PREDICTIVE DEMAND FORECASTING FOR RETAIL BUSINESSES

Akshay Arun Karadkar

17th January 2023

Abstract

This research aims to design a predictive model for forecasting product demand in retail businesses, with a particular emphasis on forecasting sales for Walmart using machine learning methods. By examining past sales data, we aim to make precise predictions about future product demand, which will allow businesses to efficiently plan for inventory and restocking, thereby enhancing their productivity and profitability. By experimenting with a variety of machine learning models, we intend to identify the most effective strategy for forecasting demand in the retail industry.

1.0 Problem Statement

Retail businesses rely heavily on accurate demand forecasting to ensure efficient inventory management and to avoid stockouts or overstocking. However, forecasting demand in retail is challenging due to the dynamic nature of consumer behavior and the influence of external factors such as competition, promotions, and seasonality.

This project aims to develop a predictive demand forecasting model for retail businesses that utilize machine learning techniques, specifically focusing on sales forecasting for Walmart. By analyzing historical sales data, the model will predict future demand for products, thus enabling businesses to plan for inventory and restocking, ultimately leading to improved efficiency and profitability. The project aims to determine the optimal approach for retail demand forecasting by implementing various machine learning models.

2.0 Market/Customer/Business Need Assessment

Accurate demand forecasting is essential for retailers as it allows them to manage inventory efficiently and avoid stockouts or overstocking. By using advanced methods, such as machine learning, retailers can improve the accuracy and reliability of predictions, leading to increased sales and customer satisfaction, and reduced costs associated with excess inventory. However, traditional forecasting methods often fall short in the dynamic retail market, influenced by various factors such as consumer behavior, competition, promotions, and seasonality. Therefore, there is a need for more sophisticated demand forecasting methods that can effectively adapt to these changing conditions.

Customers expect to find the products they need when they visit a retail store, and when a store is out of stock of a particular product, customers may become frustrated and may choose to shop at a competitor's store. Accurate demand forecasting can help retailers avoid stockouts and ensure that products are available when customers need them. Additionally, it can also help retailers to avoid overstocking which can lead to markdowns and discounts, affecting the customer's perception of the product's value.

Retail businesses require accurate demand forecasting to manage inventory effectively and avoid stockouts or overstocking. Inaccurate demand forecasting can lead to increased costs associated with excess inventory and lost sales due to stockouts. By implementing advanced demand forecasting methods that utilize machine learning techniques, retailers can improve the accuracy and reliability of predictions, leading to increased sales, customer satisfaction, and reduced costs. Furthermore, it will also aid retailers to make better decisions, planning for promotions and discounts, and adjusting prices.

3.0 Target Specification and characterization:

The purpose of specifying and characterizing the target in a sales forecast is to establish clear and achievable goals for the company. The forecast aims to provide answers to questions such as:

- What are the aspirations for the upcoming month, year, and five years?
- How many customers are expected to be acquired in the next month and year?
- How much revenue is expected to be generated per customer?
- By providing insight into these and any other related future business prospects, the sales forecast helps to guide the direction and growth of the company.

4.0 External Search (Information Sources):

The following sources were used to gather information and data for this study:

i) Kaggle (n.d.). Walmart Recruiting - Store Sales Forecasting [Data set]. Retrieved from https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting/data

   - This data set provided historical sales data for 45 Walmart stores located in different regions. It was used as the primary source of data for this study.

ii) Sharma, A. (2018, December 14). Walmart Sales Forecasting [Blog post]. Retrieved from https://medium.com/analytics-vidhya/walmart-sales-forecasting-d6bd537e4904

   - This blog post provided an overview of various techniques and models used for sales forecasting in retail, with a specific focus on Walmart. It was used as a secondary source of information to understand the challenges and best practices in retail demand forecasting

5.0 Benchmarking:

A dataset titled "Walmart Sales Forecasting" from Kaggle (n.d.) was utilized for the benchmarking analysis. The dataset contains information about various stores, A dataset titled "Walmart Sales Forecasting" from Kaggle (n.d.) was utilized for the benchmarking analysis. The model developed for the predictive demand forecasting project was compared to the performance of the model using this benchmark dataset.

