

# Assignment

# ML as a Service

---

2

Md Sadman Sakib  
StudentID:25091570

7<sup>th</sup> October 2024

Github Username	Md-SadmanSakib
Github Repor	Project Repo: <a href="https://github.com/Md-SadmanSakib/adv_mla_at2_exp">https://github.com/Md-SadmanSakib/adv_mla_at2_exp</a> API Repo: <a href="https://github.com/Md-SadmanSakib/adv_mla_at2_api">https://github.com/Md-SadmanSakib/adv_mla_at2_api</a>
API URL	API: <a href="https://sales-prediction-app-euat.onrender.com">https://sales-prediction-app-euat.onrender.com</a>

36120 - Advanced Machine Learning Application  
Master of Data Science and Innovation  
University of Technology of Sydney

## Table of Contents

1.	Executive Summary	2
2.	Business Understanding	3
a.	Business Use Cases	3
3.	Data Understanding	5
4.	Data Preparation	10
5.	Modeling	11
a.	Approach	11
6.	Evaluation	12
a.	Evaluation Metrics	12
b.	Results and Analysis	12
c.	Business Impact and Benefits	12
d.	Data Privacy and Ethical Concerns	13
7.	Deployment	15
8.	Conclusion	16
9.	References	17

## 1. Executive Summary

In the 21st century, artificial intelligence is becoming a part of our life. Now, in every aspect of life, we are relying on intelligent systems. Traditional approaches are becoming less important. After the boom of AI and automation, people are becoming more advanced in their thinking and creation. Those who are going with the trend are leveraging its benefits.

This report will elaborate on the results of two models. The first one is the predictive model, which is specialized to categorize all product sales of the superstore but targets specifically to show which products generate the most revenue on a specific date. This model helps us determine the demand for particular products or items which are most desired by consumers on specific dates. As a result, this helps the business focus on which products to display more for sales and which items should be replaced.

The second model is the forecasting model which is used to predict the sales revenue for the upcoming week from the store. Both models can benefit other stakeholder groups apart from the business itself. The competitors in the market can avail themselves of the results of this business model and distinguish their product line in order to enhance sales and, as a result, benefit from higher revenue and profits. Start-up firms planning to enter the same market can also benefit from this model and its results. It will help guide new firms in the market to determine which products are more profitable and which are not worthy of display. This way, businesses can utilize their valuable time and resources for better accomplishment in attaining higher revenues, which will result in business profits and growth in the long run.



## 2. Business Understanding

### a. Business Use Cases

The predictive model can predict the total revenue that will be generated in specific day of the super store. This model can help the business understand the demand of particular product that is mostly required to be available by customers in the store on that upcoming date. This will as a result help the business have availability of that specific item beforehand so that they do not face a lack in supply on that date. As the super stores usually store all their goods in warehouse, there are limitation to space. So, by storing product which is most needed and demanded by customers they can utilise maximum amount of storage of the certain product so that when demand arises the supply can be met. This model can help the business address issues such as remove goods from warehouse which are of no demand. The storage spaces can therefore be used to replace with profitable products of the business. Businesses such as start up firms can be of great benefit from such model and the results displayed by the model. As they can determine which products are of more demand in the market and sort their business or store shelves accordingly. The forecasting model determines the revenue generation from all stores in the upcoming week. This helps the business understand how much revenue will be generated in the upcoming week from all its sales.

The project is motivated by the challenges and limitations of warehouse space, transportation of goods, and stockout problems. Stockout problems negatively impact any company's reputation. Both models will create opportunities for businesses to optimize their operations, reduce costs, and improve customer satisfaction. Machine learning models can process large amounts of data and are able to find complex patterns. Putting these models in production will help stakeholders make decisions instantly.



## b. Key Objectives

The main objective of this project is to make the business profitable, make the retail store more reliable to the customers, and optimize inventory management. The main stakeholders are current superstores, small shoppers, and potential market competitors. The project will try to meet stakeholders' requirements with the help of advanced predictive and forecasting models.

New businesses are often the ones to have a hard time gaining customer reliability. If the start up businesses can understand the market beforehand and the demands of customers that will help them in gaining customer satisfaction which will also enhance the revenue generation of the new business. Often times we see new businesses having a hard time in reaching their profits due to no being able to gain customer attention and reliability. This model will help sort those problems as well and help the start ups enjoy the essence of profitability from the beginning. This model as a result helps the business understand its expenses and cut down on areas of expenditure which can cause a lacking in profitability of the business. In order to enhance sales products which are most sold should be stocked up in bulk quantities in order to ensure enough supply to meet demands of customers. Usually super store's stock up their goods in the warehouse and if they know which products are of more demand they can stock up accordingly. Using such advanced sales forecast businesses can predict their areas of growth and work more into those area to increase sales and consumer satisfaction. These results additionally help the firm's understand their customer's needs and requirements and deliver service accordingly. At the end a happy customer is what gains customer loyalty which as a result benefits the business from higher sales and revenue generation in the long run. Competing in the market is one way of attaining customers but when the business understands its customer and provides service accordingly that's what helps the business remain in the competitive market and retain its customer loyalty.



