

# Assignment 2

## ML as a Service

---

Sareem Sandeed  
Student ID: 24622829  
10 October 2023

GitHub [Link](#)

Heroku app: [Link](#)

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



## Table of Contents

<b>1. Executive Summary</b>	<b>2</b>
<b>2. Business Understanding</b>	<b>4</b>
a. Business Use Cases	4
<b>3. Data Understanding</b>	<b>6</b>
<b>4. Data Preparation</b>	<b>8</b>
<b>5. Modelling</b>	<b>10</b>
a. Part A/Prediction	10
b. Part B/Forecasting	10
<b>6. Evaluation</b>	<b>12</b>
a. Evaluation Metrics	12
b. Results and Analysis	12
c. Business Impact and Benefits	14
d. Data Privacy and Ethical Concerns	14
<b>7. Deployment (Link)</b>	<b>16</b>
<b>8. Conclusion</b>	<b>17</b>
<b>9. References</b>	<b>18</b>
<b>10. Appendix</b>	<b>19</b>



## 1. Executive Summary

### **Project Overview:**

This project aimed to develop two critical models for sales forecasting and revenue optimization in the retail industry. The objectives were to create a Predictive Sales Revenue Model and a Sales Revenue Forecasting Model using machine learning and time-series analysis techniques, respectively. These models play a pivotal role in enhancing decision-making processes at both the tactical and strategic levels within the retail sector.


### **Problem Statement and Context:**

The retail industry operates in a highly dynamic environment characterized by changing market conditions, consumer behaviors, and seasonal trends. Accurate sales forecasts are essential for effective inventory management, pricing strategies, and demand planning. Additionally, short-term revenue forecasts are vital for operational planning, marketing campaigns, and budget allocation. The challenge lay in developing models that could capture the complexity of these dynamics, ensuring precise predictions that can empower retail businesses to thrive in a competitive landscape.

### **Achieved Outcomes and Results:**

**Predictive Sales Revenue Model:** This model successfully harnessed machine learning algorithms to analyze historical sales data, uncover intricate patterns, and generate precise sales revenue forecasts for specific items and stores on given dates. It significantly improved decision-making for store managers, inventory managers, pricing strategists, and finance teams, leading to optimized stock levels, reduced overstocking or stockouts, and enhanced profitability.

**Sales Revenue Forecasting Model:** Utilizing time-series analysis algorithms, this model accurately predicted total sales revenue across all stores and items for the next 7 days. This short-term forecasting capability empowered senior management, marketing teams, and financial planners to make informed strategic decisions and allocate resources effectively. It enabled retailers to respond swiftly to market dynamics, optimize marketing campaigns, and enhance financial planning.



In conclusion, these models addressed the pressing needs of the retail industry by providing accurate sales revenue forecasts at different time scales. They transformed data into actionable insights, facilitating data-driven decision-making and positioning retail businesses to adapt and thrive in a constantly evolving marketplace. The success of this project underscores the significance of leveraging advanced analytics and data-driven approaches in the retail sector to gain a competitive edge.



## 2. Business Understanding

### a. Business Use Cases

#### 1. Predictive Sales Revenue Model:

- a. **Use Case:** The predictive model using Machine Learning is applied to forecast sales revenue for a given item in a specific store on a given date. This use case addresses the need for accurate sales forecasts, which are essential for inventory management, pricing strategies, and demand planning in retail.
- b. **Challenges/Opportunities:** The retail industry faces dynamic market conditions, seasonality, and ever-changing consumer behavior. Accurate sales predictions are crucial to optimize stock levels, reduce overstocking or stockouts, and enhance profitability. Machine learning algorithms can capture complex patterns in sales data, improving forecast accuracy.

#### 2. Sales Revenue Forecasting Model:

- a. **Use Case:** The forecasting model using time-series analysis is employed to predict the total sales revenue across all stores and items for the next 7 days. This use case addresses the need for short-term revenue forecasts, which are vital for operational planning, marketing campaigns, and budgeting.
- b. **Challenges/Opportunities:** Accurate short-term sales revenue forecasts enable retailers to make informed decisions about marketing spend, staffing, and resource allocation. It helps them respond quickly to changing market dynamics and optimize their strategies. Time-series analysis algorithms can capture seasonality and trends, providing valuable insights for revenue forecasting.

### b. Key Objectives

#### 1. Predictive Sales Revenue Model:

- a. **Objective:** The key objective is to build a predictive model that accurately forecasts sales revenue for specific items and stores on given dates.
- b. **Stakeholders:** Stakeholders include store managers, inventory managers, pricing strategists, and finance teams. They require accurate sales forecasts to optimize operations, pricing, and financial planning.
- c. **Addressing Requirements:** The project aims to meet stakeholders' requirements by leveraging machine learning algorithms to analyze historical sales data, identify patterns, and make precise revenue predictions. This addresses the need for data-driven decision-making.



