MAT 430 – 15819 – M01

Seminar in Applied Mathematics

Final Project

Sales Forecasting Model for Walmart

Ryan Hadginikitas

August 3, 2024

**Introduction:**

Data, in recent years, has been a new driving force in business for improving sales to contend with competition in similar markets. Data driven businesses are now twenty-three times more likely to acquire customers, six times as likely to retain those customers, and nineteen times more likely to be profitable. In fact, eighty three percent of business leaders have pursued big data projects to seize a competitive edge in the market (*10 Stats to Show You How Analytics Boost Businesses*, 2020). Big chain department stores are facing harsh competition with one another, and big data can be used in sales predictions to gain a competitive edge on their rivals. The top 3 department stores only vary in popularity by 3% (*The Most Popular Department Stores in America | Consumer | YouGov Ratings*, n.d.). Sales forecasting or sales prediction in business uses a mixture of data analysis, machine learning, and statistical modeling on historical data to make decisions about future events (Tuovila, 2020). The top three department stores in the U.S. as of fiscal second quarter of 2024 were Costco, Target, and Walmart. I began searching for sales data on the top three department stores to begin my statistical analysis for sales forecasting.

During my research, I was able to find readily available data for Walmart, a dataset consisting of 45 different stores across the country ranging from 2010 to 2013 including weekly sales for each store (Walmart Sales Forecast, n.d.). Walmart generates almost 500 billion dollars annually making it the world’s largest company by revenue (Wikipedia, 2019). Walmart takes an enormous amount of data on consumers and uses this to predict sales patterns and develop marketing strategies for promotions. The main goal of this research project is to find the best statistical regression model to predict future sales for Walmart stores using the historical data spanning three years and predict future sales using the chosen model. The Walmart dataset includes variables such as: store number, department, if the week of the year is a holiday, temperature in the region, fuel price in the region, consumer price index, unemployment in the store’s region, size of the store, date of sales, and whether there was a sale during that week.

**Model Selection:**

For this data set, I will be using time series analysis which will analyze weekly sales of approximately 8,191 dates from 2010 to 2013. Time series forecasting can be used to predict future values given a fixed time interval of historical data (CATAL et al., 2019). The foundational technique used in time series analysis is the Linear Regression Model. Regression models are used to model continuous relationships between a dependent variable and independent variables using a linear model in which the mean of the dependent variable will change for the given values of the independent variables. A multiple linear regression model takes the assumption that the relationship between the dependent and independent variables is linear, the independent variables are not too highly correlated, and the variance of the residuals is constant. The variance of residuals in this model are the differences between the observed values and the predicted values of the response variable, which is also referred to as the "goodness of fit” of the model. By using the linear regression model, I will be able to identify the relationships between the dependent variable of weekly sales and the independent variables. I will use the results of the multi variable linear regression model to detect non-linearity and outliers, as well as the model’s homoscedasticity, or if the model will be dependable at predicting weekly sales.

From the multi variable linear regression model the best predictor variables can be chosen to tune the model. I will then rerun the model to gain the highest possible coefficient of determination, or goodness of fit for the linear regression model. Once I have tuned the linear regression model to the best possible coefficient of determination, I will cross reference with a possibly better modeling technique for sales forecasting called a random forest regression model. A random forest model will take a designated amount of decision trees and takes the best fit tree to use as the model for predicting sales. A decision tree is a flowchart-like structure that consists of nodes representing tests on a variable, and they branch the outcomes of the test which then lead to the final prediction called the leaf node. The random forest model is an excellent tool that reduces outliers in our data and has higher accuracy at predicting outcomes. I will evaluate the model performance on the Walmart dataset to see if this modeling technique will be better suited than the linear regression model at predicting weekly sales for the Walmart stores.