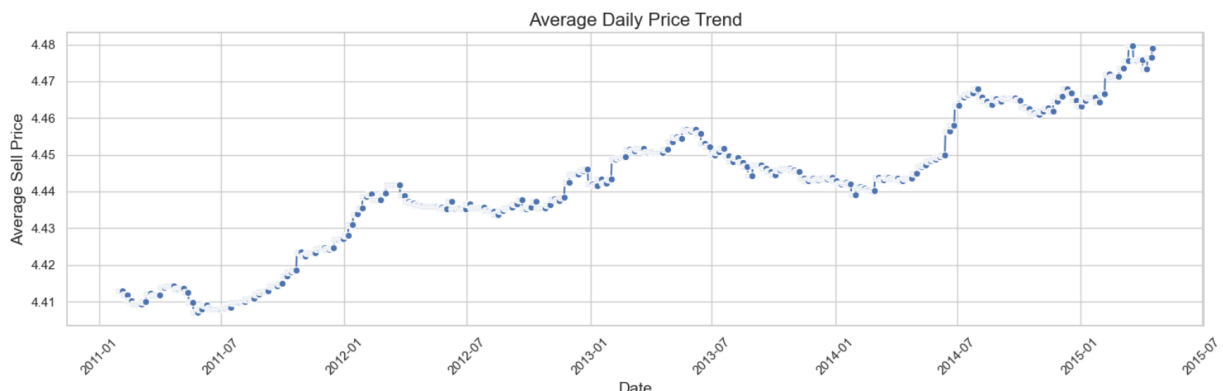
### 3. Data Understanding

This project utilized data from an American retail company across three states: California (CA), Texas (TX), and Wisconsin (WI), encompassing 10 stores in those states. Five CSV files were obtained in retail-style format. For the two use cases of prediction and forecasting, this dataset is sufficient to utilize the model. The data was collected from a website and covers the time period from 2011 to 2015. After carefully merging all these files, the resulting dataset contains 59 million rows. The training dataset comprises over 46 million rows, and the testing dataset comprises over 12 million rows. The dataset includes 13 features, each containing different unique values. Key features, sales and sell\_price were used to create the target variable, revenue, by multiplying their values. This revenue feature serves as the target for both predictive and forecasting models. The date column provides a crucial time index, benefiting both models. The number of unique values for other features in the dataset are detailed below.

Figure 1. Unique values of Each Feature

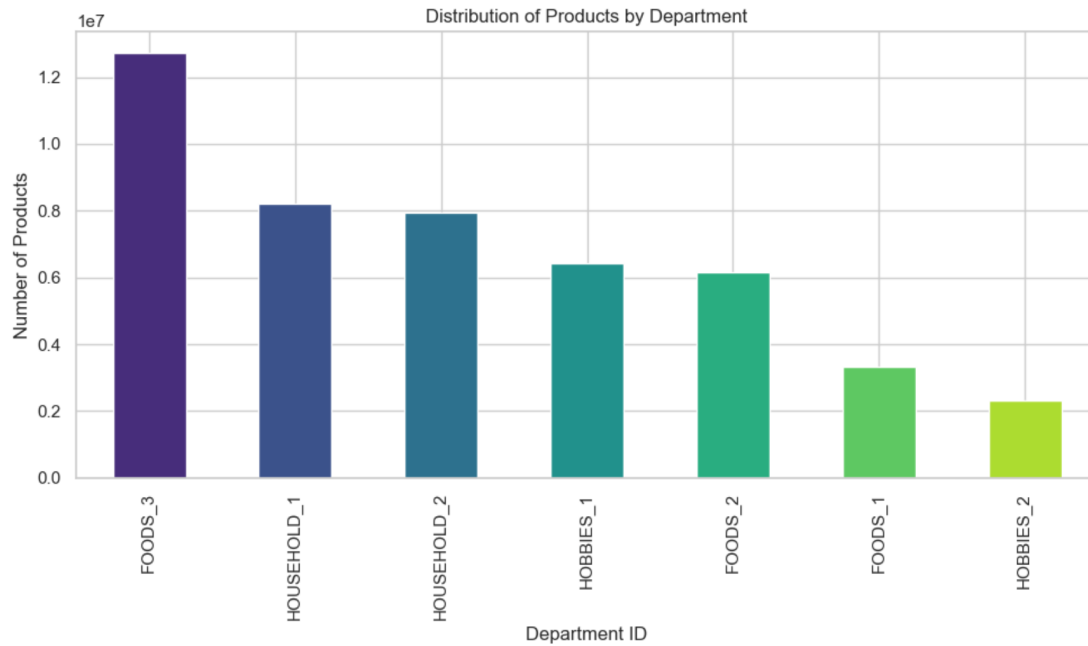
```
Column 'id' has 30490 unique values.  
Column 'item_id' has 3049 unique values.  
Column 'dept_id' has 7 unique values.  
Column 'cat_id' has 3 unique values.  
Column 'store_id' has 10 unique values.  
Column 'state_id' has 3 unique values.  
Column 'd' has 1541 unique values.  
Column 'date' has 1541 unique values.  
Column 'event_name' has 31 unique values.  
Column 'event_type' has 5 unique values.
```

