

CECS 551 Advanced Artificial Intelligence

Final Project

Fall 2022

**Artificial Intelligence (AI) approach in Retail Market Analysis and Growth**

Group 1

 11.04.2022

*Supervisors*Dr. Mahshid Fardadi   
Dr. Allen Bolourchi

FALL 2022

# Contents

1. Introduction. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2
   1. Problem Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2
   2. Dataset Description . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2
   3. Proposed Workflow . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3
      1. Phase 1. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3
2. Exploratory Data Analysis. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .4

## Correlation plots, SHAP. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .4

## Top 35% of the department sales for first 10 stores. . . . . . . . . . . . . . . . . . . 6

* + 1. Best department across the first ten stores. . . . . . . . . . . . . . . . . . . . .6

## weekly sales over CPI and unemployment. . . . . . . . . . . . . . . . . . . . . . . . . . . .8

## Impact of various types of discounts. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9

* + 1. Top 30% of the best performing stores. . . . . . . . . . . . . . . . . . . . . . . . 9
    2. Bottom 30% of the least performing store. . . . . . . . . . . . . . . . . . . . . .11

## Correlation between overall sales and holiday. . . . . . . . . . . . . . . . . . . . . . . . 12

1. Exploratory Data Analysis. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
2. Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .19
3. Source Code . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 20
4. Dashboard for Model. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .20

Abstract

In Retail Industry and chain of stores one of the biggest issues they face are supply chain management. The component of supply chain management (SCM) involved with determining how best to fulfill the requirements created from the Demand Plan.

Its objective is to balance supply and demand in a manner that achieves the financial and service objectives of the enterprise.

If we investigate the case of a retail chain stores one of the basic cases is to know the demand of products that are sold in the store. If the decision-making authority know what’s the demand of each product for a week or month, they would be able to plan the supply chain accordingly. If that is possible this would save a lot of money for them because they don't have to overstock or can plan their Logistics accordingly.

The dataset deals with an international retail business which has 30 stores spread across many geographical locations. The dataset is designed to mimic a real tech company software department and machine learning environment.

We have visualized the sales figures and sales pattern against many different features, to analyze significant factors that contribute to the change in the weekly sales across all stores.

# Introduction

In the team of six, we divided the task into two team, each having 3 individuals. Team 1 worked on data visualization of dataset1 and plotting the heatmap, correlation matrix etc., while team 2 worked on dataset 2 plotting and visualizing the dataset and creating the tableau dashboard.

## Problem Statement

The final project is designed to implement three-week sprints of the scrum process, mimicking a real tech company machine learning or software development team environment. The dataset describes the weekly sales of 35 stores across all the departments. Different exploratory data analysis methods have been applied to visualize the data. This data visualization will help in prediction of sales resulting in a more efficient operation.

## Dataset Description

The table below describes the features of the dataset for retail store. We will be considering only 10 stores for the initial analysis.

Features Description

——————————–

Store The number of stores date - MMDDYYY format

Temperature Temperature in Fahrenheit

Gas price Price per gallon

Discounts discounts discount clearance

CPI The Consumer Price Index (CPI)

Unemployment Unemployment rate in the region where store is present

IsHoliday Yes or No

Table 1: Store Feature data description.

## Proposed Workflow

We perform the analysis in three phases.

1. Exploratory Data analysis.
2. Create a machine leaning model for sales prediction.
3. Inference and recommendation to maximize the profit.

2. Exploratory Data Analysis (Dataset\_01)

Each phase is covered in individual chapters. The first phase tires to understand the data by visualizing it and defining relationship among them using correlation matrix and SHAP feature interactions.

* Rank the features based on their influence on weekly sales, identify where/when the sales are affected most by the feature and perform what-if analysis
* Understand if the seasonal change impacts the sales of certain stores.
* Draw conclusions and suggest a recommendation to optimize the sales

**Data**

Store: The store number. Range from 1–45.

Type: Three types of stores ‘A’, ‘B’ or ‘C’.

Size: Sets the size of a Store would be calculated by the no. of products available in the store ranging from 34,000 to 210,000.

We will perform detailed EDA and gather useful insights

## **2.1 Identifying key variables for the model using correlation plots, heatmaps, histograms, feature importance (SHAP)**

**2.1.1 Heatmaps**:

The heatmap above depicts the analysis by visually representing the data between store, dept, Weekly\_Sales, and IsHoliday. From the heat map, we can see from the pattern in cell colors across weekly shows that sales are maximum in Dept during weekly sales, and least in store during weekly sales.

