for forecasting, lagged features reflecting past values of 'Revenue', and cyclical time patterns were implemented. These include:

- **Revenue Lag-1:** Revenue from the previous day.
- **Revenue Lag-2:** Revenue from two days prior.
- **Revenue Lag-3:** Revenue from three days ago.
- **Sine Day Lag-1:** Sine of the day from the previous period, capturing daily cyclical patterns.
- **Cosine Day Lag-1:** Cosine of the day from the previous period, to fully represent daily cycles.
- **Sine Week Lag-1:** Sine of the week from the previous period, encapsulating weekly cyclical trends.
- **Cosine Week Lag-1:** Cosine of the week from the previous period, providing a complete weekly cycle representation.

IV. Rolling Windows:

To smooth out daily fluctuations and highlight longer-term trends in revenue, rolling window calculations were employed in our data analysis. Specifically:

- **Weekly Rolling Average:** We calculated the 7-day rolling average of revenue using the most recent daily revenue (Revenue Lag-1). This moving average helps in identifying overall trends by averaging out the noise in daily revenue data.

- **Weekly Rolling Standard Deviation:** The 7-day rolling standard deviation was also computed based on Revenue Lag-1. This measure provides insights into the variability of revenue over a week, highlighting the stability or volatility of sales.

3.3 Dataset Splitting

The dataset was divided into a training set comprising 10% of the data and a test set containing the remaining 90%. This split was strategically implemented after preprocessing to prevent data leakage. Within the training set, data was further partitioned into smaller training and validation sets to rigorously evaluate and refine the model's accuracy.

4. Project Objectives

The goal of this project was to leverage advanced machine learning techniques to predict next-day revenue across various supermarket categories. The specific objectives are detailed below:

Objective 1: Predicting Next-Day Revenue

The primary objective was to develop a predictive model capable of forecasting the next-day revenue for different product categories. This involves:

1. **Accurate Forecasting:**

   Utilized historical sales data to forecast revenue with high accuracy, enabling supermarkets to anticipate demand and adjust their operations accordingly.

2. **Category-Specific Insights:**
   Tailored predictions are made for individual categories such as fresh foods, beverages, and household items to address the unique demand dynamics of each category.

Objective 2: Comparing the Performance of different Models

To determine the most effective predictive model, a comparative analysis of various algorithms has been conducted. This includes:

1. **Model Selection:**

   We evaluated a range of models, such as linear regression, random forest, XGBoost, ANN, and LSTM to understand their strengths and limitations in the context of revenue forecasting.

2. **Performance Metrics:**
   Utilized key metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2) to quantify each model's accuracy and predictive power.

Objective 3: Identifying Key Features That Influence Revenue Figures

Knowing which factors have the most effects on revenue estimates is essential for improving model performance.

5. Exploratory Data Analysis (EDA)

In order to fully understand our data, an initial data analysis was conducted. The objective of this analysis was to get a deeper understanding into how the different variables behaved throughout the time period available, how the sales of different products compared to each other, and lastly to detect possible missing values and outliers. By performing EDA it was possible to find underlying patterns that were not apparent during the initial look at the data.

5.1 Statistical Summaries

We started the EDA with statistical summaries to describe the central tendencies, variability, and distributions of key variables within the dataset:

1. **Descriptive statistics:**

   For each variable, such as price, revenue, and quantity, summary statistics such as mean, median, standard deviation, minimum and maximum values were computed. This helped identify ranges and typical values for sales and prices across different categories.

2. **Correlation Analysis:**

   Correlation coefficients between numerical features were calculated to understand the relationships among them.

3. **Distribution Analysis:**
   The distribution of important features, like revenue and units sold, was analyzed using histograms. This helped identify the skewness of the data and the typical sales volume per category.

5.2 Visualizations

Visual analysis was conducted to identify trends, cycles, and unique distributions within the data:

1. **Time Series Trends:** Line graphs were used to visualize revenue trends over time.