## 2. Sales Revenue Forecasting Model:

- a. **Objective:** The primary goal is to create a forecasting model that predicts total sales revenue across all stores and items for the next 7 days.
- b. **Stakeholders:** Stakeholders include senior management, marketing teams, and financial planners. They need accurate short-term forecasts to make strategic decisions and allocate resources effectively.
- c. **Addressing Requirements:** The project addresses stakeholder requirements by utilizing time-series analysis algorithms to capture seasonality, trends, and short-term fluctuations. Accurate forecasting aids in optimizing marketing campaigns and financial planning.

In summary, both models serve critical business use cases in the retail industry. The predictive model enhances decision-making at the store and item level, while the forecasting model aids in short-term revenue planning and strategic resource allocation. Machine learning and time-series analysis algorithms are essential tools to address the challenges and opportunities presented by the dynamic nature of the retail sector.



### 3. Data Understanding

There were four datasets that needed to be joined to get relevant information like revenues, items prices, sales quantity, and date.

- **calendar\_events**: Major events along with date
- **calendar**: Dates with week id and date id
- **items\_weekly\_sell\_prices**: Unit sale price of each item
- **sales\_train**: Sales quantity on dates

The following snapshots shows the information about all the four datasets:

<pre>calendar_events.info()  &lt;class 'pandas.core.frame.DataFrame'&gt; RangeIndex: 167 entries, 0 to 166 Data columns (total 3 columns): #   Column      Non-Null Count  Dtype ---  - 0    date        167 non-null    object 1    event_name   167 non-null    object 2    event_type   167 non-null    object dtypes: object(3) memory usage: 4.0+ KB</pre>	<pre>calendar.info()  &lt;class 'pandas.core.frame.DataFrame'&gt; RangeIndex: 1969 entries, 0 to 1968 Data columns (total 3 columns): #   Column      Non-Null Count  Dtype ---  - 0    date        1969 non-null    object 1    wm_yr_wk    1969 non-null    int64 2    d           1969 non-null    object dtypes: int64(1), object(2) memory usage: 46.3+ KB</pre>
---	---

```
items_weekly_sell_prices.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6841121 entries, 0 to 6841120
Data columns (total 4 columns):
#   Column      Dtype
---  -
0    store_id    object
1    item_id     object
2    wm_yr_wk    int64
3    sell_price  float64
dtypes: float64(1), int64(1), object(2)
memory usage: 208.8+ MB
```

```
sales_train.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30490 entries, 0 to 30489
Columns: 1547 entries, id to d_1541
dtypes: int64(1541), object(6)
memory usage: 359.9+ MB
```



calendar_events.head()				items_weekly_sell_prices.head()				
	date	event_name	event_type		store_id	item_id	wm_yr_wk	sell_price
0	2011-02-06	SuperBowl	Sporting	0	CA_1	HOBBIES_1_001	11325	9.58
1	2011-02-14	ValentinesDay	Cultural	1	CA_1	HOBBIES_1_001	11326	9.58
2	2011-02-21	PresidentsDay	National	2	CA_1	HOBBIES_1_001	11327	8.26
3	2011-03-09	LentStart	Religious	3	CA_1	HOBBIES_1_001	11328	8.26
4	2011-03-16	LentWeek2	Religious	4	CA_1	HOBBIES_1_001	11329	8.26

sales_train.head()																	
		id	item_id	dept_id	cat_id	store_id	state_id	d_1	d_2	d_3	d_4	...	d_1532	d_1533	d_1534	d_1535	d_1536
0	HOBBIES_1_001_CA_1_evaluation	HOBBIES_1_001	HOBBIES_1	HOBBIES	CA_1	CA	0	0	0	0	...		1	1	1	0	1
1	HOBBIES_1_002_CA_1_evaluation	HOBBIES_1_002	HOBBIES_1	HOBBIES	CA_1	CA	0	0	0	0	...		0	0	0	0	0
2	HOBBIES_1_003_CA_1_evaluation	HOBBIES_1_003	HOBBIES_1	HOBBIES	CA_1	CA	0	0	0	0	...		0	0	1	0	0
3	HOBBIES_1_004_CA_1_evaluation	HOBBIES_1_004	HOBBIES_1	HOBBIES	CA_1	CA	0	0	0	0	...		8	2	0	8	2
4	HOBBIES_1_005_CA_1_evaluation	HOBBIES_1_005	HOBBIES_1	HOBBIES	CA_1	CA	0	0	0	0	...		2	0	1	3	2

calendar.head()			
	date	wm_yr_wk	d
0	2011-01-29	11101	d_1
1	2011-01-30	11101	d_2
2	2011-01-31	11101	d_3
3	2011-02-01	11101	d_4
4	2011-02-02	11101	d_5