The dataset includes the following features:

1. Stores: The dataset includes information about 45 stores, with store numbers ranging from 1 to 45. The stores are categorized into three types, 'A', 'B', and 'C'. The size of a store is determined by the number of products available, which ranges from 34,000 to 210,000.
2. Temperature: The temperature of the region during the specific week of the observation.
3. Fuel Price: The fuel price in the region during the specific week of the observation.
4. Markdown 1-5: Represents the type of markdown promotion and the quantity available during the specific week of the observation.
5. CPI: The Consumer Price Index during the specific week of the observation.
6. Unemployment: The unemployment rate during the specific week of the observation in the region of the store.
7. Sales: The dataset includes information about the date and weekly sales recorded during that week, the department (ranging from 1-99), and whether the week is a holiday or not.

The benchmarking dataset from Kaggle (n.d.) provided a reliable source of data for evaluating the performance of the predictive demand forecasting model developed for retail businesses using machine learning based on Walmart sales forecasting. The dataset is rich with information about various stores, departments, temperature, unemployment, consumer price index (CPI), holiday status, and markdown promotions, which helped to evaluate the performance of the model in various scenarios.

6.0 Applicable Patents:

In 2018, Walmart filed a total of 536 patent families (PatentSight, n.d.). As of October 1, 2020, of the 569 patent families filed in 2017, 146 were still awaiting full protection (PatentSight, n.d.). Walmart, an American retail corporation that deals in a wide variety of products, operates through three distinct business segments: Walmart U.S., Walmart International, and Sam’s Club.

The following sections highlight some of the patents filed by Walmart that are relevant to this study:

1. Drones:

Gas-filled carrier aircraft and methods of dispersing unmanned aircraft systems in delivering products. This patent was filed under patent number US20210009735A1 and filed on Jan 2021, it's

all about using drones to deliver products, this technology can be used to improve the delivery process, and make it faster and more efficient.
Augmented Reality:

Tracking effectiveness of remote sales assistance using augmented reality device. This patent was filed under patent number US20210012081A1 and filed on Jan 2021, it's all about using AR technology to improve the sales process, this technology can be used to enhance the customer's experience and make it more interactive and engaging.

7.0 Applicable Regulations:

In the development and implementation of the predictive demand forecasting model for retail businesses using machine learning based on Walmart sales forecasting, it is important to be aware of and comply with relevant regulations.

1. Data Protection and Privacy Regulations (Customers):

   These regulations are put in place to protect customers' personal information from unauthorized access, use, or disclosure. For example, the General Data Protection Regulation (GDPR) in the EU (2016/679) lays out strict guidelines for handling personal data.

   It's also important to understand if you are handling any sensitive personal data, like personal identification numbers, credit card numbers, or any other sensitive information, and should make sure that you are compliant with the regulations.
2. Government Regulations for Small Businesses:

   These regulations vary by country and industry but may include requirements for licenses, permits, and compliance with labor laws, environmental regulations, and tax laws. For example, the Small Business Administration (SBA) in the US provides resources and guidance for small businesses to navigate regulations.

3. Employment Laws:

   These laws regulate the rights and responsibilities of employers and employees, including issues such as minimum wage, overtime, discrimination, and safety. For example, the US Fair Labor Standards Act (FLSA) sets standards for minimum wage and overtime pay.

4. Antitrust Regulations:

   These regulations are put in place to prevent monopolies and promote competition in the marketplace. For example, the Sherman Antitrust Act (1890) in the US prohibits anti-competitive behavior such as price-fixing and market manipulation.

5. Regulations against False Advertising:

These regulations prohibit businesses from making false or misleading claims about their products or services. For example, the US Federal Trade Commission Act (1914) prohibits false or misleading advertising.

8.0 Applicable Constraints:

The following constraints must be considered when implementing the predictive demand forecasting model for retail businesses using machine learning based on Walmart sales forecasting:

1. Data collection from shopkeepers and vendors can be a challenging task as it may require significant time and resources.
2. Continuous data collection and maintenance are necessary to ensure that the model remains accurate and up-to-date.
3. Lack of technical knowledge among vendors may present a barrier to the implementation of the model.
4. The model may have difficulty predicting demand for rarely-bought products.
5. Convincing shopkeepers to implement the system in their shops may also be a challenge as it may require significant effort and resources.