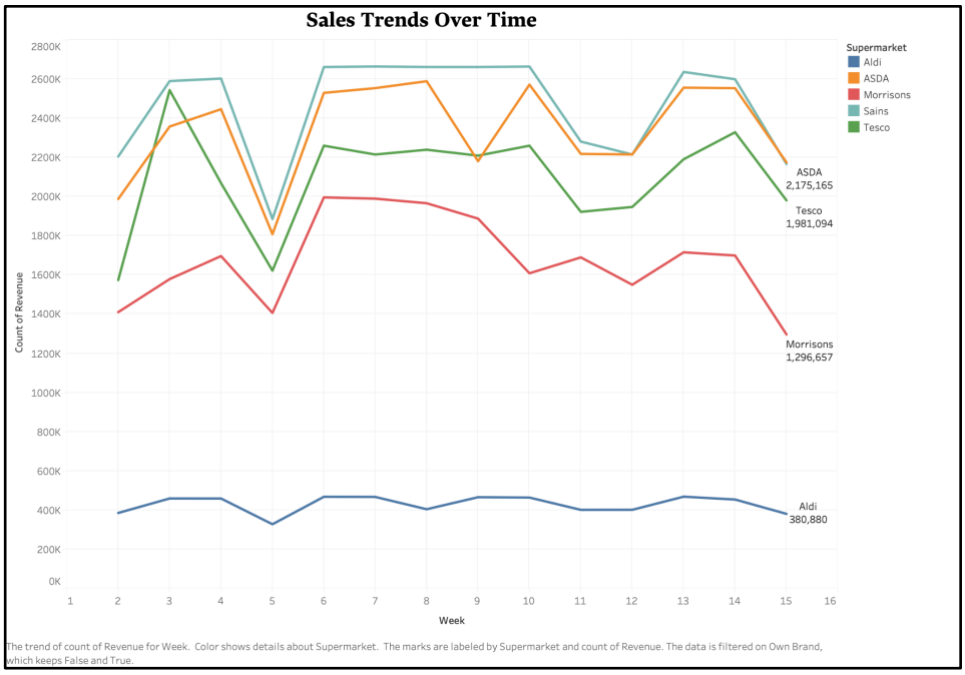

Figure 1 - Sales trends over time with respect to all 5 supermarkets

From figure 1, we can see that ASDA consistently surpasses the other supermarkets when it comes to revenue, with Tesco trailing close behind. From this, we can say that ASDA has a very strong customer base, which supports its dominant position. On the other hand, Aldi consistently posts the lowest revenue, which suggests that the customer base might spend less overall, and it also represents a solid but limited market segment.

2. **Category Comparison:** Pie charts were employed to compare the revenue contributions of different product categories.

