Semester Project - Walmart Sales Forecasting

Shalini Kothuru, Shivani Chennoju, NagaJahnavi Dhulipalla, Venkata Sai Abhigna Devarasetty

Abstract
This project's purpose is to anticipate future sales for Walmart, one of the country's largest retail companies. Accurate sales forecasting is essential for optimizing inventory levels, controlling workforce requirements, and boosting customer satisfaction. We want to address the intricacies of Walmart's business strategy, which can be influenced by a range of factors including seasonal trends, promotions, and regional market circumstances. We found and evaluated many cutting-edge solutions to this challenge, including time series and ensemble approaches. We hope to contribute to Walmart's operational success, improve its financial performance, and apply and extend our skills in data mining, machine learning, and other important sectors by enhancing the accuracy of its sales projections.

Keywords
NA as of this phase

[1] *Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, USA*

Contents

1. Problem and Data Description

1.1 Problem

The goal of our project is to forecast future sales for Walmart. Walmart is a retail chain, with many stores across the United States, and accurately forecasting sales is critical for optimizing inventory levels, managing staffing needs, and solving other business-related problems. The complexity of the Walmart business model is one of the major difficulties in solving this data mining problem. Sales can be affected by a variety of variables, such as seasonal trends, promotions, and regional market circumstances. The accuracy of the forecast may also be impacted by errors or inconsistencies in the data used for this project. Addressing these limitations would be an important aspect of our project.

1.2 Data Description

We are using four datasets for our project: ' calendar', 'sell_prices', 'sales_train validation', and 'sales_train_evaluation'.

1. Calendar dataset: It includes information on the dates on which the products are sold

Attributes Information of Calendar:

Date: Calendar Date

wm_yr_wk: Walmart internal week count(unique for a week)

weekday: represents days of the week (Starting from Saturday as 1 and ending with Friday as 7)

month, year: represent the current year and month of the calendar

d: Can be indexed, which helps in individually identifying the records

event_name_1, event_type_1, event_name_2, event_type_2: Represent if there is any festival/holiday and their type

snap_CA, snap_TX, snap_WI: This tells us whether it is a SNAP day or not (Supplemental Nutrition Assistance Program)

2. sell_prices:

It includes data on the prices of products sold per store and the day they were sold.

Attributes Information of sell_prices:

store_id: an identifier for a store where the product is sold

item_id: an identifier for a product

wm_year_week: a string representing the sale's year and week. (e.g. "2011-01" represents the first week of 2011)

sell_price: the cost of the product during that week.

3. sales_train validation:

It includes daily unit sales statistics for each product and retailer from d_1 to d_1913

Attributes Information of sales_train validation:

id: a unique identifier.

item_id: an identifier for a product

dept_id: an identifier for a department

cat_id: an identifier for a category

store_id: an identifier for a store where the product is sold.

d_1 to d_1913: is an identifier for a product sold each day.

4. sales_train_evaluation:

It has similar data to sales_train validation but this dataset is used to validate the model.

2. Data Preprocessing & Exploratory Data Analysis

2.1 Handling Missing Values

We found that none of the datasets contains null values or missing information except the calendar dataset. These columns have values only when there is an event on that particular day whereas, for other days, it has NaN values. We have replaced these NaN with 'No event'.

2.2 Exploratory Data Analysis

As we explored the data, we have derived the following insights from the given datasets. However, we don't have any scope to normalize or scale the data as it is already in a normalized state. Also, all the features are important. So we will be including all the features for the model building in the next phase.

1. Calendar dataset:

We have nearly 5.5 years of data from 29th January 2011 to 19th June 2016

We have 30 unique events from event_name_1 and 4 from event_name_2

On average, we have 5.5 holidays per year (avg holidays = total holidays/total time period = 30/5.5 i.e nearly 5.5 holidays). This indicates that we might have at least 5 upward trends in sales.