 As we can observe from the heat map, weekly\_sales are closely related to size and discounts and loosely correlated to all the features, i.e., store, dept and IsHoliday.

A picture containing chart

Description automatically generated

Fig 1:

**2.1.2 Histogram:**

A histogram is basically used to represent data provided in a form of some groups. It is an accurate method for the graphical representation of numerical data distribution. In the below histogram the bar represents the frequency of store, department, size, temperature, gas\_price, discount and weekly sales in the dataset.

Diagram

Description automatically generated

Fig 2:

**2.1.3 SHAP:**

The below plot shows the importance of the following features.

As visible from the graph unemployment, CPI, discounts is a feature of high importance, also we can see that the department is also important.

Graphical user interface, table

Description automatically generated with medium confidence

Fig 3: SHAP Value Graph

## 2.2 **Top 35% of the department sales for the first 10 stores**

2.2.1. Identify the best department across the first ten stores.

The below image represents the top 35 departments for the first ten stores along with the total sales across all weeks. As seen, department 38 is the best department as it has the highest sales total across all weeks.

Graphical user interface

Description automatically generated with low confidence

Fig 4: Best department across first 10 stores

Chart, line chart

Description automatically generated

Fig 5: Weekly sales across all weeks

The above graph represents the weekly sales pattern for the first ten stores for the top 35 departments across all weeks.

Chart, line chart

Description automatically generated

Fig 6: Weekly sales across all months

The above graph represents the weekly sales pattern for the first ten stores for the top 35 departments across all months.

## 2.3 Relationship between weekly sales over CPI and unemployment for first 10 stores

The below diagram represents the correlation heatmap of the first 10 stores. We can see that weekly sales are loosely correlated to CPI and unemployment.

Graphical user interface

Description automatically generated

Figure 7: Heatmap for various features

**Chart, line chart

Description automatically generated**

Figure 8: Line plot of CPI

The above graph shows changes in CPI with respect to weekly sales ranges.

.**Chart

Description automatically generated**

Figure 9: Line plot of Unemployment

The above graph shows changes in unemployment with respect to weekly sales ranges.

## 2.4 Investigate the impact of various types of discounts

2.4.1. Type of discount helpful in increasing the sales, considering the top 30% of the best performing stores (sales per 1000 square feet).

* Discount Promotional (DP) is the type of discount which is helpful in increasing the sales of the top 30% of best performing stores per 1000 square feet. Discount promotion plays a vital role in increasing the sales of stores across multiple departments and DP is directly proportional to the weekly sales of the top performing stores.
* It is important to understand the relationship of discount promotion, discount clearance, discount damaged goods and discount competitive vs weekly sales. The correlation coefficient is a statistical measure that quantifies the relationship between two variables, seaborn in python provides an option of correlation coefficient to draw relation between variables.
* Correlation map: From the below heatmap, it can be inferred that weekly sales of top 30% stores is directly proportional to discount promotional and possess a strong relationship. Discount competitive has the second highest coefficient measure with weekly sales, followed by discount clearance and discount damaged goods.

Graphical user interface, application, Teams

Description automatically generated

Figure 10: Heatmap correlation of discount vs weekly sales

From the above heatmap we can draw strong correlation of promotional discount, clearance discount and competitive discount.

* **Bar graph**: The graph below shows correlation of different discounts to total weekly sales for top 30% stores per 1000 square feet. It can be inferred from the graph that by providing more promotional discounts on the stores, the weekly sales of the stores grow linearly.

Chart, bar chart, histogram

Description automatically generated

Figure 11: Bar graph of Sales vs Different discounts for top 30% stores

2.4.2. Considering the bottom 30% of the least performing store (sales per 1000 square  
 feet) does the observed behavior hold true for all stores

* The observed behavior is true for all the stores when considering bottom 30% of the stores that Discount promotional(DP) is strongly co related when the DP increases the weekly sales also increases.
* Correlation map:From the below heatmap, it can be inferred that weekly sales of bottom 30% stores is directly proportional to discount promotional and possess a strong relationship. Discount competitive has the second highest coefficient measure with weekly sales, followed by discount clearance and discount damaged goods.

Graphical user interface, Teams

Description automatically generated

Figure 12: Estimated Loss of Sales in Volume

### **Bar graph:** The graph below shows correlation of different discounts to total weekly sales for bottom 30% stores per 1000 square feet. It can be inferred from the graph that by providing more promotional discounts on the stores, the weekly sales of the stores grow linearly.