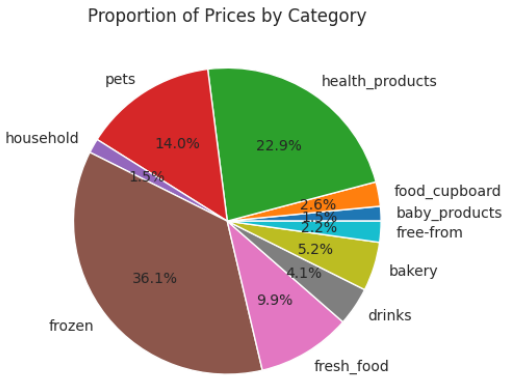

Figure 2 - Proportion of revenue by category for Aldi

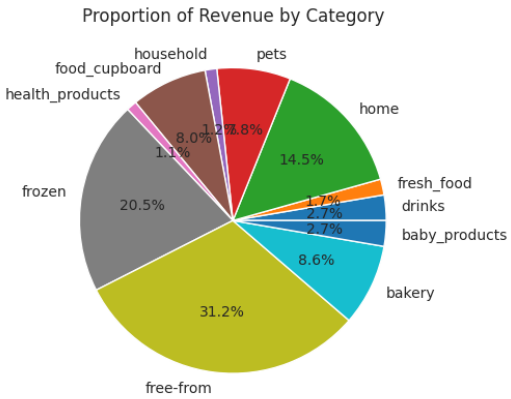

Figure 3 - Proportion of revenue by category for ASDA

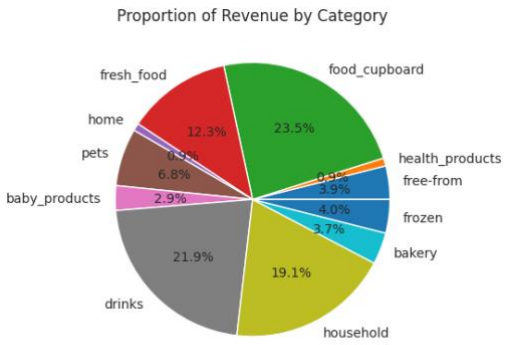

Figure 4 - Proportion of revenue by category for Morrisons

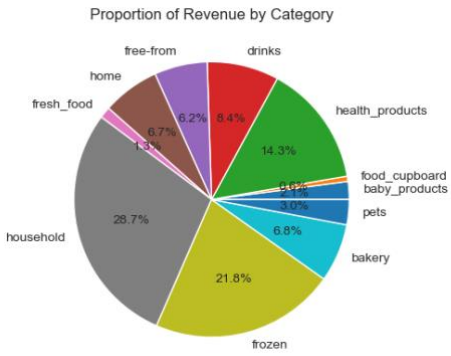

Figure 5 - Proportion of revenue by category for Sainsbury

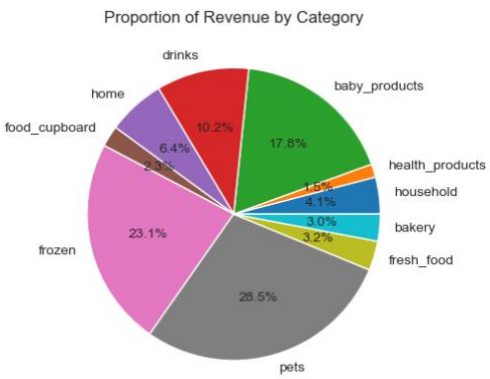

Figure 6 - Proportion of revenue by category for Tesco

- From figure 2, we notice that the frozen category from Aldi supermarket takes the lead at 36.1% out of all the other categories, which indicates that frozen products, often essential to a household, are the top selling categories among the retailer's varied assortments. Though Sainsbury, and Tesco have frozen categories among the top 5 of their highest selling categories (from figures 4 and 6), ASDA has a slightly higher rate than them.
- From figure 3, we can observe that over 31% of ASDA's revenue comes from free-from products, which may lead to the observation that their customer base is opting for dietary-specific products or being more health conscious. It is also noted that the free-from products are relatively priced lower at Aldi supermarket than the others.
- Moving to Morrisons supermarket, from figure 4, it can be seen that the variety of fresh and pantry products is the highest sold category not only within the supermarket alone but also when compared with the other 4 supermarkets as well. Along with the fresh food and food cupboard categories, Morrisons also has a high number of products within the beverages section, closely followed by the household products category.

- Lastly, Tesco's pets category from Figure 6 has the highest percentage of pet supplies sold when compared to all the other categories and the other supermarkets as well.

3. **Stacked bar graph for own brand products of supermarkets:** A stacked bar chart was developed to help view how much each category has earned within each supermarket when it comes to their own branded products.

