# Capstone Project - Walmart Sales Forecasting

# Contents

## Problem Statement

A Walmart retail store has multiple outlets across the country. They are facing Inventory Management problem. These stores are unable to match the demand with respect to supply. The basic case is to know the demand of products that are sold in the store. If the decision making authority know what is the demand of each products 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.

## Project Objective

1. Draw useful insights that can be used by each of the stores to improve in various areas.

2. Forecast the sales for each store for the next 12 weeks. We want to analyze how internal and external (Temperature, CPI, Unemployment Rate and Fuel Prices) factors are playing a role.

## Data Description

This file contains a related to the store, department, and regional activity for the given dates. It contains 6435 rows and 8 columns.

Below is brief insight to each feature:

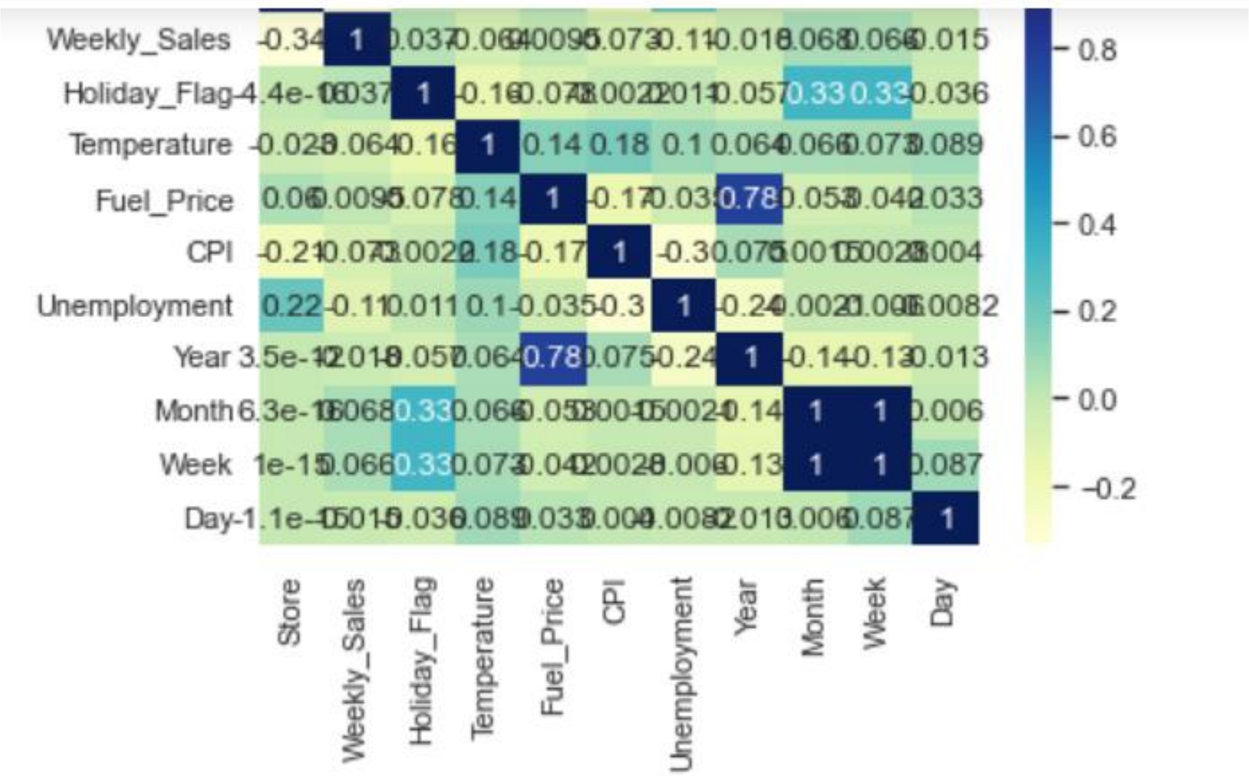
- Store : Store number
- Date : Week of Sales
- Weekly_Sales:  Sales for the given store in that week
- Holiday_Flag : If it is a holiday week
- Temperature : Temperature on the day of the sale
- Fuel_Price : Cost of the fuel in the region
- CPI Consumer : consumer Price Index
- Unemployment : Unemployment Rate during that week in the region of the store.

