Retail Sales Prediction and Forecasting for American Retailers

<Trung Kien Hoang>

<06/10/2023>

<Project code> : <https://github.com/KenUTS/American-retailers.git>

<Project API> : <https://github.com/KenUTS/Retail_api.git>

Table of Contents

1. Executive Summary

The primary aim of this business project is to construct a prediction model by employing a Machine Learning algorithm. This model will be utilized to accurately anticipate the sales revenue of a specific product within a designated retailer on a given day. One additional aim of this project is to construct a forecasting model employing a time-series analysis methodology in order to anticipate the collective sales revenue for all shops and items throughout the forthcoming 7-day timeframe. There exists a correlation between sales income and date-related factors, including the day of the month, month of the year, and day of the week. Additionally, there exists a correlation between the quantity of sales on days with activities and those without events. It is imperative to take into account the significance of this matter, as some commodities, such as food products, are frequently purchased during events. Furthermore, the consideration of date features holds significance in examining this project, as individuals tend to engage in shopping activities during weekends or designated shopping days. Customers are aware of the recurring nature of discount dates, which are consistently observed on specified days or months. These factors merit consideration in the context of this task. The achievement of certain objectives plays a significant role in enhancing sales optimization and the development of efficient marketing strategies for a retail establishment in the United States. Nevertheless, the occurrence of erroneous results might lead to substantial cost repercussions in relation to the transportation and warehousing of commodities, specifically in the context of perishable food products.

2. Business Understanding

a. Business Use Cases

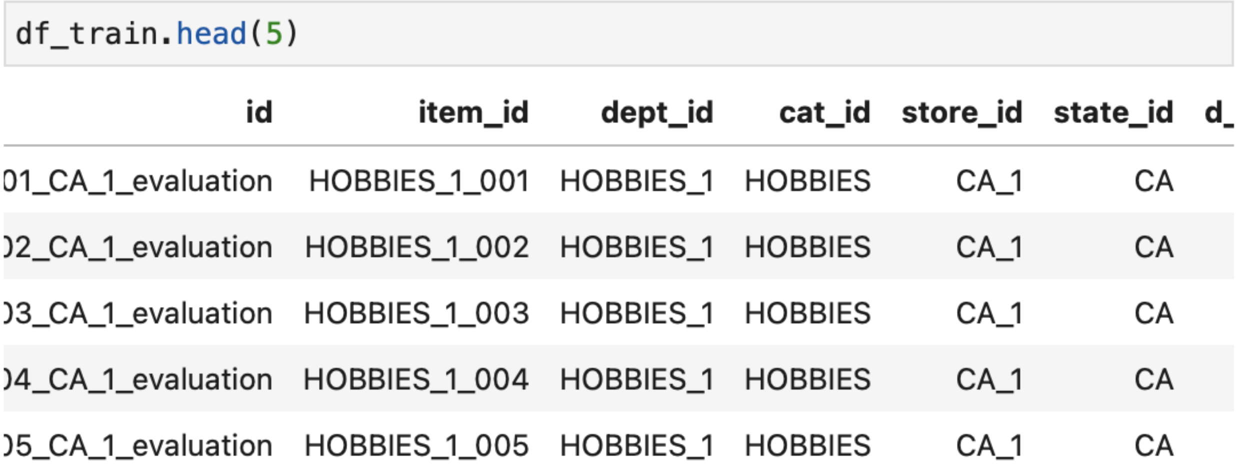
The findings will be utilized by an American retailer to inform their sales strategies, such as implementing discount promotions on days with lower revenues. Additionally, this tool can assist in the identification of stores that are experiencing poor sales performance for a given item. This enables retailers to implement appropriate solutions, such as optimizing item delivery and implementing effective marketing strategies. Similarly, in the context of a store generating substantial money from a particular item, it is advantageous for the store to prioritize the delivery of a larger quantity of its top-selling items in order to optimize its overall earnings. The attainment of precise outcomes contributes to the optimisation of sales and the formulation of effective marketing strategies for an American store. Obtaining accurate predictions for the income of individual goods poses issues due to the large number of things and the inherent difficulty in developing a model for each item. Furthermore, potential associations can be observed between sales income and variables such as the day of the week, day of the month, or month of the year. Hence, machine learning methods hold significance in this particular environment for the purpose of ascertaining said relationships.

b. Key Objectives

The primary goal of this project is to implement a predictive model that utilizes many inputs, including date, events, item, and shop identifications, in order to forecast sales income. Another goal of this project involves the implementation of a predictive model that uses date input to forecast the aggregate income for all retail establishments. The primary stakeholder for this project is a retailer based in the United States. In more precise words, a retail management team uses this methodology to enhance the efficiency of stock and personnel allocations. Store managers can employ this tool to strategize discount promotions and allocate staff resources effectively. Additionally, the supply team can utilize this platform to streamline the delivery of stocks to stores in need. In order to fulfill these objectives, the models undergo initial training and validation using the provided dataset in order to determine the optimal model performances. The model utilizes store and date inputs to assist a retail management team in identifying the store with the highest sales volume for a certain item. The model has the capability to accept item inputs, thereby assisting shop managers in identifying busy departments and facilitating staff allocation. The delivery team might utilize the algorithm to accurately predict the high-demand items in a store on a certain day, enabling them to efficiently allocate stock to stores that require replenishment.