Figure 2. Average Daily Price Trend.



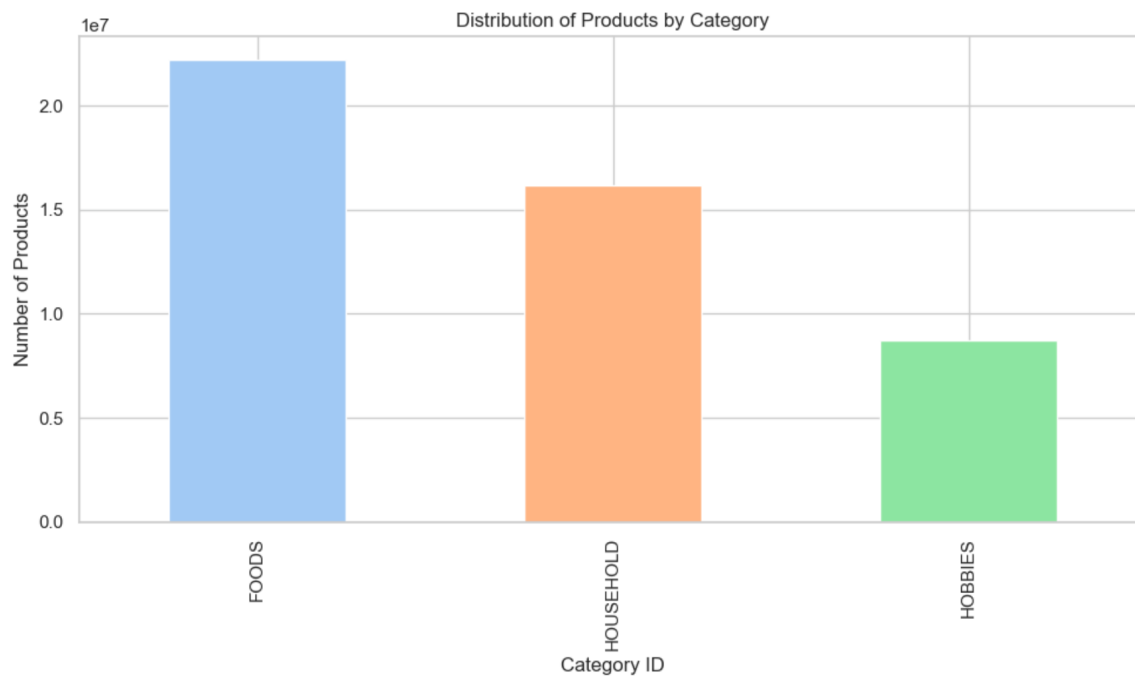
In this dataset we have some limitation. the date columns start from 2011 to 2015 we don't have the recent years data. Recent years data will capture the recent trends, and the items demands.

Figure 3. Distribution of Products by Department ID



There are seven Department ids in this retail company. Foods\_3 offered the most items, followed by Household\_1.

Figure 4. Distribution of Products by Category ID



In this dataset, the retail company has three categories they are FOODS, HOUSEHOLD, and HOBBIES. The visualization shows that stores primarily offer food and household items.