Chart, bar chart, histogram

Description automatically generated

Figure 13: Sales vs Different discounts for top 30%

## 2.5 Determining if there is any correlation between overall sales and holiday

After merging data for the first 35 stores and 10 departments in train.csv with stores\_features files to get the following data frame.

Table

Description automatically generated

Figure 14:

Weekly sales range from [ -4988.94: 406988.63]. It was divided into 10 equal groups to get a weekly sales range. From the following graph we can see that for different departments the trends of weekly sales range in comparison to the temperature. For example, in Department 06 weekly sales increases with the increase in temperature whereas for Department 10 for weekly sales increases with decrease in temperature.

Chart, line chart

Description automatically generated  
Fig 15: Weekly\_sales\_range vs Temperature Graph

In the graph below for different departments the trends of weekly sales range in comparison to the gas prices is shown. For example, in Department 01 weekly sales increases with the increase in gas prices whereas for Department 10 for weekly sales increases with decrease in gas prices.

Chart, line chart, surface chart

Description automatically generated  
Fig 16: Weekly\_sales\_range vs gas price graph

For Holidays we can see that overall sales vary more Department range (30 - 50).

Chart, line chart, histogram

Description automatically generated

Fig 17: Dept vs weekly\_sales\_range graph

**2.6** **Data preprocessing:**

Data preprocessing is a step in the data mining and data analysis process that takes raw data and transforms it into a format that can be understood and analyzed by computers and machine learning.

**Data cleaning:**

3 data frames are created because we are handling missing values with different techniques and creating a model for the same and comparing the output of the models.

**Techniques**:

1. Replace missing and NaN values with median value.
2. Filling discount NaN values with zero in the entire data frame.
3. Splitting date column into day, month and year
4. Creating a Total additional discount by adding all the discount columns

Since the date column Since the date column is in string format, typecasting it to date format and divide them to individual columns for day, month and year.

One hot Encoding is performed on columns such as **type**, **department** and **store** as it is necessary to convert the categorical data variables to be provided to machine and deep learning algorithms which in turn improve predictions as well as classification accuracy of a model.

How weekly sales are related to other features:

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

Fig 18:

Key observations from the above graphs:

* When the gas\_price is in the range of 2.5 - 3.75 there is more sales, for >3.75 the sales is comparatively lower.
* Type A and Type B have more weekly sales than type C stores.
* There are more sales when there is a holiday than non-holiday days.
* When the temperature is 30 units of temperature there is less sales than >30 units of temperature.
* Highest sale is for the store whose size is 125000 sq ft.

**Data Normalization:**

Data Scaling is a data preprocessing step for numerical features. Many machine learning algorithms like KNN algorithm, linear and logistic regression, etc. require data scaling to produce good results.

**MinMax Scaler:**

In this method the features are made equal to zero and the maximum of features equal to one. MinMax Scaler shrinks the data within the given range, usually of 0 to 1. It transforms data by scaling features to a given range.

Below is the image after normalizing the data using MinMax Scaler.

Graphical user interface

Description automatically generated with medium confidence

Fig 19:

**2.7 Feature Selection:**Feature Selection is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data. It is the process of automatically choosing relevant features for your machine learning model based on the type of problem you are trying to solve.

**Method 1: OLS**The ordinary least squares (OLS) method is a linear regression technique that is used to estimate the unknown parameters in a model. The method relies on minimizing the sum of squared residuals between the actual and predicted values. The OLS method can be used to find the best-fit line for data by minimizing the sum of squared errors or residuals between the actual and predicted values.

Table

Description automatically generated

Fig 20:

**Results after OLS:**

After 8 iterations of OLS fitting the mode, the linear regression model achieved an accuracy of 75.28.

**Method 2: PCA**

Principal Component Regression (PCR) is a regression technique that serves the same goal as standard linear regression — model the relationship between a target variable and the predictor variables.

The idea is that the smaller number of principal components represents most of the variability in the data and (presumptively) the relationship with the target variable. Therefore, instead of using all the original features for regression, we only utilize a subset of the principal components.

Cross-validation and visual analysis are generally used to calculate the number of main components (k). K is essentially a hyperparameter that needs to be tuned. We evaluate the RMSE scores that occur as we iterate over an increasing number of principal components to include in regression modeling. The training set performance of the PCR increases with more principal components, as expected, when we look at the plot of training set cross-validation RMSE vs. the number of principal components employed. The baseline standard linear regression model with all of the original features' RMSE benchmark is represented by the green line.

After determining the best number of principal components to use (i.e., M=17), we proceed to run PCR on our test dataset. Accuracy 78.02 % after performing PCR was not satisfactory.