2. sell_prices:

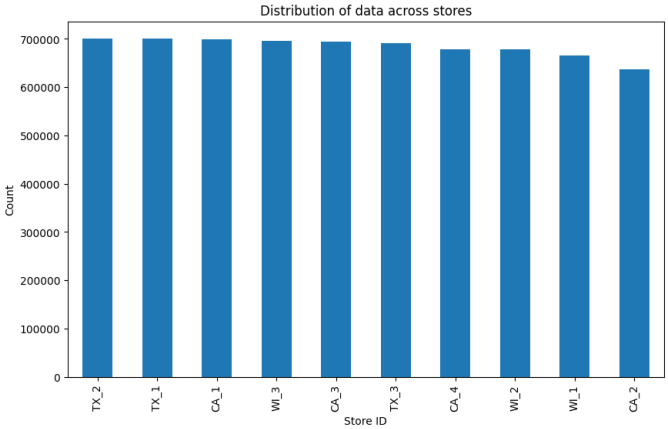
This dataset includes sales information of items sold in 4 different stores in California ('CA_1', 'CA_2', 'CA_3', 'CA_4') and 3 different stores in Texas('TX_1', 'TX_2', 'TX_3') and 3 different stores in Wisconsin('WI_1', 'WI_2', 'WI_3'). In total 10 different stores.

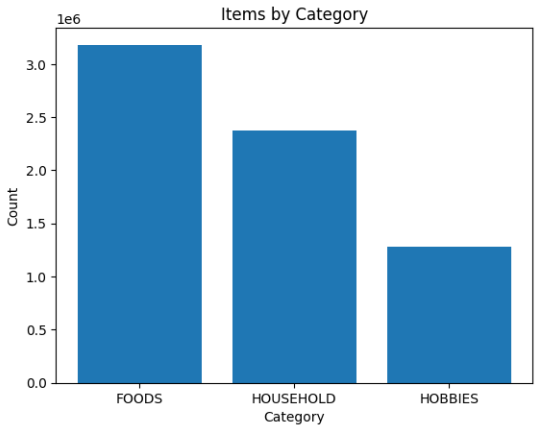
There are three different categories foods, household, and hobbies, and there is more number of items for sale from the foods category followed by household and hobbies.

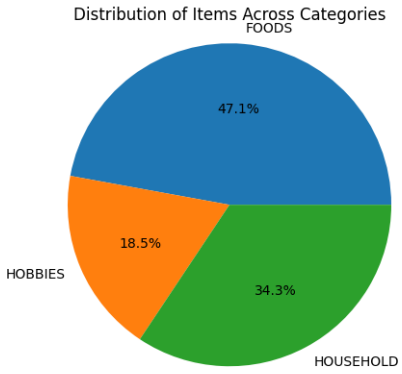
By understanding the data, it is found that the item having the highest average sale price is HOUSEHOLD_1_060, with an average sale price being 29.94.

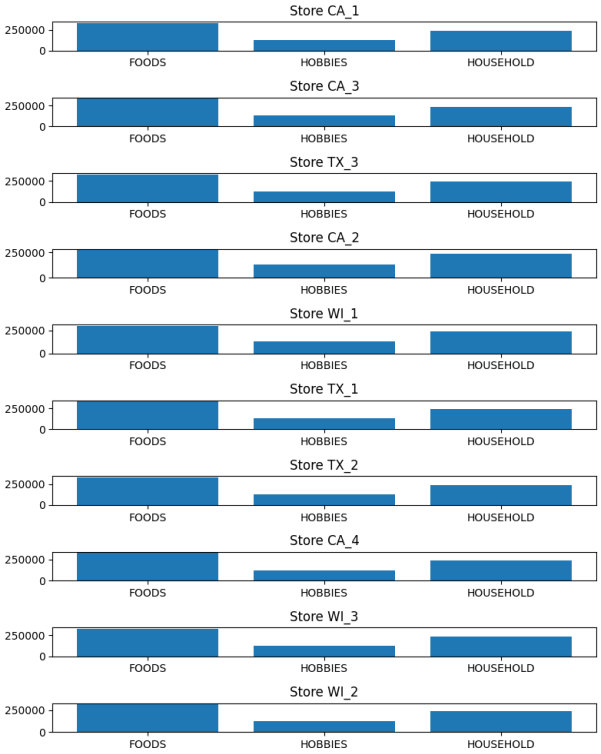
After finding the items with the highest average sell price per store, it can be seen that the items related to hobbies are expensive in stores in California('CA_1', 'CA_2', 'CA_3' and 'CA_4') and also in Washington('WI_1', 'WI_2' and WI_3) and items related to the household are expensive in stores of Texas('TX_1', 'TX_2', 'TX_3').