Figure 5. Distribution of Products by State

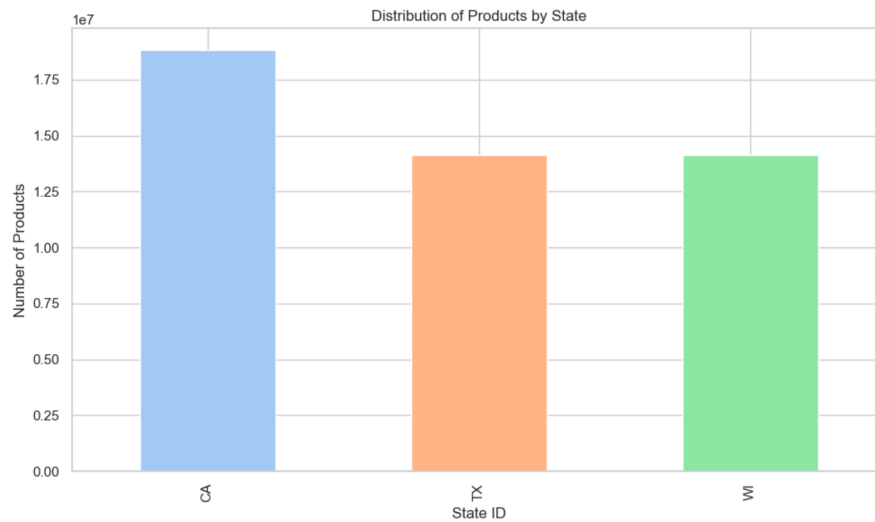
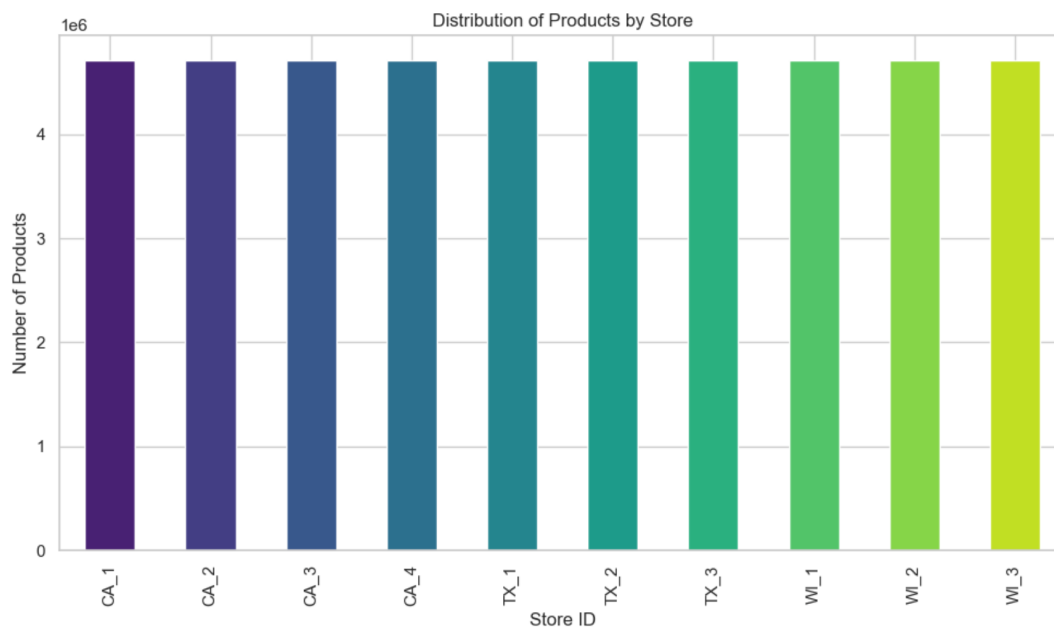


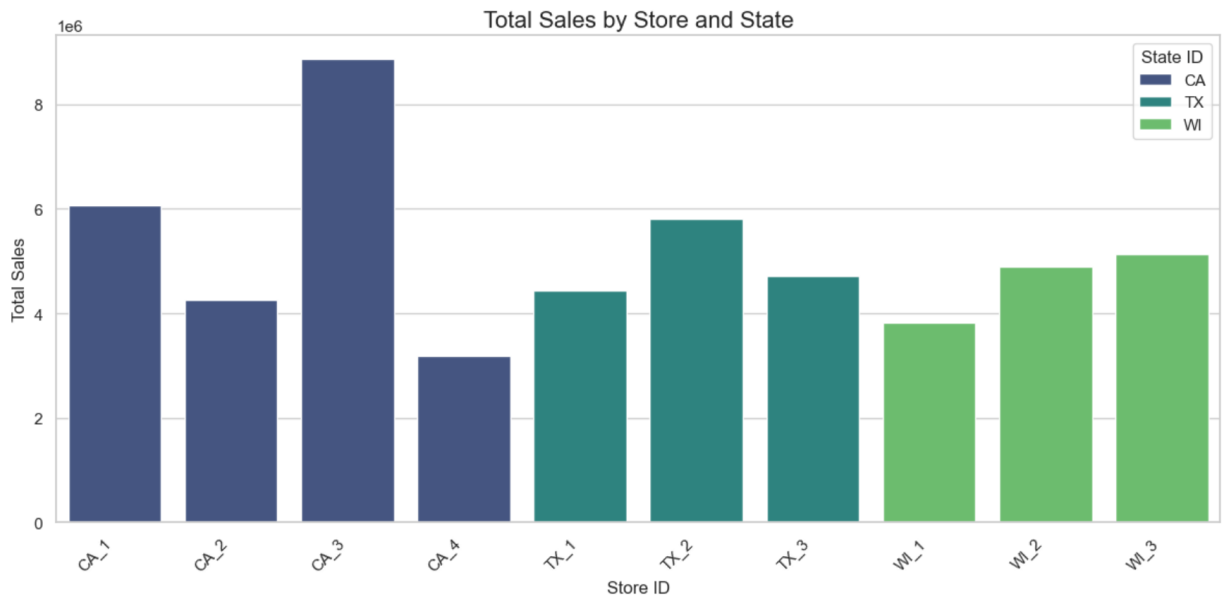
Figure 6. Distribution of Products by Store