**Method 3: RandomForest Regressor**

Feature selection using random forest regressor technique improves performance, reduces overfitting and increases interpretability. It is an embedded technique that is a combination of filter and wrapper methods. A random extraction of the observations from the dataset and a random extraction of the features are used to build each of the 4–12 hundred decision trees that make up a random forest.

Every tree also consists of a series of yes-or-no questions depending on a single or several attributes. The dataset is split into two buckets at each node (i.e., each question) by the three, with each bucket containing observations that are more similar to one another and distinct from those in the other bucket. The significance of each feature is therefore determined by how "pure" each of the buckets is. For regression the measure of impurity is variance.

**Table

Description automatically generated**

**Fig 21:**

**Linear Regression:**

A variable's value can be predicted using linear regression analysis based on the value of another variable. The dependent variable is the one you want to be able to forecast. The independent variable is the one you're using to make a prediction about the value of the other variable.

Following scatter plot shows actual vs predicted output after Linear Regression:

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Fig 22: Fig 23:

We get best accuracy with linear regression after feature selection with random forest regressor:

Approach 1: 80 - 20 Test Train split on Stores 1 – 35

|  |  |
| --- | --- |
| **Approach** | **Accuracy** |
| Linear Regression | 91.098 |
| Random Forest Regression | 96.759 |
| Ridge Regression | 91.091 |
| XGBoost | 96.945 |

Approach 2: Train on Store 1- 10 and Test on Store 11- 35

|  |  |
| --- | --- |
| **Approach** | **Accuracy** |
| Linear Regression | 89.379 |
| Random Forest Regression | 95.701 |
| Ridge Regression | 89.368 |
| XGBoost | 97.146 |

**ARIMA:**

ARIMA is an acronym that stands for AutoRegressive Integrated Moving Average. It is a class of models that captures a suite of different standard temporal structures in time series data.

Here we will be performing the below steps to calculate time series.

1. Visualize the weekly series

Chart, histogram

Description automatically generated

Fig 24:

2. Make sure the variance of weekly sales is non stationary

3. Test stationary of data using ADF Test

Chart, histogram

Description automatically generated

Fig 25:

Weak evidence against null hypothesis, time series has a unit root, indicating it is non-stationary.

Chart

Description automatically generated

Fig 26:

Strong evidence against the null hypothesis, reject the null hypothesis. Data has no unit root and is stationary.

4.  Convert data to stationary using first order differencing and calculate correlation function.

Chart, line chart

Description automatically generated

Fig 27:

1. Predict weekly sales over a time period. Here we are predicting weekly sales from May 2010 to Sept 2010 the blue mark shows forecast vs actual values in blue.

Chart, histogram

Description automatically generated

Fig 28:

**2.8 Ensemble modeling**

Ensemble Modeling is the process of running two or more related but different analytical models and then synthesizing the results into a single score or spread in order to improve the accuracy of predictive analytics and data mining applications.

## **Classification and Prediction**

The problem required us to merge the following dataframes

Train

Stores

Store\_feature

train\_data\_expanded

Table

Description automatically generated

Fig 29:

Adjustment made to the dataset

1. train\_data\_expanded = train\_data\_expanded.fillna(0)
2. train\_data\_expanded['IsHoliday'] = train\_data\_expanded['IsHoliday'].astype('str').map({'True':0,'False':1})
3. train\_data\_expanded['Type'] = train\_data\_expanded['Type'].astype('str').map({'A':0,'B':1,'C':2})

1.Filled Nan values with 0

2.for Conditioned input for below prediction models we the Encoded Categorical data to  numerical features

RandomForestClassifier

XGBClassifier

LogisticRegression

## **Ensemble Modeling:**

Ensemble modeling is the process of running two or more related but different analytical models and then synthesizing the results into a single score or spread in order to improve the accuracy of predictive analytics and data mining applications.

1. **Averaging method:**

 It is mainly used for regression problems. The method consists of building multiple models independently and returning the average of the prediction of all the models

mean\_squared\_error = 0.02

Graphical user interface, text

Description automatically generated

Fig 30:

1. **Max voting:**

It is mainly used for classification problems. The method consists of building multiple models independently and getting their individual output called ‘vote’. The class with maximum votes is returned as output