**Hypothesis:**

To find the best possible model to predict weekly sales, I must first find my hypothesis to test. For my research, the null hypothesis (H0) is that the store number, department, holiday week, temperature, fuel price, consumer price index, unemployment, size, date of sale, and discount does not have a significant impact on weekly sales for Walmart Stores. My alternative hypothesis (H1) is that the store number, department, holiday week, temperature, fuel price, consumer price index, unemployment, size, date of sale, and discount have a significant impact on weekly sales. To test this hypothesis, we assume that the data extracted from the Walmart stores is complete and accurate. There are several null values in my dataset, as well as some columns that were irrelevant and have been removed for the purpose of my hypothesis testing. Given the timeframe of the course, the sample size of 45 different stores is suitable for the model analysis I will be conducting but is a limitation on predicting weekly sales for Walmart with the highest possible accuracy. If there was a larger dataset including all stores, it would take a significantly longer time to run and test the models. The sample of 45 stores will still provide enough data to make a reasonable assumption for predicting Walmart weekly sales since we have a time span of three years’ worth of data.

When gathering the dataset on the sales for the Walmart stores, there were four different files that needed to be merged. The features dataset contained columns of data on Store, Date, Temperature, Fuel Price, Markdown 1 through Markdown 5, Consumer Price Index, Unemployment, and if there was a Holiday. The train dataset contained columns of data for Store, Department, Weekly Sales numbers, and if there was a Holiday. The test dataset contained columns of data for Store, Department, and whether there was a Holiday. The stores dataset contained columns of data for Store, Type, and Size. I used Python programming to import each of these csv files and merge them into one larger csv file. The Holiday data was listed in this dataset as either True or False within the dataset column. The linear regression model and the random forest model cannot use what is called a ‘string’ format for the analysis, but we can still recreate a true or false statement with 0 and 1. I assigned 0 to be false and 1 to be true and converted the column to those integers. In python I created what is known as a ‘Boolean’ and defined the 0 and 1 in the column to be a qualitative variable instead of a quantitative variable. I noticed when looking at the dataset that the Markdown columns had some missing values listed as NaN. I knew my model would not run appropriately with these NaN values in the rows of data and must be converted or deleted. Deleting the NaN variables would create a problem when running my models as there would be missing data if I chose to use Markdowns as a predictor variable in my model. I chose then to replace the NaN values with the integer 0. The reason I chose to replace the missing values with 0 lies in my decision to add the five Markdown columns together into one Markdown column for the dataset. Separating the Markdown type for this analysis would not enhance the accuracy of predicting the sales for this dataset.

For the dataset I also expanded the date column into a separate Year, Month, Day, Quarter, and Week of the Year. This will allow me to separate the significance of the predictor variables for each aspect of the date during the sales. Separating the dates can also allow a visualization of weekly sales by week of the year, by the month of the year, day of the year, and sales for each year in the dataset. By first visualizing the data we can start to make assumptions about the effect of the date on the weekly sales.

Once the data set has been properly cleaned, and is ready for use, I performed correlation testing to determine if there is a correlation between the predictor variables and the response variable of Weekly Sales. A correlation matrix creates a table that tests the level of significance between two variables in the dataset. For this dataset the Pearson’s Correlation Coefficient was used to develop the correlation matrix. The Pearson Correlation Coefficient is calculated using the formula where is the correlation coefficient, is the value of the x variable, is the mean of the values of the x variable, is the value of the y variable, and is the mean of the values of the y variable. Calculating the correlation coefficient using the above formula, we will get coefficient between the range of -1 to 1. The closer the correlation to positive 1 the higher the relationship between the two variables. Once the correlation coefficients are determined we can visualize them in the heatmap that is shown below using the seaborn package in python.

A colorful squares with white and black text

Description automatically generated with medium confidence

Looking at the correlation matrix we can see that there isn’t a strong relationship between any predictor variable and weekly sales. The strongest correlation with weekly sales seems to be size at 0.24 and next is Department at 0.15. We can see that there is a negative correlation between Weekly Sales and Temperature, Fuel Price, Store, Consumer Price Index, Unemployment, Year, and Day telling us there is no correlation between these predictor variables on the response variable of Weekly Sales. It would be a safe assumption to create our original model with variables that have a positive correlation with Weekly Sales and tune the model from there.

For our original model we will use the multiple linear regression equation, where is the weekly sales of Walmart stores, through are the predictor variables of Size, Department, Holiday, Week of the Year, and Mark Down. is the y-intercept which remains constant, and through are the slope coefficients for each predictor variable. The in our equation represents the random variation of the individual data for every value of about the mean.