3. Data Understanding

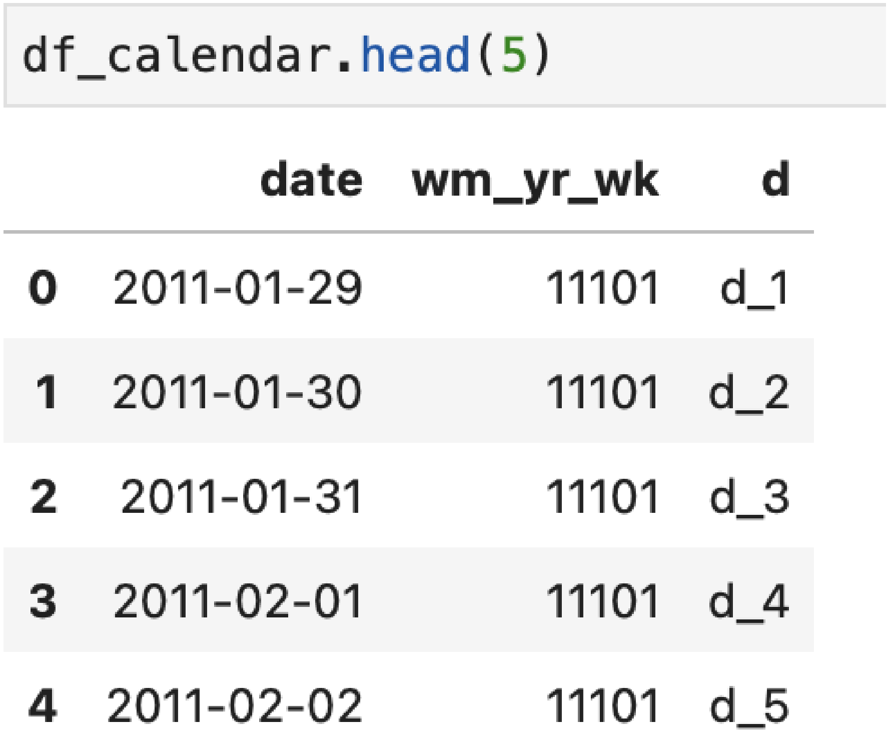
The given datasets are collected from an American retailer which contain training, evaluating, calendar, event and weekly price of items datasets.



```
df_train.head(5)
```

id	item_id	dept_id	cat_id	store_id	state_id	d_
01_CA_1_evaluation	HOBBIES_1_001	HOBBIES_1	HOBBIES	CA_1	CA	
)2_CA_1_evaluation	HOBBIES_1_002	HOBBIES_1	HOBBIES	CA_1	CA	
)3_CA_1_evaluation	HOBBIES_1_003	HOBBIES_1	HOBBIES	CA_1	CA	
)4_CA_1_evaluation	HOBBIES_1_004	HOBBIES_1	HOBBIES	CA_1	CA	
)5_CA_1_evaluation	HOBBIES_1_005	HOBBIES_1	HOBBIES	CA_1	CA	

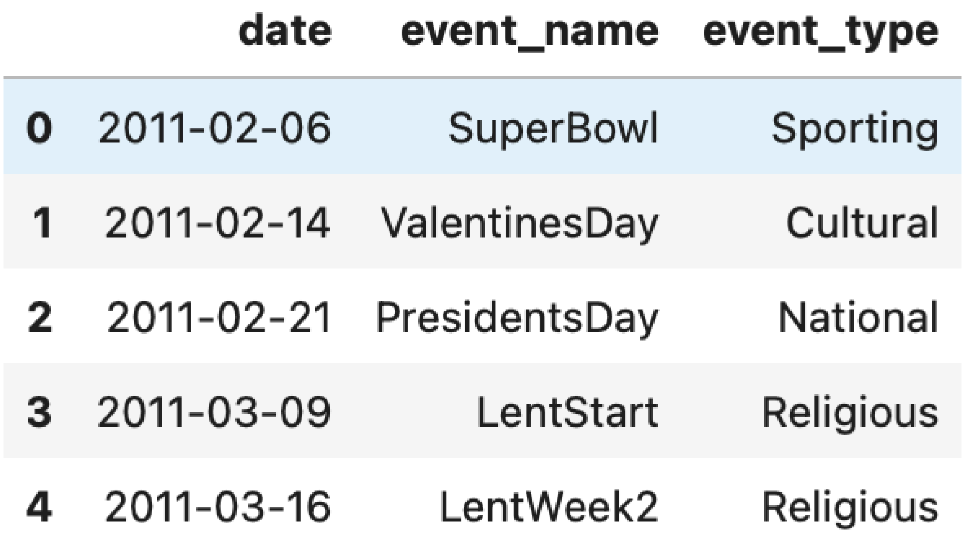
The training dataset comprises various data attributes, including item identifiers, department identifiers, category identifiers, store identifiers, state identifiers, and a collection of records spanning over 1000 days.



```
df_calendar.head(5)
```