For out model calculated log\_loss is 9.99

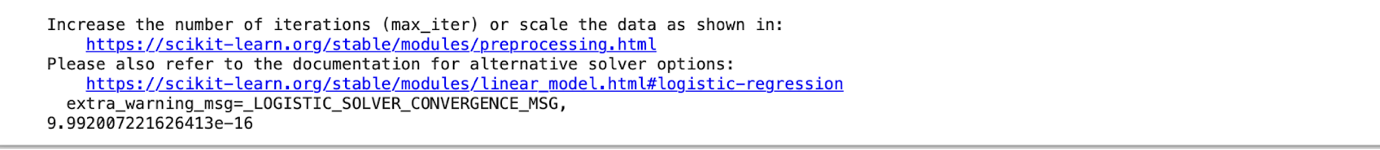


Fig 31:

1. **Bagging:**

It is also known as a bootstrapping method. Base models are run on bags to get a fair distribution of the whole dataset. A bag is a subset of the dataset along with a replacement to make the size of the bag the same as the whole dataset. The final output is formed after combining the output of all base models.

## **Recurrent Neural Network (RNN):**

A Neural Network consists of different layers connected to each other, it learns from huge volumes of data and uses complex algorithms to train a neural net.

RNN will do the following:

RNN converts the independent activations into dependent activations by providing the same weights and biases to all the layers, thus reducing the complexity of increasing parameters and memorizing each previous output by giving each output as input to the next hidden layer.

Hence these three layers can be joined together such that the weights and bias of all the hidden layers is the same, into a single recurrent layer.

They recognize patterns and clusters and use them to give us powerful insights and terrific applications of different kinds of levels. RNN or Recurrent Neural Networks, as the name suggests, is a repeating neural network. They are the kind whose output from the previous step is fed as input to the current step.

Here we have used RNN to predict the Store type:

Chart, line chart

Description automatically generated

Fig 32:

Chart, histogram

Description automatically generated

Fig 33:

While RNNs are suitable for handling temporal or sequential data, CNNs are suitable for handling spatial data (images). Though both models work a bit similarly by introducing sparsity and reusing the same neurons and weights over time (in case of RNN) or over different parts of the image (in case of CNN).

## **Convolutional Recurrent Network (CNN):**

Convolutional Neural Networks (CNNs) are designed to map image data (or 2D multidimensional data) to an output variable (1 dimensional data). They have proven so effective that they are the ready to use method for any type of prediction problem involving image data as an input.

The different layers of a CNN:

There are four types of layers for a convolutional neural network: the convolutional layer, the   pooling layer, the ReLU correction layer and the fully connected layer.

**The convolutional layer:**

The convolutional layer is the key component of convolutional neural networks and is always at least their first layer. Its purpose is to detect the presence of a set of features in the images received as input.

**The Pooling Layer:**

The pooling layer reduces the number of parameters and calculations in the network. This improves the efficiency of the network and avoids over-learning.

Thus, it reduces the number of parameters to learn and the amount of computation performed in the network.

**The ReLU Correction Layer:**

ReLU (Rectified Linear Units) refers to the real non-linear function defined by ReLU(x)=max(0, x). Visually, it looks like the following:

A picture containing diagram

Description automatically generated

Fig 34:

The ReLU correction layer replaces all negative values received as inputs by zeros. It acts as an activation function.

**The fully-connected Layer:**

The fully-connected layer is always the last layer of a neural network, convolutional or not — so it is not characteristic of a CNN.

This type of layer receives an input vector and produces a new output vector. To do this, it applies a linear combination and then possibly an activation function to the input values received.

**Text

Description automatically generated**

Fig 35:

## **Performance matrix**

Performance metrics are measure of calculation of the quality of the prediction models

Following the subcategories in performance matrix

Confusion Matrix

Precision

Recall

Roc Curve

Auc curve

F Score

Sensitivity

Now let's understand each of the categories one -by- one and let me walk us through how our code performed as per the mentioned performance matrices

**What is a confusion matrix?**

It is a matrix of size 2×2 for binary classification with actual values on one axis and predicted on another.

Like in our case we have actual values on y-axis and predicted on x-axis.

**True Positive (TP) --** model correctly predicts the positive class (prediction and actual both are positive).

For the given problem set  **13870** were made true and they were actually true also

**True Negative (TN) --** model correctly predicts the negative class (prediction and actual both are negative). In the above plot, of all stores predicted Type B  11836 were actually  type B, from test data.

**False Positive (FP**) — model gives the wrong prediction of the negative class . In the above plot,

Of all the stores predicted to be type A **, 970**  stores are  actually type B

**False Negative (FN)** — model wrongly predicts the positive class (predicted-negative, actual-positive). In the above plot , of all the stores predicted to be type B all of them were actually type b.  There is '0' false negative here.

