WalMart Sales Forecasting: Using Regression Model Predictions to Gain Business Insights

By: Ed Gatdula 10/21/21

SUMMARY

The Capstone Project objective is an understanding of Walmart stores and weekly sales relationships. This involves deducing and quantifying sales trends and behaviours by using regression models. Several ensemble and linear regression models were built using the Walmart Weekly Sales Forecasting dataset from Kaggle.com. The most important features of the XGBoost regression model provides information about which Walmart store characteristics have the largest influence on weekly sales predictions. Departments 92, 95, 38, 90, and 72 along with Store Size have the largest influence on the weekly sales predictions. The models were then compared for how accurate the predictions were. HistGradientBoost model had the best metrics with highest coefficient of determination, R2 = 0.70 and lowest mean average error, MAE = 6500.

Improving the model predictions was attempted using additional data, feature engineering, and residuals plotting. Additional consumer related data in the form of the Consumer Confidence Index data, did not improve regression models performance. Using GridSearchCV and feature engineering produced comparatively larger performance metric improvements for several regression models. Residuals scatterplots indicate the largest model errors are associated with the Christmas and Thanksgiving holidays and several stores and departments.

CONCLUSION

When considering future weekly sales and growth, based on the data provided and regression model results, the best strategy to achieve large weekly sales is to build the largest possible size stores, Type ‘C’, with departments 92, 95, 38, 90, 72. To achieve more accurate weekly sales predictions and hence a better handle on the monetary value of each store characteristic, example: Department, more feature engineering to reduce the model residuals involving Christmas and Thanksgiving is required. Further model improvement might be achieved by using surrounding area demographic info for each store.

INTRODUCTION

In 2014, Kaggle.com hosted the WalMart Sales Forecasting Recruiting Challenge. The competition consisted of building a model using a training set and then predicting weekly sales based on a test set. The training and test data sets were provided by Kaggle.com. No

additional data from outside sources was allowed. Scoring was based on a weight mean average error.

Problem Statement:

1. Exploratory Data Analysis and Regression Models produce useful business insights about the Walmart Stores Network
    a. Walmart Stores sales landscape description.
    b. What features have the largest effect on weekly sales? How can this information be used to improve business?
2. Can we improve weekly sales predictions using consumer confidence index data?
3. How much can logarithmic transform improve model performance? How does this improvement translate into revenue?

DATA DESCRIPTION

Description of columns and values from kaggle.com-

1. stores.csv

This file contains anonymized information about the 45 stores, indicating the type and size of store.

2. train.csv

This is the historical training data, which covers to 2010-02-05 to 2012-11-01. Within this file you will find the following fields:

Store - the store number
Dept - the department number
Date - the week
Weekly_Sales - sales for the given department in the given store
IsHoliday - whether the week is a special holiday week

3. test.csv

This file is identical to train.csv, except we have withheld the weekly sales. You must predict the sales for each triplet of store, department, and date in this file.

4. features.csv

This file contains additional data related to the store, department, and regional activity for the given dates. It contains the following fields:

Store - the store number
Date - the week
Temperature - average temperature in the region
Fuel_Price - cost of fuel in the region
MarkDown1-5 - anonymized data related to promotional markdowns that Walmart is running. MarkDown data is only available after Nov 2011, and is not available for all stores all the time. Any missing value is marked with an NA.
CPI - the consumer price index

Unemployment - the unemployment rate
IsHoliday - whether the week is a special holiday week
For convenience, the four holidays fall within the following weeks in the dataset (not all holidays are in the data):
Super Bowl: 12-Feb-10, 11-Feb-11, 10-Feb-12, 8-Feb-13
Labor Day: 10-Sep-10, 9-Sep-11, 7-Sep-12, 6-Sep-13
Thanksgiving: 26-Nov-10, 25-Nov-11, 23-Nov-12, 29-Nov-13
Christmas: 31-Dec-10, 30-Dec-11, 28-Dec-12, 27-Dec-13

5. Consumer Confidence Indicator

Consumer Confidence Data from:
https://data.oecd.org/leadind/consumer-confidence-index-cci.htm

Location - Country ID
Indicator - Economic metric ID
Frequency - time period of measurement reporting
Time - date
Value - value of economic indicator
Flag Codes - Code ID

DATA WRANGLING

Stores.csv Summary:

1. There are three store types. Appear to be sorted by Size (square footage). However, boxplots show one or two outliers in each type. Maybe outlier stores should be reclassified? To be answered in exploratory data analysis.
    A. Following stores identified as outliers in their respective Type:
    Type 'A': Stores 33, 36. Located outside lower bounds of Type 'B' distribution. Falls into Type 'C' size distribution bounds.
    Type 'B': Stores 5, 3. Located outside lower bounds of Type 'B' distribution. Falls into Type 'C' size distribution bounds.
    Type 'C': Store 30. Located outside upper bounds. However, not large enough to qualify for B, C Type.

2. Grouping the data by store type and computing median of each group
    A type store median size: 202406.0
    B type store median size: 114533.0
    C type store median size: 39910.0

Features.csv Data Wrangling Summary:

1. df_train shape: (8190, 11). 8,190 entries. 11 Features.
    A. Categorical: Store, IsHoliday