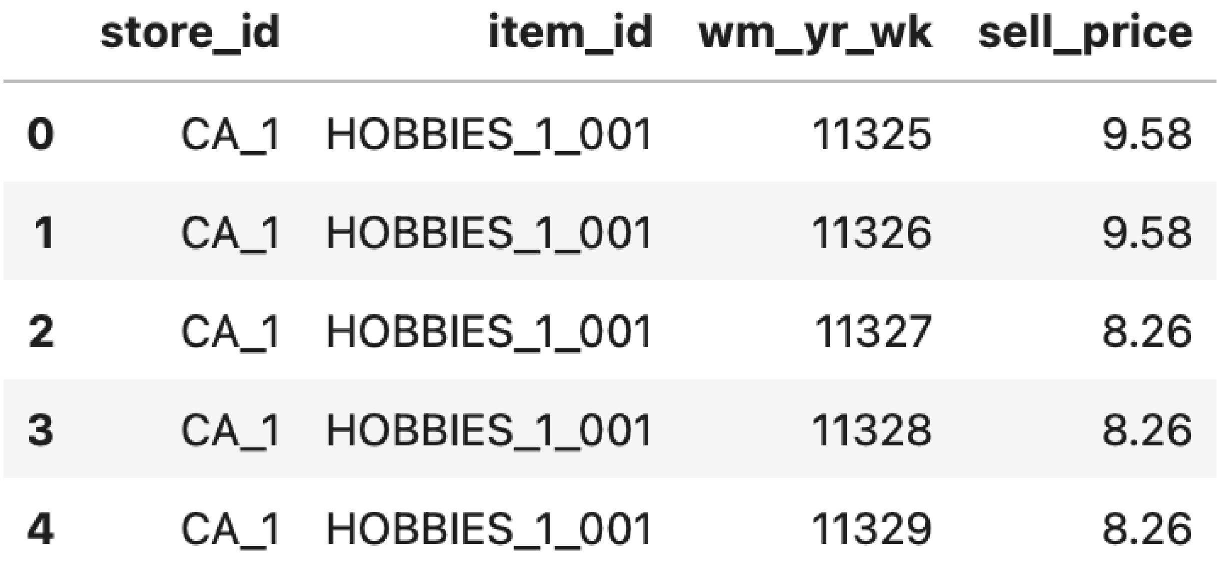
	date	wm_yr_wk	d
0	2011-01-29	11101	d_1
1	2011-01-30	11101	d_2
2	2011-01-31	11101	d_3
3	2011-02-01	11101	d_4
4	2011-02-02	11101	d_5

The dataset pertaining to the calendar comprises dates represented in the format of year-month-date, together with information regarding the number of weeks elapsed since the start and the commencement of the day.



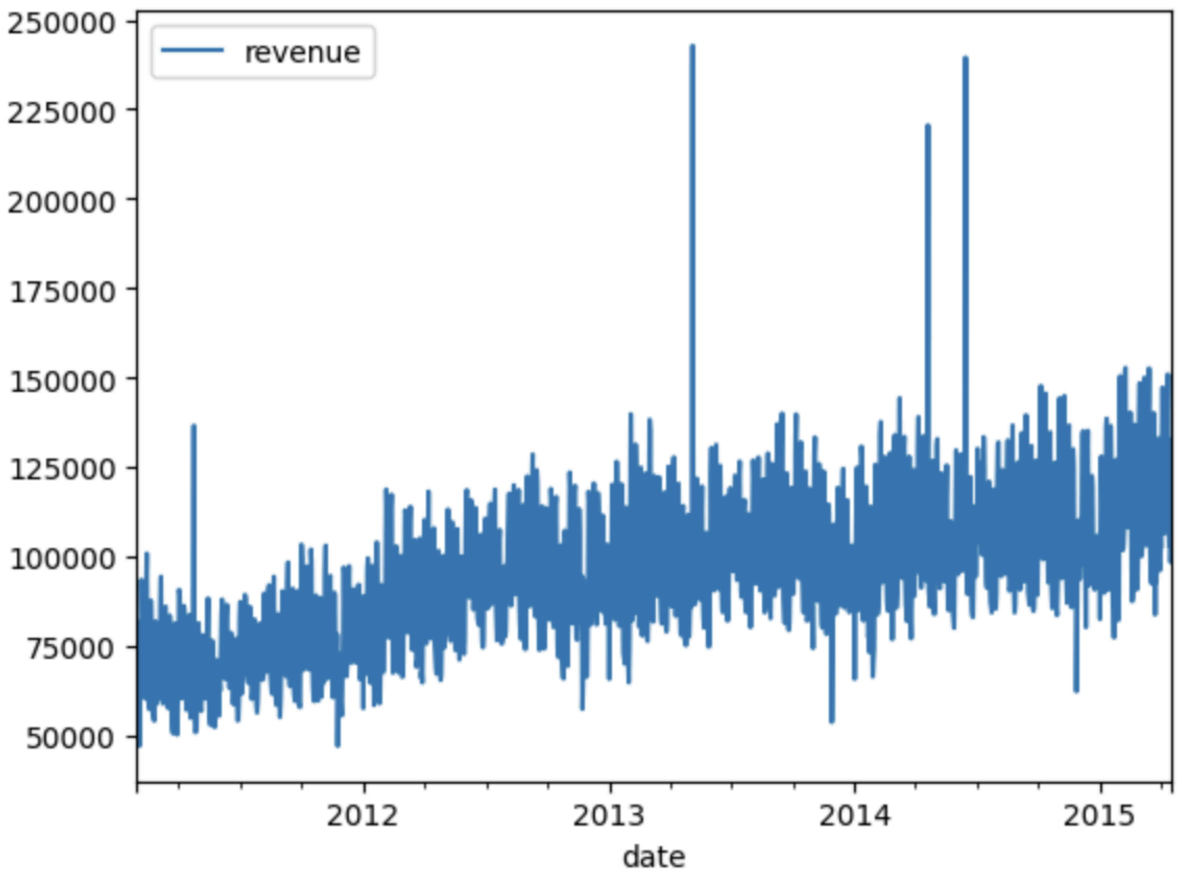
	date	event_name	event_type
0	2011-02-06	SuperBowl	Sporting
1	2011-02-14	ValentinesDay	Cultural
2	2011-02-21	PresidentsDay	National
3	2011-03-09	LentStart	Religious
4	2011-03-16	LentWeek2	Religious

