**DSC630 PA - Project Proposal**

Bellevue University

Winter Term- 2021

Walmart Sales Forecasting

Nitin Mahajan | Ganesh Kale

**Abstract**

**Sales Forecasting** is the process of using a company’s sales records over the past few years to predict the short-term or long-term sales performance of the company in the future. This is one of the pillars of proper financial planning. Sales forecasting is a globally conducted corporate practice where number of objectives are identified, action-plans are chalked out as well as budgets and resources are allotted to them. Here in this project, we are going to build the **Sales Forecast Model** that would learn from the past sales records, events and predict the accurate sales so company will be ready to source appropriate resources before the actual event happens.

The Sales Forecast Model will be machine learning model built using python, trained, and tested on Walmart sales data, the CRSP-DM methodology is used complete this project. Sales Forecast Model is trained using different Machine Learning algorithms such as Linear Regression, Decision Tree, Random Forest, and Gradient Boosting (XGBoost) and model with highest accuracy score is finalized to predict the sales forecast in real time.

The factors considered to build the model and findings about the data with the detailed use cases and technical information is provided in the paper.

# INTRODUCTION

## **Background**

Walmart, Inc. is part of the retail and wholesale business and is based in Bentonville, Arkansas. The President, Chief Executive Officer, and Director is C. Douglas McMillon. Walmart operates Walmart, Walmart Neighborhood Market, Wal-Mart, Walmart.com, and Sam’s Club. Retail companies commonly have issues with predicting sales accurately throughout the days, months, and years ahead. There are many varying factors that can cause issues with predicting sales such as holidays, economic factors, temperature, fuel prices, Consumer Price Index (CPI), and unemployment. Sales are the lifeblood of business. With an accurate sales forecast in hand, one can plan wisely. If the varying factors are not predicted correctly, then there could be staffing issues at stores, financial implications, and the business could become obsolete if customer satisfaction goes down.

Business Sales Executives often find themselves scrambling for answers when it comes to sales forecasting during business reviews with their leaderships team. The Sales Forecast Model will help sales executives to find such answers upfront and be ready with numbers and predictions to share with leaderships team. This model would help individual stores to upscale their customer satisfaction by stocking the right products at right time and decrease overstocking and wastage of food products.

## **Problem Statement**

The goal of this analysis is to predict future sales for the Walmart stores based on the varying features and events mentioned in the introduction.

In addition, Walmart runs several promotional markdown events throughout the year. These markdowns precede prominent holidays, the four largest of which are the Super Bowl, Labor Day, Thanksgiving, and Christmas. The weeks including these holidays are weighted five times higher in the evaluation than non-holiday weeks.

* Build the Machine Learning model that would learn from past records, events and predict the accurate outcomes.
* Predict the Sales forecast for Store and its departments on specific week of the year considering if it is before holiday or after holiday.
* Modeling the effects of markdowns on the holiday weeks in the absence of complete/ideal historical data and predict the sales.

1. **Methods**

The goal of the project is to predict Sales value based on different features, here we used Walmart sales records from different stores for different year. To understand the factors impacting Sales Value we followed different methodologies such as understanding the features correlating with sales value, machine learning algorithms to understand what percentage of features account for sale value and Time Series forecasting. We have explained below each of the methods used to achieve the result in detail.

* 1. **Feature Correlations**:

To understand the features driving the sales value, we should understand first are there any correlation with features and target variable, here target variable is Sales Value. Each feature from the data is compared with Sales value using different visualizations such as bar chart, line chart and heat maps chart and evaluated for any potential impact if sales value. Also created and visualized correlation matrix of all the key features to see correlation among them and with the target value. The higher correlation coefficient value signifies stronger correlation.

* 1. **ML Algorithms**:

There are several machine learning algorithms that can be used to predict the value based on historical data. These ML algorithms are used to train the model and evaluated using Weighted mean absolute error. The model with lowest WMAE score is baselined. Following are the list of ML algorithms are used to train the model –

* Linear Regression
* Decision Tree
* Random Forest
* Gradient Boosting Machine
* ARIMA - Auto Regressive Integrated Moving Average
  1. **Time Series Forecasting:**

The data here we are dealing with is sales history records by date, it is a time series data because it is a series of data points measured at consistent time intervals. We used Auto ARIMA model on this time series to predict the sales values.

Brief about ARIMA model – it is a very popular statistical method for time series forecasting. ARIMA stands for Auto-Regressive Integrated Moving Averages. ARIMA models work on the following assumptions –

* The data series is stationary, which means that the mean and variance should not vary with time. A series can be made stationary by using log transformation or differencing the series.
* The data provided as input must be a univariate series, since arima uses the past values to predict the future values.

# results

Expecting to extract features which have significant impact on predicting the sales by checking the correlations among them and impacting the sales values.

Using this data an appropriate machine learning algorithms are trained and evaluated for their accuracy and different evaluation matrix and model with highest accuracy score will be baselined that can predict future sales.

# Conclusion

Based on the exploratory analysis performed on different features, correlation among features and training result of ML algorithms, following inferences or conclusions can be drawn:

* 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. Notables are Thanksgiving and Christmas 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
* Based on the trained models for predicting the future sales, Gradient Boosting Machine with tuned hyperparameters performs the best.
* The data will be trained on several more machine learning algorithms such as Linear regression, decision tree, Random Forest and ARIMA and accuracy score will be calculated to baseline the model in next milestone.

# Acknowledgements

We would like to express special thanks to each member of our team who helped to complete this project starting from identifying project goal to finalizing the data sources, approach to resolve the problem and deciding on machine learning algorithms.

We also would like to thank you our professor for valuable feedback and guidance provided to complete this project.

# References

* Data Source: <https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting/overview>
* ARIMA Model Info - <https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/>
* <https://www.analyticsvidhya.com/blog/2018/08/auto-arima-time-series-modeling-python-r/>