The bar chart shows the uniform distribution of all products across all store IDs, with each store having the same 3049 unique products and their average sales entries per store: 4710705. This indicates that the company maintains consistent product offerings in all its stores. CA (California) has 4 stores because of that it has total highest number of products.

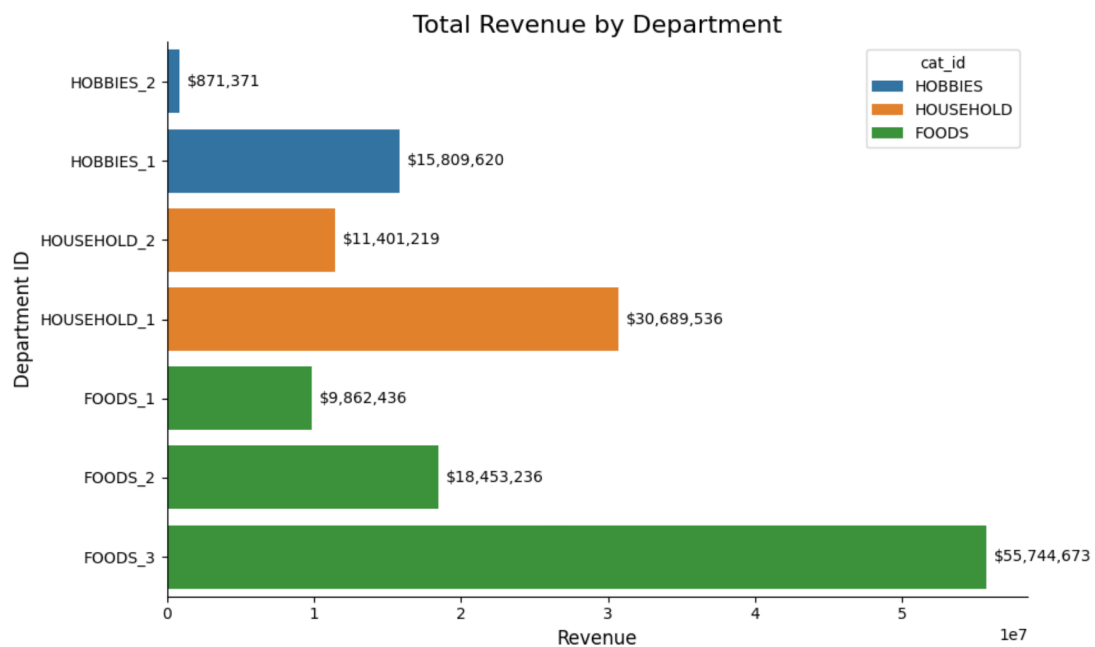


Figure 7. Totals sales by Store



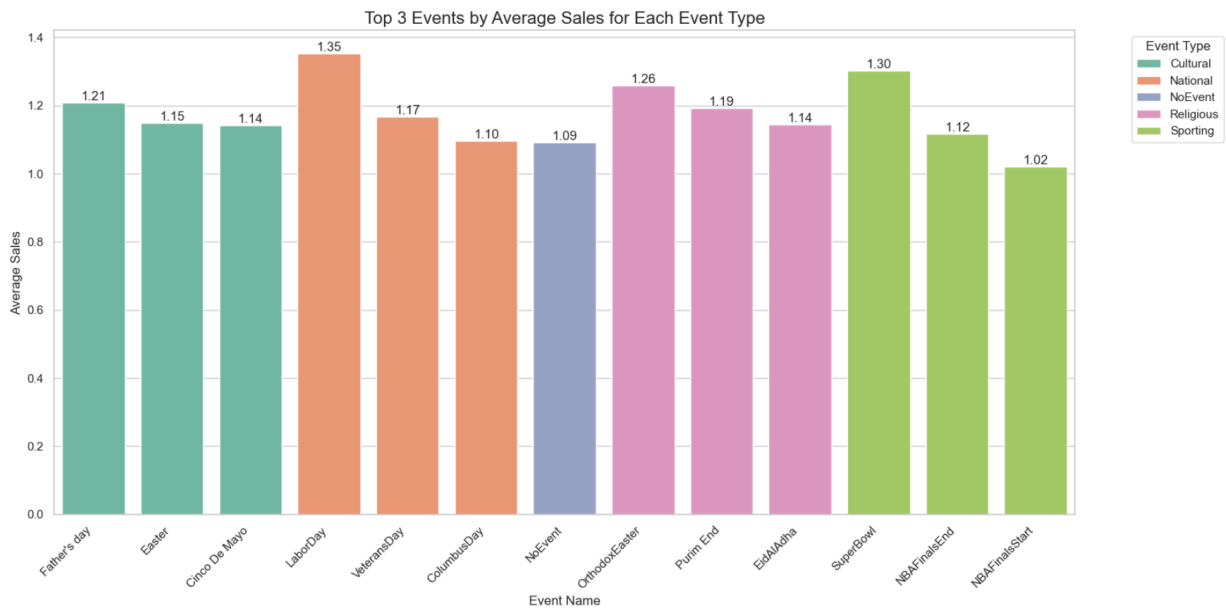
CA\_3 has the highest sales, followed by TX\_2 and WI\_3. California has 4 stores whereas other states have three stores each.