## Data Pre-processing Steps and Inspiration

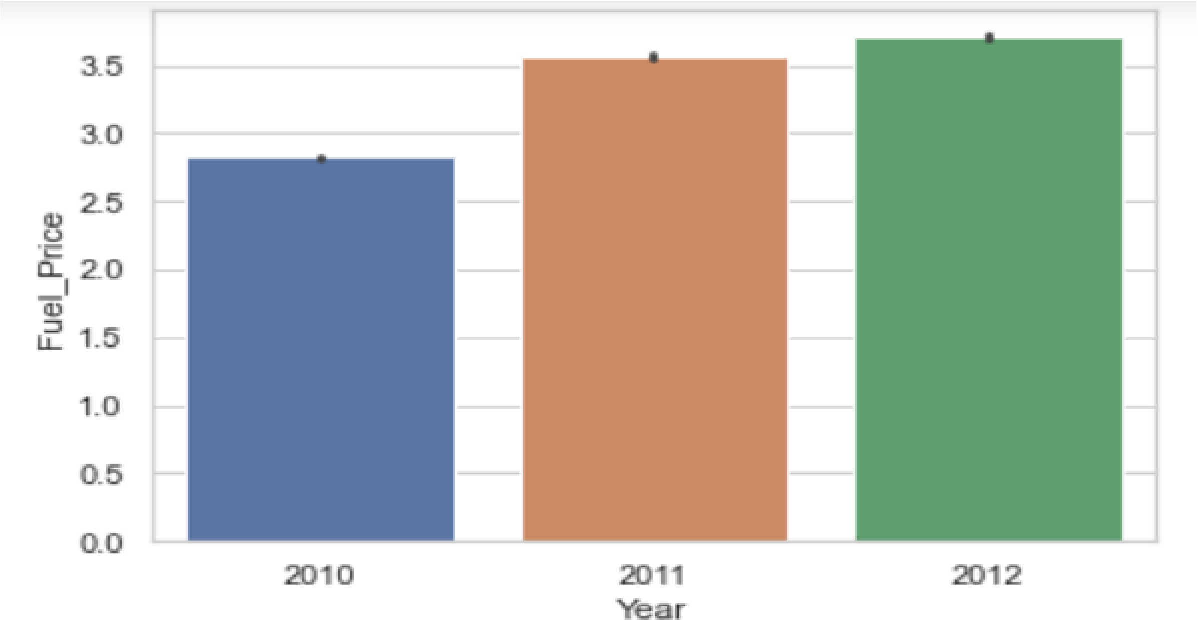
Zero null Values were identified in dataset.

There were no negative weekly sales. Hence, it can be concluded that data is free from outliers.

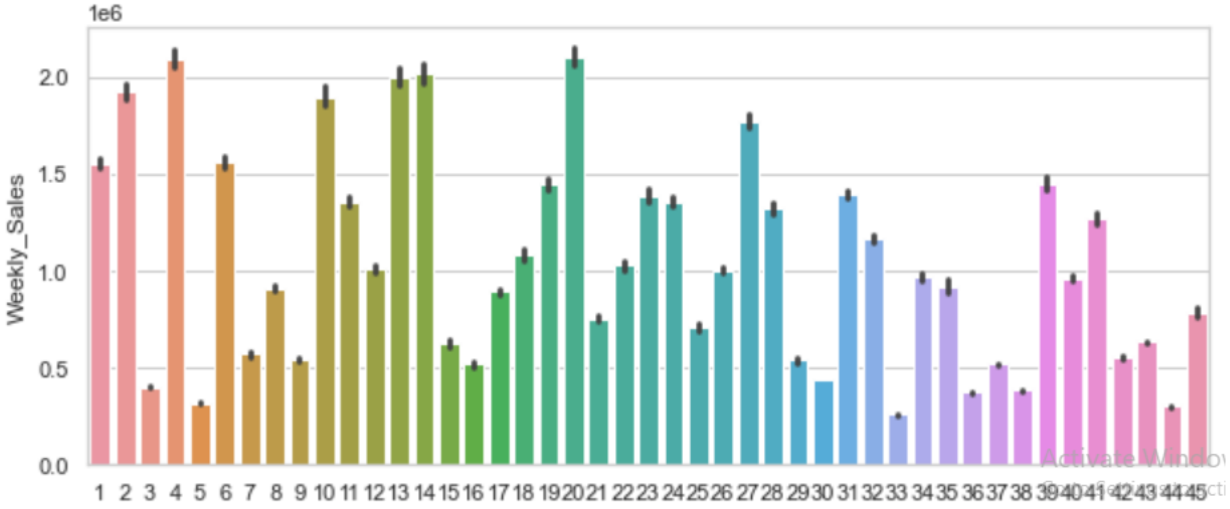
Below correlation heatmap revealed strong positive correlation between year and Fuel Price.