## 4. Data Preparation


The following steps were taken for data preparation for **predictive** modelling:

1. Checking the null values, duplicates in each of the datasets.
2. Checking data types of all columns.
3. Converting wide form to long form(shown below) for sales\_train dataset
4. Merging all the datasets.
5. Dropping columns that are correlated and already have the same information in different columns can lead to overfitting and inaccurate predictions.
6. Splitting data store wise as I wanted to model at store level.
7. Final store level data was split into train and validation data.

sales_train_melted.head()									
		id	item_id	dept_id	cat_id	store_id	state_id	day	item_sales
0	HOBBIES_1_001_CA_1_evaluation	HOBBIES_1_001	HOBBIES_1	HOBBIES	CA_1	CA	d_1		0
1	HOBBIES_1_002_CA_1_evaluation	HOBBIES_1_002	HOBBIES_1	HOBBIES	CA_1	CA	d_1		0
2	HOBBIES_1_003_CA_1_evaluation	HOBBIES_1_003	HOBBIES_1	HOBBIES	CA_1	CA	d_1		0
3	HOBBIES_1_004_CA_1_evaluation	HOBBIES_1_004	HOBBIES_1	HOBBIES	CA_1	CA	d_1		0
4	HOBBIES_1_005_CA_1_evaluation	HOBBIES_1_005	HOBBIES_1	HOBBIES	CA_1	CA	d_1		0

The following steps were taken for data preparation for **forecasting** modelling:

1. Checking the null values, duplicates in each of the datasets.
2. Checking data types of all columns.
3. Converting wide form to long form for sales train dataset
4. Merging 3 datasets
5. Dropping columns that are not needed.
6. Finally have a table with date as index and corresponding aggregated revenue at national level (shown below)
7. At last, data was split into train and validation data.



	date	revenue
0	2011-01-29	81650.61
1	2011-01-30	78970.57
2	2011-01-31	57706.91
3	2011-02-01	60761.20
4	2011-02-02	46959.95

**Features engineering** was done on **date** to get year, month, day and days of the week.

**One-Hot Encoding:** One-hot encoding is used for categorical features like year, month, day, and day\_of\_week.

**Target Encoding:** Target encoding is chosen for item\_id due to 3049 unique items. One-hot encoding would not be suitable in such scenario..

**Train-Validation Split (80:20):** It provides a sufficiently large training dataset for the model to learn patterns and relationships in the data. At the same time, it reserves a substantial portion for validation, which is crucial for assessing the model's generalization performance on unseen data.

■ ■ ■

## 5. Modelling

### a. Part A/Prediction

I chose to use LightGBM Regressor as the predictive model due to its advantages in handling complex tabular data and efficient gradient boosting capabilities. For predictive part, since the datasets were too big, modelling was done at store level. Since there are 10 stores, so 10 models were created.

Reasons for Choosing LightGBM:

1. Efficiency: LightGBM is known for its speed and efficiency, making it suitable for handling large datasets, which is common in retail sales forecasting.
2. Gradient Boosting: It uses gradient boosting, an ensemble learning technique known for its ability to capture complex patterns in data and produce accurate predictions.
3. Categorical Feature Support: LightGBM has native support for categorical features, which is essential in this retail scenario where store locations, item categories, and dates may be categorical variables.

Hyperparameter- No hyperparameter tuning was done as given the large datasets, it requires high computational power and hence limited my capability to explore further. Since the data was too big doing grid search or random search for model hyperparameter was too computation intensive.

Other models, such as Random Forest, SDG were trained but the LightGBM ran faster comparatively on large datasets. While Random Forest were giving better results but it was taking too much time.

### b. Part B/Forecasting

For the experiment, which aims to forecast sales revenue for the next 7 days for the retail giant, ARIMA time series forecasting models were considered and evaluated. For doing this there was a requirement to have only two columns, date as index and aggregated revenue for all stores on day basis.

Validation data: last 28 days

Train data: Remaining days



ARIMA (Auto Regressive Integrated Moving Average):

Why Chosen: ARIMA is a classic and widely used time series forecasting model that can capture trends and seasonality in data. It is a suitable starting point for time series forecasting experiments. It runs fast and less computational.

Hyperparameters Tuned: Order parameters (p, d, q) were tuned.

Values Tested: Different combinations of p, d, and q were tested to find the best order for the ARIMA model.

Rationale: Tuning the ARIMA order is essential to capture the time series patterns effectively.

The models selected for this experiment were chosen based on their suitability for time series forecasting, with each model having its unique strengths.

Prophet and LSTM were computationally more intensive were taking a lot of time to train, therefore, ARIMA was considered over these.