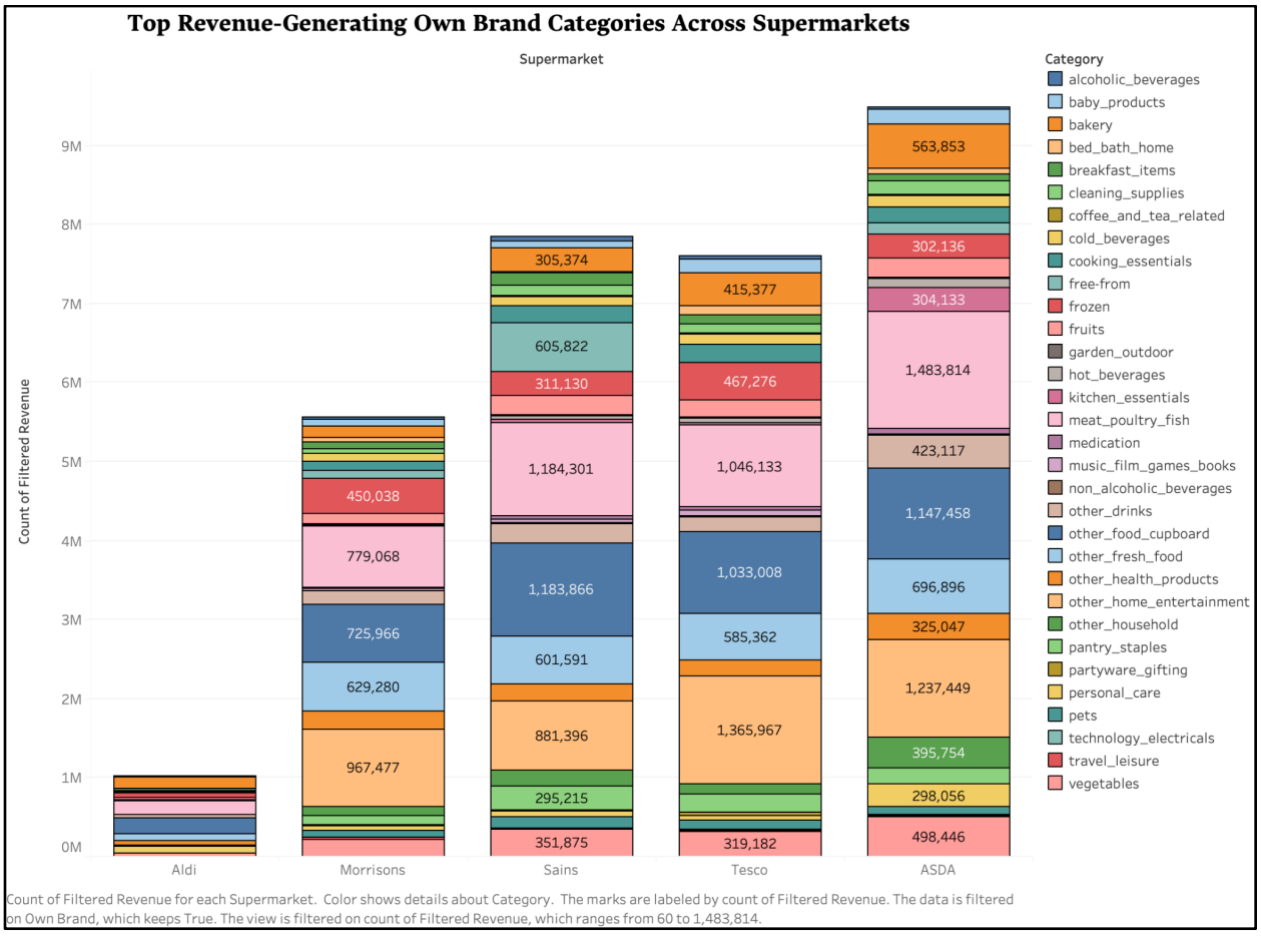

Figure 7 - Top Revenue generating own brand categories across all 5 supermarkets.

- From figure 7, we can see that ASDA supermarket's own brand products generate a higher revenue when compared with the other supermarkets, and the quantity of products present within each of those categories is also varying.

- The meat, poultry, and fish categories from all 3 Sainsbury, ASDA, and Tesco generated the highest revenue within their own branded products. Aldi supermarket did have the same category as its competitors but was not able to generate the same amount of revenue.
- In ASDA and Sainsbury, meat, poultry, and fish turned out to be their highest revenue generating categories, while in Tesco and Morrions, it was their home entertainment that brought them the highest revenue.
- Trailing behind is the food cupboard category among all the 5 supermarkets, which was the third most revenue generating category. Sainsbury had the highest number of food cupboard products, such as canned foods, sold among the 5, closely followed by ASD, Tesco, Morrisons, and Aldi.
- Overall, when it came to the supermarket's own products, ASDA dominated when compared with the others when it came to revenue generation as well as the quantity of the products.

6. Feature Engineering

Feature engineering is a critical step in the predictive modeling process, involving the creation, transformation, and extraction of data features. Key components include feature creation from existing data, transforming and imputing missing or invalid features. This process was essential for adapting the raw supermarket data into a format suitable for machine learning algorithms, ultimately enhancing their ability to predict next-day revenue. **[4]**

6.1 Encoding Categorical Variables

Given the categorical nature of product category features in our dataset, appropriate encoding techniques were necessary:

1. **One-Hot Encoding:**

   This method was employed for nominal categorical variables with no intrinsic ordering, such as product categories. By transforming these categories into binary vectors, we ensured that models could utilize this categorical data without imposing any ordinal assumptions.

2. **Label Encoding:**

   For ordinal categorical variables, such as product categories, label encoding was applied. This technique assigns each unique category a numerical value.

