**Capstone Project – Walmart**

**Table of Contents**:

1. Problem Statement

2. Project Objective

3. Data Description

4. Data Pre-processing Steps and Inspiration

5. Choosing the Algorithm for the Project

6. Motivation and Reasons for Choosing the Algorithm

7. Assumptions

8. Model Evaluation and Techniques

9. Inferences from the Same

10. Future Possibilities of the Project

11. Conclusion

12. Reference

**Problem Statement:**

A retail store that has multiple outlets across the country are facing issues in managing the inventory - to match the demand with respect to supply

**Project Objective:**

•To find out the impact of various factors on weekly sales, including unemployment , Fuel Price, temperature, and Consumer Price Index (CPI) etc .

•Identify Best-performing and worst-performing stores based on historical data.

•To find out which is the best predictive modelling techniques to forecast sales for each store over the next 12 weeks

**Data Description:**

Dataset Overview:

* Number of Rows: 6435
* Number of Columns: 8

Columns and Data Types:

* **Store**: Numeric identifier for each store (int64)
* **Date**: Date of the week of sales (object)
* **Weekly Sales**: Sales for the given store in that week (float64)
* **Holiday Flag:** Indicator for a holiday week (int64)
* **Temperature**: Temperature on the day of the sale (float64)
* **Fuel Price**: Cost of fuel in the region (float64)
* **CPI**: Consumer Price Index (float64)
* **Unemployment**: Unemployment Rate (float64)

**Data Pre-processing Steps and Inspiration:**

See if there is any null value present or not .

We also extracted year, month , week and day from the date column so we can obtain some meaningful insights from it.

Handling the outliers with interquartile technique and also summarized the data to understand the data very well.

DATA INSIGHTS:

* STORE 4 was the Top performing among all the stores followed by store number 20,14,2,13 .
* Top 5 worst performing stores were store number 33,44,5,36,38 and from that store 33 had a least of sales.
* Sales difference between maximum sales and minimum sales for store was 2.438750e+08
* Top 5 days when Sales was maximum were on 9th September 2011 , 25th November 2011, 4th June 2012, 23rd December 2011 , 12th October .
* Rate of unemployment was maximum in 2010 ,while the store 38,44,4 was most affected due to unemployment rate .
* Store number 36,14,35 was most affected due to CPI
* Year wise Fuel rate kept on increasing.
* Sales during first Quarter is mostly stable, after June the sales started to go down and in the last quarter, we can suddenly see there is rise in the sales .
* The data has some outliers or extreme values, indicated by the spikes in the residual component
* Weekly sales show the seasonal trends this could be because during the holiday season, consumers tend to spend more on gifts, decorations, and other seasonal items, leading to an increase in sales.

**Choosing the Algorithm for the Project :**

* Overall, the data showed a clear seasonal and trend pattern.
* Time series model such as ARIMA (Auto Regressive Integrated Moving Average) , Sarima (Seasonal ARIMA) , Facebook Prophet could be appropriate for predicting the future sales.

**Motivation and Reasons for Choosing the Algorithm:**

Data consists of various irregularities and complex seasonality so Arima model did not perform well

So, we predicted the model using sarima and fb prophet model as this model we well suited for the data

Model was able capture both seasonal and non-seasonal patterns.

But from that Sarima model performed better as compared to fb prophet because the RMSE value for sarima was less as compared to Fb prophet

So, Sarima model was the best model .

**Assumptions :**

* Data was assumed to be stationary i.e. Mean and Variance do not vary with time
* The selection of model order which was made was assumed to be best according to data (i.e. p, d, q) and seasonal order were also correct (P ,Q).
* Model will be Stable over the time

**Model Evaluation and Techniques:**

* For model evaluation technique used was finding RMSE values of both the model and whose RMSE value was less is considered the best model .
* RMSE value of SARIMA was 69038.896 and that of Fb prophet was 98612.661
* As a result, Sarima model was chosen.

**Inferences from the Same :**

* The prediction data shows a similar seasonal pattern, with a small difference in the magnitude of the sales .
* The SARIMAX model seems to have captured the seasonal and trend components of the data, as the prediction follows the general pattern of the actual data.
* However, there are some differences between the actual and predicted data. This could be due to the presence of an outlier or extreme value in the data, which the model might not have captured fully.
* Overall, the SARIMAX model seems to be a good fit for the data

**Future Possibilities of the Project**:

However, you could explore the possibility of personalized sales prediction, where the model predicts sales for individual products or brands.

We could explore the possibility of incorporating external factors, such as weather, holidays, and events, into your model.

We could try other time series models or ensemble methods or deep learning models to see if we can achieve better accuracy.

**Conclusion:**

Data cleaning and preprocessing were crucial in improving the accuracy of the time series models.

Various techniques were used to gather important information for predicting future sales.

ARIMA, SARIMA, and Facebook's Prophet were tested as potential time series models for predicting future sales.

SARIMA was chosen as the best performing model due to its lowest RMSE value.

**Reference:**

* Dataset:- Kaggle
* Facebook. (2017). Prophet: Forecasting at scale. Available at: https://research.fb.com/publications/prophet-forecasting-at-scale/