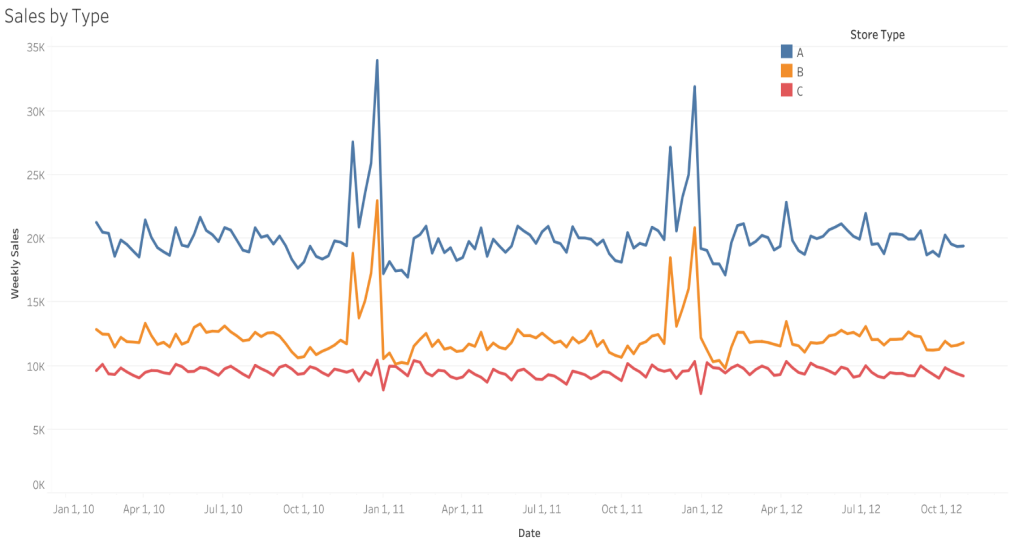
B. Numerical: Temperature, Fuel_Price, MarkDown1, MarkDown2, MarkDown3, MarkDown4, MarkDown5, CPI, Unemployment

2. MarkDown1, MarkDown2, MarkDown3, MarkDown4, MarkDown5 have a significant amount of NaN values. Boxplot and histogram plots of each MarkDown feature show large counts of outliers and non-normal distribution.

EXPLORATORY DATA ANALYSIS

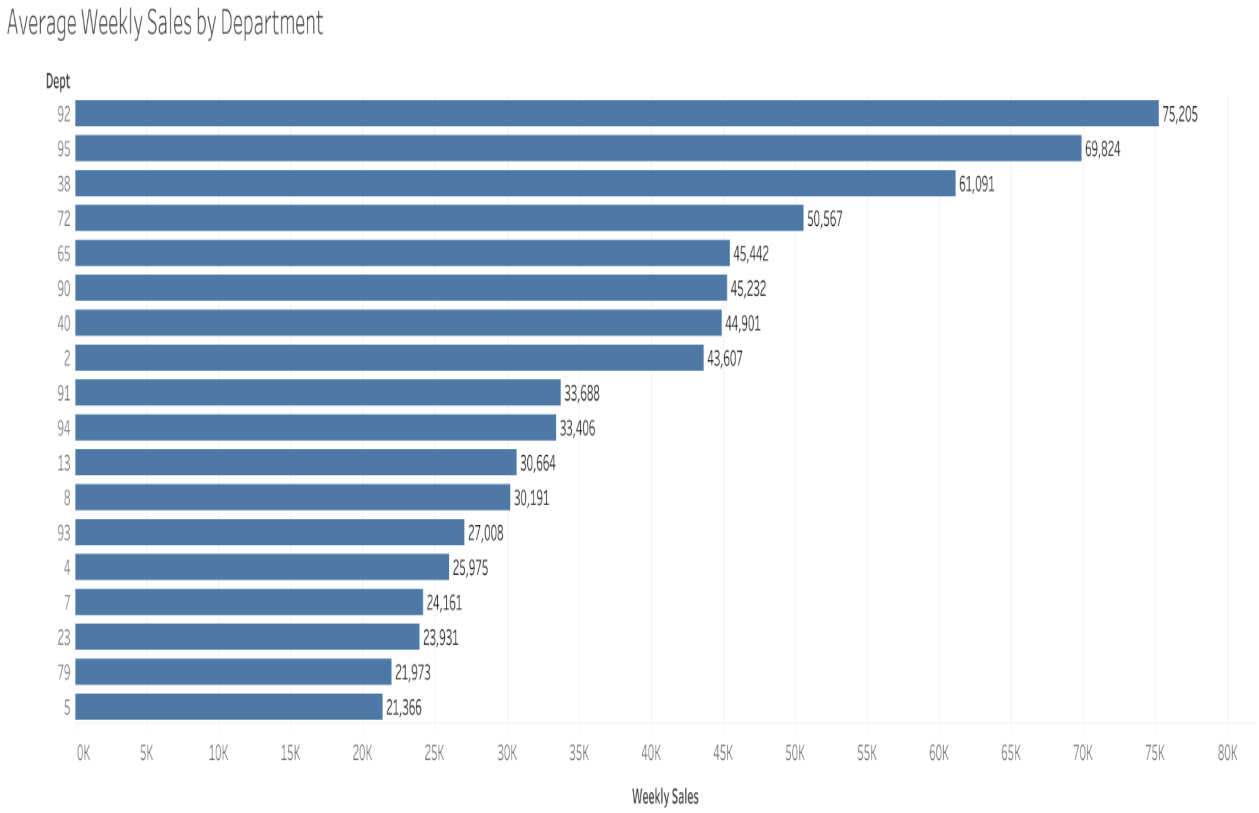
Each training data set feature was examined for errors, trends, and correlation between weekly sales. The analysis for each feature is summarized as below:

- TYPE



1. 'Type' feature is based on total weekly_sales amount. Types A, B, and C have distinct weekly_sales ranges.
2. There are 22 Type A Stores. There are 17 Type B Stores. There are 6 Type C Stores.
3. Stores evaluated for 'Type' assignment. Should following stores, 36, 33, 5, 3, 30 be reassigned to Class C? Based on Weekly Sales, yes.

