Logo

Description automatically generated

CECS 551 ARTIFICIAL INTELLIGENCE

Artificial Intelligence (AI) Approach in Retail Market Analysis and Growth

<Sprint 2>

***Group: 6 Supervisor***

* Ankit Ramrakhyani (029371899) Dr. Mahshid Fardadi
* Aparna Popade (026735460) Dr. Allen Bolourich
* Sarthak Jariwala (029341388)
* Shruthi Venkatchalam (029395663)
* Devansh Goel (029362552)

~ FALL 2022 ~

**Contents**

[1. Introduction 4](#_Toc119693805)

[2. Dataset 1 5](#_Toc119693806)

[2.1 Problem Statement 5](#_Toc119693807)

[2.1.1 Analyze the dataset\_01 5](#_Toc119693808)

[2.2 Dataset Description 5](#_Toc119693809)

[2.3 Final Dataset Visualization: 6](#_Toc119693810)

[2.3.1 Machine Learning Models 7](#_Toc119693811)

[2.4 Colab Link: 14](#_Toc119693812)

[3. Dataset 2 14](#_Toc119693813)

[3.1 Problem Statement 14](#_Toc119693814)

[3.2 Dataset Description 14](#_Toc119693815)

[3.3 Final Dataset Visualization 15](#_Toc119693816)

[Exploratory Data Analysis 15](#_Toc119693817)

[Exploratory Data Analysis to prove that Seasonality Exists 16](#_Toc119693818)

[Sales\_function by store 21](#_Toc119693819)

[Feature Engineering 31](#_Toc119693820)

[3.4 Colab Link 36](#_Toc119693821)

[4. Conclusion 36](#_Toc119693822)

**Abstract**

The project entails analyzing inventory data of two datasets of around more than 30 stores of an international retail business. It is designed to implement two-week sprints of the scrum process, mimicking a real tech company software development and machine learning work environment. The purpose of the analysis is to use the inventory data to improve sales, resulting in a more efficient operation.

Dataset 01: The task is to predict the department-wide sales for each store in a week using regression. For the second part, the task is to predict the type of store using classification.

Dataset 02: For Sprint2, the target was to visualize the daily, monthly, and total sales in STATE1 and STATE2. We also had to identify the department with the highest and lowest sales across all stores. Then we had to determine the product sales and availability over time in each store, and average price on each category. The final task was to Visualize the monthly and total sales for each store and each category. Identify the most sold products in each state in each department and then finally investigate the average sales of highly sold products for a given price of each month and on weekdays, and average sales on event types, for example, cultural, national, etc.

# **1. Introduction**

For Dataset 01, the task is to predict the department-wide sales for each store in a week using regression. For the second part, the task is to predict type of the store using classification.

For dataset 02, we came up with methods to carryout data preprocessing and Exploratory Data Analysis. For bonus points, we took efforts to perform downcasting the dataset size. We have also performed hyperparameter optimization using Long short-term memory (LSTM) layer in Tensorflow.

During the relevant sprint, we delivered the dataset to the team members and made an effort to complete the project's second milestone on time. The accompanying table shows what each team member provided as a result.

|  |  |  |
| --- | --- | --- |
| **Name** | **Dataset** | **Role and Contribution** |
| Aparna Popade | 1 | Code for Dataset 1, Worked on Report and presentation slides |
| Ankit Ramrakhyani | 1 | Creator of the Presentation slides and Dataset 1 |
| Shruthi Venkatchalam | 2 | Report editor and creator of the Presentation slides and Dataset 2 |
| Sarthak Jariwala | 2 | Code for the dataset 2 |
| Devansh Goel | 2 | Editor, creator for report and Presentation for both the datasets |

# **2. Dataset 1**

For dataset 1, we have created the Colab Notebook for data visualization.

## **2.1 Problem Statement**

### **2.1.1 Analyze the dataset\_01**

For the first part, we designed a prediction model to forecast the weekly sales across first ten stores and used the same model to make predictions for store\_11\_35. We began with Linear Regression Model. We performed the analysis with other models like Ridge and Ada Booster.

For the second part, we designed multiclass classfication to predict the store type using features mentioned. Classification algorithms used are 3 ensemble methods, RNN and CNN.