The event dataset comprises dates in a specific format, together with the corresponding names and types of events. It is possible for a single date to have numerous events associated with it.



	store_id	item_id	wm_yr_wk	sell_price
0	CA_1	HOBBIES_1_001	11325	9.58
1	CA_1	HOBBIES_1_001	11326	9.58
2	CA_1	HOBBIES_1_001	11327	8.26
3	CA_1	HOBBIES_1_001	11328	8.26
4	CA_1	HOBBIES_1_001	11329	8.26

The weekly pricing data includes the unique identification for each item, the identifier for the store, the number of weeks since the start, and the corresponding price for each item.



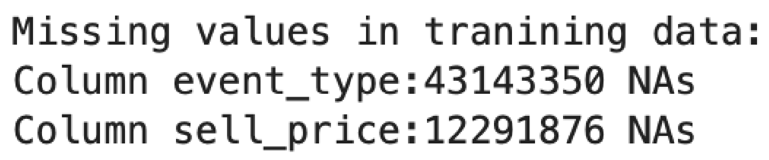
The graph presented above illustrates a gradual increase in total income from 2011 to 2015. Moreover, it is evident that there is a noticeable seasonal tendency present in this dataset, which will be further investigated in the subsequent stages of this project. Furthermore, there exist certain outliers characterized by exceptionally high income figures, which might perhaps be attributed to the specific event date.

4. Data Preparation

The initial stage in data preparation for modeling is the consolidation of all datasets into a unified final dataset. The training data undergoes a melt process in which the columns representing the date are transformed into a new column labeled 'date'. After that, there are combining processes that are perform by following:

1. Combining calendar and events datasets.
2. Combine the melted dataset vs combined datasets above.
3. Combine the dataset above with items' weekly price dataset.

Once the final dataset is generated, the missing records are examined inside the dataset.



```
Missing values in tranining data:
Column event_type:43143350 NAs
Column sell_price:12291876 NAs
```

The empty entries inside the weekly sales price records are populated with the mean cost of identical items across various establishments. Once the null values have been addressed, unused columns such as sale, day, wm_yr_wk, sell_price, and event_type are eliminated in order to conserve memory during the training of the model. For data splitting, as a result of the substantial size of the final dataset, it has been partitioned into seven smaller subsets, taking into consideration the various departments across all stores. The subsets are partitioned into training and validation datasets, with a ratio of 8:2 for the purpose of training and evaluating models. Finally, a stratify technique is performed to ensure that the distributions of items in training and validating data.

For this project, a novel variable denoted as "revenue" is generated through the multiplication of the sales figure by the price of each item. A newly added column, titled "event," has been developed to verify the presence of an event on a certain day. A pipeline is also built to transform features such as scaler all numerical columns, one hot encoding for store identification and label encoding for all item identifications. Furthermore, date records are converted from timestamp to new columns as day of month, month of year and day of week.

5. Modeling

In order to predict the sales income for a particular item within a specific retailer on a given date, three machine learning algorithms, including linear regression, decision tree, and XGBoost, were trained. Linear regression is a frequently employed approach for regression tasks because of its widespread usage. Decision tree models are chosen in cases where all characteristics are categorical, as this model is capable of capturing the intricacies of the dataset. Finally, XGboost is chosen as the ultimate decision due to its lightweight nature, making it well-suited for handling this extensive dataset. To determine the ideal hyperparameters, a range of values for the maximum depth of the tree is specified in order to identify the most suitable value for the decision tree algorithm.

