Project Report

on

Capstone Project

(Walmart)

Table of Contents

- Problem Statement
- Project Objective
- Data Description
- Data Pre-processing Steps and Inspiration
- Choosing the Algorithm for the Project
- Motivation and Reasons for Choosing the Algorithm
- Model Evaluation and Techniques
- Inferences from the Same
- Future Possibilities of the Project
- Conclusion
- References

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.

I am provided with the weekly sales data for their various outlets. Use statistical analysis, EDA, outlier analysis, and handle the missing values to come up with various insights that can give them a clear perspective on the following:

a. If the weekly sales are affected by the unemployment rate, if yes - which stores are suffering the most?
b. If the weekly sales show a seasonal trend, when and what could be the reason?
c. Does temperature affect the weekly sales in any manner?
d. How is the Consumer Price index affecting the weekly sales of various stores?
e. Top performing stores according to the historical data.
f. The worst performing store, and how significant is the difference between the highest and lowest performing stores.

Project Objective

The purpose of this study is to predict the weekly sales for Walmart based on available historical data (collected between 2010 to 2013) from 45 stores located in different regions around the country. Each store contains a number of departments and the main deliverable is to predict the weekly sales for all such departments.

The data has been collected from Kaggle and contains the weekly sales for 45 stores, the size and type of store, department information for each of those stores, the number of weekly sales, and whether the week is a holiday week or not. There is additional information in the dataset about the factors that might influence the sales of a particular week. Factors like Consumer Price Index (CPI), temperature, fuel price, promotional markdowns for the week, and unemployment rate have been recorded for each week to try and understand if there is a correlation between the sales of each week and their determinant factors.

The main focus of this research is to predict Walmart's sales based on the available historic data and identify whether factors like temperature, unemployment, fuel prices, etc affect the weekly sales of particular stores under study. This study also aims to understand whether sales are relatively higher during holidays like Christmas and Thanksgiving than normal days so that stores can work on creating promotional offers that increase sales and generate higher revenue.

Additionally, the application of big data analytics will help analyze past data efficiently to generate insights and observations and help identify stores that might be at risk, help predict as well as increase future sales and profits and evaluate if the organization is on the right track.

Data Description

The dataset for this study has been acquired from a past Kaggle competition hosted by Walmart, this can be found here: https://www.kaggle.com/c/Walmart-recruiting-store-sales-forecasting/data. It contains historic weekly sales information about 45 Walmart stores across different regions in the country along with department-wide information for these stores. The main goal of this study is going to be to predict the department-wide weekly sales for each of these stores.

Feature Name	Description
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 Price Index
Unemployment	Unemployment Rate

Data Pre-processing Steps and Inspiration

Studies have previously been performed to predict sales for retail industry corporations based on the availability of relevant historic data. Several authors from the Fiji National University and The University of the South Pacific analysed the Walmart dataset to predict sales ("Walmart's Sales Data Analysis - A Big Data Analytics Perspective," 2017). Tools like Hadoop Distributed File Systems (HDFS), Hadoop MapReduce framework, and Apache Spark along with Scala, Java, and Python high-level programming environments were used to analyse and visualize the data. Their study also aimed at trying to understand whether the factors included in the dataset have any impact on the sales of Walmart.

In 2015, Harsoor and Patil (Harsoor & Patil, 2015) worked on forecasting Sales of Walmart Stores using big data applications: Hadoop, MapReduce, and Hive so that resources are managed efficiently. This paper used the same sales data set that has been used for analysis in this study, however, they forecasted the sales for the upcoming 39 weeks using Holt's winter algorithm. The forecasted sales are visually represented in Tableau using bubble charts.

Michael Crown (Crown, 2016), a data scientist, performed analysis on a similar dataset but instead focused on the usage of time series forecasting and non-seasonal ARIMA models to make his predictions. He worked on ARIMA modelling to create one year of weekly forecasts using 2.75 years of sales data, with features of the store, department, date, weekly sales, and holiday data. Performance was measured using normalized root-mean-square error (NRMSE).

Forecasting has not been limited to just business enhancement. Several researchers have tried to utilize machine learning and statistical analysis to build predictive models that can accurately predict the weather, monitor stock prices and analyse market trends, predict illnesses in a patient, etc.

Adopting a similar approach to Kassambara (kassambara, 2018), this study also studies the interaction effects between the multiple independent variables in the dataset like unemployment, fuel prices, CPI, etc. and tries to find if there is a relationship between a combination of these factors and weekly sales.

Choosing the Algorithm for the Project

Trying to find and implement the most effective model is the biggest challenge of this study. Selecting a model will depend solely on the kind of data available and the analysis that has to be performed on the data (UNSW, 2020).

Several models have been studied as part of this study that were selected based on different aspects of our dataset; the main purpose of creating such models is to predict the weekly sales for different Walmart stores and departments, hence, based on the nature of models that should be created, the following four machine learning models have been used:

- Linear Regression
- Lasso Regression
- Gradient Boosting Machine
- Random Forest

Each of these methods have been discussed briefly in the upcoming report. For each of the models, why they were chosen, their implementation and their success rate (through WMAE) have been included.