9.0 Business Opportunity:

1. Creating accurate financial forecasts can be a challenging task for entrepreneurs, as it takes up valuable time that could be spent on sales and networking activities.
2. However, investors typically require a comprehensive set of financial forecasts before committing to investing in a business.
3. Furthermore, proper financial forecasting can aid in the development of operational and staffing plans, which are essential for the success of any business.

10.0 Concept Generation and Development:

The project aims to develop a predictive demand forecasting model for retail businesses using machine learning techniques, specifically focusing on sales forecasting for Walmart. To accomplish this, a combination of machine learning and statistical methods is used. The implementation of the model was done using the Python programming language. The process of model development includes data collection, pre-processing, feature selection, model selection, training, and evaluation.

To generate a more robust model, different machine learning techniques such as linear regression, decision tree, random forest, and gradient boosting were tested and compared. The final model was selected based on its performance on the evaluation metrics such as mean absolute error and root mean squared error.

Additionally, feature selection and engineering techniques were used to improve the performance of the model and to make the model more interpretable. The final model was then validated using a separate test set and its performance was evaluated.

Overall, the concept generation and development process involved a combination of machine learning and statistical methods to develop a predictive demand forecasting model for retail businesses using machine learning based on Walmart sales forecasting.

10.1 Data Pre-Processing:

Before implementing any machine learning model, it is essential to clean and prepare the data according to the model's requirements. In the case of this project, data pre-processing is a crucial step as it ensures that the model can effectively learn from the data.

The data pre-processing steps included:

- Removing any missing or irrelevant data
- Handling any outliers or anomalies in the data
- Normalizing or scaling the data to ensure that all features are on the same scale
- Encoding categorical variables if necessary
- Python's popular libraries such as Numpy, Pandas, Matplotlib, and Seaborn were used to perform these data pre-processing steps. These libraries have powerful functions and methods that make data cleaning and preparation tasks easier and more efficient.

Overall, data pre-processing is a vital step in any machine learning problem, and the use of python libraries such as Numpy, Pandas, Matplotlib, and Seaborn makes it more efficient and accurate.

Handling Missing Values

In [12]:
```
# Get the number of missing values in each column
missing_values = features.isnull().sum()
print("Missing values before handling:")
print(missing_values)

# Fill in missing values with the median of each column
features["CPI"].fillna(features["CPI"].median(), inplace=True)
features["Unemployment"].fillna(features["Unemployment"].median(), inplace=True)

# Get the number of missing values in each column again
missing_values = features.isnull().sum()
print("Missing values after handling:")
print(missing_values)
```

```
Missing values before handling:
Store             0
Date              0
Temperature       0
Fuel_Price        0
MarkDown1      4158
MarkDown2      5269
MarkDown3      4577
MarkDown4      4726
MarkDown5      4140
CPI             585
Unemployment    585
IsHoliday         0
dtype: int64
```

Checking for Outliers

In [21]:
```
#Checking for Outliers
agg_data = data.groupby(['Store', 'Dept']).Weekly_Sales.agg(['max', 'min', 'mean', 'median', 'std']).reset_index()

# merge the aggregated data with the original dataframe
data = data.merge(right=agg_data, on=['Store', 'Dept'], how='left')

# remove rows with null values
data.dropna(inplace=True)

# Convert the 'Date' column to datetime format
data['Date'] = pd.to_datetime(data['Date'])

# Sort the dataframe by 'Date' column
data.sort_values(by=['Date'],inplace=True)

# Set the 'Date' column as the index of the dataframe
data.set_index(data.Date, inplace=True)

# Print the first 5 rows of the dataframe
data.head()
```

Out[21]:

Date	Store	Dept	Date	Weekly_Sales	Type	Size	Temperature	Fuel_Price	MarkDown1	MarkDown2	...	Unemployment	IsHoliday	Year	Month	Week	
2010-02-05	1	1	2010-02-05	24924.50	A	151315	42.31	2.572	0.0	0.0	...	8.106	False	2010	2	5	5