For the forecasting task, the objective is to develop forecasting models utilizing a time-series analysis technique to predict the aggregate sales income for all retailers and items during the upcoming 7-day period. The ARIMA, SARIMA and XGboost models have been chosen for training the models. In the context of XGboost, a series of experiments were conducted with varying numbers of estimators, namely 50, 100, and 200. Additionally, experiments were conducted with different learning rates, namely 0.1, 0.01, 0.05, and 0.004. Furthermore, the maximum depth of the decision trees used in the experiments was varied, with values of 2, 5, and 10 being considered. The ARIMA model is widely utilized in time series analysis, although it is limited in its ability to capture seasonal patterns. Hence, the SARIMA model is selected for the purpose of investigating the seasonal component in this experiment. In addition, the random_state parameter is consistently assigned a value of 42 for all procedures, including data partitioning and data modeling.

1. Predictive models

1.1. Approach 1

The Logistic Regression Classifier is trained in training, validating, and testing datasets because it is a standard model in binary classification. No hyperparameters were adjusted in this experiment. Once the features and target variables have been established, the model is trained using a trending dataset. Subsequently, the trained model is employed to make predictions on the target variable validating the dataset.

1.2. Approach 2

The second model is Decision Tree with, the optimal hyperparameter for the decision tree is identified as a maximum tree depth of 14 and minimum sample of leaf of 1 in order to mitigate the issue of data overfitting. No other hyperparameters were adjusted in this experiment. Once the features and target variables have been established, the model is trained using a trending dataset. Subsequently, the trained model is employed to make predictions on the target variable in validating the dataset.

1.3. Approach 3

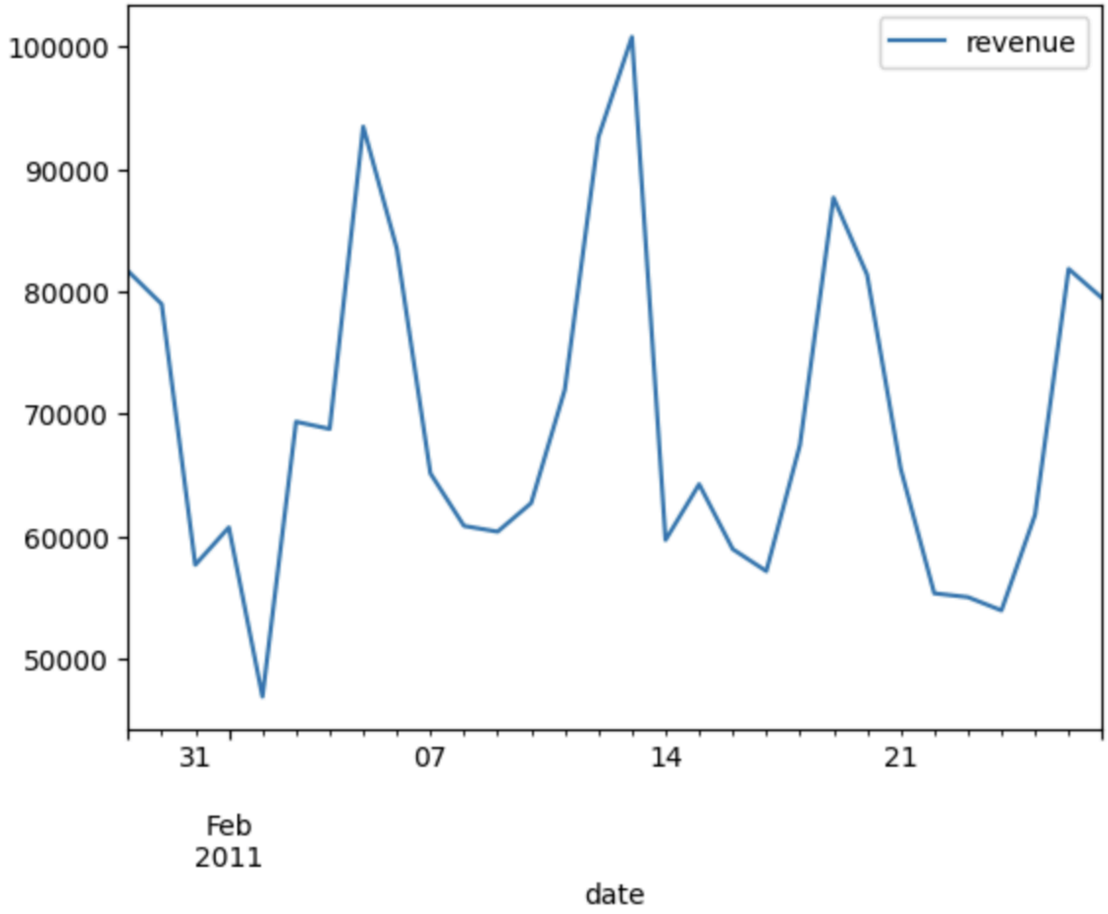
The third model is XGboost which is especially designed to be trained and validated on datasets. No hyperparameters were adjusted in this experiment. Once the features and target variables have been established, the model is trained using a trending dataset. Subsequently, the trained model is employed to make predictions on the target variable in validating the dataset.