## 6. Evaluation

### a. Evaluation Metrics

The metric used for assessing model performance is the Root Mean Squared Error (RMSE) score. RMSE, which stands for Root Mean Squared Error, is a commonly used metric in machine learning and statistics to evaluate the accuracy or performance of a predictive model, particularly regression models.

RMSE is in the same unit as the target variable, making it easy to interpret. This makes it more understandable to stakeholders. Also, RMSE allows for straightforward model comparison for all 10 stores consistently.

### b. Results and Analysis

#### 1. Predictive modelling:

The metric used for assessing model performance is the Root Mean Squared Error (RMSE) score. RMSE is in the same unit as the target variable, making it easy to interpret. This makes it more understandable to stakeholders. Also, RMSE allows for straightforward model comparison for all 10 stores consistently.

	baseline_rmse	train_rmse	val_rmse
store_CA_1	7.242093	7.233376	7.276859
store_CA_2	5.735208	5.721615	5.789264
store_CA_3	9.695821	9.700983	9.675144
store_CA_4	4.415324	4.409242	4.439569
store_TX_1	5.785319	5.787674	5.775888
store_TX_2	7.192938	7.199415	7.166971
store_TX_3	6.388319	6.383801	6.406360
store_WI_1	4.808941	4.805001	4.824669
store_WI_2	6.594577	6.585164	6.632098
store_WI_3	6.373680	6.362419	6.418527

Consistency: Across most stores, the RMSE values are relatively consistent, indicating that the model's performance is similar across different store locations.

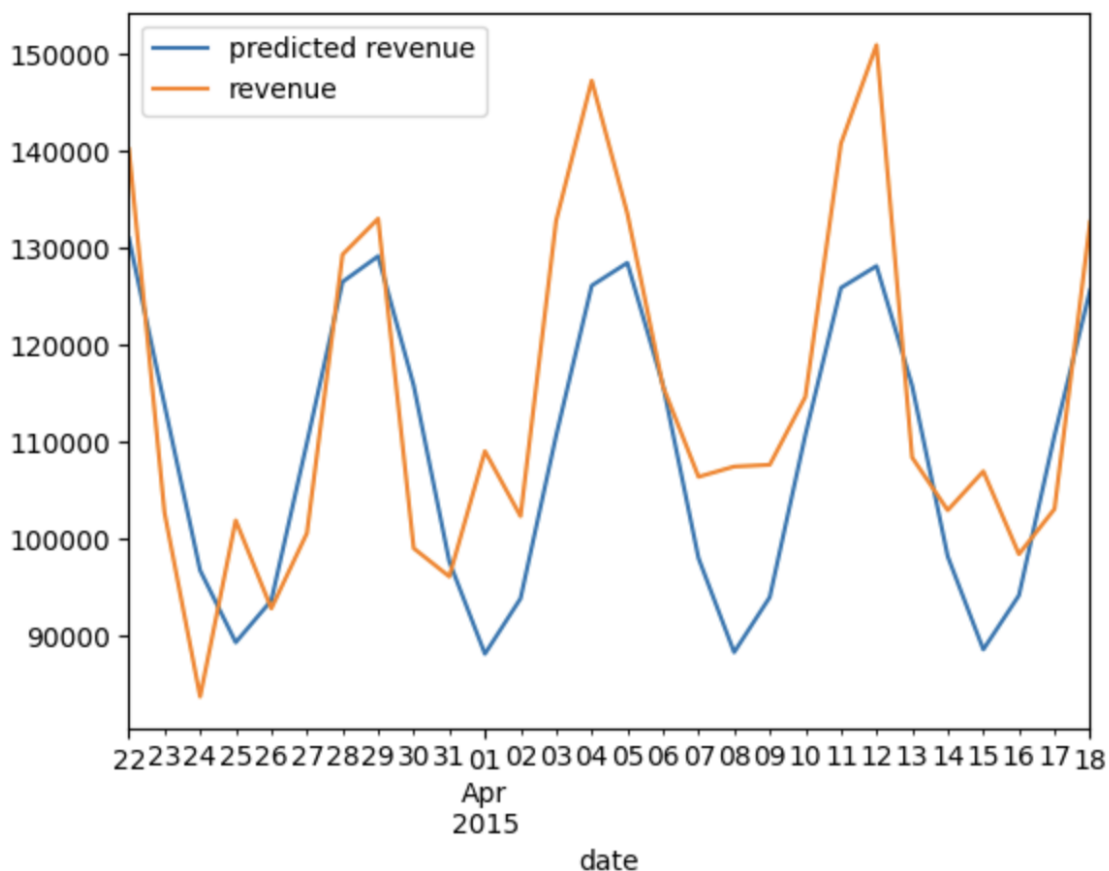
Generalization: RSME is similar across baseline train and validation, therefore model has been able to general very well.