Motivation and Reasons for Choosing the Algorithm

With growing technology and increasing consumer demand, Walmart can shift its focus on the e-commerce aspects of the business. Taking inspiration from Amazon's business model, Walmart can grow its online retail business massively and gather huge profits. With already established stores and warehouses, it is easier for the organization to create a nationwide reach, limiting the presence of physical stores and helping their customers save on fuel costs by delivering goods at their doorstep. It also makes it a lot easier to identify consumer buying patterns. An important aspect of this study is also to try and understand customer buying behaviour based on regional and departmental sales. This customer segmentation can help the organization in creating and communicating targeted messages for customers belonging to a particular region, establishing better customer relationships, focusing on profitable regions, and identifying ways to improve products and services in specific regions or for specific customers.

A huge constraint of this study is the lack of sales history data available for analysis. The data for the analysis only comes from a limited number of Walmart stores between the years 2010 and 2013. Because of this limited past history data, models cannot be trained as efficiently to give accurate results and predictions. Because of this lack of availability, it is harder to train and tune models as an over-constrained model might reduce the accuracy of the model. An appropriate amount of training data is required to efficiently train the model and draw useful insights. Additionally, the models created have been developed based on certain preset assumptions and business conditions; it is harder to predict the effects of certain economic, political, or social policies on the sales recorded by the organization. Also, it is tough to predict how the consumer buying behaviour changes over the years or how the policies laid down by the management might affect the company's revenue; these factors can have a direct impact on Walmart sales and it is necessary to constantly study the market trends and compare them with existing performance to create better policies and techniques for increased profits.

Model Evaluation and Techniques

Linear Regression:

Regression analysis helps find a relationship between two or more variables, an independent variable (predictor) and a dependent variable (target), in a dataset. Regression analysis specifies how the value of a target variable changes with a change in the value of the predictor variable when all other variables in the dataset are constant and evaluate the accuracy of this relationship (Panchotia, 2020).

Starting with the most basic and straightforward model for this analysis, linear regression aims at finding relationships between two linear variables, a predictor and a target. For the purpose of this study, multiple linear regression helps in predicting the future value of a numeric variable based on past data (Bari et al., n.d.).

Linear regression is most suitable for linear data, i.e., data without the presence of outliers as these disrupt the linear nature of the dataset, hence resulting in a high error rate and low model performance. It is imperative to deal with outliers (influential points) in the training dataset before creating linear regression models as the presence of such outliers affects the regular distribution of points, i.e., slope of the model, resulting in inaccuracies.

Lasso Regression:

Lasso (least absolute shrinkage and selection operator) regression, also known as regularized linear regression, is another regression analysis model that handles data that suffer from multicollinearity and is primarily suitable when there are several predictor variables in the data. Lasso regression is an extension of linear regression in which the loss function is modified to minimize the complexity of the model by limiting the sum of the absolute values of the model coefficients. (Singh, 2019)

The algorithm creates a penalty against complexity by adding a regularization term such that with increasing value of the regularization parameter, the weights are reduced (Kumar, 2020). As the value of the regularization parameter (here alpha) goes up, it reduces the absolute weight values shrink and the loss function is curtailed. Lasso is also considered as an effective feature selection method as when the loss function decreases because of regularization, some of the features from the dataset are removed as their weights become zero. One

of the key differences between linear and several other regression models is that while linear regression does not exactly tune the parameters, the other models allow for tuning of the hyperparameter, in this case, lambda.

Gradient Boosting Machine:

Gradient boosting is a sequential technique, based on ensemble learning, which combines the decisions from multiple machine learning models to improve the overall performance of the model. In boosting, the model outcomes for any instance in the sequence are weighed based on the results of the last instance. The correct predictions are assigned a lower weight, while the incorrect predictions are assigned a higher weight. The model will then focus on higher weight points as it might go wrong with the lower weight points. After many iterations, a combined result is generated, using the accuracy of all the models. Gradient Boosting Machine uses 3 types of parameters: Tree-Specific Parameters, Boosting Parameters, and other miscellaneous parameters (Jain, 2016).

Using GBM provides multiple advantages: it usually provides better predictive accuracy compared to other models, it provides various hyperparameter tuning options, is time-efficient for larger datasets, and handles missing data (Guide, n.d.).

Random Forest:

The random forest regression operates by making multiple and different regression decision trees at the time of training. Each tree predicts a decision based on the criteria it picked. The random forest then makes a prediction by taking the average of individual predictions (Bakshi, 2020).

Random forest usually has good accuracy compared to other linear models and scales well with new features or samples. This regression model can handle missing data and outliers which makes it time-saving and easy to use (Keboola, 2020). This model is powerful because it performs well on various problems, including attributes with non-linear relationships.