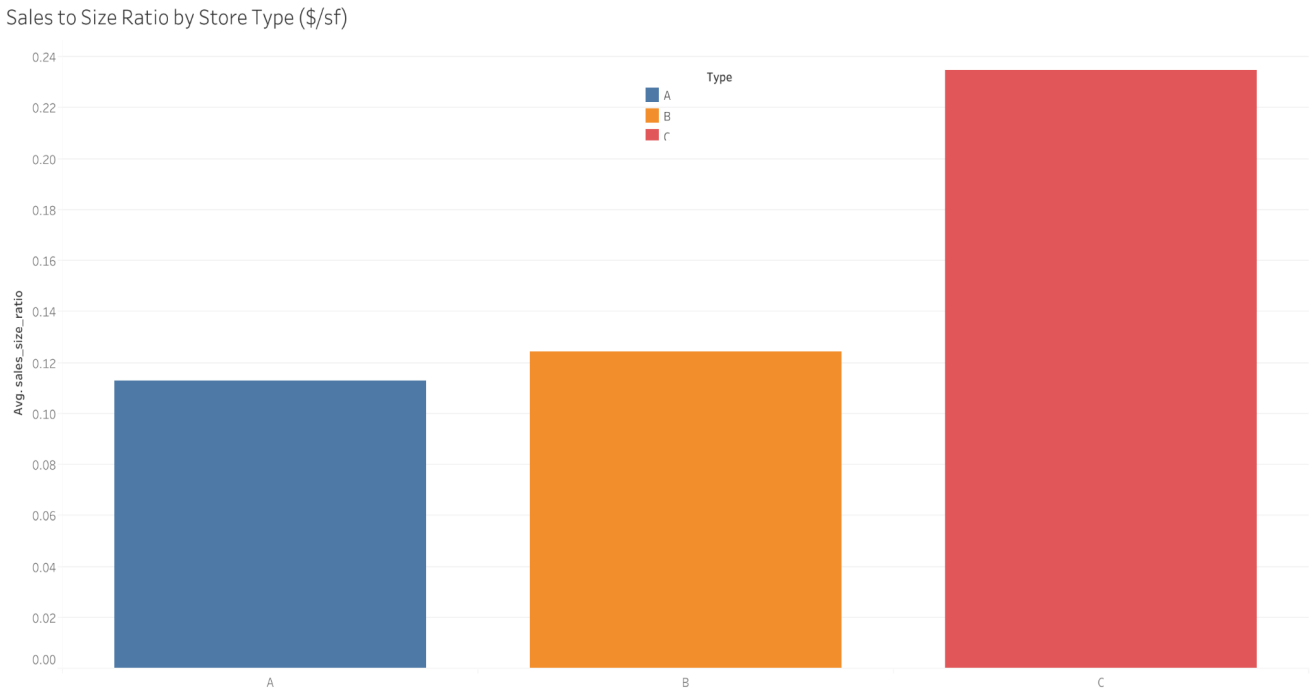
- DEPARTMENT



1. Cumulative sales grouped by Department over date range shows strong peaks.
2. Top Five Departments in Average Aggregated Weekly Sales:

Dept	Weekly_Sales
92	$75,205
95	$69,824
38	$61,091
72	$50,567
90	$45,442

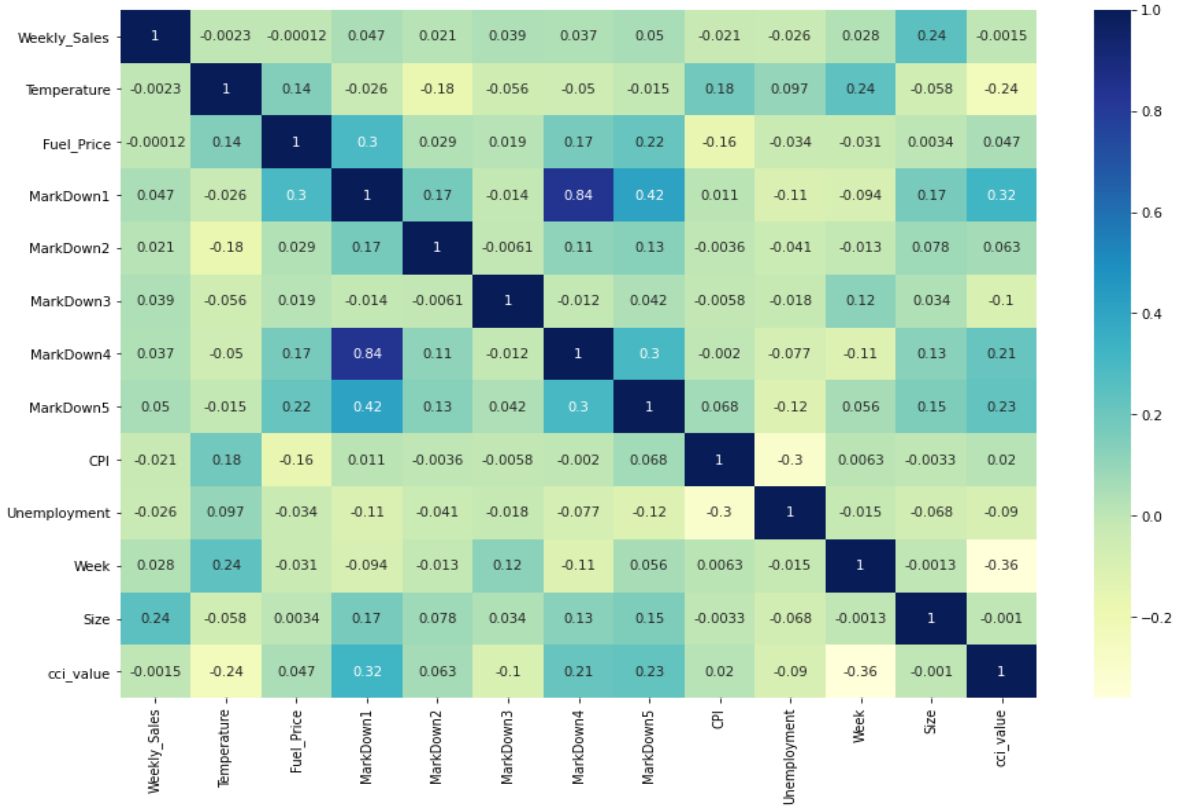
- ISHOLIDAY
    1. Several departments show considerable difference in 'Weekly_Sales' dependent on 'IsHoliday' Feature
- SIZE
    1. Larger Store size correlated with large weekly_sales? Yes, regression plot shows increased weekly_sales for larger store size.
    2. Larger Store size correlated with large weekly_sales? Yes, regression plot shows increased weekly_sales for larger store size.



