WALMART SALES FORECASTING

Project

EDUBRIDGE LEARNING

OVERVIEW

Sales forecasting is the process of estimating the future sales of a product or service. It is a crucial part of any company and its business plan, as it helps businesses make informed decisions about resource allocation, marketing strategy, and investment. Walmart is one of the largest retail companies in the world, and it uses a variety of sales forecasting methods, such as time series analysis, regression analysis, etc., to inform its business decisions. In this article, we will apply regression analysis for sales forecasting on a dataset provided by Walmart

WHAT ARE WE BUILDING

In this project, we will use the weekly sales data provided by Walmart It consists of historical sales data for 45 Walmart stores located in different regions. Each store contains a number of departments, and our objective is to predict the department-wide sales for each store. We will apply regression analysis to forecast the sales for each department based on multiple factors, such as temperature, fuel price, store type, CPI, employment index, etc.

HOW WE HAVE BUILD THIS MODEL

- We handled missing values present in the dataset during the data cleaning stage.
- We have performed exploratory data analysis (EDA) using various visualization techniques to identify underlying patterns and correlations that can help us derive insights.
- Further, we trained and developed multiple regressor ML models, such as KNN Regressor, Decision Tree Regressor, and Random Forest Regressor,Regressor model and compare their performance based on the model's accuracy and RMSE.

PRE-REQUSITES & REQUIREMENTS

PRE-REQUISITES

- Python
- Data Visualization
- Descriptive Statistics
- Machine Learning
- Data Cleaning and Preprocessing

REQUIREMENTS

- Pandas
- Numpy
- Matplotlib
- Seaborn
- Sklearn

DATASET FEATURE DESCRIPTION

SALES.CSV

This file consists of information about the 45 stores, indicating the type and size of the store.

TRAIN.CSV

It contains each department's historical weekly sales data. It contains the following features -

- Store - the store number.
- Dept - the department number.
- Date - the week.
- Weekly_Sales - sales for the given department in the given store.
- IsHoliday - whether the week is a special holiday week.

FEATURES.CSV

It contains additional features regarding the store, as mentioned below -

- Store - the store number.
- Date - the week.
- Temperature - the average temperature in the region.
- Fuel_Price - the cost of fuel in the region.
- MarkDown1-5 - anonymized data related to promotional markdowns. It is only available after Nov 2011 and is not available for all stores all the time.
- CPI - the consumer price index.
- Unemployment - the unemployment rate.
- IsHoliday - whether the week is a special holiday week.

BUILDING WALMART FORECASTING SYSTEM

We started the project by importing all necessary libraries to load the dataset, perform EDA, and build ML models. Then We loaded all the datasets – features, store, and train, in pandas dataframes and join them together

	Store	Dept	Date	Weekly_Sales	IsHoliday	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	Type	Size
0	1	1	2010-02-05	24924.50	False	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.106	A	151315
1	1	1	2010-02-12	46039.49	True	38.51	2.548	NaN	NaN	NaN	NaN	NaN	211.242170	8.106	A	151315
2	1	1	2010-02-19	41595.55	False	39.93	2.514	NaN	NaN	NaN	NaN	NaN	211.289143	8.106	A	151315
3	1	1	2010-02-26	19403.54	False	46.63	2.561	NaN	NaN	NaN	NaN	NaN	211.319643	8.106	A	151315
4	1	1	2010-03-05	21827.90	False	46.50	2.625	NaN	NaN	NaN	NaN	NaN	211.350143	8.106	A	151315

EXPLORATORY DATA ANALYSIS

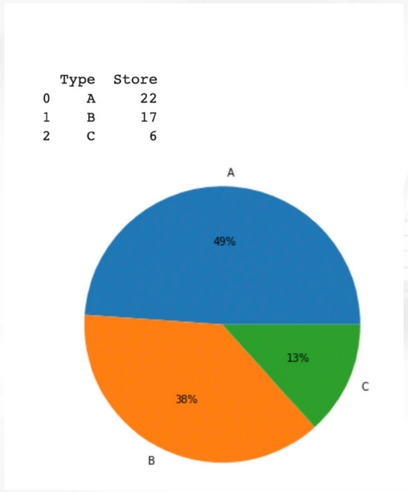
We explored summary statistics of the numerical features in the dataset. As you can see in the below figure, the average weekly sales are 16K, the average CPI is 171, and the average size of the store is 136727.