2. Forecasting models

2.1. Approach 1

The first forecasting model is the Autoregressive Integrated Moving Average (ARIMA) model, which is specifically tailored for training and validation on datasets. The ideal parameter values for p, d, and q are determined as (1,1,1). There were no adjustments made to any other hyperparameters in this experiment. After determining the features and target variables, the model is trained using a dataset that exhibits current trends. Following this, the model that has been trained is utilized to produce predictions on the target variable inside the validation dataset.

2.2. Approach 2



The second forecasting model is the Seasonal Autoregressive Integrated Moving Average (SARIMA) model, which is specifically tailored for training and validation on datasets. The ideal parameter values for seasonal days is 7 days. There were no adjustments made to any other hyperparameters in this experiment. After determining the features and target variables, the model is trained using a dataset that exhibits current trends. Following this, the model that has been trained is utilized to produce predictions on the target variable inside the validation dataset.

2.3. Approach 3

The third forecasting model is XGboost with the best hyperparameters as 'n_estimators'=100, 'learning_rate'=0.04, and 'max_depth'=2. which is especially designed to be trained and validated on datasets. No other hyperparameters were adjusted in this experiment. Once the features and target variables have been established, the model is trained using a trending dataset.

Subsequently, the trained model is employed to make predictions on the target variable in validating the dataset.

6. Evaluation

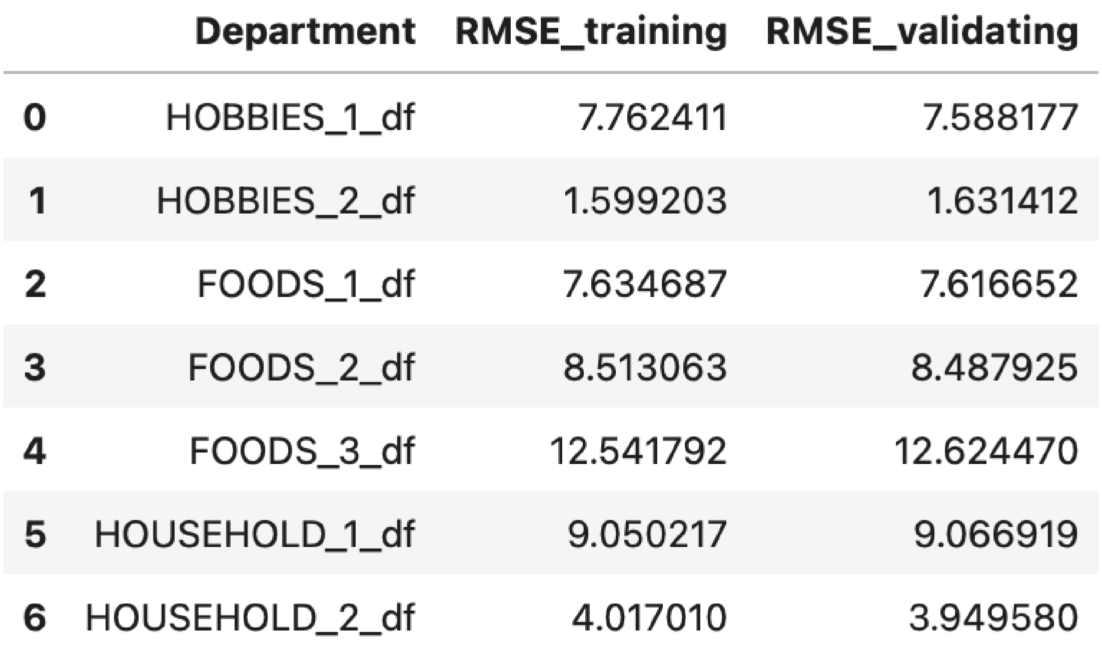
1. Evaluation Metrics

For the regression tasks, The RMSE metric score is utilized to evaluate the performance of different models. It is the mathematical square root of the mean of the squared differences between the predicted values and the actual values. Furthermore, a model's fit to a dataset improves as the root mean square error (RMSE) decreases.