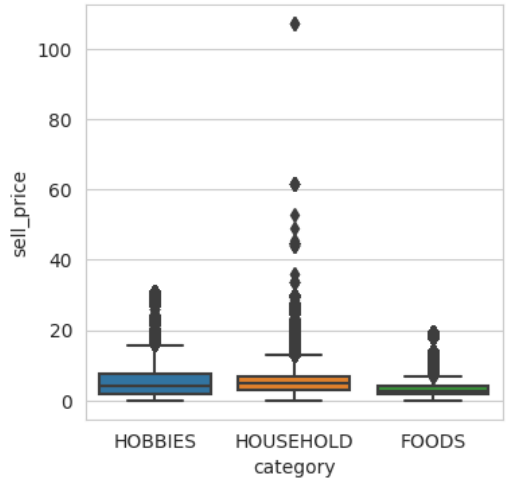
It can be seen that there are many price variations in the items belonging to foods and households, but there is less price variation for items related to hobbies.











3. sales_train validation:

We have more sales in California when compared to Texas and Wisconsin.

There are three different categories foods, household and hobbies, and there are more number of items for sale from foods category followed by household and hobbies.

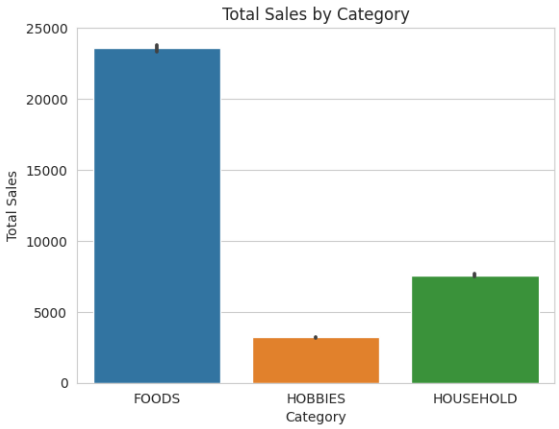
When compared with the different stores in different states, the sales in CA_3 have the most sales.

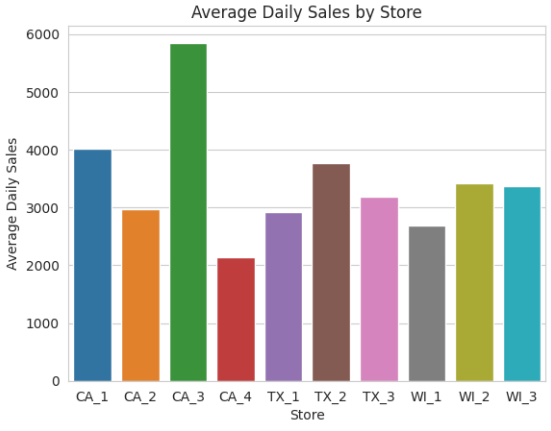
In California the sales during the starting period were around 10,000 and then gradually at the end of the period they increased to approx 25,000. Texas and Wisconsin sales are almost the same at the end of the period but at the starting point, Texas has more sales compared to Wisconsin.

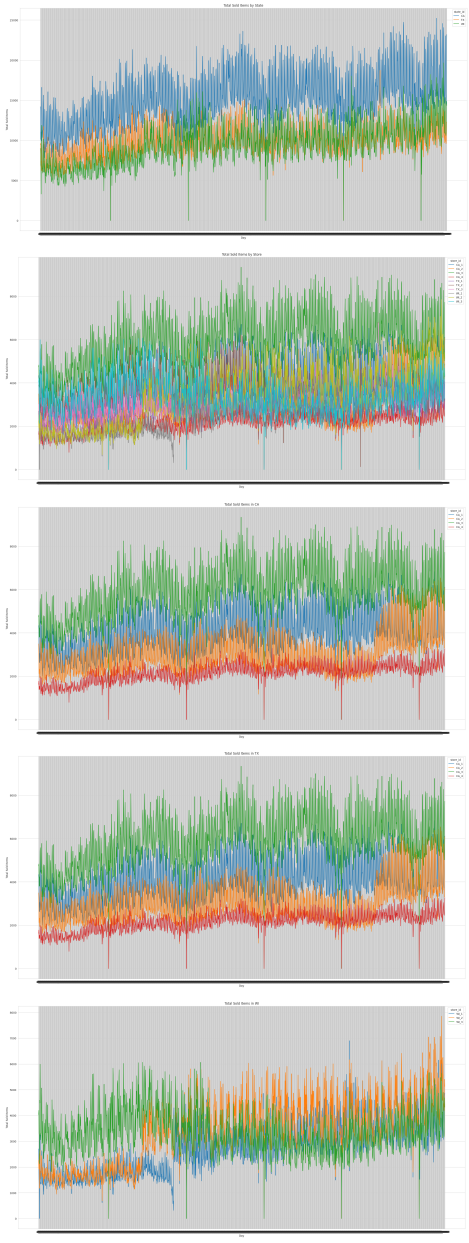
In California state the highest sales occurred in the CA_3 store and the lowest was in the CA_4 store.