## 2. Forecasting modelling:

The metric used for assessing model performance is the Root Mean Squared Error (RMSE) score. RMSE is in the same unit as the target variable, making it easy to interpret. This makes it more understandable to stakeholders.

- RSME for train: 11651.05
- RSME for validation: 12426.31
- Mean of revenue in train: 92288.90
- Mean of revenue in validation: 114249.01

Predicted revenue of 28 days comparison with actual value given below:



The RMSE values indicate the level of error in your ARIMA model's predictions. In this case, the RMSE for the training set (11651.05) is lower than the RMSE for the validation set (12426.31). This suggests that the model might be overfitting to some extent.

Comparing RSME to mean revenue, it can forecast with good accuracy.

### c. Business Impact and Benefits

The experiment involved training a LightGBM model to predict sales revenue. The model achieved reasonable RSME on train and validation data. These scores indicate that the model can predict revenue on a given day for an item at a store with good confidence.

The experiment involved training a ARIMA time series model to predict sales revenue for next 7 days. The model achieved reasonable RSME on train and validation data.

Impact of business:

1. Accurate results from the predictive or forecasting model will enable the business to make data-driven decisions, leading to improved profitability, efficient operations, and enhanced customer satisfaction.
2. Bad pricing decisions resulting from incorrect forecasting may lead to missed opportunities for profit during high-demand periods or unnecessary discounts during low-demand periods.
3. Misallocated marketing budgets based on inaccurate forecast can result in ineffective campaigns and missed chances to boost brand visibility and ROI.
4. Inaccurate sales forecast can disrupt financial planning, budgeting, and revenue target setting, potentially leading to financial instability.
5. Inadequate stock levels due to inaccurate forecast can result in customer dissatisfaction, leading to lost sales and reduced loyalty.
6. Inaccurate forecast can damage the credibility of the business, especially when widely communicated to stakeholders, investors, or partners.


Next Steps: The specific actions to be taken should align with the business objectives and the desired level of accuracy. Potential actions include hyperparameter tuning, feature engineering, and exploring alternative modeling techniques to enhance model performance.

### d. Data Privacy and Ethical Concerns

#### **Data Privacy Concerns:**

**Data Granularity:** The project deals with sales revenue data at the item, store, and date levels. While this data may not contain explicitly sensitive personal information, it is still important to protect proprietary and confidential business data. Aggregating data at a higher level may help mitigate privacy risks.

**Customer Data:** Depending on the dataset used, there might be indirect customer-related data (e.g., transaction timestamps). Care should be taken to anonymize or remove personally identifiable information (PII) if present.



Data Security: Ensuring the security of the dataset is crucial to prevent unauthorized access or data breaches. Robust data security measures must be in place during data collection, storage, and model training.

**Ethical Concerns:**

Bias: Machine learning models may inadvertently learn biases present in historical data. For example, biased pricing or inventory decisions could result from biased historical sales data. It's important to monitor and mitigate bias in model predictions.

Transparency: Lack of model transparency can be an ethical concern. Stakeholders should have an understanding of how the model makes predictions, especially when these predictions impact business decisions.

Fairness: Model predictions should not discriminate against specific groups, whether intentionally or unintentionally. Fairness considerations should be addressed to ensure equitable outcomes.

**Steps to Ensure Data Privacy and Ethical Concerns:**

Data Anonymization: Any personally identifiable information (PII) or sensitive customer data is carefully anonymized or removed from the dataset to protect privacy.

Access Control: Access to the data and models is restricted to authorized personnel only. Role-based access control ensures that only individuals with legitimate reasons can access the data.

Data Encryption: Data at rest and during transmission is encrypted to prevent unauthorized access or data leaks.

Bias Mitigation: Regular audits are conducted to identify and mitigate bias in the model. Techniques such as re-sampling, re-weighting, or fairness-aware algorithms are used to ensure fair predictions.

Data Retention Policies: Clear data retention policies are established, specifying how long data is retained and when it should be securely deleted.

By implementing these steps, the project aims to protect data privacy, mitigate ethical concerns, and ensure that the use of machine learning models is ethical, transparent, and fair. These measures are crucial to maintain trust with stakeholders and users of the predictive and forecasting models.







## 7. Deployment ([Link](#))

Steps in deploying a machine learning model using FastAPI, Docker, and Heroku:

- Model Export: Save the trained model and preprocessing steps for deployment.
- FastAPI Application: Create a FastAPI app with defined endpoints.
- Model Loading: Load the model during app initialization.
- API Integration: Define logic for handling API requests.
- Docker Containerization: Package the app into a Docker container.
- Build and Push Docker Image: Create and upload the Docker image.
- Heroku Setup: Set up a Heroku account and app.
- GitHub connection: Connect GitHub repo with Heroku app
- Heroku Deployment: Deploy the Dockerized app to Heroku.
- Monitoring and Logging: Implement performance monitoring.