6.2 Feature Selection Rationale

The selection of relevant features was guided by both statistical analysis and domain knowledge:

1. **Time-Related Features:** Recognizing the temporal dynamics of supermarket sales, features like day of the week, and month were included.
2. **Lagged Features:** To incorporate the time series nature of the data, lagged features such as previous day's revenue and lags of cyclic patterns were created.

3. **Price and Volume Interactions:** Interaction terms between price and units sold were considered to capture the effect of pricing strategies on sales volume.

6.3 Applying Feature Engineering Techniques

The application of these feature engineering techniques involved several steps:

1. **Data Transformation Scripts:** Python scripts were developed to automate the encoding and generation of new features, ensuring consistency and efficiency in processing.
2. **Exploratory Data Analysis Revisited:** After feature engineering, a second round of EDA was conducted to assess the impact of the new features.
3. **Feature Importance Evaluation:** Using initial model fits, feature importance was evaluated to prune irrelevant or redundant features, optimizing model complexity and performance.

6.4 Recategorization of Product Categories

6.4.1 Overview

During the data preprocessing stage, we identified a critical requirement for recategorization of product categories. This process was vital to reducing variability in category naming, which previously included overly specific or slightly different naming for similar items, thereby complicating any comparative analysis across the dataset**.**

6.4.2 Methodology

We first developed a systematic approach by identifying common keywords within product names that could signify broader category affiliations. Using Python and natural language processing libraries, we designed a script to automatically assign new category labels based on the identified keywords. This automation was crucial for handling the extensive dataset efficiently.

6.4.3 Impact and Benefits

The recategorization yielded several significant benefits:

1. **Improved Data Clarity and Usability:** The new categorization greatly enhanced the clarity of the dataset, making it more navigable and easier to analyze. It allowed for more accurate comparisons across product types and simplified the aggregation of sales data.
2. **Enhanced Analytical Accuracy:** By standardizing product categories, we reduced the noise in our data, leading to more reliable insights from our predictive models.

7. Models

7.1 Machine Learning Models

7.1.1 Linear Regression

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The method assumes that the dependent variable can be approximated by a linear combination of the independent variables, adjusted for a random error. The primary goal of linear regression is to identify the best linear model that minimizes the discrepancies between observed values and values predicted by the model. **[5]**

Equation:

$$Y = \beta_0+\beta_1X_1+\beta_2X_2+...+\beta_nX_n+\epsilon \quad -(1)$$

Where,

Y: This is the dependent variable.

β_0: This is the intercept of the regression line, representing the predicted value of Y when all independent variables are zero.

β_1, β_2, ..., β_n: These are the coefficients for the independent variables.

X_1, X_2, ..., X_n: These are the independent variables or predictors.

Role in this project

Linear regression is a statistical modeling technique used to predict a continuous outcome variable, in this case, next-day revenue for each category in supermarkets, based on predictor variables. The model assumed a linear relationship between the dependent variable (revenue) and independent variables. It aimed to fit a linear equation to observed data that minimized the difference between the observed values and the values predicted by the linear equation.

7.1.2 Random Forest

A Random Forest is an ensemble learning method that operates by constructing multiple decision trees during the training phase and outputting the class mean prediction (regression) of the individual trees. Random forests aim to reduce the overfitting tendency of decision trees by averaging multiple trees trained on different parts of the same training set, thereby improving predictive accuracy and robustness. It is particularly effective in handling large datasets with high dimensionality and complex data structures. **[6]**

Equation:

$$Y = 1/B \sum^{B}_{b=1} T_b(x) \quad – (2)$$

Where,

Y: This is the predicted outcome for the input x

B: The total number of decision trees in the forest

T_b: The prediction made by the b-th decision tree in the forest

x: The vector of input features used for making predictions

Role in this project:

Random Forest model leverages an ensemble of decision trees to forecast sales outcomes. Each tree in the forest was built from a random subset of data and features, and the final revenue prediction was made by averaging the predictions from all trees. This methodology not only helped in capturing more complex patterns in the data but also reduced the risk of overfitting, making it highly suitable for handling the diverse and multidimensional data typically found in supermarket sales datasets.