## **2.2 Dataset Description**

The Table below shows features of dataset 1 of a 35 Different Stores.

|  |  |
| --- | --- |
| **Features** | **Description** |
| Store | The Store Number |
| Type | Store Segregated into Three Types A, B And C |
| Size | Size Of the Store |
| Dept | Department ID |
| Date | MM/DD/YYYY Format |
| Is holiday | Yes/no |
| Weekly sales | Sales per week |
| Temperature | Temperature In Fahrenheit |
| Gas price | Price per gallon in $ discount |
| Promotional | Type Of Discount |
| Discount clearance | Type of discount |
| Discount damaged good | Type of discount |
| Discount competitive | Type of discount |
| Discount employee | Type of discount |
| CPI | The Consumer Price Index (CPI) |
| Unemployment present | The unemployment rate in the region where the store is |

## **2.3 Final Dataset Visualization:**

### **2.3.1 Machine Learning Models**

#### **Feature Engineering**

Table

Description automatically generated

***Figure 1****: multiple coloumns and their values against them*

***Inference****: It is visually clear that the coloumns containing “discounts” have the highest values compared to all other coloumns.*

Discount columns have maximum null values. We have checked previously from correlation plot that, discount\_promotional & discount\_competitive are highly correlated. We will drop one of these to avoid multi-collinearity. Secondly, we will add columns which gives us insights if discount was given or not. Then, we will replace null values of these discount columns with zero. Missing values in CPI, Unployement are very less, it would be best to drop them.

Chart, scatter chart

Description automatically generated

***Figure 2:*** *Effect of Temperature on Sales for Dept 92*

***Inference:*** *The graph of Weekly sales vs temperature is not linear. It seems like the temperature increases weekly sales increases but as diminishing effect. So, after certain temperature sales go down with an increase in temperature. The reason behind that could be people like to shop when it’s nice weather and prefer not to go out in extreme heat or cold. We will add a Squared term for temperature to capture the diminishing return.*

*Data is preprocessed & adding a few more features, we will remove Date & Store ID from data frame as they will not add any value for linear regression.*

#### **Regression**

##### *Linear Regression Without PCA*

Table

Description automatically generated

***Figure 3****: Linear Regression Results*

***Inference****: It can be seen that R2 for the model turns out to be 0.74. Now because the model is performing in moderate conditions, it will predict the sales. It can also be seen that “Dept” has impacted the sales.*

*From above table the R2 for model is 0.74. Model is performing well to predict the sales. Dept has high impact on sales as p-value for almost all Dept is significant (p < 0.05). As expected, Temperature square has negative coefficient. Features that do not have impact on Weekly sales ( p-value > 0.05) :  
gas\_price  
IsGiven\_Clearance  
Dept\_19  
Dept\_27  
Dept\_28  
Dept\_36  
Dept\_19  
Dept\_39  
Dept\_51  
Dept\_60  
Dept\_77  
Type\_c*

***Insights:*** *Customer's shopping behavior doesn't change with increase in gas prices. Customers do care about if promotional offers & discounts on damaged goods. Also amount of offer given improves sales. It doesn't matter if discount is given on clearance items but if the discount given is more than customers prefer to buy that. Most of the departments are different than Dept\_1. But few departments don’t change behavior of purchase for customers. Getting exact names of department can improve insights.*

*Table

Description automatically generated*

***Figure 4****: Linear Regression Results*

***Inference****: Removing those features didn’t change R2 for the model.*

##### *Linear Regression With PCA*

Background pattern

Description automatically generated ***Figure 5****: The number of components needed to explain variance*

***Inference:*** *Including 8 features can explain varince of more than 95% from prediction label.*

##### *ARIMA*

Chart, histogram

Description automatically generated

***Figure 6****: Visual data of the sales in one department of one store from Feb 2010 to Oct 2012*

***Inference:*** *It is evident from the graph above that there are 4 general spikes in the graph. They are around November, December, February, and April.*

Graphical user interface, text, application, email

Description automatically generated

***Figure 7****: Graphical presentation of the change in the values of “trend, seasonal, residual” over a duration of multiple months for almost 2 years.*