	Weekly_Sales	Temperature	CPI	Size
count	421570.000000	421570.000000	421570.000000	421570.000000
mean	15981.258123	60.090059	171.201947	136727.915739
std	22711.183519	18.447931	39.159276	60980.583328
min	-4988.940000	-2.060000	126.064000	34875.000000
25%	2079.650000	46.680000	132.022667	93638.000000
50%	7612.030000	62.090000	182.318780	140167.000000
75%	20205.852500	74.280000	212.416993	202505.000000
max	693099.360000	100.140000	227.232807	219622.000000

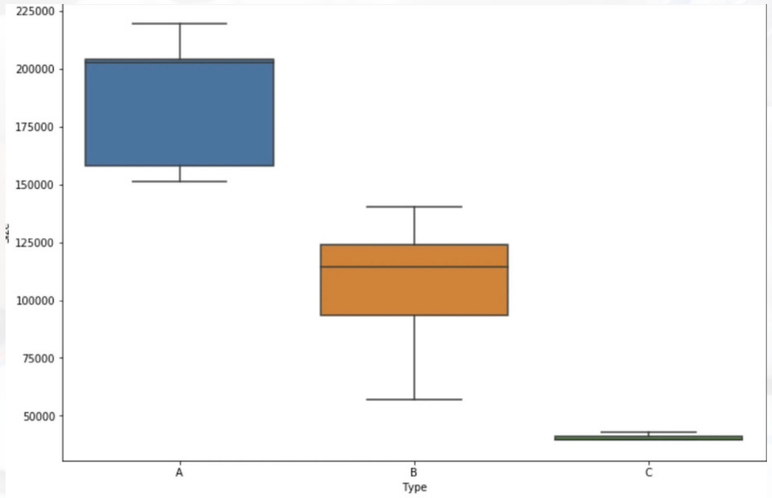
DATA ANALYSIS

There are a total of 3 store types - A having a share of 49% of total stores, B having a share of 38% of total stores, and C having a share of 13% of total stores.

	Type	Store
0	A	22
1	B	17
2	C	6

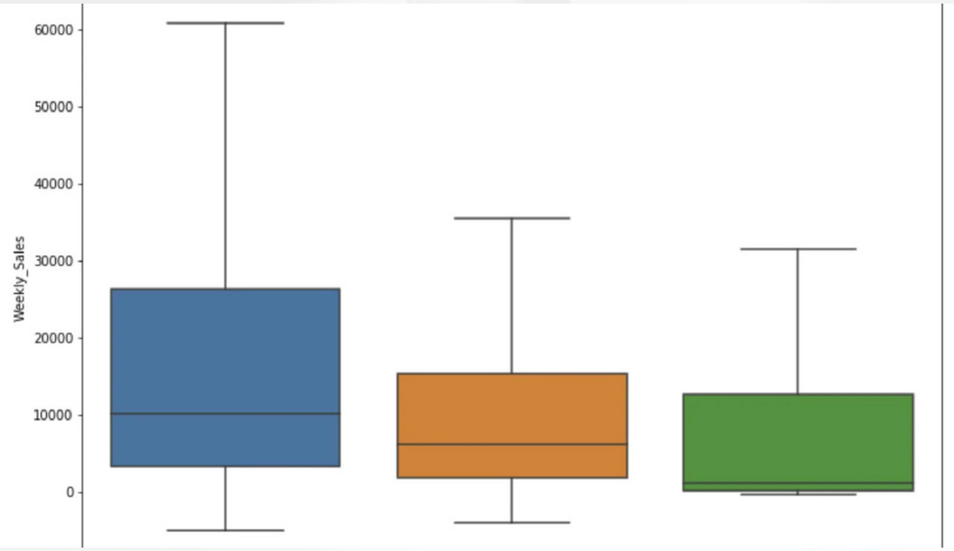


We explored the size distribution of stores for each store type. As shown below, store type A are the largest ones, store type B are smaller than A, and store type C are the smallest ones.