**Precision** -

Out of all the values predicted positive how many are actually positive

So, we have 13870 Stores predicted positive , and all of them actually turned out to be A

Thus, precision is **1.**

**Recall -**

Out of the total positive, what percentage are predicted positive. It is the same as TPR (true positive rate).

**Accuracy** -

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations.

**F1 Score –**

It is the harmonic mean of precision and recall. It takes both false positive and false into account. Therefore, it performs well on an imbalanced dataset.

Below is the confusion matrix as asked for given dataset.

Chart, waterfall chart

Description automatically generated

Fig 36: Confusion Matrix

**ROC curve**

An **ROC curve** (**receiver operating characteristic curve**) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

* True Positive Rate
* False Positive Rate

**AUC: Area Under the ROC Curve**

**AUC** stands for "Area under the ROC Curve." That is, AUC measures the entire two-dimensional area underneath the entire ROC curve from (0,0) to (1,1).

AUC provides an aggregate measure of performance across all possible classification thresholds. One way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example.

AUC  = 0.96

Graphical user interface, text

Description automatically generated

Fig 37:

# **Dataset\_02**

# 3.1 Exploratory Data Analysis (Dataset\_02) Datasets contain 2 files:

1) Calendar.csv

2) Sell\_prices.csv

We have data from California, Wisconsin, Texas. And we have items which are sold in each state containing numerous stores in them.

Yearly sales of 2011 shows that there is an increase in sales in month of April, July, December 2011.

Chart, line chart

Description automatically generated

Fig 38:

**Which Item in each group has highest and lowest sales?**

Ans: Food\_3 has highest sales and Food\_1 has lowest sales. Hobbies\_2 has highest sales and Hobbies\_1 has lowest sales. Household\_2 has highest sales and Household\_1 has lowest sales. Out of all Household\_2 is highest selling of 7,115,264.

Chart, bar chart

Description automatically generated with medium confidence

Fig 39:

**Which state has maximum sales?**

Ans: California has maximum sales in all the years. And Wisconsin has least sales for all the years.

Graphical user interface, chart, line chart

Description automatically generated

Fig 40:

**What do we interpret from Pie-Chart?**

Ans: Out of 3 item categories, household account for 43.04%, Food account for 34.28%, and Hobbies account for 22.69%.

A picture containing background pattern

Description automatically generated

Fig 41:

**What are our sales averages on Event days?**

Ans: Average sales is $4 Million. But most of our sales are being done on normal routine days. In terms of event most sales are done on National Events.

Chart

Description automatically generated with low confidence

Fig 42:

This chart shows that daily sales based on weekdays are very similar. Average sales in all the states are equal.

Chart, bar chart

Description automatically generated

Fig 43:

**What are our findings from this graph?**

Ans: We get to know that sales gradually increase from year 2011 to 2016 gradually. And most of the sales amount is made from household and food categories.

Chart, bar chart

Description automatically generated

Fig 44:

**Which is the most sold item in all the stores?**

Ans: Household\_2\_446 is the most sold item across all the stores.

Chart, box and whisker chart

Description automatically generated

Fig 45:

**How does our sales flare over the period of year?**

Ans: We interpret from the graph that there is sudden increase in graph from February to May, then the sales decrease for every store in all the states.

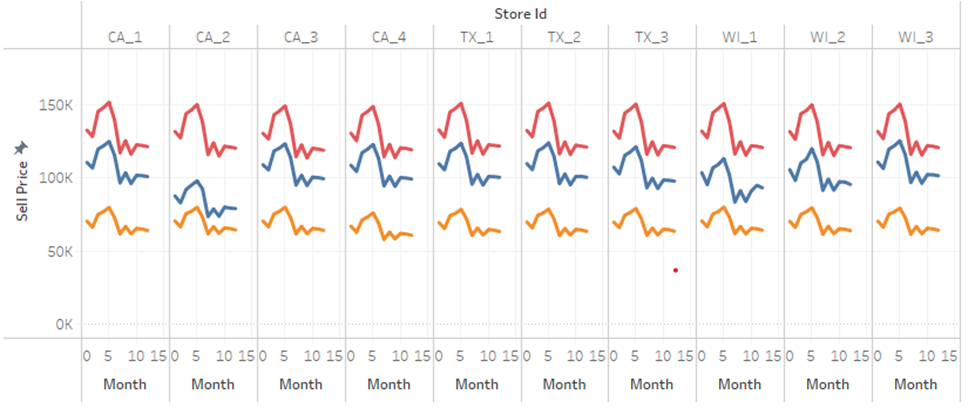


Fig 46:

**What does this graph represent?**

Ans: We see that most sales are done on normal days. One of the reasons can be that normal days are more than event days.