2. Results and Analysis

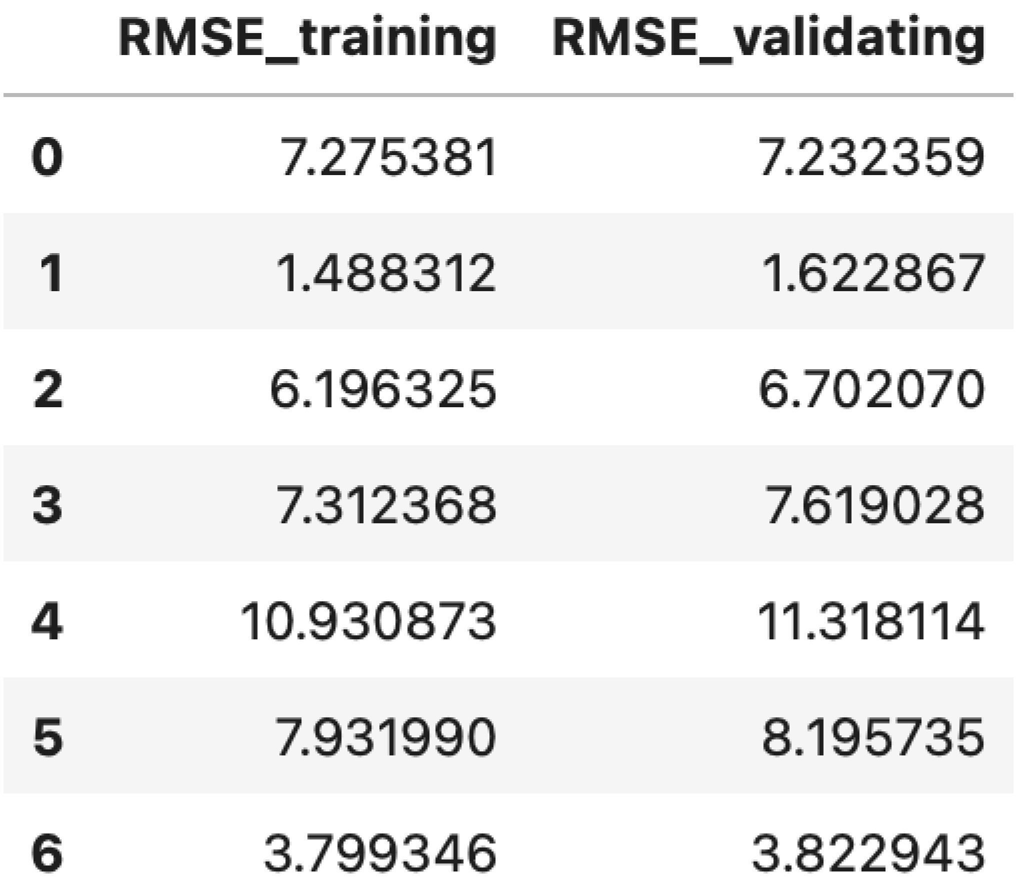
2.1. For predictive models

RMSE scores for the linear regression model for training and validating sets:



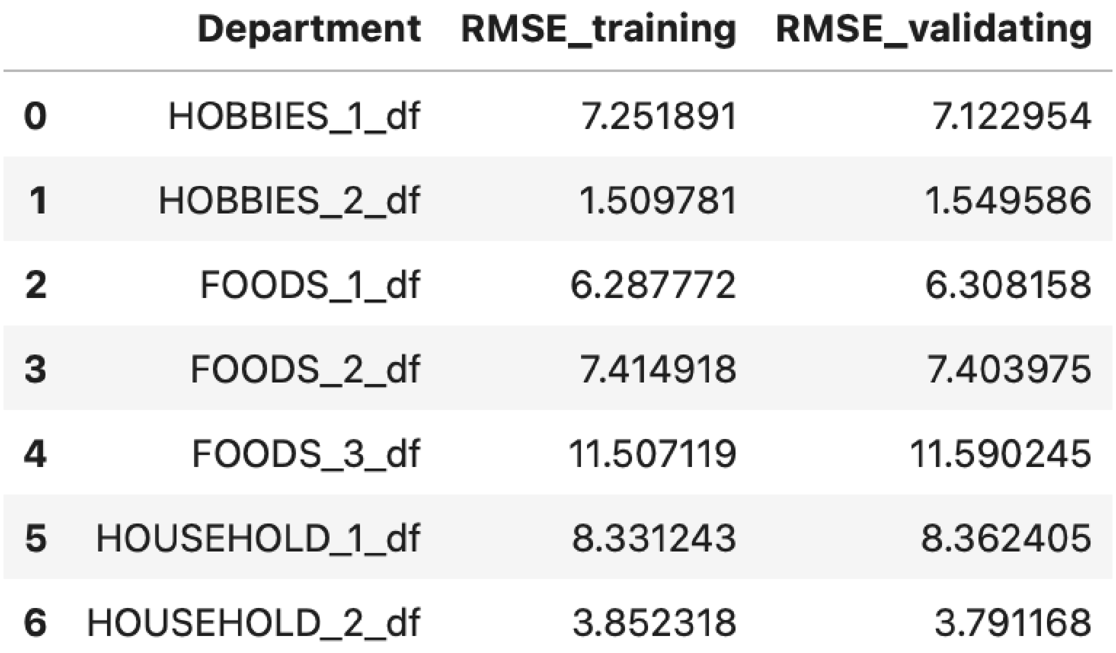
	Department	RMSE_training	RMSE_validating
0	HOBBIES_1_df	7.762411	7.588177
1	HOBBIES_2_df	1.599203	1.631412
2	FOODS_1_df	7.634687	7.616652
3	FOODS_2_df	8.513063	8.487925
4	FOODS_3_df	12.541792	12.624470
5	HOUSEHOLD_1_df	9.050217	9.066919
6	HOUSEHOLD_2_df	4.017010	3.949580