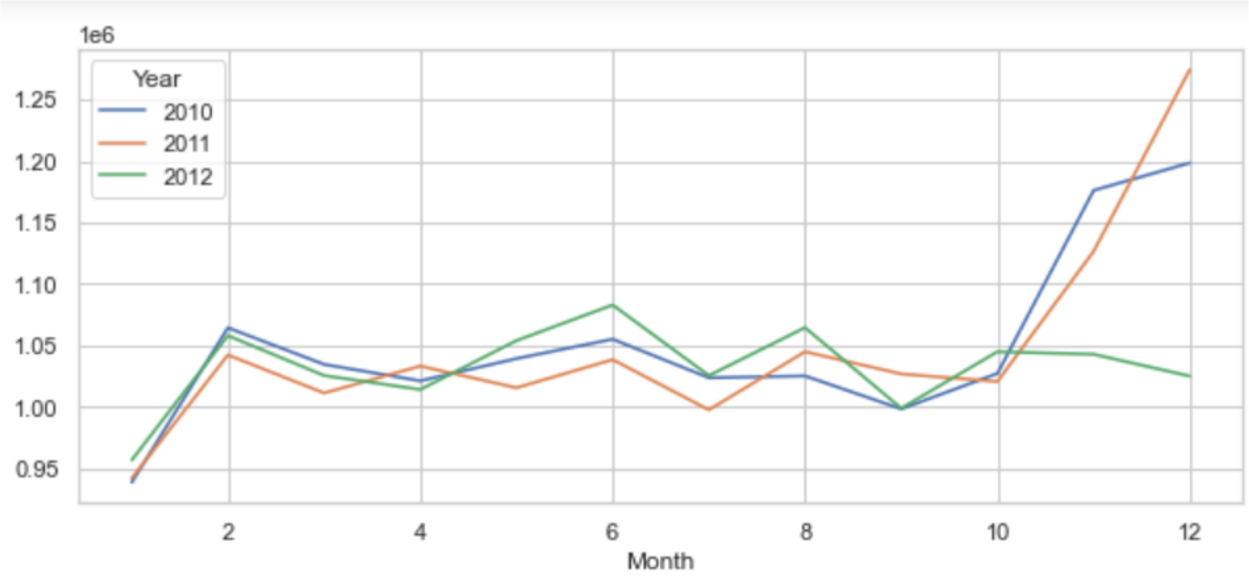
Below is better analysis of same:

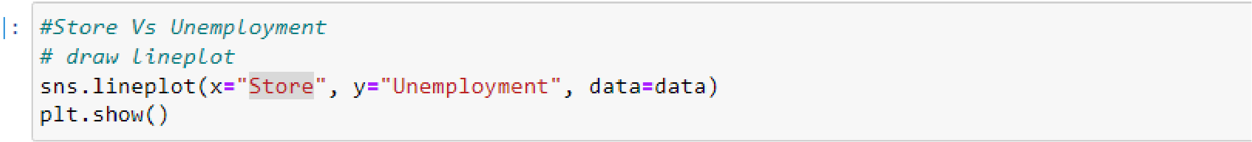


Below chart showcases Weekly Sales against store

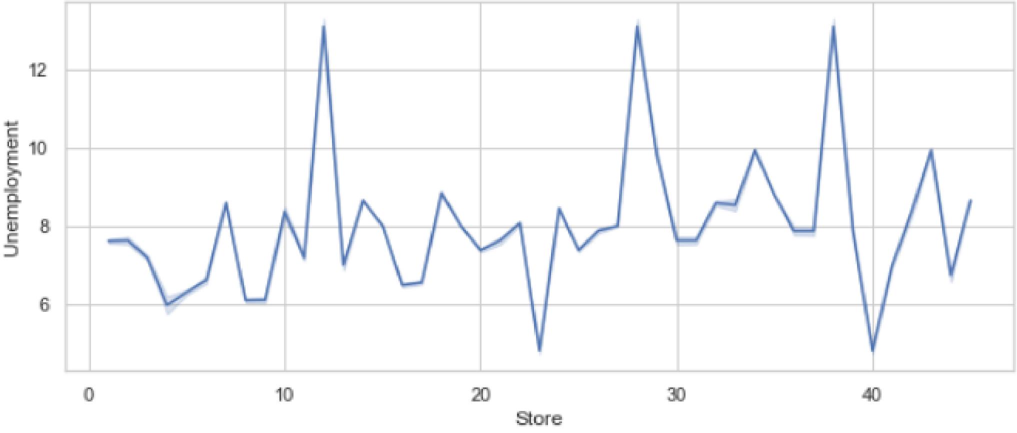


Below show sales are high after 10th month as it is festive season





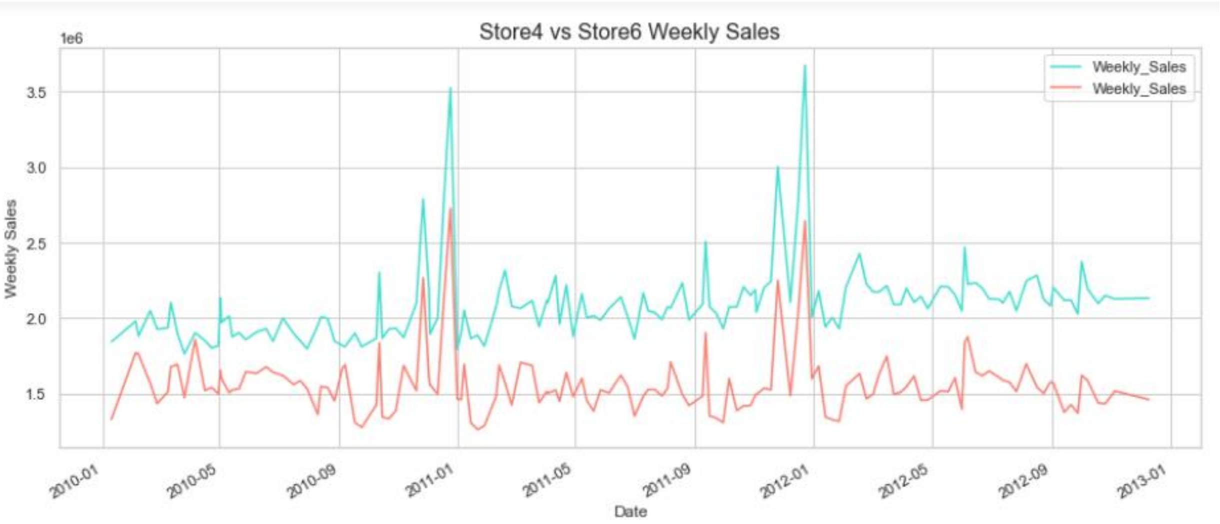
```
#Store Vs Unemployment
# draw lineplot
sns.lineplot(x="Store", y="Unemployment", data=data)
plt.show()
```



## Choosing the Algorithm for the Project

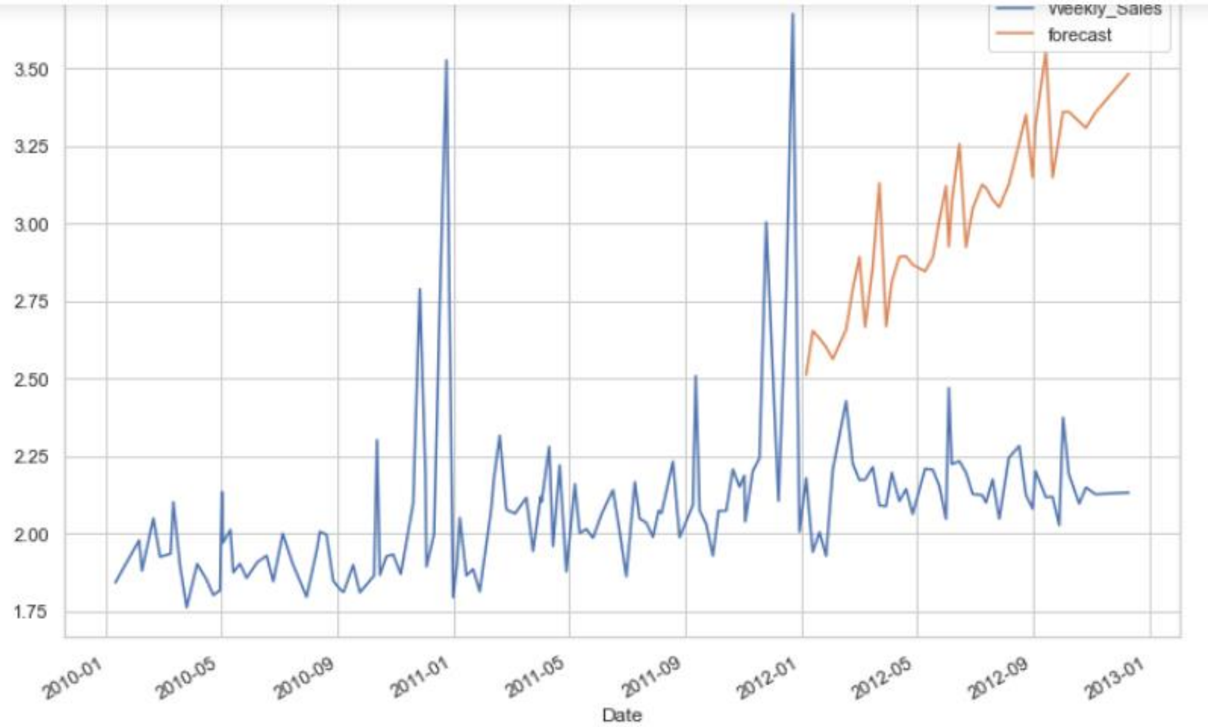
We will explore multiple algorithm and decide best among those. Data supplied to us is Time series Data. As there is only one variable dependent on time so we can conclude it as univariate time series. We will consider all stores as well as individual stores (couple of them in this example) and perform a detailed time-series analysis on it. Additionally, we will also explore regression models.