7.1.3 Extreme Gradient Boost (XGBoost)

XGBoost (Extreme Gradient Boosting) is an advanced implementation of gradient boosting algorithms, designed to be highly efficient, flexible, and portable. It uses a gradient boosting framework to build sequential models that correct the residuals of previous models, thereby improving prediction accuracy incrementally. XGBoost employs several regularization techniques to prevent overfitting, making it robust, and it is well-known for its ability to handle sparse data,

scale to large datasets, and achieve high performance on various types of predictive modeling tasks. **[7]**

Equation:

$$Obj(\Theta) = \sum^{n}_{i=1} l(y_i, \hat{y}_i) + \sum^{k}_{k=1} \Omega(f_k) \quad – (3)$$

Where,

$l(y_i, \hat{y}_i)$ is a differentiable convex loss function that measures the prediction error between the actual output and the predicted output from the ensemble of trees.

$\Omega(f_k)$ represents the regularization term, which penalizes the complexity of the model.

K is the number of trees

Θ represents the parameters of the trees.

Role in this project:

XGBoost was particularly suited due to its capability to model complex patterns and interactions deeply. Its robustness to overfitting, even with sparse data and numerous features, made it an excellent choice for ensuring reliable forecasts that were crucial for pricing strategies in supermarkets.

7.2 Deep Learning Models

7.2.1 Artificial Neural Networks (ANN)

An artificial neural network is a collection of simple, interconnected algorithms that process information in response to external input. They are capable of learning from data through a process that adjusts the weights of connections, optimizing the network to perform tasks such as classification, regression, and pattern recognition without being explicitly programmed to perform the task. **[8]**

Equation:

$$Y = f(\sum^{n}_{i=1} w_i x_i + b) \quad – (4)$$

Where,

x_i is the input signals

w_i is the weights

b is the bias

f is the activation function

y is the output

Role in this project:

ANN was employed to model complex nonlinear relationships in the data that simpler models might have missed. The network learned to integrate signals from various inputs, such as historical sales data, pricing trends, and product categories. By adjusting weights through training, ANN optimized predictions for each category's revenue.

7.2.2 Long Short-Term Model (LSTM)

Long Short-Term Memory (LSTM) networks are a type of recurrent neural network (RNN) capable of learning order dependence in sequence prediction problems. Unlike standard feedforward neural networks, LSTMs have feedback connections that make them capable of processing entire sequences of data (e.g., a time series). This makes LSTMs well-suited to tasks where context from the input data is essential for making predictions. The key innovation of LSTMs is their ability to remember information for long periods, which is achieved through a complex gating mechanism that regulates the flow of information to be remembered or forgotten during the learning process. **[9]**

Role in this project:

AN LSTM network was particularly beneficial for capturing temporal dynamics and dependencies across sales data. By processing sequences of daily sales figures, the LSTM model was able to learn patterns that influenced revenue fluctuations over time, such as seasonal trends, and weekly cycles. This capability allowed the LSTM to forecast future sales with a higher degree of accuracy compared to models that do not account for time dependencies.

7.3 Time Series Models

7.3.1 Autoregressive Integrated Moving Average (ARIMA)

In time series analysis, ARIMA modeling is widely regarded as one of the most sophisticated techniques for time series forecasting within the realm of statistical learning tools. This methodology comprises three fundamental components: the Autoregressive (AR) part, the Moving Average (MA) part, and the pivotal requirement of ensuring the stationarity of the underlying time series. **[1]**

Equation:

$$Y_t = c + Ø_1 Y_{t-1} + Ø_2 Y_{t-2} + ….. + Ø_p Y_{t-p} + Θ_1 \epsilon_{t-1} + Θ_2 \epsilon_{t-2} + …. + Θ_q \epsilon_{t-q} + \epsilon_t \quad – (5)$$

Where,

Y_t is the value of the time series at time t

c is a constant term

$Ø_1, Ø_2, ...Ø_p$ are the autoregressive parameters representing the relationship between the current observation and the p previous observations.

$Y_{t-1}, Y_{t-2}, ..., Y_{t-p}$ are the lagged values of the time series

$Θ_1, Θ_2, ..., Θ_q$ are the moving average parameters representing the relationship between the current observation and the q previous error terms.

$\epsilon_{t-1}, \epsilon_{t-2}, ...\epsilon_{t-q}$ are the lagged error terms.

ϵ_t is the error term at time t

Role in this project:

The ARIMA model plays a crucial role in our project by providing a powerful framework for forecasting future revenue, in this case, next-day revenue, based on historical data from supermarkets. By fitting the model to historical data, we are able to generate forecasts for the very next day, allowing supermarkets to anticipate demand and plan inventory accordingly. We are also able to examine how past sales performance impacts future sales by using the autoregressive and moving average components of the model.