Average weekly sales to size ratio is largest for Type ‘C’ stores. This means Type C stores generate the most revenue on average per unit area.

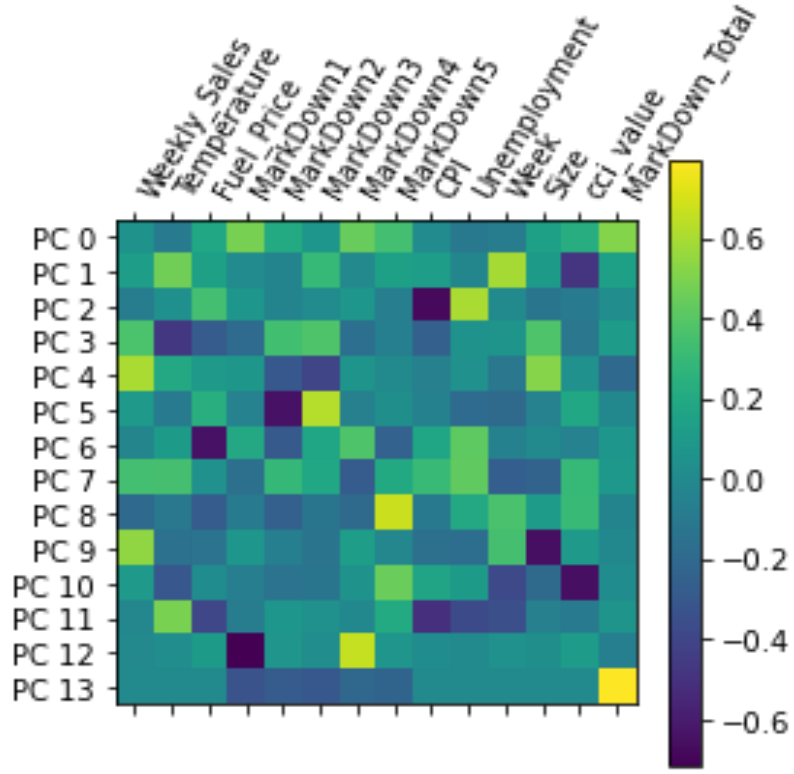
- CCI_VALUE
    3. For each Dept, a scatterplot of Weekly_Sales vs Date colored by cci_value does not indicate positive correlation.
- TEMPERATURE
    1. for each store, scatterplot of aggregated weekly_sales for each date vs. temperature. no positive correlations for any store.
- FUEL_PRICE
    1. Scatterplot of weekly_sales vs fuel prices does not indicate correlation between the two.
- MARKDOWNS
    1. Scatterplots of weekly_sales vs MarkDown1, MarkDown2, MarkDown3, MarkDown4, MarkDown5 do not indicate strong positive correlations
- CPI
    1. CPI steadily increases over date range. begins roughly 167 and ends slightly above 176.
    2. scatterplot of weekly_sales summed over all stores vs CPI values does not indicate correlation

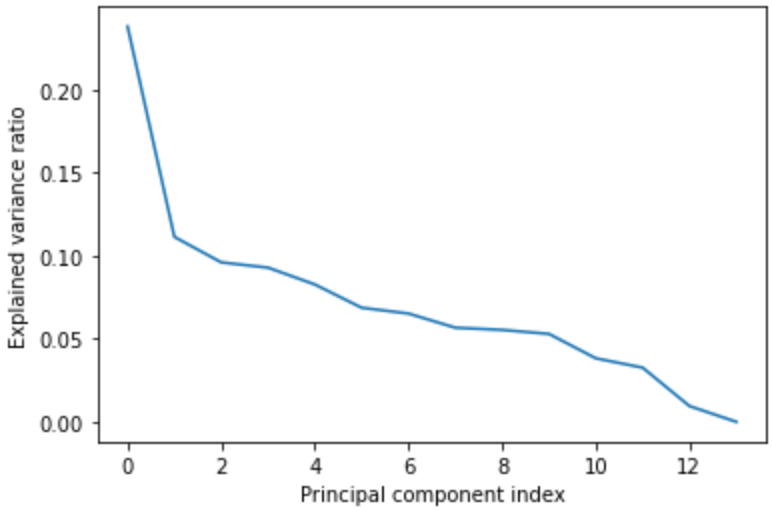
- UNEMPLOYMENT
    1. Scatterplots of weekly_sales vs unemployment rate does not indicate positive correlations.
- CORRELATION HEATMAP:
    1. 'Size' Feature has largest correlation value, 0.24, with Weekly_Sales. This is not indicative of a strong correlation.
    2. All other numerical features correlation values fall in range from -0.0034 to 0.047.
    3. Correlation Heatmap reinforces deductions regarding correlations between weekly_sales and features from scatterplot visualizations.



PRINCIPAL COMPONENT ANALYSIS (PCA)

1. First principal component is responsible for approx. 20% of variance.
2. Second principal component is responsible for approx. 11% of variance.
3. Remaining explained variance is spread relatively evenly over remaining principal components.


Weekly_Sales
Temperature
Fuel_Price
MarkDown1
MarkDown2
MarkDown3
MarkDown4
MarkDown5
CPI
Unemployment
Week
Size
cci_value
MarkDown_Total
PC 0
PC 1
PC 2
PC 3
PC 4
PC 5
PC 6
PC 7
PC 8
PC 9
PC 10
PC 11
PC 12
PC 13
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6


Explained variance ratio
0.20
0.15
0.10
0.05
0.00
0
2
4
6
8
10
12
Principal component index

Experiment Procedure:

Part I.

Training Data

- Base Regression Models
- No consumer confidence index
- Measure cross-validation performance metrics (R2, MSE, MAE)

Part II.