Using the programming language R, a multiple linear regression was run using the predictor variables with the highest correlation number to our independent variable of Weekly Sales. Again these dependent variables were Size, Department, Holiday, Week of the Year, and Markdown in price. After running the model, we can summarize our findings using the ANOVA table, or rather Analysis of Variance which is attached below.

A screenshot of a computer program

Description automatically generated

Using this ANOVA table from R, we can see that the model produced a coefficient of multiple determination of 0.08275. The multiple R-squared is the percentage variance in Weekly Sales when explained by the predictor variables in the model. It is calculated by subtracting 1 from the division of the sum of squares error by the total sum of squares or . SSE is the measure of discrepancy between the data and the linear regression model. SST is the sum of all squared differences between the observed predictor variables and the overall mean. The multiple R-squared tells us if the predictor variables used in this model were a good fit at predicting Weekly Sales. The closer our multiple R-squared is to 1, the better the model fits. Since we have and extremely low multiple R-squared this tells us that this model is not a good fit for predicting Weekly Sales. To further visualize that this model does not accurately predict Weekly Sales, I have plotted the residuals verse fitted values for the model as well as a QQ plot. Residuals in regression are a measure of the error between a predicted value and the actual value. The fitted values are the predicted values of Weekly Sales for the predictor variables in the model. The residual vs. fitted values plot below is used to see if our model has a linear relationship, if there is homoscedasticity, and if there are any outliers in our model.

A graph of red dots

Description automatically generated

From this plot we can see there is general linearity in our model, but there are a large number of outliers, and the error is not constant across the dependent variables as the plots do not seem to be randomly scattered and develop a generic pattern. As well as interpreting the residuals vs fitted plot, I have produced a QQ plot below,

A graph of a line graph

Description automatically generated with medium confidence

This quantile-quantile plot shows that our data points are heavily tailing upward from the reference distribution showing a positive skew. This tells us that there is some deviation, and the variables are not normally distributed. This further proves that our linear regression model is not a good fit for predicting Weekly Sales.

To confirm my theory that multiple linear regression is not the right model to predict Weekly Sales for the dataset of Walmart stores, I performed a variable selection technique known as backwards selection. I will include all the predictor variables from my dataset first and use them in my linear regression model to find how well the model will predict Weekly Sales. From this model I will evaluate the p-value, or the probability under the assumption of no effect, to tell if the predictor variables are statistically significant to the model. For my backwards selection, I will compare the p-values to a significance level of 0.05, which will tell us there is a 5% level of uncertainty of the p-value being statistically significant to the model. The p-values of the predictor variables that are less than 0.05 will be kept in our model, and we will remove one at a time the highest p-value from the previous model until all our predictors in the model have a p-value below the level of significance of 0.05. After generating multiple linear regression models, I was left with a final model using predictor variables of Store, Department, Consumer Price Index, Size, Year, Week of Year, and Markdown. From the ANOVA table attached we can see that the multiple R-squared is still very low at 0.08577, further proving my theory that multiple linear regression was not a good model fit for predicting Weekly Sales using time series analysis.

A screenshot of a computer

Description automatically generated

After researching some of the best time series analysis models, I decided to try a random forest regression model for sales predictions. A random forest model creates multiple decision trees, often established by the user, and each one is built using randomly chosen subsets of data. The model then takes the outputs of all these decision trees and makes an overall prediction for any unseen data points. This technique in decision tree modeling helps to reduce variance and improve accuracy in the prediction. Random forest modeling works well with larger datasets and can model nonlinear relationships between the dependent and independent variables. The random forest model tends to not overfit data since it is running so many random trees independently and averages them out. The other advantage of the random forest model is that can handle missing values in the dataset and handle qualitative variables well.