***Inference:*** *from the image above, we can see that plot for “trend” falls for a couple of months but then gradually grown over several month ahead. As for “residual” it always stays at 1 over the entire duration of months.*

Chart, histogram

Description automatically generated

***Figure 8:*** *Plotted graph of weekly sales and prediction sales over the months from Feb 2010 to May 2012*

***Inference:*** *The predication follows the trend for Weekly\_Sales. Model is performing good for future sale prediction*

Chart, line chart

Description automatically generated ***Figure 9:*** *Sales for first 10 stores From May 2012 to Oct 2012*

***Inference:*** *These are the predictions for test dataset*

Chart

Description automatically generated

***Figure 10:*** *Sales**for first 10 stores from feb 2010 to oct 2012*

***Inference:*** *These are the predictions for whole dataset*

Chart, line chart

Description automatically generated

***Figure 11:*** *Sales**for stores 11 to 35 from May 2010 to Oct 2012*

***Inference:*** *Predicting the Weekly Sales for stores using ARIMA Model trained on first 10 stores*

A picture containing chart

Description automatically generated ***Figure 12:*** *Sales**for stores 11 to 35 from May 2010 to Oct 2012*

***Inference:*** *These are the predictions for whole dataset*

##### *Regression Result*

Performance of linear regression is better because properties of the linear-regression models are well understood and can be trained quickly.

*Graphical user interface, application

Description automatically generated*

***Figure 13:*** *Regression results for all the models*

***Inference:*** *We can see that linear regression performance is better as compared to other models. Even after performing hyper-parameter tuning on ARIMA, Ridge and Ada boosting, the accurancy is low. For ARIMA, prediction is based on sum of sales across all departments of first 10 stores. The performance of ARIMA is better to predict future sales of those 10 stores but not for stores 11-35.*

#### **Ensemble Models**

##### *Random Forest Classifier*

Chart

Description automatically generated

***Figure 14:*** *Confusion Matrix for the Random Forest Classifier*

***Inference:***  *Accuracy is 100%*

##### *XGBoost*

Chart

Description automatically generated

***Figure 15:*** *Confusion Matrix for the XGBoost*

***Inference:*** *Accuracy is 100%*

##### *LightGBM*

Chart

Description automatically generated

***Figure 16:*** *Confusion Matrix for the LightGBM*

***Inference:*** *Accuracy is 100%*

##### *Recurrent neural network (RNN)*

Chart

Description automatically generated

***Figure 17:*** *Confusion Matrix for the RNN*

***Inference:*** *Accuracy is 44%*

##### *Convolutional Neural Networks (CNN)*

A picture containing application

Description automatically generated

***Figure 18:*** *Confusion Matrix for the CNN*

***Inference:*** *Accuracy is 52.45 %*

#### **Plot the relevant graphs and tabulate the performance metrics**

It is not possible to plot ROC, AUC, Precision Recall and F1 score for multiclass classification.

We tabulated the results under classification result.

##### *Classification Result*

Graphical user interface, application

Description automatically generated

Ensemble models performs better as compared to neural network model as neural network models are non-linear and have a high variance.

## **2.4 Colab Link**

<<https://colab.research.google.com/drive/1paontoradBN5KuUtzlcBBbmquBtWpXcH?usp=sharing>>

# **3.** **Dataset 2**

For dataset 2 we have deployed the tableau dashboard.

## **3.1 Problem Statement**

1. Visualize the daily, monthly, and total sales in STATE1 and STATE2. Identity the department with the highest and lowest sales across all stores. You may make a comparison of “State” with “Item Categories”.
2. Determine the product sales and availability over time in each store, and average price on each category.
3. Visualize the monthly and total sales for each store and each category. Identify the most sold products in each state in each department.
4. Investigate the average sales of highly sold products for a given price of each month and on weekdays, and average sales on event types, for example, cultural, national, etc.