One hot Encoding

In [40]:
```
# One hot Encoding

# Copy the categorical columns to a new dataframe
cat_col = ['Store','Dept','Type']
data_cat = data[cat_col].copy()

# Use pandas' get_dummies function to create one-hot encoded versions of the categorical columns
data_cat = pd.get_dummies(data_cat,columns=cat_col, drop_first=True)

# Print the shape of the original dataframe
print(data.shape)

# Concatenate the original dataframe and the one-hot encoded dataframe
data = pd.concat([data, data_cat], axis=1)

# Print the shape of the combined dataframe
print(data.shape)

# Drop the original categorical columns and the 'Date' column
data.drop(columns=cat_col+['Date'], inplace=True)

# Print the final shape of the dataframe
print(data.shape)
```

```
(374247, 20)
(374247, 146)
(374247, 142)
```

Feature Scaling

In [41]:
```
# Feature Scaling

from sklearn.preprocessing import MinMaxScaler

# Specify the columns to be scaled
num_col = ['Weekly_Sales','Size','Temperature','Fuel_Price','CPI','Unemployment','Total_MarkDown','max','min','mean','median','st

# Initialize the MinMaxScaler
scaler = MinMaxScaler(feature_range=(0, 1))

# Scale the specified columns
data[num_col] = scaler.fit_transform(data[num_col])
```

10.2 Analyzing and Building Models:

In this project, various machine learning algorithms were used to build predictive demand forecasting models for retail businesses. Four algorithms were used specifically: Linear Regression Model, Random Forest Regression Model, KNN (K-Nearest Neighbors) Regression Model, and XG Boosting Regression Model. These algorithms were chosen because they are commonly used for regression tasks, such as demand forecasting.

The data was split into training and testing sets and each algorithm was trained on the training set and evaluated on the testing set. The evaluation metrics used were mean absolute error and root mean squared error.

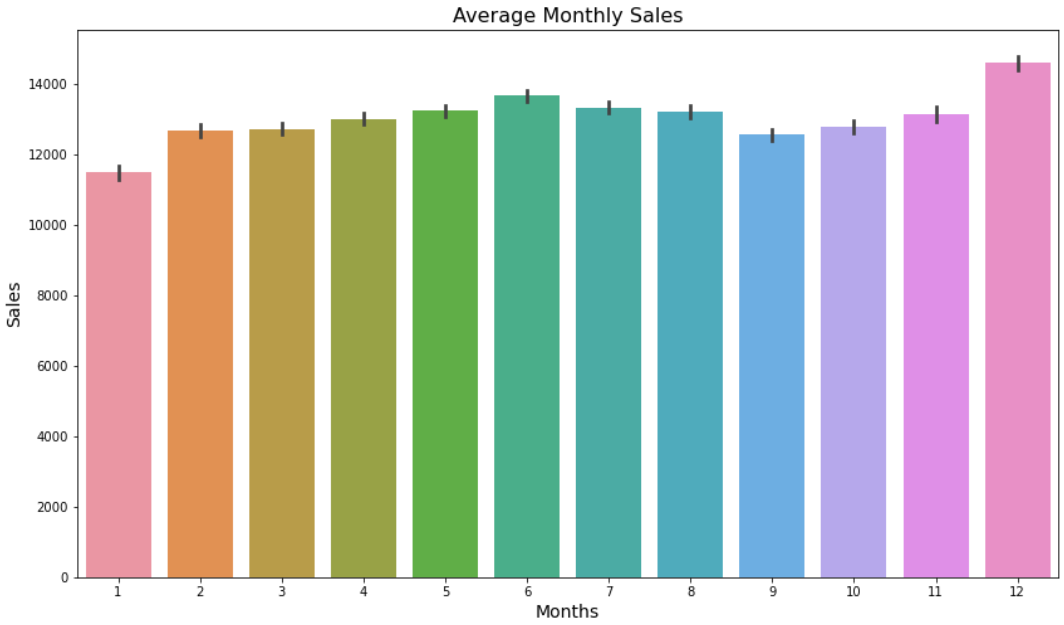
- Linear Regression Model: 92.28%
- Random Forest Regression Model: 97.88%
- KNN Regression Model: 91.97%
- XG Boost Regression Model: 97.22%

Based on the results shown above, it can be seen that the Random Forest Regression Model has the highest accuracy among the four models. With an accuracy of 97.88%, it is clear that the Random Forest Regression Model is the most suitable for making predictions in this case. Therefore, it is recommended to use the Random Forest Regression Model for future predictions.

