**Project documentation:**

Sprint1:

Firstly, required libraries and data are imported. Given files in the dataset which consists of information about 35 stores have been merged. Along with merging the given files, all the anomalies were taken care of. A pair plot (where each variable in the data will be shared in the y axis across a single row and in the x axis across a single column) was created.

1. using correlation heatmap, we visualized the strength of relationship between variables, and we created various subplots and finally identified the key variables.
2. For the first 10 stores we grouped the data by using attributes (store, department)

and picked top 35 departments (where weekly sales are highest) out of the 100 departments (Assumption).

We added additional columns (month year and week year) into the data frame just to make the data clear, this has to be done to avoid records having the same week and month number from being merged into one record.

B.1->Identified the best department in each store by sorting the values according to

Weekly sales.

1. Here we investigated the relationship between CPI and weekly sales and unemployment over a line plot(which is a way to display data along a number line)
2. With the help of heatmap we analyzed the impact of various types of discounts namely discount promotional, discount clearance, discount damaged good, discount competitive over weekly sales.
3. Using heat map the products which were highly impacted by the external factors (temperature, gas price, holiday) were identified. The correlation between overall sales and is holiday has been identified. Wherever gas prices are less we are observing better weekly sales across all the stores. Is holiday and weekly sales are inversely proportional across the stores.

Sprint2:

DATASET 01:

First things first, we imported the necessary libraries and loaded the csv file from the google drive. Then we appended all the files together where the new data frame has 12 columns.

1. we took only the first ten stores for designing the prediction model.
2. The scikit-learn library along with the Linear Regression function in Python has been used to create the linear regression model. The training dataset has been further divided into the training and validation datasets; the validation dataset is primarily created to test models based on different parameters to see what parameters result in higher accuracy. After implementing the linear regression model and generating predictions on both datasets (training and validation), the WMAE (weighted mean absolute error) was calculated (based on the previously defined function in the data cleaning and pre-processing section). Considering that the WMAE values for the validation data is extremely high, linear regression cannot be considered as a very efficient model for this analysis based on this model.
3. Regression, also known as regularized linear regression, is another regression analysis model that handles data that suffer from multicollinearity and is primarily suitable when there are several predictor variables in the data. Lasso regression is an extension of linear regression in which the loss function is modified to minimize the complexity of the model by limiting the sum of the absolute values of the model coefficients. One of the key differences between linear and several other regression models is that while linear regression does not exactly tune the parameters, the other models allow for tuning of the hyperparameter. Just like the linear regression model, the lasso regression model does not perform well on the validation data, giving a relatively higher WMAE value. Hence, the model cannot exactly be considered as an efficient model for the study.
4. Gradient boosting is a sequential technique, based on ensemble learning, which combines the decisions from multiple machine learning models to improve the overall performance of the model. In boosting, the model outcomes for any instance in the sequence are weighed based on the results of the last instance. The correct predictions are assigned a lower weight, while the incorrect predictions are assigned a higher weight. The model will then focus on higher weight points as it might go wrong with the lower weight points. After many iterations, a combined result is generated, using the accuracy of all the models. Using GBM provides multiple advantages: it usually provides better predictive accuracy compared to other models, it provides various hyperparameter tuning options, is time-efficient for larger datasets, and handles missing data. Using the XGBRegressor object from the xgboost library under ‘scikit-learn’, the training and validation datasets were used to create a basic gradient boosting model and the WMAE was calculated.
5. Arima, predicting sales using Arima is done. ARIMA, short for ‘Auto Regressive Integrated Moving Average’ is actually a class of models that ‘explains’ a given time series based on its own past values, that is, its own lags and the lagged forecast errors, so that equation can be used to forecast future values.

Model Performance: out of the all models we tried , ARIMA model with our selected features proved to be the finest for predicting future sales. It had the best RSME,MAE,MAPE values when compared.

|  |  |
| --- | --- |
| Model | MAPE values |
| Linear Regression | 0.15 |
| ARIMA | 0.13 |
| XGBoost | 0.09 |

2)As was given only first 10 stores were considered.

a) we used classification algorithms and performed hyper parameter tuning for the deep learning models.

1. Ensemble models (3 statistical methods
2. Recurrent neural network
3. Convolution neural network

RNN works on the principle of saving the output of a particular layer and feeding this back to the input in order to predict the output of the layer.

**Deep learning part:** In the first step we encoded the categorical variable is\_a\_holiday and then generated matrix of features (x) and (y).in the next step predictor and target variables were defined and then feature scaling is done. ANN is initialised which was followed by the addition of first layer, second layer and the output layer. The next is compiling and fitting ANN an then prediction is done for single observation.

DATASET 02:

1. Data pre-processing
2. Exploratory data analysis

In the first step merging of files is done ([[using wm\_yr\_wk]]\nand then that/ price.scsv with data\_train.csvs [[using item\_id]])) and EDA is performed on the merged dataframe.

(Graphs are attached in the presentation).

Feature engineering observations: weekly sales are in high in areas where the income median of the individuals is high. People tend to buy more in those areas where we have moderate to mild temperatures.

Sprint3:

**1. Model Deployment using Streamlit**

The ARIMA model developed in the last sprint, was deployed using streamlit in the local environment.

The model is trained with data until december of 2016 and the sales are predicted for the next N days that are passed as inputs to the model.