## **3.2 Dataset Description**

* **calender.csv**

|  |  |
| --- | --- |
| **date** | date |
| **weekday** | categorical |
| **wday** | weekdays in numeric month – month in numeric year – year in numeric |
| **d** | each day assigned in sequential order |
| **event name 1, event name 2** | the name of events |
| **event type 1, event type 2** | the type of events |
| **snap STATE1 and STATE2** | snap is a nutritional program for low-income families. |

* **data\_test.csv and data\_train.csv**

|  |  |
| --- | --- |
| **id** | product id |
| **item id** | items |
| **dept id** | department |
| **cat id** | category |
| **store id** | store id |
| **state id** | state id |
| **d 1 - d 1941** | day 1 to day 1941 |

* **price.csv**

|  |  |
| --- | --- |
| **store id** | store id |
| **item id** | item id |
| **sell price** | selling price |

## **3.3 Final Dataset Visualization**

### **Exploratory Data Analysis**

Graphical user interface

Description automatically generated

***Figure 20:*** *Block diagram of the sum of sales across whole in USA in different states, stores, categories, departments*

***Inference:***  *It can be inferred from the diagram that foods\_3 across ca\_3, ca\_1, tx\_2, tx\_1, wi\_3, tx\_3 stores has the highest sales in the us whereas other sales in other departments have the lowest sales in.*

### **Exploratory Data Analysis to prove that Seasonality Exists**

Chart, line chart

Description automatically generated

***Figure 21:*** *Sales**in the state of California over a period of over 6 yrs.*

***Inference:***  *It can be interpreted that there is a general growth in sales from 2011 to 2016 with slight occasional drops between 2012 and 2015.*

Chart, line chart

Description automatically generated

***Figure 22:*** *multi graph image**with graphs of Trend, Seasonal and Resid.*

***Inference:*** *There is a trend of general growth in the graph of Trend. In the seasonal graph there is a repetitive fluctuation over the course of 6 years. As for Resid the lowest plot goes below -1000 and the highest plot goes above 500 multiple times over the period of 6 years.*

**Chart, line chart

Description automatically generated**

***Figure 23:*** *Sales**in the state of Wisconsin over a period of over 6 yrs*

***Inference:*** *It can be interpreted that there is a general growth in sales from 2011 to 2016 with slight occasional drops.*

**Chart, line chart

Description automatically generated**

***Figure 24:*** *multi graph image**with graphs of Trend, Seasonal and Resid.*

***Inference:*** *There is a trend of general growth in the graph of Trend. In the seasonal graph there is a repetitive fluctuation over the course of 6 years. As for Resid the lowest plot goes below -500 and the highest plot goes to 1000 in the period of 6 years*

Chart, line chart

Description automatically generated

***Figure 25:*** *Sales**in the state of Texas over a period of over 6 yrs*

***Inference:*** *It can be interpreted that there is a general growth in sales from 2011 to 2016 with slight occasional drops in 2012 and 2015.*

**Chart, line chart

Description automatically generated**

***Figure 26:*** *multi graph image**with graphs of Trend, Seasonal and Resid.*

***Inference:*** *There is a trend of general growth in the graph of Trend. In the seasonal graph there is a repetitive fluctuation over the course of 6 years. As for Resid the lowest plot goes to -500 and the highest plot goes above 500 in the period of 6 years*

**Chart, line chart

Description automatically generated**

***Figure 27:*** *Sales**in the category of Foods over a period of over 6 yrs*

***Inference:*** *It can be interpreted that there is a general growth in sales from 2011 to 2016 with slight occasional drops.*

Chart, line chart

Description automatically generated

***Figure 28:*** *multi graph image**with graphs of Trend, Seasonal and Resid.*

***Inference:*** *There is a trend of general growth in the graph of Trend. In the seasonal graph there is a repetitive fluctuation over the course of 6 years. As for Resid the lowest plot goes below -1000 and the highest plot goes above 1000 over the period of 6 years*

Chart, line chart

Description automatically generated

***Figure 29:*** *Sales**in the category of Hobbies over a period of over 6 yrs*

***Inference:*** *It can be interpreted that there is a general growth in sales from 2011 to 2016 with one major drop between 2012 and 2015.*

**Chart, line chart

Description automatically generated**

***Figure 30:*** *multi graph image**with graphs of Trend, Seasonal and Resid.*

***Inference:*** *There is a trend of general growth in the graph of Trend. In the seasonal graph there is a repetitive fluctuation over the course of 6 years. As for Resid the lowest plot goes below -200 and the highest plot goes above 200 over the period of 6 years*

