**CIND820 Capstone Project: Walmart Sales Forecasting**

**Initial Results and Code**

**GitHub Link: https://github.com/Hasib147/CIND820-Project**

**By: Hasib Chowdhury (500439135)**

**Supervisor: Dr. Uzair Ahmad**

**Date of Submission: June 27, 2022**

**Exploratory Data Analyses:**

For this Walmart Sales forecasting model, the data is consistent with the variables, the following table shows a summary of the values in the original dataset:

A close-up of a document

Description automatically generated with medium confidence

Out of the 23 variables in the dataset, about 8 of them are categorical and the rest are numerical including the target variable (weekly sales) which we will use to predict the sales of Walmart for that week. This dataset as mentioned earlier in the literature review & data description have over 420,000+ rows and 23 columns over the 3 year time frame (2010 to 2012).

The following is the summary of the normalized dataset (normalized only numerical values):

A close-up of a document

Description automatically generated

**Data Preparation:**

Chart, bubble chart

Description automatically generated

The above graph represents the corrplot of the different types of variables in the dataset, the temperature and isHoliday along with Fuel\_price and isHoliday are weak correlations whereas the other variables have moderate to strong correlations (depending on the size of the circle).

Out of the 23 variables in the dataset, the last 3 categorical variables (day, week, and year) were not taken into account when normalizing the dataset and are also variables that are not as important as the other ones such as Temperature, Fuel Price, or the CPI since they mostly affect the weekly sales.

Chart, scatter chart

Description automatically generated

The above graph shows the scatterplot of the Fuel Price and Temperature, most of the values being between 20-80 for the temperature and 2.5-4.0 for the fuel price. There are a few outliers, one of them being (0, 3.0) and some being near the 100 degrees temperature point.

**Modelling:**

The model that I have decided to use for this dataset is the stepwise regression model, which I think will be the right fit for this project. There are 2 approaches to this model, one is the forward selection and the other is the backward elimination algorithm. The forward selection approach starts with no variables and adds each new independent variable one by one depending on the values of the AIC (goes from large values to small values depending on the model’s input variables) testing for statistical significance. The backward elimination method begins with a full model loaded with multiple independent variables and then removes one independent variable at a time to test its statistical significance to overall results depending on the value of the AIC (goes from small to large, opposite to the forward selection).

**Screenshot of the original dataset model:**

Forward Selection:

Text, table

Description automatically generated with medium confidence

Backward Elimination:

Table

Description automatically generated

**Screenshot of the normalized dataset model:**

Forward Selection:

Text

Description automatically generated

Backward Elimination:

Text, table

Description automatically generated

**Evaluation:**

The stepwise regression model for the Walmart dataset is the following:

For forward selection: (original dataset – CIND820\_clean\_data):

Weekly\_Sales = 0.08945\*Size + 0.1275\*MarkDown3 - 19.66\*CPI + 30.57\*Temperature -240.1\*Unemployment + 0.07627\*MarkDown5 - 748.6\*Fuel\_Price + 0.02755\*MarkDown4 + 0.01738\*MarkDown2 + 9526

For Backward Elimination: (original dataset – CIND820\_clean\_data):

Weekly\_Sales = 30.57\*Temperature - 748.6\* Fuel\_Price - 19.66\*CPI -240.1\*Unemployment 0.08945\*Size + 0.01738\*MarkDown2 + 0.1275\*MarkDown3 + 0.02755\*MarkDown4 + 0.07627\*MarkDown5 + 9526

For forward selection: (normalized dataset – CIND820\_clean\_data2):

Weekly\_Sales = 0.08945\*Size + 0.1275\*MarkDown3 - 19.66\*CPI + 30.57\*Temperature -240.1\*Unemployment + 0.07627\*MarkDown5 - 748.6\*Fuel\_Price + 0.02755\*MarkDown4 + 0.01738\*MarkDown2 + 0.388

For Backward Elimination: (normalized dataset – CIND820\_clean\_data2):

Weekly\_Sales = 30.57\*Temperature - 748.6\* Fuel\_Price - 19.66\*CPI -240.1\*Unemployment 0.08945\*Size + 0.01738\*MarkDown2 + 0.1275\*MarkDown3 + 0.02755\*MarkDown4 + 0.07627\*MarkDown5 + 0.3888

Markdown1 is redundant (not considered in the model) and markdowns 2-5 are more relevant and observed in the intercept as well as the whole dataset. Size is the most important variable due to the values and might have more departments in specific locations to attract more clients and also the products/services that location may have.

**K-Fold Cross Validation:**

**Cross-validation for the original dataset:**

Graphical user interface, text, email

Description automatically generated

From the R-squared value above, this tells us the error value is low and also 378,000 rows out of the 420,000 were taken into account when doing the 10-fold cross validation (9:1 train-test split)

**Cross-validation for the normalized dataset:**

Text

Description automatically generated

The cross-validation for the normalized dataset used the same number of sample sizes compared to the original and what was observed was that the R-squared value (measure of the correlation between the predictions made by the model and the actual observations) remained the same, however the RMSE (root mean square error) has dropped in value which tells us that the model can be more closely related to the actual observation (this was done through trial & error using k-fold and the accuracy of the result may be off from the predicted values)

References

*K-Fold Cross Validation in R (Step-by-Step).* Statology (November 4, 2020). Retrieved June 27, 2022 from <https://www.statology.org/k-fold-cross-validation-in-r/> .