Training Data

- Base Regression Models
- Add consumer confidence index
- Measure cross-validation performance metrics (R2, MSE, MAE)

Experiment Procedure:

Part III:

Training Data

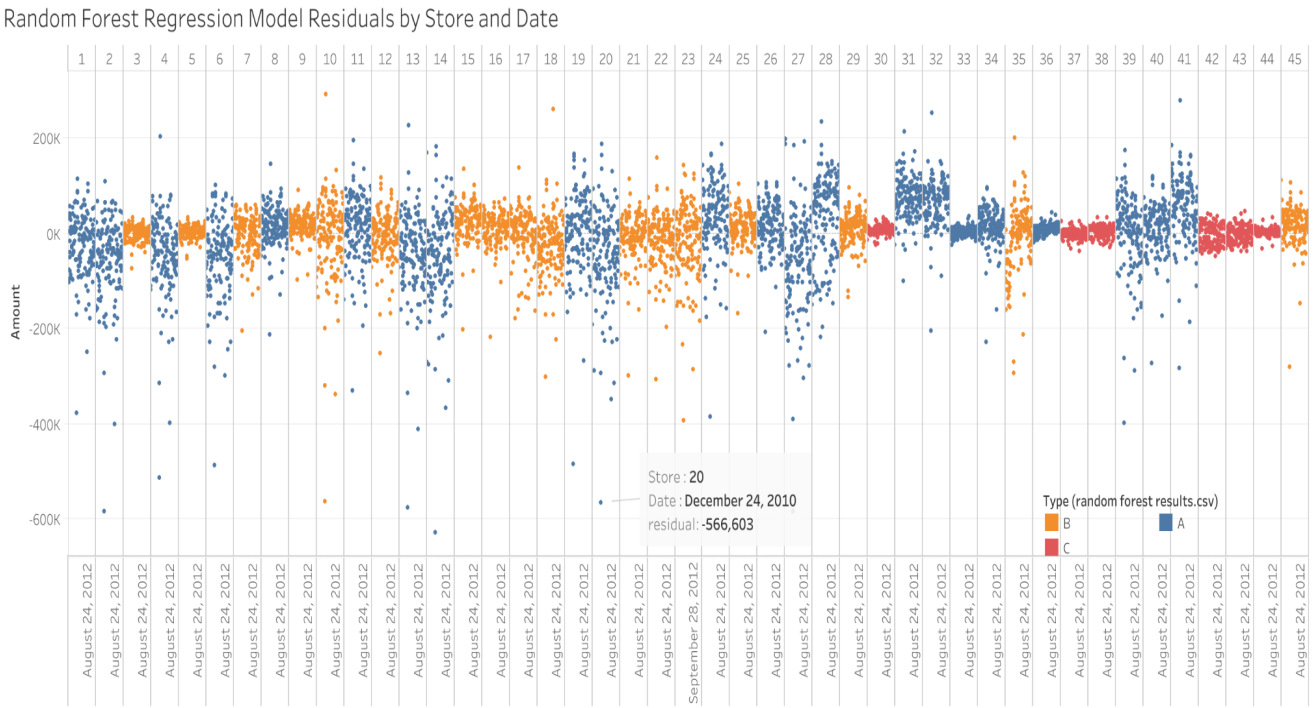
- Hyperparameter Tuning Using GridSearchCV on Regression Models
- No consumer confidence index
- Measure cross-validation performance metrics (R2, MSE, MAE)

Part IV:

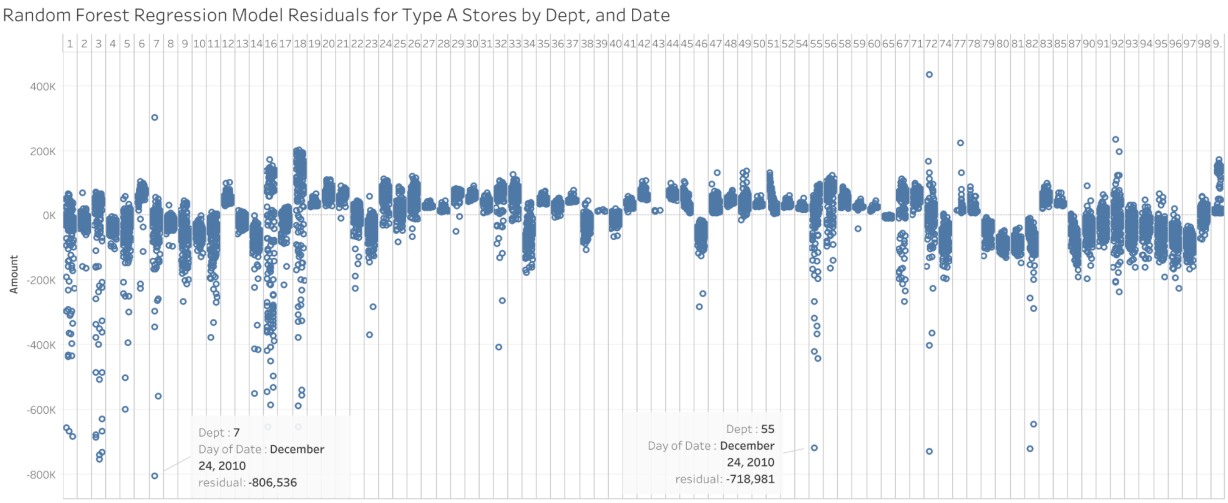
Training Data

- Apply logarithmic function to MarkDown features
- Fit modified training data on ElasticNet and HistGradientBoost models
- No consumer confidence index
- Measure cross-validation performance metrics (R2, MSE, MAE)