**Chart, line chart

Description automatically generated**

***Figure 31:*** *Sales**in the category of Household over a period of over 6 yrs*

***Inference:*** *It can be interpreted that there is a general growth in sales from 2011 to 2016 with some drops between 2012 and 2015.*

**Chart, line chart

Description automatically generated**

***Figure 32:*** *multi graph image**with graphs of Trend, Seasonal and Resid.*

***Inference:*** *There is a trend of general growth in the graph of Trend. In the seasonal graph there is a repetitive fluctuation over the course of 6 years. As for Resid the lowest plot goes below -500 and the highest plot goes above 500 over the period of 6 years*

### **Sales\_function by store**

Chart, line chart

Description automatically generated

***Figure 33:*** *Sales**in the store CA\_1 over a period of over 6 yrs*

***Inference:*** *It can be interpreted that there is a general growth in sales from 2011 to 2016 with two major drops between 2012 and 2014.*

Chart, line chart

Description automatically generated

***Figure 34:*** *multi graph image**with graphs of Trend, Seasonal and Resid.*

***Inference:*** *There is a trend of general growth in the graph of Trend. In the seasonal graph there is a repetitive fluctuation over the course of 6 years. As for Resid the lowest plot goes below -400 and the highest plot goes above 200 over the period of 6 years*

Chart, line chart

Description automatically generated

***Figure 35:*** *Sales**in the store CA\_2 over a period of over 6 yrs*

***Inference:*** *It can be interpreted that there is a growth in sales from 2011 to 2016 with two major drops between 2012 and 2013 and 2015 and 2016.*

Chart, line chart

Description automatically generated

***Figure 36:*** *multi graph image**with graphs of Trend, Seasonal and Resid.*

***Inference:*** *There is a trend of growth in the graph of Trend in the end before plateauing between 2012 and 2015. In the seasonal graph there is a repetitive fluctuation over the course of 6 years. As for Resid the lowest plot goes below -400 and the highest plot goes above 400 over the period of 6 years*

Chart, line chart

Description automatically generated

***Figure 37:*** *Sales**in the store CA\_3 over a period of over 6 yrs*

***Inference:*** *It can be interpreted that there is a growth in sales from 2011 to 2016 with major drops between 2012 and 2016,*

Chart, line chart

Description automatically generated

***Figure 38:*** *multi graph image**with graphs of Trend, Seasonal and Resid.*

***Inference:*** *There is a trend of growth in the graph of Trend up till 2014 and then plateauing between 2014 and 2016. In the seasonal graph there is a repetitive fluctuation over the course of 6 years. As for Resid the lowest plot goes below -400 and the highest plot goes above 400 over the period of 6 years*

Chart, line chart

Description automatically generated

***Figure 39:*** *Sales**in the store CA\_4 over a period of over 6 yrs*

***Inference:*** *It can be interpreted that there is a growth in sales from 2011 to 2016 with major drops between 2012 and 2016.*

Chart, line chart

Description automatically generated

***Figure 40:*** *multi graph image**with graphs of Trend, Seasonal and Resid.*

***Inference:*** *There is a trend of growth in the graph of Trend. In the seasonal graph there is a repetitive fluctuation over the course of 6 years. As for Resid the lowest plot goes below -100 and the highest plot goes above 100 over the period of 6 years*

Chart, line chart

Description automatically generated

***Figure 41:*** *Sales**in the store TX\_1 over a period of over 6 yrs*

***Inference:*** *It can be interpreted that there is a growth in sales from 2010 to 2016 with major drops between 2013 and 2016.*

Chart, line chart

Description automatically generated

***Figure 42:*** *multi graph image**with graphs of Trend, Seasonal and Resid.*

***Inference:*** *There is a trend of growth in the graph of Trend. In the seasonal graph there is a repetitive fluctuation over the course of 6 years. As for Resid the lowest plot goes below -100 and the highest plot goes above 100 over the period of 6 years*

Chart, line chart

Description automatically generated

***Figure 43:*** *Sales**in the store TX\_1 over a period of over 6 yrs*

***Inference:*** *It can be interpreted that there is a growth in sales from 2010 to 2016 with major drops and peaks between 2013 and 2016.*