TimeSeries Analysis was conducted by extracting data for store4 and store6.

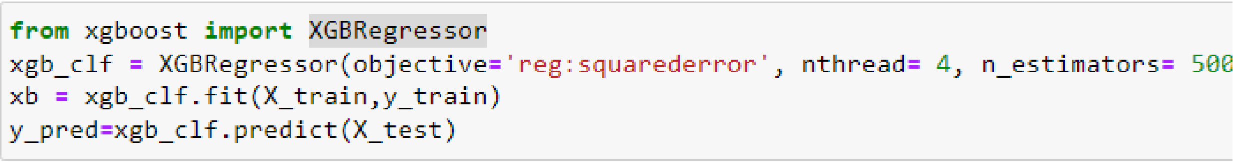


Both the stores have almost the same trends and spike just the magnitude is different. Further from above graphs, we can infer that sales spike up in festive season (Christmas, New year etc). If you look at entire year, you can conclude it as seasonal as well as stationary.

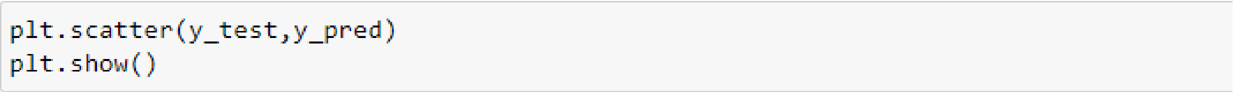
SARIMAX(2, 1, 2) did not yield satisfactory result.



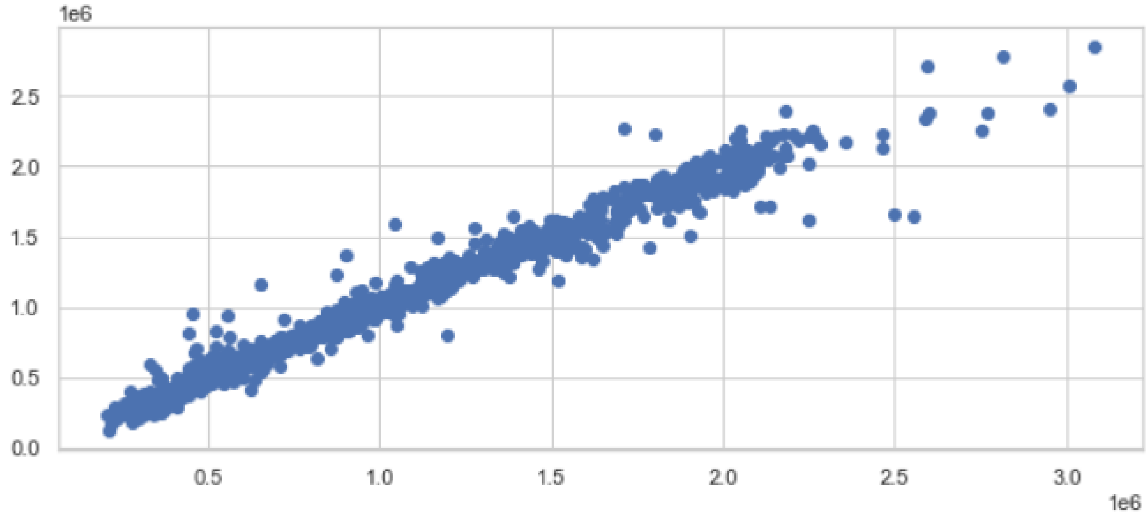
Upon evaluation of regression models like KNN, Decision tree, RandomForest and XGB , highest and reliable accuracy of 97.14 was found in XGB.



```
from xgboost import XGBRegressor
xgb_clf = XGBRegressor(objective='reg:squarederror', nthread= 4, n_estimators= 500
xb = xgb_clf.fit(X_train,y_train)
y_pred=xgb_clf.predict(X_test)
```



```
plt.scatter(y_test,y_pred)
plt.show()
```