8. Results

Models	R-square	MAE	RMSE
Linear Regression	Train: 0.88 Test: 0.85	Train: 5.31 Test: 5.37	Train: 14.98 Test: 15.78
Random Forest	Train: 0.93 Test: 0.90	Train: 2.34 Test: 2.39	Train: 10.98 Test: 11.08
XGBoost	Train: 0.94 Test: 0.91	Train: 3.87 Test: 3.76	Train: 13.68 Test: 13.55
ANN	Train: 0.92 Test: 0.90	Train: 3.65 Test: 3.60	Train: 11.93 Test: 11.68
LSTM	Train: 0.92 Test: 0.85	Train: 3.65 Test: 5.17	Train: 11.93 Test: 16.43
ARIMA	Train: 0.70 Test: 0.67	Train: 6.30 Test: 6.40	Train: 8.44 Test: 7.40

Table 1: Model Outputs

8.1 Analysis:

- With a R-square of 0.91, XGBoost performed the best overall, closely followed by Random Forest and ANN with a R-square value of 0.90.
- Though XGBoost has the highest R-square value, it also has the highest RMSE and MAE value when compared to the other two models. Random Forest, although has the second best R-square value, has the lowest RMSE and MAE value, closely followed by ANN.
- Therefore, we arrived at the conclusion that using ANN and Random Forest models might greatly improve short-term revenue and give supermarkets an effective tool for pricing.

8.2 Future Goals:

- Introducing more factors to the models, like consumer demographics, market trends, and economic indicators, may increase their accuracy and robustness.
- Additionally, we are planning to incorporate weather data to provide additional insights into how the products are being sold and how many customers the supermarkets were able to attract on a specific day.

9. Streamlit

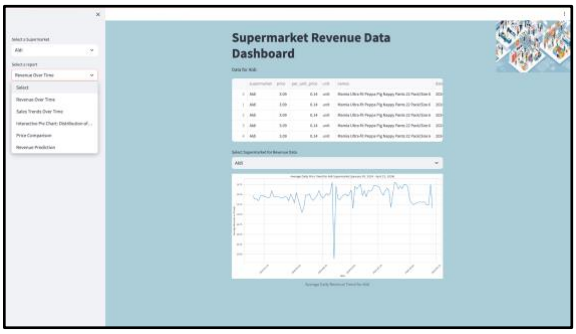

Figure 8 - Supermarket Dashboard with Revenue over time

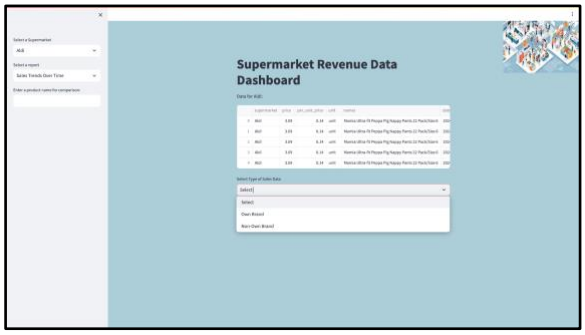

Figure 9 - Selecting a supermarket for Revenue Trends over Time

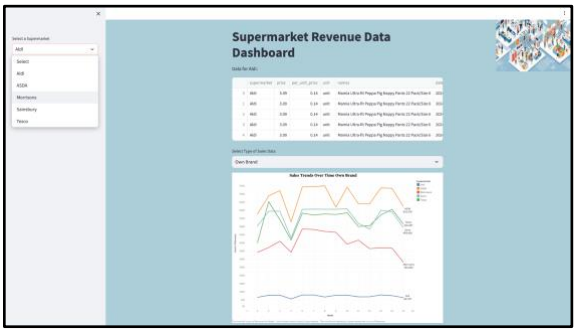

Figure 10 - Viewing all the supermarkets revenue trends

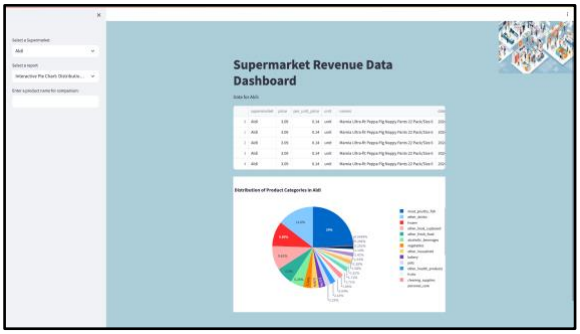

Figure 11 - Provided an interactive pie chart to visualize the categories.