**Chart, line chart

Description automatically generated**

***Figure 44:*** *multi graph image**with graphs of Trend, Seasonal and Resid.*

***Inference:*** *There is a trend of growth in the graph of Trend till mid of 2013 and then drops to 3750 around mid of 2014 then there is a slight growth. In the seasonal graph there is a repetitive fluctuation over the course of 6 years. As for Resid the lowest plot goes below -200 and the highest plot goes above 200 over the period of 6 years*

Chart, line chart

Description automatically generated

***Figure 45:*** *Sales**in the store TX\_2 over a period of over 6 yrs*

***Inference:*** *It can be interpreted that there is a growth in sales from 2011 to 2016 with major drops and peaks between 2013 and 2016.*

Graphical user interface, chart, line chart

Description automatically generated

***Figure 46:*** *multi graph image**with graphs of Trend, Seasonal and Resid.*

***Inference:*** *There is a trend of growth in the graph of Trend. In the seasonal graph there is a repetitive fluctuation over the course of 6 years. As for Resid the lowest plot goes below -100 and the highest plot goes above 100 over the period of 6 years*

Chart, line chart

Description automatically generated

***Figure 47:*** *Sales**in the store WI\_1 over a period of over 6 yrs*

***Inference:*** *It can be interpreted that there is a growth in sales from 2011 to 2016 with major drop and peaks between 2012 and 2013.*

Chart, line chart

Description automatically generated

***Figure 48:*** *multi graph image**with graphs of Trend, Seasonal and Resid.*

***Inference:*** *There is a trend of growth in the graph of Trend. In the seasonal graph there is a repetitive fluctuation over the course of 6 years. As for Resid the lowest plot goes below -500 and the highest plot goes above 250 over the period of 6 years*

Chart, line chart

Description automatically generated

***Figure 49:*** *Sales**in the store WI\_1 over a period of over 6 yrs*

***Inference:*** *It can be interpreted that there is a growth in sales from 2011 to 2016 with major drop and peaks between 2012 and 2014.*

Graphical user interface, chart

Description automatically generated

***Figure 50:*** *multi graph image**with graphs of Trend, Seasonal and Resid.*

***Inference:*** *There is a trend of growth in the graph of Trend. In the seasonal graph there is a repetitive fluctuation over the course of 6 years. As for Resid the lowest plot goes below -500 and the highest plot goes above 250 over the period of 6 years*

Chart, line chart

Description automatically generated

***Figure 51:*** *Sales**in the store WI\_1 over a period of over 6 yrs*

***Inference:*** *It can be interpreted that there is a major fluctuation in sales from 2011 to 2016 with major drop around 2015 and peak in the mid-2012..*

**Chart, line chart

Description automatically generated**

***Figure 52:*** *multi graph image**with graphs of Trend, Seasonal and Resid.*

***Inference:*** *There is a trend of fluctuation in the graph of Trend with peak in 2012 and drop in around 2015. In the seasonal graph there is a repetitive fluctuation over the course of 6 years. As for Resid the lowest plot goes below -200 and the highest plot goes above 400 over the period of 6 years*

### **Feature Engineering**

#### ARIMA without any external features

Text

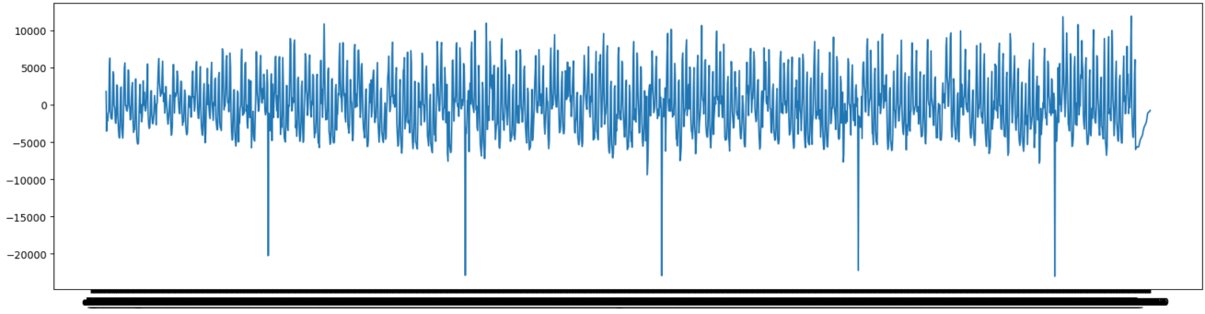
Description automatically generated with low confidence

Chart, line chart

Description automatically generated

***Figure 53:***

***Inference:*** We can observe that Test Statistic is not below the 1% of the Critical value. So, Series is not stationary.