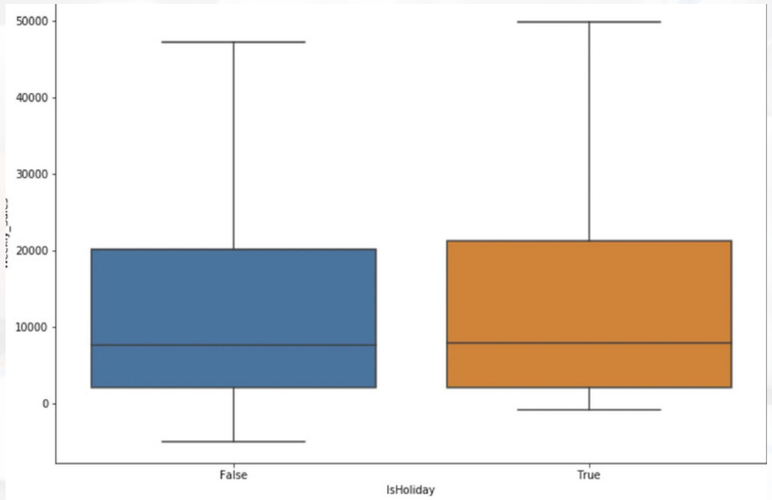


DATA ANALYSIS

We explored the distribution of weekly sales for each store type using box plots. As shown below, weekly sales are on the higher side for A compared with B and C.

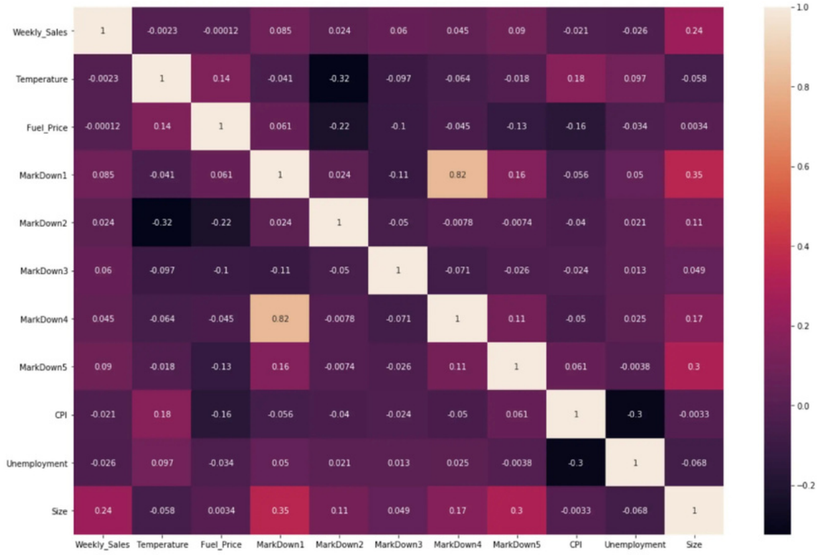


We explored how holidays affect the sales of each store. As you can see in the below figure, there is no major impact observed of holidays on weekly sales figures.



CORELATION

We analyzed the correlation between numerical features and weekly sales using a heatmap. As you can see below, weekly sales have the highest correlation of 0.24 with the size of the store, which is in line with our previous findings. The remaining features are mostly uncorrelated with each other, except MarkDown1 has a correlation of 0.84 with Markdown2.



DATA PREPROCESSING

In this stage, we performed the following steps on the dataset -

- We impute NULL values with 0 in MarkDown1-5 features.
- Created new features year, month, and day from date.
- Before developing the ML models, we need to hot encode categorical features and standardize numerical features in the dataset. Further, we split the input data into training and testing data with an 80:20 ratio.