Chart, line chart

Description automatically generated

Chart, scatter chart

Description automatically generated

The actual (black dots) and predicted (blue line) values over time. High level trend of predicted values.

**2A. PRODUCT SEGMENTATION** based on demand variability (ABC analysis)

**What exactly is ABC analysis?**

ABC Analysis is a commonly used practice for inventory grouping that divides inventory into categories based on two factors, the cost per unit and the quantity held in stock. ABC Analysis divides the inventory into three major groups, which allows different inventory management techniques to be applied to different segments of the inventory in order to increase the revenue and decrease the cost.

Each category A, B, and C consists of part of the total quantity of the items and represents a part of the total value of the items in the warehouse.

**Category A** items generally represent approximately 15%-20% of an overall inventory by item, but represent 80% of value of an inventory. Items in this category are goods which annual consumption value is the highest.

**Category B** items represent 30%-35% of inventory items by item type, and about 15% of the value. The items in this category are goods with medium consumption value.

**Category C** items represent 50% of actual items but only 5% of the inventory value. The items in this category are goods with the lowest consumption value.

**FOLLOWING STEPS ARE INVOVLED IN THE ANALYSIS**

Step 1: Calculate the annual spend for all the parts by multiplying the unit cost with the annual unit demand  
Step 2: Sort inventory in decreasing order of annual spend  
Step 3: Calculate the cumulative annual spend and the percent spend  
Step 4: Divide the inventory into classes  
Step 5: Analyse the classes and make appropriate decisions

**ABC analysis in Action:**

* *First things first , all the required libraries were imported.*

**Text

Description automatically generated with medium confidence**

* *Reading the data from the csv file (sell\_prices.csv file)*
* *Defining three classes based on the quantity percentages* **( *A-60% , B-25% ,  C-15% )***

**Graphical user interface, text, application

Description automatically generated**

* *Taking a subset of data , we need to use the price and quantity of each item.**Here we considered all the four columns of the dataset inorder to create a subset and named it as* ***data\_sub***
* *Create the column of total sales per item and order it by the cumulative selling price*
* *Next step , running cum sell\_price of the cumulative selling price  was created*
* *Column of tot\_sum (total sum was created.*
* *In the next step  a new column 🡪 running percentage was created by performing a simple division operation (run\_cumsellprice/tot\_ sum)*
* *A column class was created where ABC segmentation was applied on running percentage.*

**Graphical user interface, application

Description automatically generated with medium confidence**

* *Total cost per class*

Text

Description automatically generated

* *Percentage of total cost per class*

Text

Description automatically generated

* *Following graphs show the ABC analysis*

1. *ABC analysis 🡪 prices per unit*

Chart, line chart

Description automatically generated

1. *ABC analysis 🡪 cumulative prices per unit*

Chart, line chart

Description automatically generated

**2B. How stable is the customers’ demand? (Coefficient of Variation)**

To understand which products will bring planning and distribution challenges, compute the coefficient of variation of the yearly distribution of sales of each reference.

Coefficient of Variation:

The coefficient of variation (CV) is the ratio of the standard deviation to the mean and shows the extent of variability in relation to the mean of the population. The higher the CV, the greater the dispersion.

CV = Standard Deviation / Mean

The coefficient of variation is useful as it is dimensionless (i.e., independent of the unit in which the measurement was taken) and thus, comparable between data sets with different units or widely different means.

**Demand Variability**How stable is your customers’ demand?

* Average Sales: µ
* Standard Deviation:
* Coefficient of Variation: CV = σ/µ

For SKUs with **a high value of CV**, you may face **unstable customer demand** that would lead to workload peaks, forecasting complexity and stock-outs.

Code

* Filter on the first year of sales for HOBBIES SKU’s
* Calculate Mean, Standard deviation and CV of sales
* Sorting (Descending) and Cumulative sales calculation for ABC analysis
* Normalization

Data Pre – processing:

Graphical user interface

Description automatically generated with low confidence

Import Sales Evaluation:

Text

Description automatically generated

A picture containing scatter chart

Description automatically generated

Import Calendar:

Graphical user interface, text

Description automatically generated

Import Pricing:

Text

Description automatically generated

A picture containing text

Description automatically generated

Statistical Analysis:

A picture containing graphical user interface

Description automatically generated

Diagram

Description automatically generated with medium confidence

Turnover Sales:

Graphical user interface, application, Word

Description automatically generated

A picture containing chart

Description automatically generated

Coefficient of Variation : CV = σ/µ

A picture containing scatter chart

Description automatically generated

Product Segmentation:

The inventory is divided into three main groups by analysis, allowing different inventory management strategies to be applied to various inventory segments in an effort to boost revenue and cut costs.

Text, letter

Description automatically generated

Text

Description automatically generated

Chart, line chart

Description automatically generated

Segmentation by Demand Variability

| Coefficient of Variation = f (% TO)

Graphical user interface, text, application

Description automatically generated

Chart, scatter chart

Description automatically generated

Distribution for Important Products with low variability | Example of item with low COV (<2)

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Chart, histogram

Description automatically generated

Distribution for Important Products with high variability | Example of product with High COV

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Chart, histogram

Description automatically generated

Distribution for Low rotation Products with very high variability | Example of item with very high COV (>2)

Text, letter

Description automatically generated

Text

Description automatically generated

Chart

Description automatically generated with medium confidence

Normality Test:

Graphical user interface, text, application

Description automatically generated

Chart, scatter chart

Description automatically generated