RMSE scores for the decision tree model for training and validating sets:



	RMSE_training	RMSE_validating
0	7.275381	7.232359
1	1.488312	1.622867
2	6.196325	6.702070
3	7.312368	7.619028
4	10.930873	11.318114
5	7.931990	8.195735
6	3.799346	3.822943

RMSE scores for the XGboost model for training and validating sets:



	Department	RMSE_training	RMSE_validating
0	HOBBIES_1_df	7.251891	7.122954
1	HOBBIES_2_df	1.509781	1.549586
2	FOODS_1_df	6.287772	6.308158
3	FOODS_2_df	7.414918	7.403975
4	FOODS_3_df	11.507119	11.590245
5	HOUSEHOLD_1_df	8.331243	8.362405
6	HOUSEHOLD_2_df	3.852318	3.791168

The results indicate that both the Decision Tree and XGboost models outperform other models in terms of their lower RMSE scores and reduced overfitting across all departments. In the context of model deployment, it is worth noting that XGboost requires a substantial installation size of approximately 200 MB. Consequently, when considering the prediction of total revenue sale, the Decision Tree model emerges as the optimal choice. Furthermore, the categories pertaining to Hobbies 2 and Household 2 exhibit greater compatibility with this particular model.

2.2. For forecasting models

The root mean square error (RMSE) values obtained for the validation data are 21485 for the ARIMA model, 13996 for the SARIMA model and 23164 for the XGboost model. This implies that the SARIMA model exhibits higher levels of predicting accuracy compared to the ARIMA and XGboost model.

3. Business Impact and Benefits

For the final predicting model, the model utilizes the difference of mean square values, around 7, to make predictions about sales revenue. One potential approach for identifying and predicting the sales revenue of an item across several stores is to employ a suitable methodology or model. The potential inaccuracies in the results may have a detrimental effect on the appropriate allocation of stocks across stores, resulting in increased financial losses in terms of delivery and storage.

For the forecasting model, the findings obtained from this experiment can be utilized to provide empirical evidence in favor of the practice of predicting future sales revenues across all retail establishments. However, the presence of inaccurate findings has a significant impact on the efficacy of the firm sales strategy.

4. Data Privacy and Ethical Concerns

There are some ethical concerns related to data collection and usage:

1. Privacy and Informed Consent: These datasets are collected from an American retailer so it is important to ensure that they are duly notified in the event that their data is utilized for further research endeavors.
2. Fairness: The prediction of the model is contingent upon the availability of commodities within a retail establishment. Consequently, if some items are predicted to have low sales within a particular store, their distribution will be restricted. The potential consequence of this situation is the restriction of availability for these things within a particular area, which could result in customers perceiving a sense of inequity if they primarily patronize a specific store.

7. Deployment

After the model is trained, the best models are saved in the `models/` folder with the 'joblib' formats. It will then be deployed by using the FastAPI web framework. The model is deployed by local host or online API using Heroku.

Local Host(http://localhost:8080/)

To deploy those models into local laptop, it is required to install Docker Desktop and run with following commands:

1. Build the image from the Dockerfile: docker build -t sgd-fastapi:latest .
2. Run the built image with port 8080 mapped to 80: docker run -dit --rm --name ass_2_fastapi -p 8080:80 sgd-fastapi:latest

After that, users could access "http://localhost: 8080/docs" to interface with my models.

For Heroku API: "https://secure-shelf-18349-28dcdd5de471.herokuapp.com/docs"

In the context of a predictive model, there exist three essential inputs: the date, specified in the format of year-month-day, the item, and the store. Additionally, there is an optional input denoting the occurrence of an event, which can be represented by a binary value of either 0 or 1. Users are needed to input the right information in the specified forms.

In the context of a forecasting model, it is necessary to provide a specific input in the form of a date, which includes the year, month, and day. Users are needed to input the right information in the specified forms.

8. Conclusion

In conclusion, the objective of this project is to construct two distinct models that will be implemented as application programming interfaces (APIs) in a production environment. This study presents the initial implementation of a prediction model utilizing a Machine Learning algorithm to effectively forecast the sales income of a particular item within a designated retailer on a specific day. The second forecasting model employs a time-series analysis method to predict the aggregate sales income for all stores and items in the upcoming 7-day period. Decision Tree is a machine learning algorithm commonly employed for predictive activities, whereas the SARIMA model is typically selected for forecasting jobs. Both models demonstrate the performance of satisfaction in relation to the business objectives. However, it is crucial to evaluate the models using the testing dataset.

9. References

- Include a list of references used throughout the project report.

Instructions: Include a list of references used throughout the project report, following the appropriate citation style.

Note: The CRISP-DM steps (Cross-Industry Standard Process for Data Mining) provide a framework for structuring the project report, but feel free to adapt the template to match the specific requirements and guidelines of your project or organization.