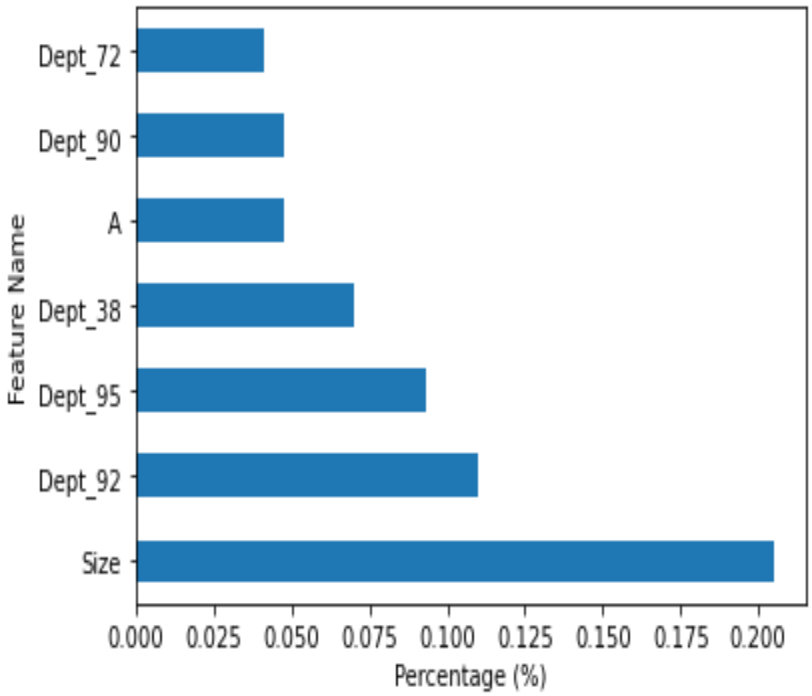
RESULTS



A scatterplot of the random forest model residual indicates that the largest weekly sales prediction errors are associated with Type ‘A’ stores.

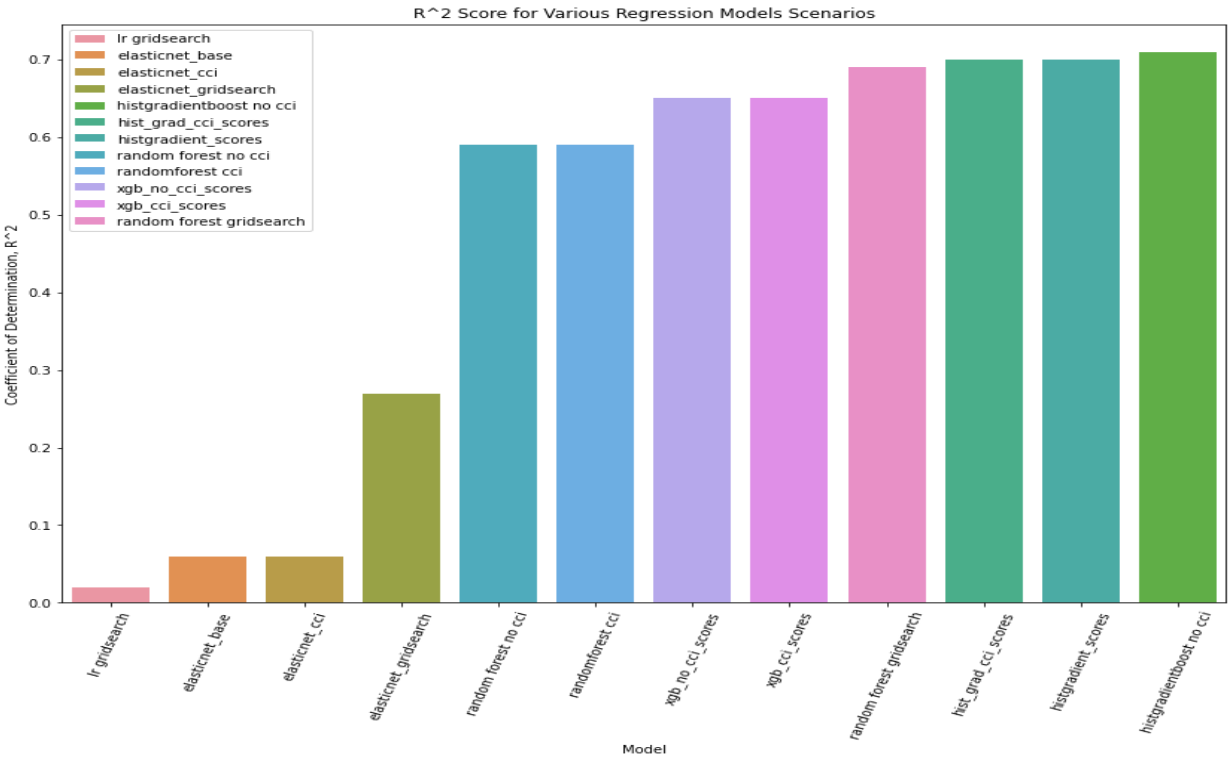


When broken down by Department, a scatterplot of Type ‘A’ store random forest residuals shows that Departments 1, 3, 7, 55, 72, and 82 have residual outliers associated with Christmas and Thanksgiving holidays.