### **Integration and real-world considerations:**

- Data Integration: Ensure seamless data flow.
- Security: Protect the API with encryption and authentication.
- API Versioning: Maintain compatibility.
- Error Handling: Provide clear error messages.
- Testing: Thoroughly test the deployed API.
- Deployment Automation: Use CI/CD for efficiency.

### **Challenges and considerations:**

- Scalability: Plan for increased usage.
- Cost Management: Optimize resource usage.
- Version Errors: Errors while containerisation due different version of python libraries.
- Slug memory error: Due to 500MB limit of slug in Heroku





## 8. Conclusion

Business key findings, insights, and outcomes from the datasets are attached in appendix at the end of this report.

### **Key Findings and Outcomes:**

**1. Model Performance:** The project achieved good and generalise model performance across 10 stores prediction algorithms. While for forecasting the sales revenue for next 7 days, the results are quite convincing with good confidence level.

**2. Impact on Draft Choices:** The models contribute to NBA teams by providing accurate predictions for draft selections, potentially leading to better team performance and resource allocation.

**3. Enhanced Fan Engagement:** The project successfully engages fans and analysts by offering data-driven insights into player performance, enriching discussions, and supporting fantasy sports and betting activities.

### **Success in Achieving Goals:**

The project has succeeded in its primary goals, benefiting both NBA teams and fans. It provides valuable predictions for draft selections and fosters fan engagement, aligning with the project's objectives.

### **Future Work and Recommendations:**

**1. Hyperparameter Tuning:** Explore further hyperparameter tuning to potentially improve model performance, especially for Random Forest and LSTM or Prophet models.

**2. Business Integration:** Integrate the predictive model's outputs into key business processes, such as inventory management, pricing, marketing, and demand planning.

**3. Interpretability:** Develop methods to make model outputs more interpretable for stakeholders, helping them understand the reasoning behind sales predictions and forecasting.