Inferences from the Same

Studies have previously been performed to predict sales for retail industry corporations based on the availability of relevant historic data. Several authors from the Fiji National University and The University of the South Pacific analysed the Walmart dataset to predict sales ("Walmart's Sales Data Analysis - A Big Data Analytics Perspective," 2017). Tools like Hadoop Distributed File Systems (HDFS), Hadoop MapReduce framework, and Apache Spark along with Scala, Java, and Python high-level programming environments were used to analyse and visualize the data. Their study also aimed at trying to understand whether the factors included in the dataset have any impact on the sales of Walmart.

In 2015, Harsoor and Patil (Harsoor & Patil, 2015) worked on forecasting Sales of Walmart Stores using big data applications: Hadoop, MapReduce, and Hive so that resources are managed efficiently. This paper used the same sales data set that has been used for analysis in this study, however, they forecasted the sales for the upcoming 39 weeks using Holt's winter algorithm. The forecasted sales are visually represented in Tableau using bubble charts.

Kassambara (kassambara, 2018), in his article, throws light on the implementation of interaction effects with a multiple linear regression in R. Taking a basic multiple regression model as a base where he tries to predict sales based on advertising budgets spent on YouTube and Facebook, he tries to create an additive model based on two relevant predictors (budget for YouTube and budget for Facebook). His model assumes that the effect on sales of YouTube advertising is independent of the effect of Facebook advertising
and subsequently creates a regression model. With an R2 score of 0.98, he observes that there is an interactive relationship between the two predictor variables (YouTube and Facebook advertising) and this additive model performs better than the regular regression model.

Adopting a similar approach to Kassambara (kassambara, 2018), this study also studies the interaction effects between the multiple independent variables in the dataset like unemployment, fuel prices, CPI, etc. and tries to find if there is a relationship between a combination of these factors and weekly sales.

Future Possibilities of the Project

With growing technology and increasing consumer demand, Walmart can shift its focus on the e-commerce aspects of the business. Taking inspiration from Amazon's business model, Walmart can grow its online retail business massively and gather huge profits. With already established stores and warehouses, it is easier for the organization to create a nationwide reach, limiting the presence of physical stores and helping their customers save on fuel costs by delivering goods at their doorstep. It also makes it a lot easier to identify consumer buying patterns.

An important aspect of this study is also to try and understand customer buying behaviour based on regional and departmental sales. This customer segmentation can help the organization in creating and communicating targeted messages for customers belonging to a particular region, establishing better customer relationships, focusing on profitable regions, and identifying ways to improve products and services in specific regions or for specific customers.

Another aspect that would be worth exploring with this study is identifying trends with sales for each of the stores and predicting future trends based on the available sales data. Time series forecasting can be utilized (ARMA and ARIMA modelling) to predict future sales for each of the stores and their respective departments.

Conclusion

The main purpose of this study was to predict Walmart's sales based on the available historic data and identify whether factors like temperature, unemployment, fuel prices, etc affect the weekly sales of particular stores under study. This study also aims to understand whether sales are relatively higher during holidays like Christmas and Thanksgiving than normal days so that stores can work on creating promotional offers that increase sales and generate higher revenue.

Pertaining to the specific factors provided in the study (temperature, unemployment, CPI, and fuel price), it was observed that sales do tend to go up slightly during favourable climate conditions as well as when the prices of fuel are adequate. However, it is difficult to make a strong claim about this assumption considering the limited scope of the training dataset provided as part of this study. By the observations in the exploratory data analysis, sales also tend to be relatively higher when the unemployment level is lower. Additionally, with the dataset provided for this study, there does not seem to be a relationship between sales and the CPI index. Again, it is hard to make a substantial claim about these findings without the presence of a larger training dataset with additional information available.

Interaction effects were studied as part of the linear regression model to identify if a combination of different factors could influence the weekly sales for Walmart. This was necessary because of the presence of a high number of predictor variables in the dataset. While the interaction effects were tested on a combination of significant variables, a statistically significant relationship was only observed between the independent variables of temperature, CPI and unemployment, and weekly sales (predictor variable). However, this is not definite because of the limitation of training data.

References

- Bakshi, C. (2020). Random forest regression. https : / / levelup . gitconnected . com / random-forest-regression-209c0f354c84
- Bari, A., Chaouchi, M., & Jung, T. (n.d.). How to utilize linear regressions in predictive analytics. https://www.dummies.com/programming/big-data/data-science/ how-to-utilize-linear-regressions-in-predictive-analytics/
- Baum, D. (2011). How higher gas prices affect consumer behaviour. https : / / www . sciencedaily.com/releases/2011/05/110512132426.htm
- Brownlee, J. (2016). Feature importance and feature selection with xgboost in python. https : / / machinelearningmastery . com / feature - importance - and - feature - selection-with-xgboost-in-python/
- Crown, M. (2016). Weekly sales forecasts using non-seasonal arima models. http : / / mxcrown.com/walmart-sales-forecasting/