DATA PREPROCESSING

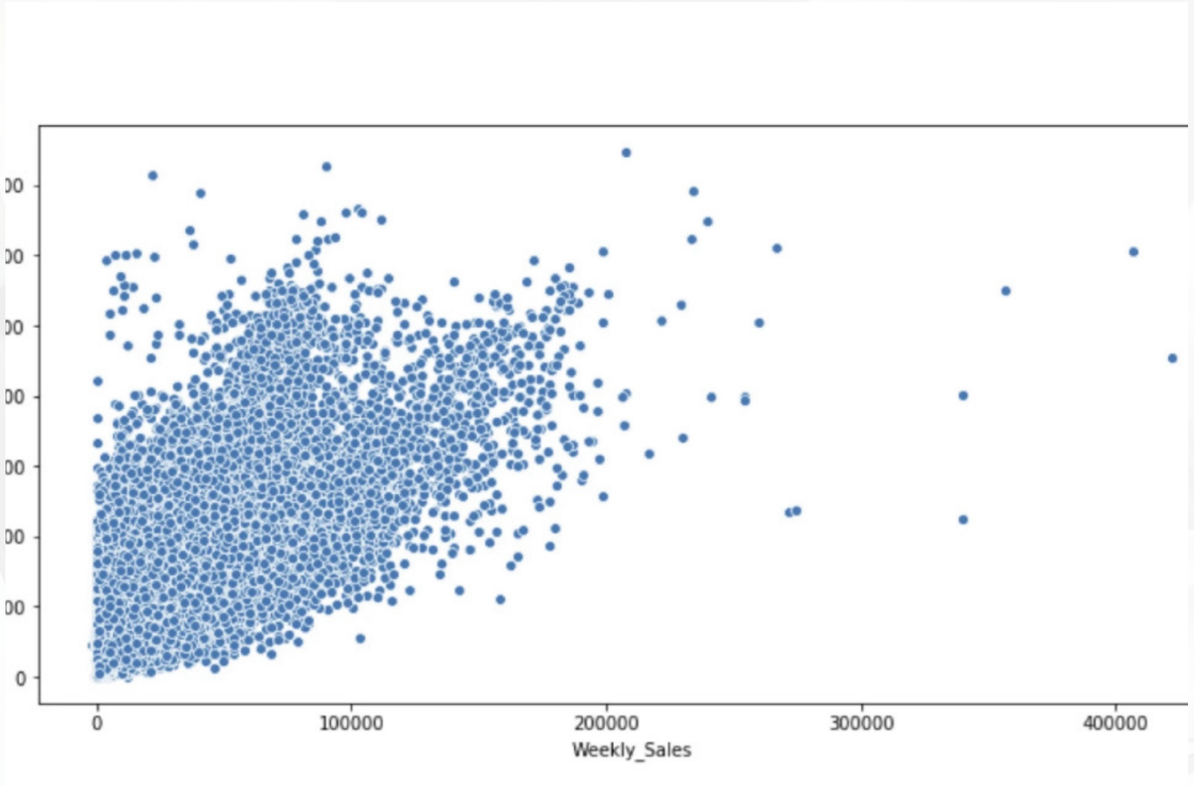
Developing the ML Models

- First, we trained a KNN regression model and evaluate its performance. In this project, we will use the coefficient of determination (accuracy score or r-squared score), mean absolute error (MAE), and root mean squared error (RMSE)scores to compare and evaluate the performance of the ML models.

```
MAE is - 8764.248232903195
RMSE is - 14669.238560915983
Accuracy Score is - 0.578131555334959
```

As we can see in the above figure, the r-squared score is only 0.57. Let's plot the scatter plot between observed and predicted values of weekly sales in the test data.

DATA PREPROCESSING



As you can see in the above figure, predicted and observed values have low correlation, and points in the plot are spread out. Later we train a Decision Tree Regressor to check whether we get any improvement in the r-squared score or not.