Figure 8. Total Revenue by Department



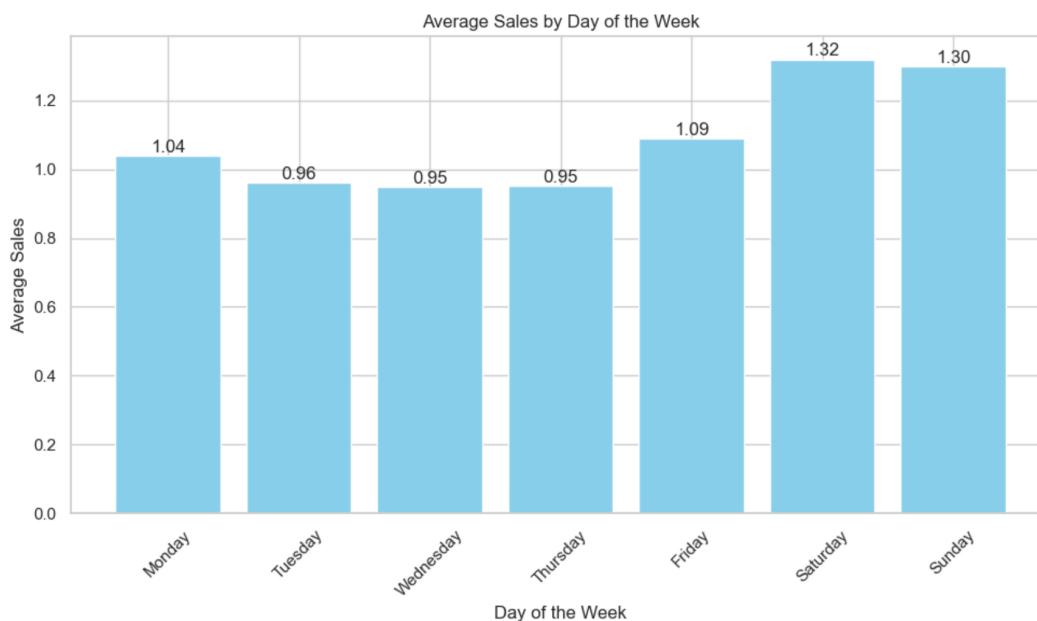
FOODS\_3 department offered the most items and generated the most revenue.

Figure 9. Average sales for Each Event Type



LaborDay has the highest average sales, followed by Super Bowl and NBA Finals End. From the graph, it can be said that National and Sporting events tend to generate higher average sales compared to Cultural, Religious, and NoEvent days.

Figure 10. Average Sales by Day of the Week



Stores have the highest sales on Saturdays and Sundays. Weekday sales are lower, though Mondays and Fridays are a little better than Tuesdays, Wednesdays, and Thursdays.

## 4. Data Preparation

After merging all the CSV files, the dataset became huge. Training this large amount of data is computationally expensive and time-consuming. In the EDA section, work was done with the full dataset. When merging the dataset, a few features had missing values. The 'sell\_price' column in the training dataset had approximately 26% missing values. These missing values were filled with the mean value after observing the data distribution pattern. Event name and Event type showed the highest number of missing values. By looking at the data visualization, weekdays (Monday to Friday) and days without occasions were presented with null values. These values were filled with "no event" for better analysis. To capture the pattern of the item\_id, no rows were removed.

After handling the missing values, the next crucial step was data splitting for training and validation. The data spans from 2011 to 2015. To capture the recent trends of the items and their prices, the timeline was trimmed from 2014 to 2015. After selecting this timeline, stratified splitting was used so that each of the training, validation, and testing datasets get proper item\_ids. This approach ensured that all datasets contained all 3,049 unique items.