***Figure 54:***

***Inference:*** We can observe that Test Statistic is way less than 1% of critical value. So, we can conclude that the above Series is 99% stationary.

**Chart, line chart

Description automatically generated**

***Figure 55:*** *Figure of Autocorrelation Fucntion*

***Inference:*** Here, the function cross upper confident value between 1 and 2. Hence chose 2 as p for ARIMA

**Chart, line chart

Description automatically generated**

***Figure 56*** *Image of a graph of Partial Autocorrelation Function*

***Inference:*** Partial Autocorrelation function drops to 0 when value is between 1 and 2. choose 2 as q value

Shape

Description automatically generated

***Figure 57***

***Inference:*** *It can be inferred that the majority of the graph point stays between 10000 and -10000 with 5 major drops that go to around -30000*

A picture containing chart

Description automatically generated

***Figure 58:*** *The graph of the forecast and the confidence interval among different d\_xxxx*

***Inference:*** *it can be inferred that forecast plot stays between 10000 and -10000 with some peaks reaching around 20000. The confidence interval goes through major fluctuation all over the graph and peaks above 20000 and drops to around -20000*

Chart

Description automatically generated

***Figure 59***

***Inference:*** *It can be inferred that there are multiple peaks and drops all over the graph. Original Diff peaks at around 6000and drops in at around -4000. Predicted Diff peaks at around 8000 and drops at around -6000 to -8000.*

Arrow

Description automatically generated

***Figure 60*** *Original vs Predicted: RMSE*

***Inference:*** *There are multiple peaks and drops in this plot graph with predicted peaking at around 45000 and dropping at below 20000 and actual peaking at around 40000 and dropping at around 20000*

Chart, histogram

Description automatically generated

***Figure 61*** *Arima with external features*

***Inference:*** *It can be inferred that there are multiple peaks and drops all over the graph.*

Chart

Description automatically generated

***Figure 62*** *LSTM without any external features*

***Inference:*** *It can be inferred that there are multiple peaks and drops all over the graph. Most of the plotting of Prediction stays within the range of 2 to 6 and the plotting of Time Series is all over the place ranging from 0 to 12.*

Chart, histogram

Description automatically generated

***Figure 63*** *LSTM with external features*

***Inference:*** *It can be inferred that there are multiple peaks and drops all over the graph.*

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Food: RMSE score** | **Hobbies: RMSE score** | **Household: RMSE score** |
| **ARMIA without external features** | 5309.79 | 5634.12 | 4541.43 |
| **ARIMA with external features** | 4987.34 | 4679.90 | 4652.76 |

|  |  |
| --- | --- |
|  | **RMSE Score** |
| **LSTM without external features** | 5920.76 |
| **LSTM with external features** | 4029.46 |

## **3.4 Colab Link**

*<https://colab.research.google.com/drive/1VFrbGlotNN4AmAYxldNU2q-FCKogQsXQ?usp=sharing>*

# **4. Conclusion**

After the successful completion of *Sprint 2* we understood the business scenario from the given dataset for an organization.

We observed few key findings for two different datasets.

***Data Set 01***: Linear regression performs best amongst all other regression models. All ensemble performs better than neural networks.

***Data Set 02:*** As of now we can conclude that seasonality does exist in this dataset. Moreover, the columns such as event\_name and eventt\_type can be removed since its of no significance.

The ARIMA predictions have resulted in different RMSE score for ones with external variables and ones without external variables.

The ARIMA prediction without external features is not stationary due to its high value of test statistics. We have made it stationary by moving average.

The ARIMA prediction with external features is 99% stationary due to its low value of test statistics.

he ARIMA prediction with external features is 99% stationary due to its low value of test statistics.