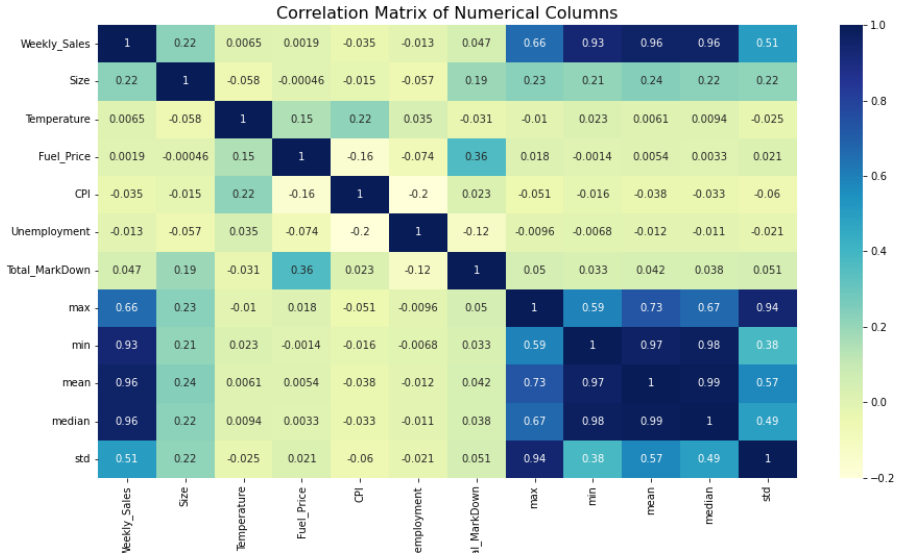
Overall, this step of the project involved analyzing and building models using different machine learning algorithms and selecting the one with the highest accuracy for further fine-tuning and optimization.