Chart

Description automatically generated  
Fig 47:

## **3.2 Modeling:**

* Large Files
  + sales\_train\_evaluation is 116 MB and has 30490 rows, 1947 columns
  + sell\_prices are 193 MB and has 1048575 rows, 4 columns
  + calendar is 101 KB and has 1969 rows, 14 columns

Downcasting: Shrink Pandas DataFrames with precision safe schema inference. Pandas-downcast finds the minimum viable type for each column, ensuring that resulting values are within tolerance of original values.

Chart, bar chart

Description automatically generated

Fig 48: File sizes after down casting

**EDA:**

**A picture containing chart

Description automatically generated**

Fig 49:

From this graph, the variation in mean sales can be seen vs the sales in USA

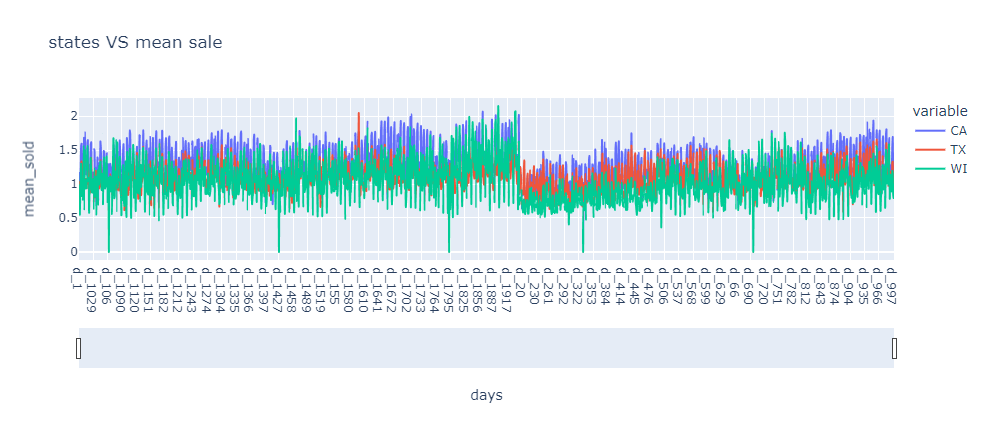


Fig 50 :

The variation in sales state wise vs mean sales can be observed from this graph

Chart, bar chart

Description automatically generated

Fig 51:

The variation in total sales according to the day can be observed. As can be seen, sales are higher on Friday, Saturday and Sunday. A new feature called is\_high\_sale\_day is introduced corresponding to this.

Chart, bar chart

Description automatically generated

Fig 52:

The variation in total sales according to the month can be observed. As can be seen, sales are higher in the months of March, April, and May. A new feature called is\_high\_sale\_months is introduced corresponding to this. This corresponds to seasonality in the data.

Chart, bar chart

Description automatically generated

Fig 53:

As observed in this graph, there is no significant change in sales based on event\_type1. Therefore event\_type1 can be dropped.