In predictive modeling, one-hot encoding was applied to the month and day features. Item\_id, store\_id, and other features were encoded and saved in joblib for future use. Unnecessary features were removed from the training, validation, and test datasets used for the forecasting model. Additionally, the data within all three datasets was rearranged chronologically, as time sequence is crucial for forecasting models.



## 5. Modeling

In this project, the LightGBM algorithm was used for the predictive model. This algorithm, developed by Microsoft, works well with large datasets. For time series forecasting, Facebook's Prophet model was used. This model is designed to handle seasonality and holidays, which are important factors in the retail sector. In this sector, accurate prediction and reliable forecasting play an important role. These techniques can help the management team with inventory, pricing strategies, warehouse storage design, and can solve supply chain issues. The combination of these two powerful algorithms aims to provide comprehensive insights in decision-making and creating robust strategies to tackle upcoming shortage issues.

### a. Approach

**Predictive Model:** In the first approach to creating a predictive model, Random Forest was initially planned to be implemented. However, tuning the large amount of data and the model itself becoming bigger made it difficult to use in production phases. After considering all this, LightGBM was chosen. LightGBM is a gradient boosting framework. It uses a technique called gradient boosting and has the capability to handle a large number of features while also providing better accuracy. Unlike other tree-based algorithms, it uses leaf-wise tree growth that provides better results. In the hyperparameter section, `num_leaves`, `max_depth`, `n_estimators`, `learning_rate`, and `min_child_samples` were introduced. LightGBM performed well in both validation and testing datasets. Both RMSE and MAE values are quite low.

**Forecasting Model:** Time series analysis was performed using Facebook Prophet. This time series model is known for its lightweight nature and works well with holidays and seasonality. This model is specially designed to capture holiday patterns. It is easy to implement and also works with large amounts of data. This model's hyperparameter tuning is comparatively easier than other time series models. The hyperparameters used during the training phase are `changepoint_prior_scale`, `seasonality_prior_scale`, and `seasonality_mode`. Daily, weekly, and yearly seasonality were used. The model performed well in both validation and testing datasets after applying hyperparameter tuning.

## 6. Evaluation

### a. Evaluation Metrics

The main goal of this project is to predict how much money specific products will generate and how much total revenue will be generated over time. Both predictive and forecasting models use two evaluation metrics to assess their performance: RMSE (Root Mean Square Error) and MAE (Mean Absolute Error). These metrics were chosen to see how much the model's predictions deviate from the actual values and to understand how well the model is performing. Lower scores in both metrics indicate solid model performance. RMSE is particularly effective at identifying large errors in revenue forecasts, while MAE provides a straightforward average of all errors, regardless of their direction.

### b. Results and Analysis

Figure 11. Achieved Result

	Model Name	Validation RMSE	Test RMSE	Validation MAE	Test MAE
Predictive Model	LightGBM	8.83	9.68	4.58	4.69
Forecasting Model	Prophet	7404	4619	4530	3715


In the predictive task, the LightGBM model's validation RMSE is around 8.83, and MAE is around 4.58. 5-fold cross-validation was performed, and the same results were obtained over time. The model is performing consistently. The constant result of this model indicating the reliability for individual store and item predictions.

The forecasting model's validation RMSE is high compared to the test result. However, cross-validation showed consistent results. The forecasting model is based on the total sales revenue across all stores. The target column revenue value is relatively high. Lower MAE and RMSE in the test set indicate that the model is able to capture the overall trend. These are promising signs for its potential effectiveness in production.

### c. Business Impact and Benefits

The two developed models offer significant benefits to the superstore and potential stakeholders. The predictive model will generate more revenue during specific times. This helps the store stock enough of the right products, avoid running out of popular items, and reduce having too much of the wrong products.

The forecasting model shows consistent results in cross-validation and captures overall trends. This helps the business make smart decisions about staffing, plan for busy and slow periods, and allocate resources effectively.




Both models offer useful information. Based on these models, existing superstores can improve their product selection and also add new items when other products are in less demand. By doing this, they can introduce new items to their customers. The forecasting model will give them an early overview of how much revenue they can generate, based on which they can reschedule their management and operations.

The warehouse issues can be solved and rearranged with both of these models. These models save money and boost business by lowering storage costs, speeding up delivery of popular items, and reducing wasted stock. This will create a better reputation among their customers by continuously supplying popular items.


#### d. Data Privacy and Ethical Concerns

**Data Privacy Implications of the project:** Sales strategy is a confidential aspect of a business, so when the model allows other businesses in the market to look into private data of sales forecasts, it can cause a negative impact on the business. As the business's privacy is exposed to the whole market where other firms can easily avail themselves of a business's marketing and sales strategy and implement it in their own business, it is a matter of privacy breach. A particular business's personal strategy to generate higher revenue is exposed to its rivals and competitors, which is a disadvantage for the business. The rival firms can end up using these same strategies in their business to enhance sales and generate greater revenue, which will be a disadvantage for the main firm. The models process sensitive data including daily sales figures, product-specific revenue, and inventory levels, which collectively reveal the business's operational strategies.