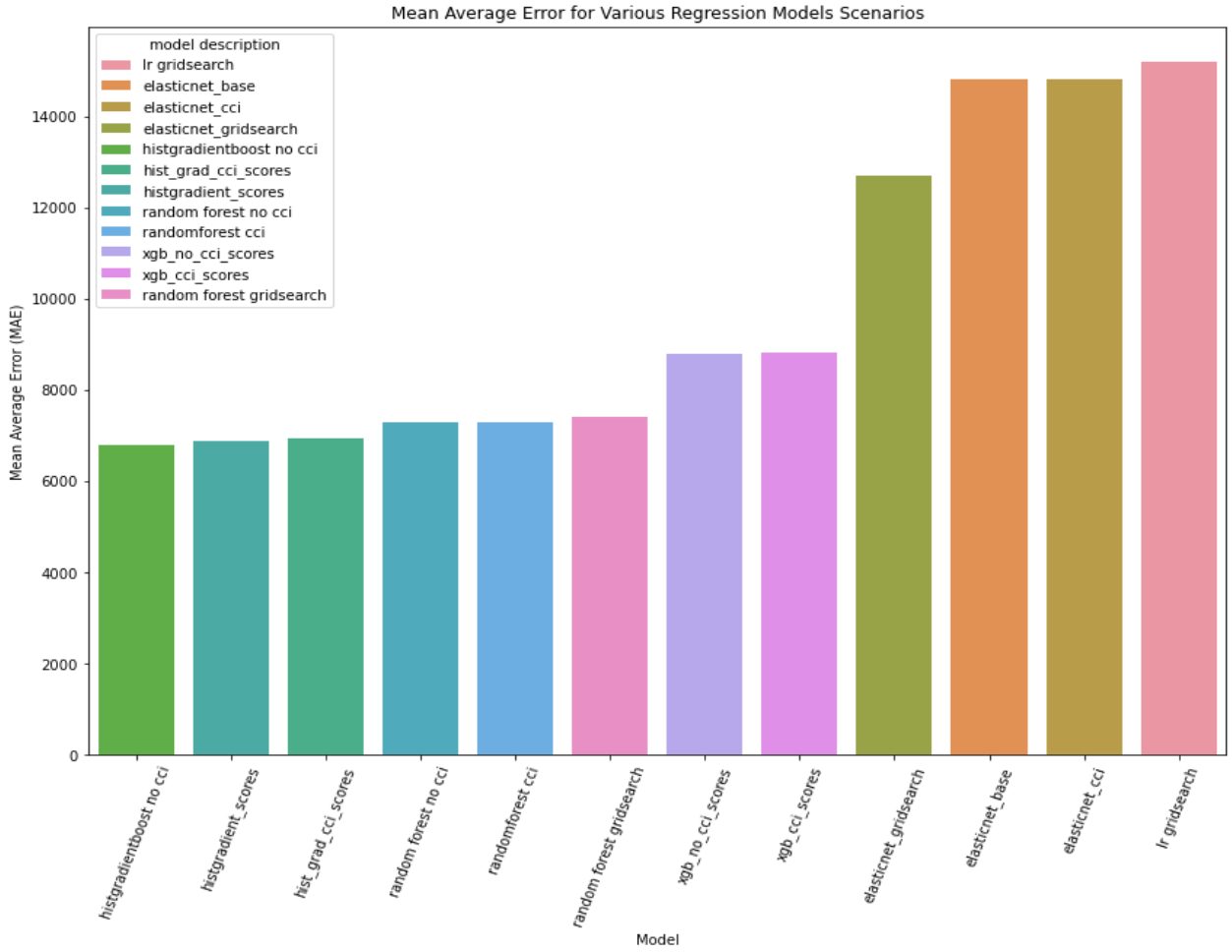


The XGBoost Regressor indicates that the Size characteristic has the largest, ~0.210%, influence on weekly sales predictions followed by Departments 92, 95, 38, 90, and 72.

