DSC630 PA - Project Summary

Bellevue University

Winter Term: 2021-2022

Walmart Sales Forecasting

Nitin Mahajan | Ganesh Kale

Executive Summary

What is Sales Forecasting - It 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.

Why is Sales Forecasting Required – Sales Forecasting is one of the pillars of proper financial planning. It 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 to improve sales. With an accurate sales forecast in hand, companies can plan wisely and if the varying factors are not predicted correctly, then there could be staffing issues at stores, financial implications, and the business could become obsolete.

Who and how would it benefit - Business Sales Executives often find themselves scrambling for answers when it comes to sales forecasting during business reviews with their leaderships team. The Sales Forecasting will help sales executives to find such answers upfront and be ready with numbers and predictions to share with leaderships team.

With help of sales forecasting, the individual stores can upscale their customer satisfaction by stocking the right products at right time and decrease overstocking and wastage of food products.

How to solve this problem – The sales forecast can be predicted by building the predictive model using statistical algorithms and machine learning techniques and such techniques are used to identify the likelihood of future outcomes based on historical data. The **Sales Forecast Model** that learns from the past sales records, events, demographic details, and predict the accurate sales so company is ready to source appropriate resources before the actual event happens.

Explanation of Solution – To build the sales forecast model, historical data from 45 Walmart stores located in different regions are collected for past 3 years (2010 to 2012). The gathered for this were having store ids, department ids, weekly sales, and date, the week was having any holidays, average temperature in that region, fuel price, unemployment rate and consumer price index etc. The exploratory analysis was performed to understand what are the factors that influence weekly sales and would help predict sales values. With the help of analysis, we figured out that stores type A is having highest median weekly sales than other stores, the thanksgiving week and Christmas weeks were having highest weekly sales for all three years, the median weekly sales for holiday weeks and non-holiday weeks are almost same but max weekly sales during holidays are quite higher than non-holiday weeks. Also performed correlation test to understand what the factors are those strongly correlated with sales value and departments and store size are the top ones. This analysis also helped to remove unwanted features that would not contribute sales prediction.

To build the model, several machine learning algorithms are used. The ML regressors are used to predict the continuous values based on multiple predictor variables, here weekly sales value is predicted using predictor variables such as store, department, size, temperature, isHoliday etc.

These different models are then tested with test data and evaluated the model's accuracy and the model which have best accuracy which is 97.7% and least errors are baselined as Sales Forecast Model.

How would model benefit – The baselined model which produced best accuracy score can be used to predict the sales forecast for any given store and its departments. The model can be deployed to production system and simple application can be built to predict sales.

The Application would accept values such store number, department number, week of the year, size of the store, is holiday in the week, average temperature, unemployment rate in that week, fuel price etc. and based on these values it would predict the sales value for that store and departments.

TECHNICAL REPORT

1. INTRODUCTION

1.1. 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.

1.2. 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.

2. 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.

2.1. **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.

2.2. **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 RMSE score and best accuracy score is baselined. Following are the list of ML algorithms are used to train the model –

- KNN Regression
- Decision Tree
- Random Forest
- Gradient Boosting Machine
- ARIMA - Auto Regressive Integrated Moving Average

2.3. 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.

3. 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.

4. 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

- Department, Store size and Type have moderate correlation with the weekly sales.
- Markdowns 1-5, Temperature. Fuel price, CPI and Unemployment are very weakly correlated with 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 is trained on several more machine learning algorithms such as KNN regression, decision tree, Random Forest and XGBoost and accuracy score is calculated to baseline the model.
- Comparing the accuracy of different models, it turns out that XGBoost regressor with accuracy score 97.7% and Root Mean Squared Error 3463 is the best model for this project and is baselined.

5. 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.

6. 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/