I began with the same predictor variables from our original linear regression model which were chosen using our correlation matrix. With these variables selected, I then split my data into training and testing sets using a 70% and 30% split respectively. The training set will be run and compared to my test set so we can compare how well the model performed at predicting weekly sales. If our training model has performed close to our test set, we can verify that this model is a good fit. I ran the random forest model through Python programming language, using 200 decision trees on both my training set and my test set. After my models were created, I ran metrics of Out of Bag Score, the mean squared error, and the R-squared. Out of Bag Score is a way of validating the random forest model. It is the rows of data that are not used for the training set and have been left out. Once the train model has been run, the leftover rows will be given to the model as unseen data. This model will take the unseen data and see how well it can accurately predict the outcome and produce a score accordingly. The Out of Bag Score can be used to represent the accuracy of our model. Our training model produced an OOB score of 0.97327 which tells us that the model correctly predicts Weekly Sales 97% of the time. The R-squared value obtained from the Random Forest Model will show us whether our model is a good fit for predicting Weekly Sales. The Random Forest model on the training set produced an R-squared value of 0.99624 which is a drastic improvement from the multiple linear regression model we ran earlier. This tells us the random forest model is a good fit for predicting Weekly Sales. The last analysis I conducted was the Mean squared error or MSE. This metric is used to measure the average squared difference between the predicted and the actual values in the dataset. It measures how close the regression line is to the actual data points in the model. Our MSE value for this model was 2163428.0447 which is extremely high. Normally for such a large dataset, there would be more than 200 trees run for the random forest model, and as we increase the number of trees to run the MSE gradually decreases. The number of trees that would be needed to lower the MSE to a reasonable number would take excess memory than my computer would be able to handle and would take quite a long time to run. For the sake of this analysis, we are assuming that if we were to run thousands of trees the MSE would drop to a reasonable number. When compared to our random forest test set, which had an OOB score of 0.96388, and R-squared of 0.99509, and a MSE of 274199.819 we can see that our training set is the best model to use to predict Weekly Sales given the highest correlated predictor variables to Weekly Sales.

I further compared the Random Forest Model that I created to the Linear Regression Model to run a feature importance on the predictor variables on predicting Weekly Sales. This plot took the attributes Gini Index, a measure of how impure a dataset is. It measures the probability of a random instance being misclassified when chosen randomly. When running a plot of the predictor variables for the Random Forest we can see below that each predictor variable has a positive importance.

A graph with different colored squares

Description automatically generated

When running the same feature importance visualization on the linear regression model, we can see that not all features are positive leading to further assumptions that the linear regression model was not suitable for our goal of predicting Weekly Sales using time series analysis.

A graph with a bar and text

Description automatically generated with medium confidence

**Limitations**

One of the biggest limitations of this research was the small amount of data available for both the number of stores, as well as the span of time the data was recorded. We could have had higher accuracy predicting weekly sales if we had the data on every Walmart store in the country as well as a greater span of time. When looking at variables that might have a positive correlation with weekly sales, it is difficult to find a relationship between economic factors such as consumer price index and unemployment. These economic factors can have varying degrees of effect on different Walmart sales given geographic region.

**Conclusion**

The goal of this research project was to find the best mode to predict the Weekly Sales of the Walmart stores based on past records including data such as store number, department, past weekly sales, whether the week fell on a holiday, temperature in the stores region, fuel price in the stores region, the consumer price index in the region, the size of the store, and whether there was a sale. Once the optimal model has been chosen, we can then predict the following two years of weekly sales for the Walmart stores. Relationships between these variables on weekly sales were identified using the correlation matrix and implemented through a multiple linear regression model. From this model, the summary of analysis was obtained to determine that the linear regression model created was not a suitable fit for the prediction of sales. Modifying the multiple linear regression model through backwards selection, we also determined that the final regression model was also not a good fit for the goal of predicting weekly sales. Using the same predictors as the original linear regression model, a random forest model was created, and yielded a high model fit and accuracy at determining the weekly sales given the predictors from our dataset. We ran a feature importance on the variables using the Gini Index and found that all the variables have a positive effect on the weekly sales prediction, with the department having the greatest performance importance on the model, followed by the size of the store. It was further proved that the linear regression model was not a good fit as the department and size have a negative impact on the prediction of weekly sales. We can also conclude that there is a nonlinear relationship between the target variable of weekly sales and the predictor variables chosen in the random forest model. Finally, a visualization of the weekly sales for the two years following the dataset can be found below using the random forest model.

A blue line graph with numbers

Description automatically generated