In Texas state, all stores have the approx same sales towards at end of the period. While at the starting the TX_2 store has the highest sales.

In Wisconsin, the store WI_3 has the highest sales at the start but it gradually decreased at the end of the period whereas it is quite opposite in the sales of the store WI_2.







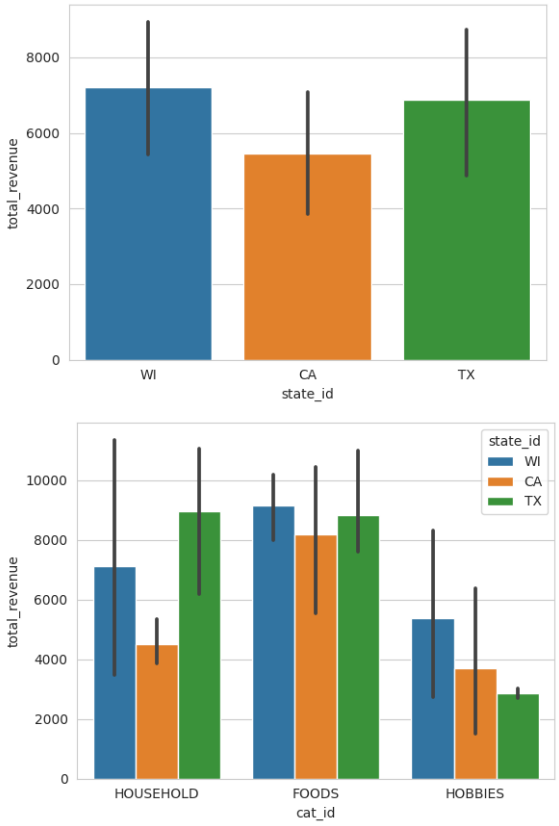
4. sales_train_evaluation: This dataset will be used for validation after building the model.

5. Insights from merged datasets:

We have merged sell_price and sales_train_validation data to extract more insights from the combined data about the revenue.

The total revenue generated by Wisconsin is highest and followed by Texas and California.

The total revenue generated by the category food sold in Wisconsin is highest among all the three categories and in states and hobbies is the least among all the states.



The next sections will be completed as we progress in our project.

3. Algorithm and Methodology

We have implemented the following algorithms:

1. ARIMA (AutoRegressive Integrated Moving Average)

A time series model that forecasts future values using past values and trends. It consists of three elements: autoregression (AR), differencing (I), and moving average (MA). Can be used to detect seasonality and trends in data, but only with stationary data. Following hyperparameter tuning, our optimal p,q,d values are 1,1,1.

2. SARIMAX((Seasonal AutoRegressive Integrated Moving Average with eXogenous factors)

An ARIMA extension that accounts for seasonal patterns and allows for the inclusion of exogenous variables. This model can handle both short-term and long-term seasonal patterns and includes seasonal P,D,Q values. Can be used to assess the impact of external factors such as promotions or holidays on sales. After hyperparameter tuning, our optimal values are (1,0,1),(0,1,1,7).

3. PROPHET

A simple time series model developed by Facebook that employs a decomposable additive model. Can handle trend

changes, seasonality, and holiday effects. Allows for the inclusion of exogenous variables and includes a built-in mechanism for dealing with missing data and outliers. To run the model, the ds(date) and y(attribute to be forecasted) columns are required.

4. XGBOOST

An ensemble of decision trees is used to predict in a machine learning model. It is capable of handling non-linear relationships between predictors and targets. Can automatically handle missing values and outliers. Can be used to assess the impact of a variety of sales predictors, including promotions, weather, and economic indicators.

5. LIGHTGBM

LightGBM is a fast and high performing gradient boosting framework based on decision trees which grows leaf-wise and reduces memory usage with better accuracy. We have performed hyperparameter tuning using Optuna to get the best parameters and trained the LGBM Regressor model using these parameters which resulted in better performance of the model.