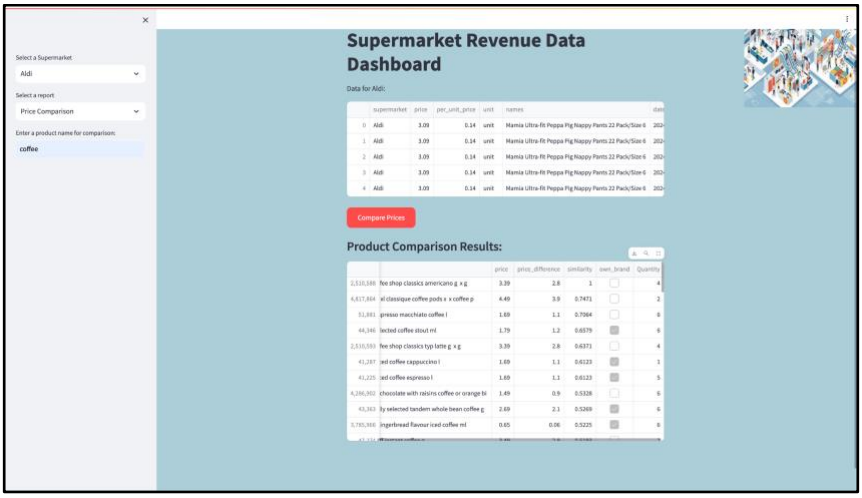

Figure 12 - Price comparison for a specific product

In our project, we have developed a comprehensive Streamlit application that serves as an interactive platform for analyzing supermarket data.

9.1 Application Features:

- The application offers a simple and intuitive interface where users can select supermarkets and report types they wish to analyze.
- Users can select the supermarket of interest, and the application dynamically loads the corresponding dataset. This allows for flexible analysis across different datasets, including revenue data and price comparisons.
- The application also includes an option to view revenue trends over time for selected supermarkets.
- Users can explore the distribution of product categories within a supermarket through our interactive pie chart (Figure 11). This feature helps in understanding which categories contribute most to the revenue of a supermarket.
- A significant feature of our application is the price comparison tool (Figure 12). It allows users to input a product name to compare prices across brands within the supermarket. This feature uses TF-IDF vectorization and cosine similarity to find similar products and displays a comparison of prices, helping users identify price discrepancies and make informed purchasing decisions.

Our Streamlit application effectively integrates various analytical tools and visualizations to provide a comprehensive platform for supermarket data analysis. It is designed to assist stakeholders in making data-driven decisions by providing real-time insights into revenue trends, price comparisons, and product category distributions. This application demonstrates the practical application of data analytics techniques in the retail industry, enhancing business operations and strategic planning.

References

1. U.S. Bureau of Labor Statistics, Consumer Price Index for All Urban Consumers: All Items in U.S. City Average [CPIAUCSL], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/CPIAUCSL, October 10, 2023.
2. Liu, Y. (2022/04/18). Grocery sales forecasting. Paper presented at the 215-219. https://doi.org/10.2991/aebmr.k.220404.040 https://www.atlantis-press.com/proceedings/cike-22/125972906
3. Fredén, D., & Larsson, H. (2020). Forecasting daily supermarkets sales with machine learning
4. Feature engineering (2024).
5. Gareth James, Daniela Witten, Trevor Hastie, & Robert Tibshirani. (2021). An introduction to statistical learning (Second Edition ed.). Springer.
6. Breiman Leo. (2001). Random forests | machine learning. Springer, Volume 45, 5-32. https://link.springer.com/article/10.1023/a:1010933404324
7. Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), DOI: 10.1145/2939672.2939785
8. Jeske, S.What is artificial neural network (ANN) - definition from MarketMuse blog. MarketMuse Blog. Retrieved May 5, 2024, from https://blog.marketmuse.com/glossary/artificial-neural-network-ann-definition/
9. Brownlee, J. (2017, -05-23T19:00:16+00:00). A gentle introduction to long short-term memory networks by the experts. https://machinelearningmastery.com/gentle-introduction-long-short-term-memory-networks-experts/