DATA PREPROCESSING

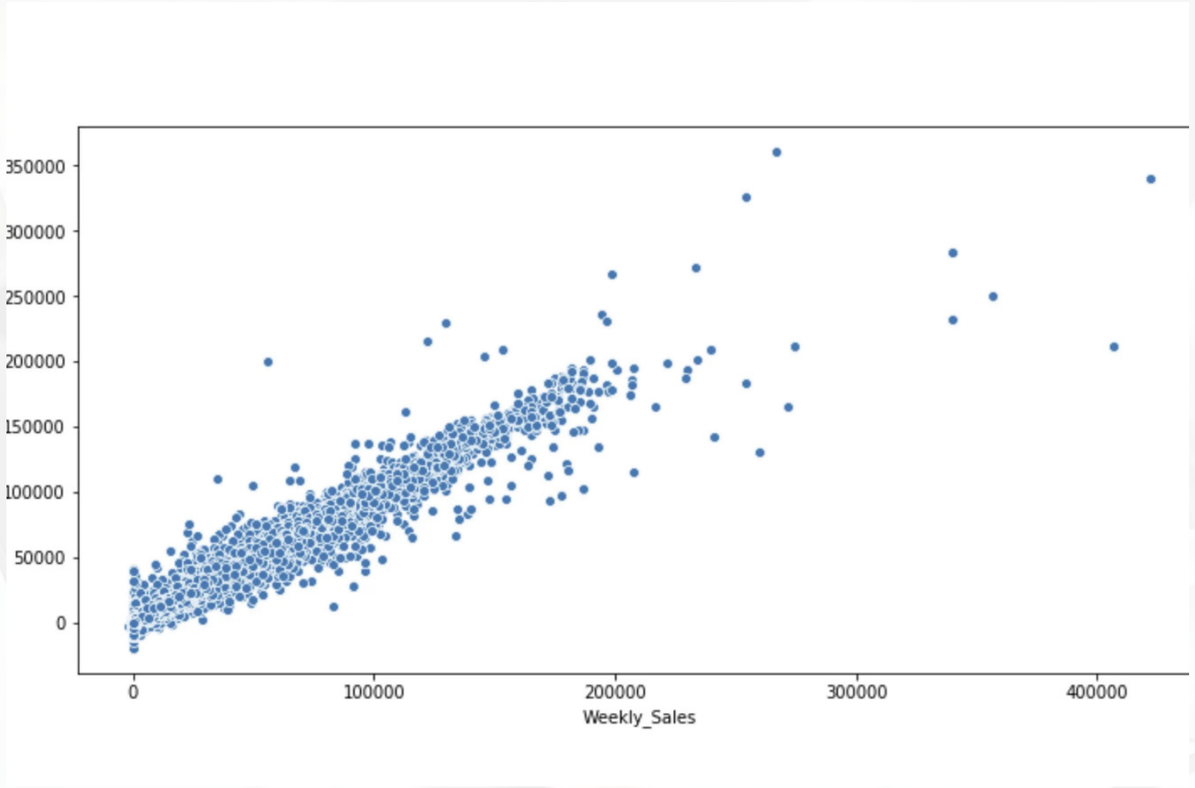
```
MAE is - 1853.4500077092766
RMSE is - 4612.982820038718
Accuracy Score is - 0.9582817988
```

As we can see in the above figure, with Decision Tree Regressor, we get a huge improvement in both RMSE and r-squared score. We get an r-squared score of 0.95. Let's train a Random Forest Regressor and check whether we get any further improvement

We were able to gain slight improvement in RMSE and r-squared score. With Random Forest, we got an r-squared score of 0.96.

```
MAE is - 1939.5245433047946
RMSE is - 4310.870896561282
Accuracy Score is - 0.9635672517975025
```

SCATTER PLOT



LWe plot the scatter plot between observed and predicted values of weekly sales in the test data. As you can see, points in the plot are very tightly distributed, and observed and predicted values of weekly sales have a strong correlation.

CONCLUSION

- We examined the Walmart store's sales forecasting dataset by applying various statistical and visualization techniques.
- We trained and developed four ML models. We also concluded that for this problem Decision Tree Regressor works the best

THANK YOU

FOR YOUR ATTENTION