Methodology Our team began by performing data preprocessing and exploratory data analysis (EDA) to gain insights into the Walmart sales dataset. We then implemented various time series models, including ARIMA, SARIMAX, and LightGBM, and utilized hyperparameter tuning to optimize the models' performance. We also combined multiple datasets, such as calendar, sell prices, and sales_train_validation, to enhance our analysis. With a training dataset spanning from 2011 to 2016 for all 30490 products, we predicted the sales for the next 28 days using the implemented algorithms. Finally, we compared the predicted results using sales_train_evaluation to obtain the RMSE values for all models. Our approach involved a combination of data science techniques and domain expertise to effectively forecast Walmart sales and identify key trends and factors driving product demand.

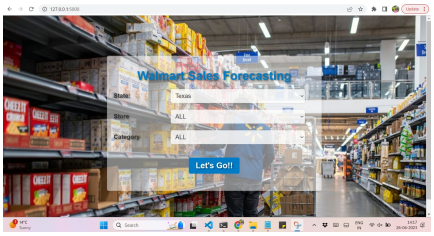
4. Experiments and Results

We have implemented 5 models. Out of all the algorithms implemented LightGBM performed better while XGBoost is the worst performing algorithm to forecast the sales. The table below lists the root mean squared error obtained for different models.

Models	RMSE
SARIMAX	1.769427337470288
ARIMA	1.6992314983921402
PROPHET	2.063853111133504
XGBOOST	2.7637794
LIGHTGBM	0.20426240252043837

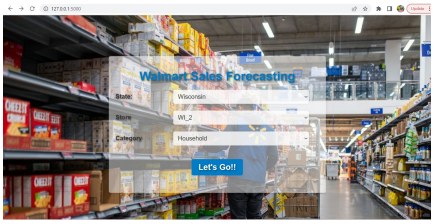
We have developed an application to forecast sales using the Flask framework and HTML. Below are the results for 2 different inputs as follows-

Scenario 1 : If we want to see the sales of all categories(FOODS, HOUSEHOLD, HOBBIES) in all stores located in Texas, we will give input as TEXAS for the state field and 'ALL' for the remaining fields.



Results obtained for the above input that shows sales of all categories in TX_1, TX_2 and TX_3.

Scenario 2 : Forecasting sales in a particular state in a particular category.



Results obtained for the above input that shows sales of HOUSEHOLD products in WI_2

5. Deployment and Maintenance

Along with the M5 forecasting project, we also created a website using Flask and HTML to share our work with a larger audience. Users can easily interact with our models and browse the Walmart sales dataset on the website.

As part of our long-term strategy, we want to move the website and our models to a cloud platform for simpler access and maintenance. This guarantees that users can still access the website and that the models are continually updated with the most recent information. In order to enhance our models and give users better insights, we also plan to incorporate user feedback into the website.

6. Summary and Conclusions

The M5 forecasting project was a difficult but rewarding experience in which we attempted to forecast future sales of 30490 products across diverse stores. We have used five different forecasting algorithms, ARIMA, SARIMAX, PROPHET, XGBOOST, and LIGHTGBM, to accomplish this.

We have gained a thorough understanding of the challenges and complexities involved in using big data in the retail industry by working with vast amounts of data from Walmart. By examining such sizable datasets, we have gained valuable knowledge about customer behavior, inventory management, and sales forecasting. Because of this experience, we have also gotten better at data wrangling, exploratory data analysis, and machine learning, all of which are crucial for success in the field of data science. Because of this project, we have the skills and knowledge necessary for success in the data science sector, as well as a solid foundation for working with large, complex datasets.

In conclusion, using a variety of algorithms and evaluating their effectiveness has allowed us to better understand the data and determine the most effective forecasting strategy. The accuracy and effectiveness of the models have increased because of the use of sophisticated techniques like hyperparameter tuning, which has given rise to invaluable insights into the potential future sales of the retail goods.

Acknowledgments

We owe a huge debt of gratitude to Indiana University Bloomington, our professor, and the Teaching Assistants for the priceless knowledge and abilities they have shared with us throughout this data mining course. Their commitment and diligence allowed us to complete the Walmart Sales forecasting project successfully and gain useful experience. We are now proficient with data mining tasks and have a solid understanding of all machine learning models. Utilizing superior data sets for assignments and projects has improved our knowledge. We will always be grateful for the advice, support, and lessons they have given us. From the bottom of our hearts, thank you!

References

References

1. https://www.kaggle.com/competitions/m5-forecasting-accuracy /overview
2. https://www.sciencedirect.com/science/article /pii/S016920702100187
3. https://towardsdatascience.com/m5-forecasting-accuracy-24d7f42130de