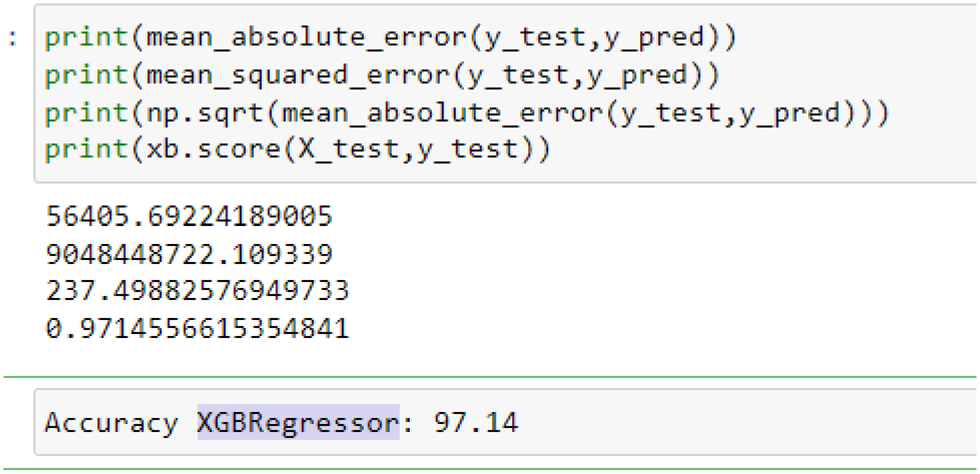
## Motivation and Reasons for Choosing the Algorithm

We concluded on this by taking averages of top n models. As here available data is less, so loss difference is not extraordinary.

## Model Evaluation and Techniques

We primarily evaluated Model on basis of

- mean_absolute_error
- mean_squared_error
- root mean_squared_error
- model score



```
: print(mean_absolute_error(y_test,y_pred))
  print(mean_squared_error(y_test,y_pred))
  print(np.sqrt(mean_absolute_error(y_test,y_pred)))
  print(xb.score(X_test,y_test))
```

```
56405.69224189005
9048448722.109339
237.49882576949733
0.9714556615354841
```

```
Accuracy XGBRegressor: 97.14
```

## Inferences from the Same

- Size of the store is the highest contributing predictor in the model out of all.
- Each store has a unique prediction power. They can be separately analyzed to get prediction for each individual store
- The Sales are very high during November and December and go down in January. So it’s better to employee more staff as casual employee in November and December and encourage permanent staff to take leaves during January.
- The predicted sales data can be used to analyze the sales pattern and accordingly adjust the staff in the store.
- When we implement the project to department level it helps to plan the inventory and staff from a centralized station to every store, which will further help in better planning and cost cutting for inventory management, supply chain management and human resource.
- The low selling stores should look forward to increasing their size and capacity to store more items and consumer products.
- Special discount coupons can be distributed during low selling periods to attract more customers
- Sales are likely to fluctuate during holidays. Special offers can be given during festive season accompanied with suitable marketing to keep the sales high during holidays as well

## Future Possibilities of the Project

Next steps would be

- To check into the store that have poor prediction and check deep what makes those bad.
- To further improve the predictive model using the ensembling method to combine models and come with better model.
- Take the data to Department level and to predict the Department level sales which would help to solve the inventory management issues and supply chain management.

## Conclusion

We can conclude that on basis of evaluation of multiple models and timeseries analysis, we were able to evaluate multiple features that had impact on weekly sales. We also narrowed down to XGB regress or model for future forecasting of Weekly sales.

## References

https://www.youtube.com/watch?v=kdVMiW5b9Xo

https://www.analyticsvidhya.com/

http://www.kaggle.com