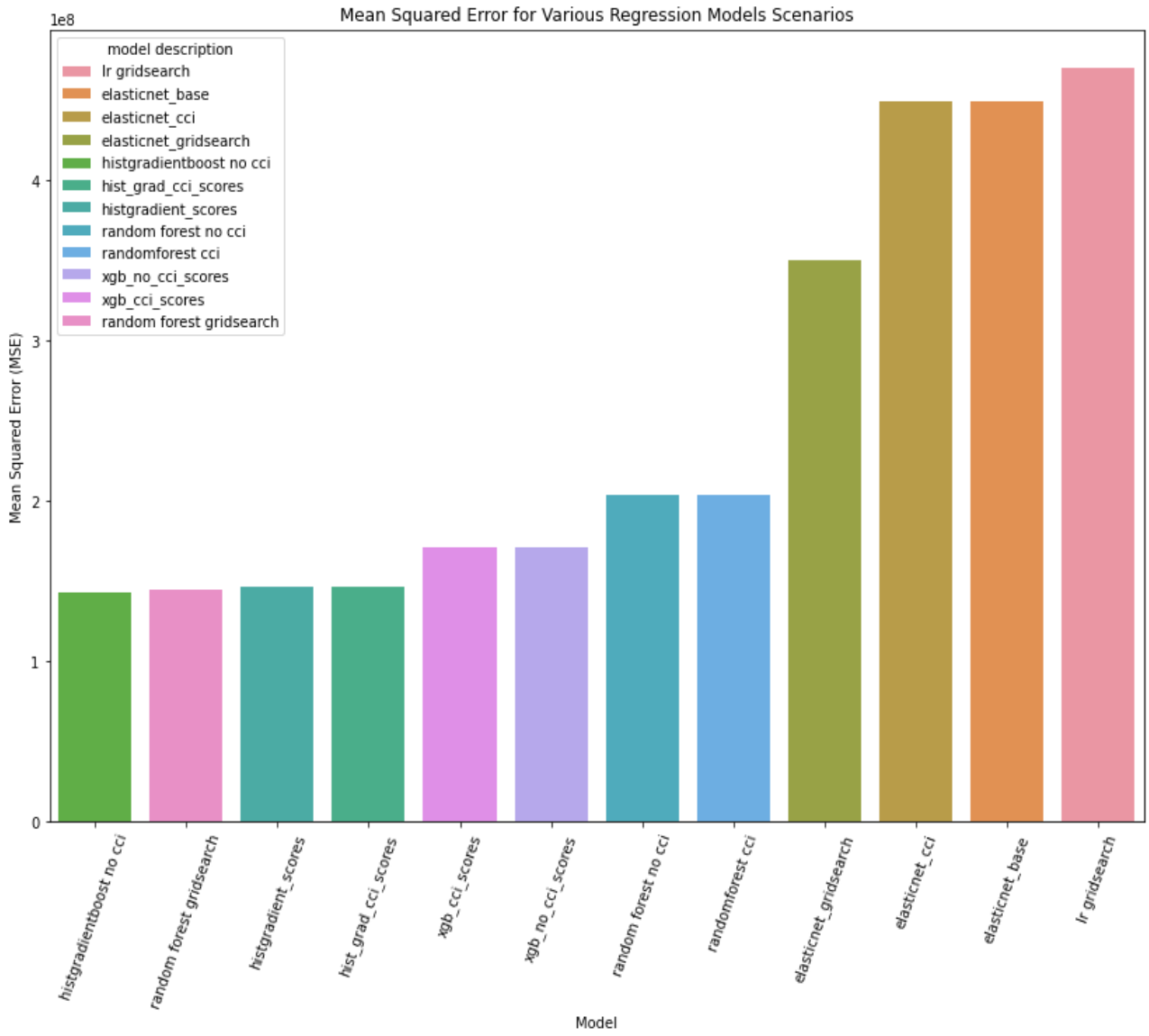
DISCUSSION



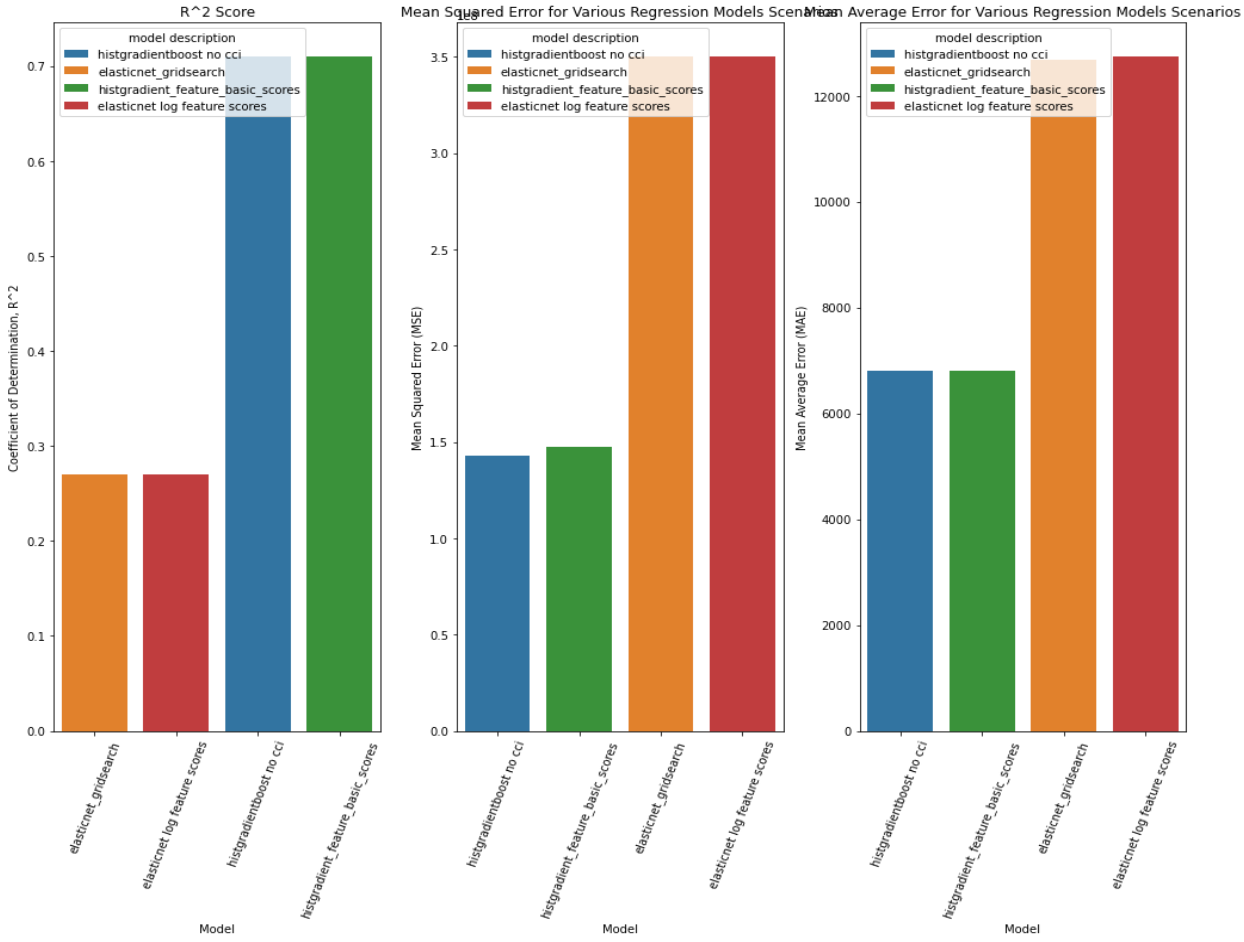
HistGradientBoost regressor produces best coefficient of determination, R2, score. GridSearchCV tuning and consumer confidence index data slightly decrease HistGradientBoost R2 score. GridSearchCV produces the best R2 improvement for RandomForest regressor and ElasticNet models.



Consumer confidence data does not decrease Mean Average Error (MAE). MAE are lowest for the HistGradientBoost regressor. GridSearchCV and Consumer Confidence Index data do not decrease the HistGradientBoost regressor MAE. GridSearchCV decreases ElasticNet MAE, whereas Consumer Confidence Index data does not.



Mean Squared Error (MSE) barplot depicts equivalence between HistGradientBoost models and GridSearchCV RandomForest model.



Regarding feature engineering, minimal difference is observed before and after logarithmic transformation in the R2, MAE, MSE scores.

CONCLUSION -

The Capstone Project objective is an understanding of Walmart stores and weekly sales relationships. This involves deducing and quantifying sales trends and behaviours by using regression models. Several ensemble and linear regression models were built using the Walmart Weekly Sales Forecasting dataset from Kaggle.com. The most important features of the XGBoost regression model provides information about which Walmart store characteristics have the largest influence on weekly sales predictions. Departments 92, 95, 38, 90, and 72 along with Store Size have the largest influence on the weekly sales predictions. The models were then compared for how accurate the predictions were. HistGradientBoost model had the best metrics with highest coefficient of determination, R2 = 0.70 and lowest mean average error, MAE = 6500.

Improving the model predictions was attempted using additional data, feature engineering, and residuals plotting. Additional consumer related data in the form of the Consumer Confidence Index data, did not improve regression models performance. Using GridSearchCV and feature engineering produced comparatively larger performance metric improvements for several regression models. Residuals scatterplots indicate the largest model errors are associated with the Christmas and Thanksgiving holidays and several stores and departments.