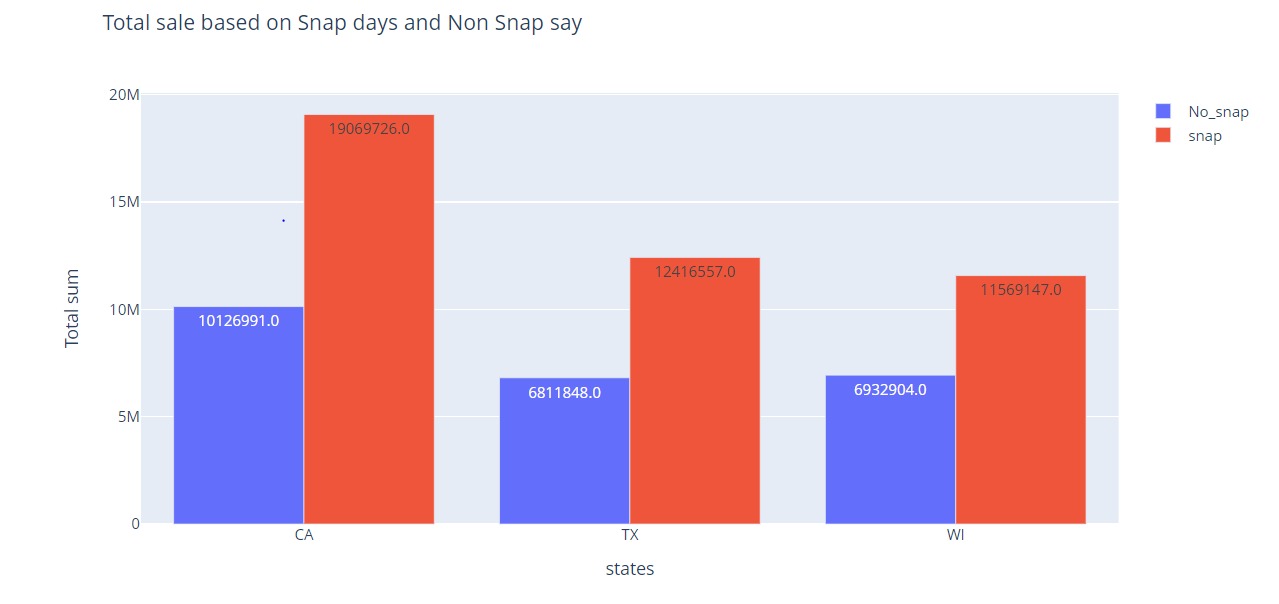


Fig 54:

The variation in sales in the three states on Snap and non-Snap days can be seen here. Clearly, Snap days have higher sales than non-snap days.

Chart, bar chart

Description automatically generated

Fig 55:

The average sales for each category for each year are plotted. As observed, the sales in all three categories are increasing year-on-year.

**Feature Engineering:**

Weather data was obtained for the 4 Zip Codes in California, 3 Zip Codes in Texas, and 3 Zip Codes in Wisconsin. This data includes Precipitation, Snow and Temperature. The days on which Precipitation or Snow is high, sales are down compared to other days.  This weather data was added as one of the features.

Median Income was also fetched from the internet for all 4 California Zip Codes, 3 Texas Zip Codes, and 3 Wisconsin Zip Codes. The median income was also added as one of the features.

**Models:**

**ARIMA**

ARIMA is an acronym for Auto Regressive Integrated Moving Average. It is a model used for analyzing and forecasting time series data.

**Chart, line chart

Description automatically generated**

Fig 56: Prediction for HOBBIES

Chart, line chart

Description automatically generated

Fig 57: Prediction for FOOD

Chart, line chart

Description automatically generated

Fig 58: Prediction for HOUSEHOLD

Chart, line chart

Description automatically generated

Fig 59: Prediction for HOBBIES for 10 days relative to the average sale

Chart, line chart

Description automatically generated

Fig 60: Prediction for FOODS for 10 days relative to the average sale

Chart, line chart

Description automatically generated

Fig 61: Prediction for HOUSEHOLD for 10 days relative to the average sale

|  |  |  |
| --- | --- | --- |
| Category | RMSE without External Features | RMSE with External Features |
| HOBBIES | 492.16 | 550.25 |
| FOOD | 3624.14 | 2877.21 |
| HOUSEHOLD | 1122.09 | 1015.2 |

**LSTM**

Long Short Term Memory networks(LSTMs) are a special kind of Recurrent Neural Networks, capable of learning long-term dependencies.

**Chart, line chart

Description automatically generated**

Fig 62: Prediction without external features

Chart, line chart

Description automatically generated

Fig 63: Prediction with external features

|  |  |
| --- | --- |
| RMSE without external features | RMSE with external features |
| 2.025 | 1.687 |

**Model Comparison**:

|  |  |  |
| --- | --- | --- |
| Model | RMSE without external features | RMSE with external features |
| ARIMA | 1746.13 | 1480.90 |
| LSTM | 2.025 | 1.687 |

4 Conclusion

* Following inferences and conclusions can be drawn from the analysis of the data:
* Type 'A' stores are more popular than 'B' and 'C' types.
* Type 'A' stores outclass the 'B' and 'C' types in terms of size and the average weekly sales.
* Weekly Sales are affected by the week of year. Holiday weeks witnessed more sales than the non-holiday weeks.
* Size of the store is a major contributing factor in the weekly sales.
* Sales are also dependent on the department of the store as different departments showed different levels of weekly sales.
* Among the trained models for predicting the future sales, XG Boost with tuned hyperparameters performs the best.

# Source Code (Dataset\_01)

<https://colab.research.google.com/drive/1F1kAHjIxbO-LwYp8MZizqiOY_HHxH-UR#scrollTo=1CwY5eXgPp2c>

# Dashboard for Model (Dataset\_02)

<https://public.tableau.com/app/profile/varun.shah1550/viz/Group_1_Shah_Sharma_Patil_Patel_Lokhande_Katkamwar/Dashboard1?publish=yes>