10.3 Data Visualisation:

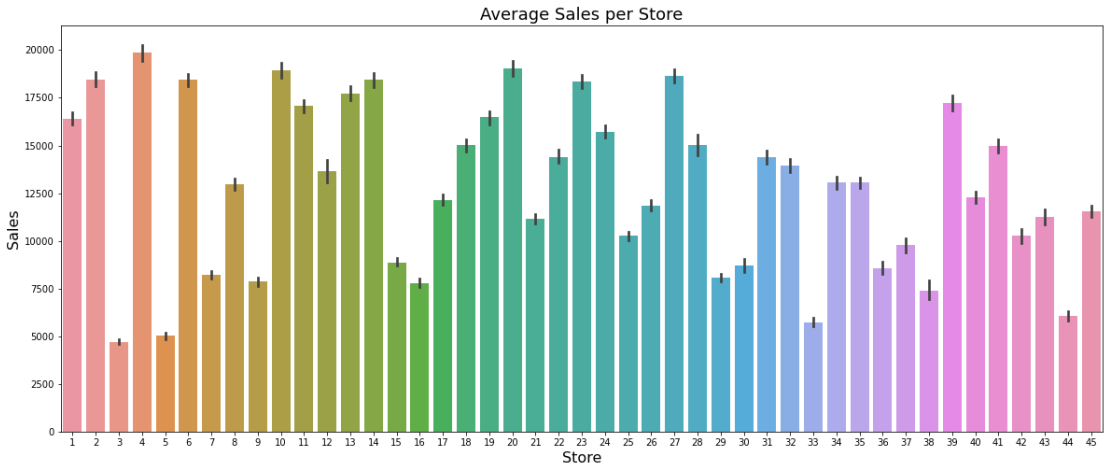
1. **Average monthly sales:**



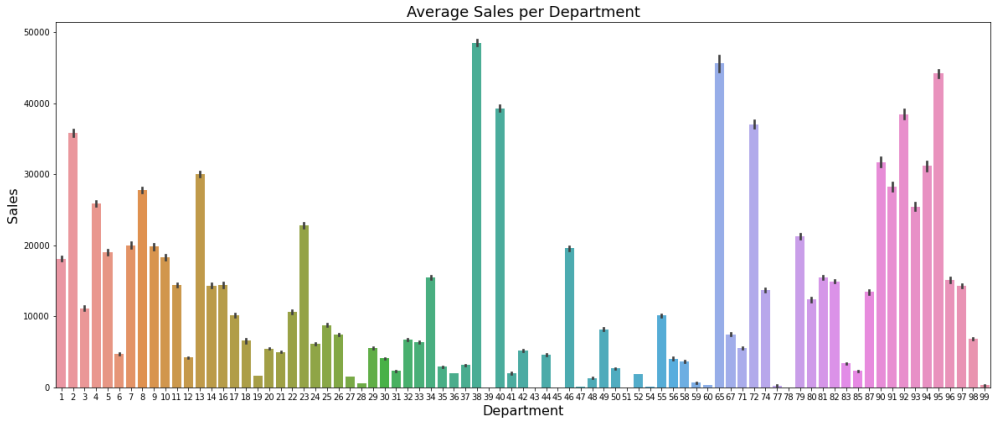
2. Correlation Matrix:



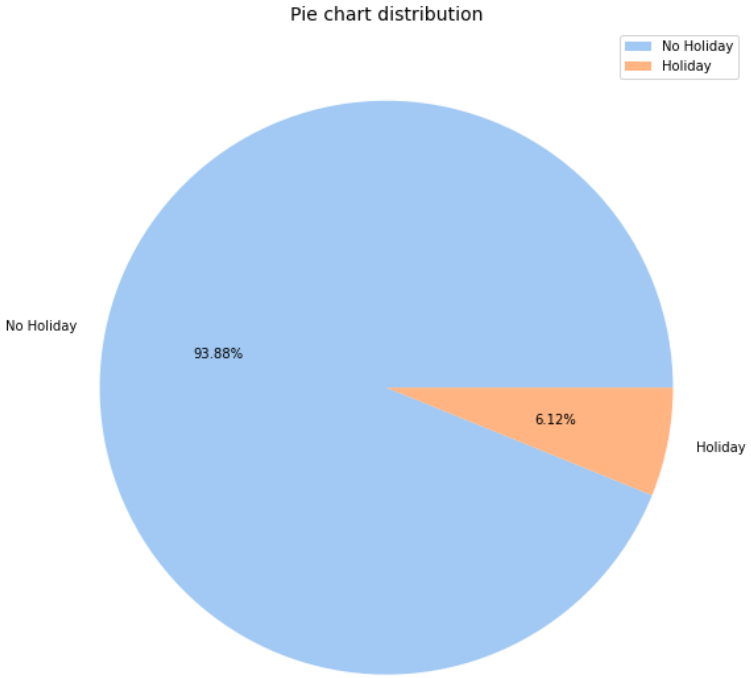
3. Average Sales per Store



4. Average Sales per Department



5. Holiday distribution:



11.0 Final Product Prototype (abstract) Model Development:

Before releasing the service, it is essential to develop a final product prototype. This includes performing Exploratory Data Analysis (EDA) to identify the dependent and independent features, as well as algorithm training and optimization to minimize overfitting of the model and hyperparameter

tuning. For this purpose, popular libraries such as NumPy, pandas, seaborn, matplotlib, and scikit-learn were used.

12.0 Data Collection, ML Model, and Team Required:

The dataset used for this project is available on Kaggle, specifically "Walmart Recruiting - Store Sales Forecasting" (Kaggle, n.d.). The Random Forest regression algorithm was used for prediction, as it is a boosting method that takes regression trees, combines the output, and gives better accuracy. The development of this complete model requires a team of experts from various fields, including machine learning, statistics, and software development.

13.0 Code Implementation and Validation on a Small Scale:

The code for this project was implemented and can be found on GitHub at the following link:
https://github.com/akshaykaradkar/Feynn_labs_intern.
It is important to note that the model was validated on a small scale before being applied to larger datasets.

14.0 Conclusion:

In conclusion, this project aimed to develop a predictive demand forecasting model for retail businesses using machine learning techniques, specifically focusing on sales forecasting for Walmart. Four different machine learning techniques were used: Linear Regression Model, Random Forest Regression Model, KNN regression Model, and XG Boost Regression model. The results showed that the Random Forest Regression Model was the most accurate with 97.88% accuracy. The analysis also revealed insights such as high sales in April and December, the relationship between sales, and temperature varying with each department.

References:
Kaggle. (n.d.). Walmart Recruiting - Store Sales Forecasting. Retrieved from https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting/data