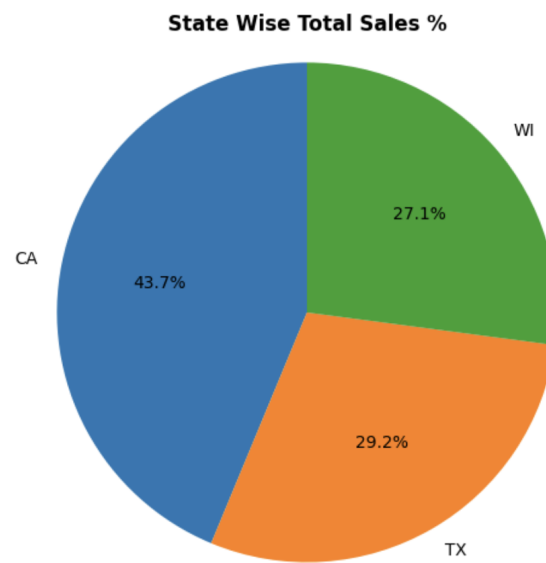


## 9. References

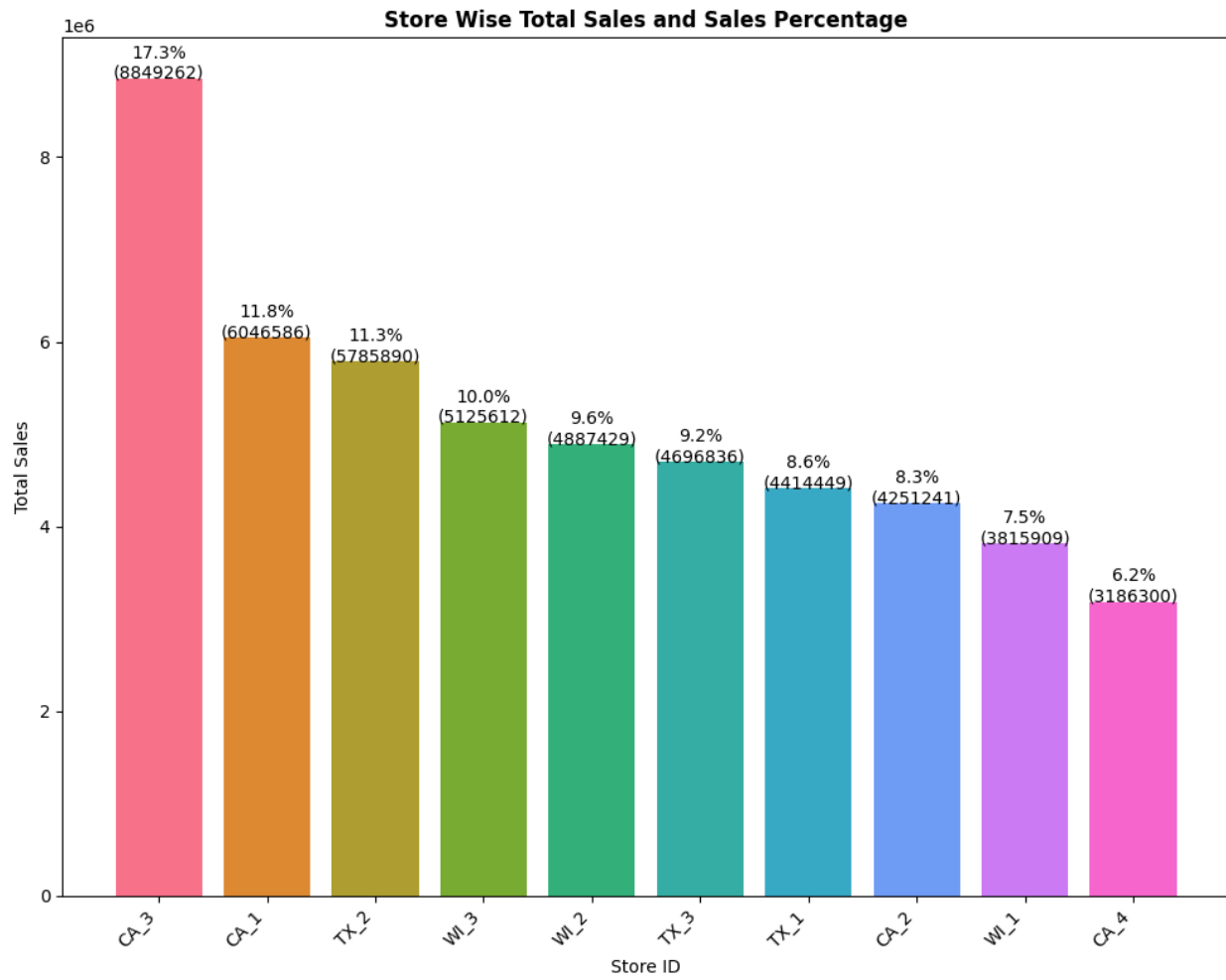
1. Cookiecutter: Better Project Templates—Cookiecutter 2.3.0 documentation. (n.d.). Retrieved September 8, 2023, from <https://cookiecutter.readthedocs.io/en/stable/>
2. How to use pyenv to manage Python versions. (2022, November 28). The Teclado Blog. <https://blog.teclado.com/how-to-use-pyenv-manage-python-versions/>
3. Poetry—Python dependency management and packaging made easy. (n.d.). Retrieved September 8, 2023, from <https://python-poetry.org/>
4. Sklearn. Ensemble. Gradientboostingclassifier. (n.d.). Scikit-Learn. Retrieved September 8, 2023, from <https://scikit-learn/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html>
5. Sklearn. Ensemble. Randomforestclassifier. (n.d.). Scikit-Learn. Retrieved September 8, 2023, from <https://scikit-learn/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>
6. Sklearn. Linear\_model. Logisticregression. (n.d.). Scikit-Learn. Retrieved September 8, 2023, from [https://scikit-learn/stable/modules/generated/sklearn.linear\\_model.LogisticRegression.html](https://scikit-learn/stable/modules/generated/sklearn.linear_model.LogisticRegression.html)
7. Hotz, N. (2018, September 10). What is crisp dm? *Data Science Process Alliance*. <https://www.datascience-pm.com/crisp-dm-2/>