**Ethical Concerns:** As far as ethical concerns are considered, the usage of AI models to predict the inventory to be stored in the warehouse and revenue generated by higher sales could have an impact on its employees. As work is done using technology, the requirement for human labor declines. This can result in many employees losing their jobs and, as a result, increase the rate of unemployment in the economy. This will hamper the economy in many ways as people's purchasing power reduces, which eventually will affect the Gross Domestic Product (GDP) of the nation in the long run. Another issue which can occur is if the model's prediction for sales and revenue generation works but eventually causes a loss to the business. If the product that is in high demand is available in many other stores in the market, that will lead to an oversupply of a specific product in the market that will eventually lead to a reduction in the price of that product, as when supply is higher, the price of the product is automatically lower. In that way, the business can, on the other hand, face losses. The model may contain inherent biases that could unfairly impact certain product categories or customer segments, potentially perpetuating existing market inequalities.



**Steps to Ensure Data Privacy and Ethical Considerations:** To tackle the issue of data privacy, the business should limit the information shared by the model. As a result, store all its private data and information in software, which is not available to anyone except the business itself for future needs. This way the business can keep its private data safe and away from its rival businesses in the market. Implementing robust encryption, two-way password protection protocols, and strict access controls for all sensitive data significantly enhances data protection. In order to ensure employees do not lose their jobs and face unemployment, the employees in those sectors where they are no longer required can be trained for other sectors. They be trained to operate these models, which helps them in technological advancement and faster work output overall. As far as the greater availability of a product in many stores across the market leads to reduction in price and total revenue, ways to sustain consumer loyalty towards the business by ensuring better after-sales service to customers be explored. Establishing an independent ethics board to oversee the project's development and deployment ensures ongoing ethical considerations are addressed proactively.



## 7. Deployment

After training the model, the next crucial part is deployment. Before starting the deployment, the first stage is noting down all the package versions used during the training. Setting up an environment is the most important part. Mismatches of the versions can take a long time to fix. Model artifacts were pushed to GitHub in the model folder. In the app folder, a Python file was created. In that file, the application front-end interface was created with the Streamlit application, a popular app to create front-end interfaces. After this procedure, the app was run locally to see how its interface looked.

The next part was to dockerize the app and create the Docker image before deploying it to Render. Port selection mismatches happened if the Docker file did not provide the information. Before pushing to GitHub, testing was done several times on the local computer to see how the application looked. In the last step, the dockerized file and requirements file were pushed to GitHub. In the Render account, the web service connected with the Git account. Render automatically started building the application. Continuous monitoring happened on the log site to solve problems and restart if necessary. This was part of monitoring the integration. A robust error handling system was implemented that managed unexpected inputs or system failures gracefully.

There were many challenges faced during app deployment. The project started with Random Forest, but the file size was too big for Render's free tier version to accept. That was the key lesson learned from this project. In production phases, it was learned that the deployment model had to be of minimum size if a free tier version was used. Version mismatches would prevent the creation of any application. Before installing and using any version of packages, it's important to write down their versions; otherwise, it can take time to solve problems later. Debugging had to be done properly to solve every issue faced during dockerization. Implementing a monitoring system to track the model's performance over time and detect any degradation was considered.

For future deployment efforts, optimize the model size from the start if planning to use free-tier cloud services. This will solve file size problems. Secondly, keep track of all package versions used during any projects. This prevents time-wasting ,version mismatch issues during deployment.





## 8. Conclusion

The project has successfully developed two powerful models that have been implemented in the production site. These two models offer significant insights for superstore operations and strategic decision-making. Both models showed consistent performance in their training phases. The predictive model effectively identifies product revenue on specific dates. This enables businesses to optimize their inventory management and product display strategies. The forecasting model provides accurate weekly revenue predictions that will help the store allow for better resource allocation and operational planning.

The project has met the primary objective of enhancing business profitability. By utilizing both models, stakeholders facing challenges such as warehouse space limitations and product shortages can solve them instantly. These models will help them stay one step ahead of their rival companies. The model's user interface is designed in a way that non-technical stakeholders can easily use it without any problems.

Future steps should focus on updating the models with new and unseen data. The models were trained on the previous year's data. To improve accuracy for recent years, these models will be trained with new data.

Customer buying behavior is not implemented here. Insights into which customers are more likely to buy which products will give this application more robust results.

The potential for real-time data integration needs to be explored to provide more timely insights and recommendations.

The user interface needs to be updated to a better interface where real-time product revenue will be shown to stakeholders along with economic indicators.



## 9. References

■ ■ ■