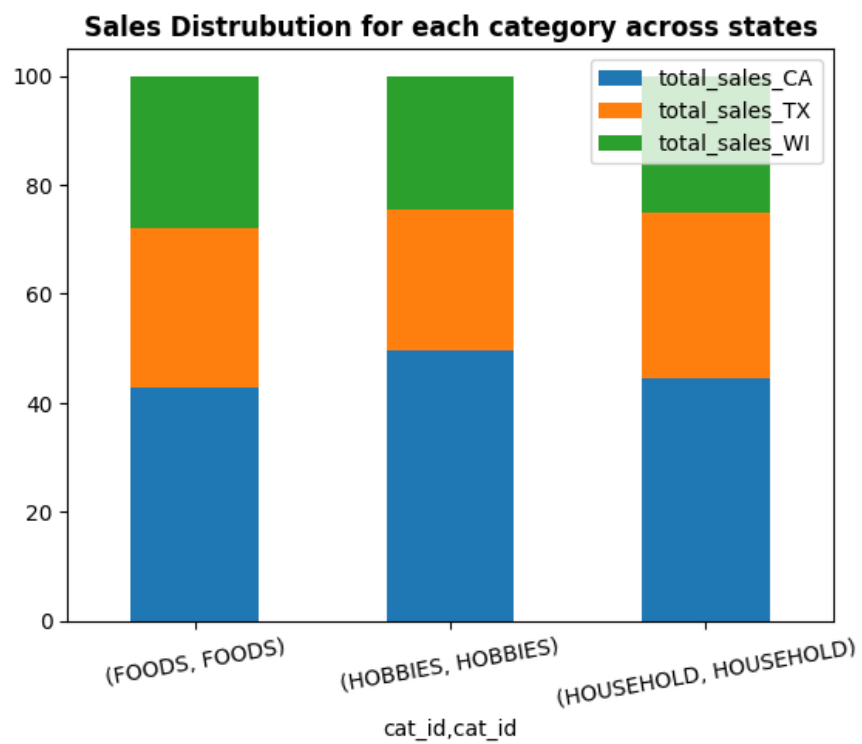
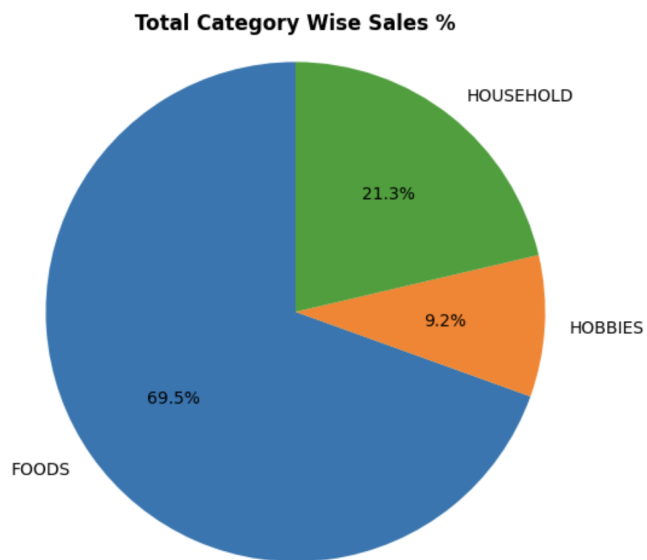
## 10. Appendix

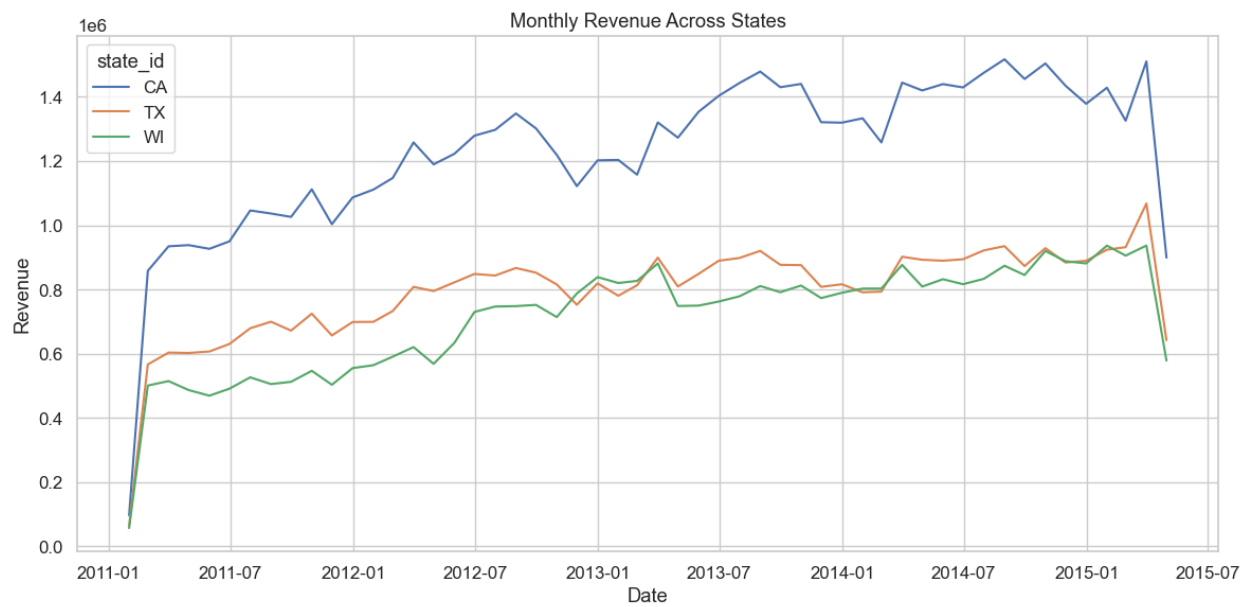
- GitHub [Link](#)
- Heroku app: [Link](#)
- 



●







A

