Walmart Sales Kaggle Competition Revisited

Project Background:

In Spring 2014, Kaggle hosted the Walmart Recruiting - Store Sales Forecast competition. The goal was to predict weekly sales using the historical sales data of forty-five Walmart stores. Scoring was based on a weighted mean absolute error.

In addition to features directly related to individual stores (i.e. store size, dept), the training data included features thought to affect weekly sales. Gasoline prices, temperature, unemployment, and the consumer price index (cpi) were included.

The Consumer Confidence Index (CCI) survey measures consumer sentiment in the United States. It is often used as a gauge of consumers' willingness to spend.

Would adding CCI data to the competition training data improve or worsen weekly sales predictions? Does the effect vary with the regression model used?

Problem Statement:

1. Exploratory Data Analysis and Regression Models produce useful business insights about the Walmart Stores Network
    a. Walmart Stores sales landscape.
    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 Sources-

1. Kaggle.com-

   Stores.csv

   Features.csv

   Train.csv

   Test.csv

2. Consumer Confidence Index Data-

   https://data.oecd.org/leadind/consumer-confidence-index-cci.htm

Data Wrangling Summary -

- Training Data: 421,570 entries covering 5 Feb 2010 thru 26 Oct 2012
- 17 Features including target, Weekly Sales
- 5 Categorical Features

    4 Nominal - ‘IsHoliday’, 'Dept', 'Store', 'Type'

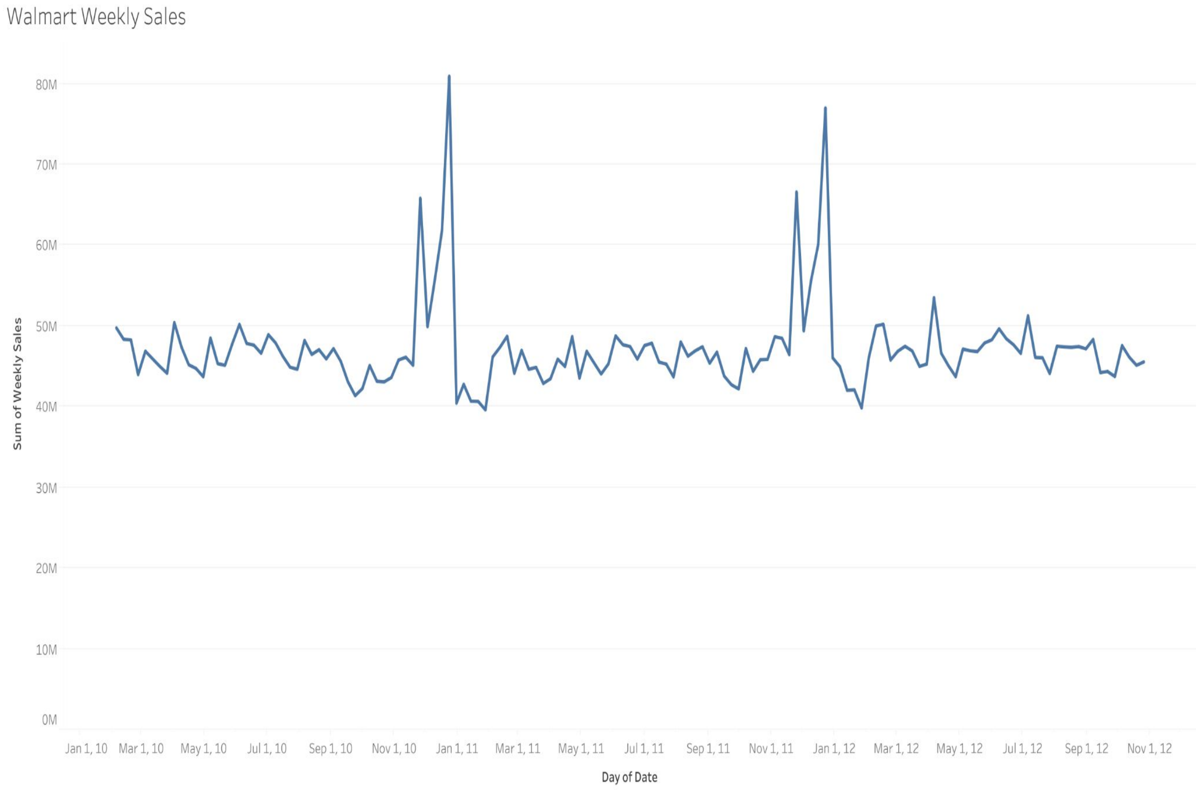
    1 Ordinal - ‘Week’

- 12 Numerical Features

    ‘Weekly_Sales’, 'Temperature', 'Fuel_Price', 'MarkDown1', “'MarkDown2', 'MarkDown3', 'MarkDown4', 'MarkDown5', 'CPI', 'Unemployment', 'Size’, 'cci_value'

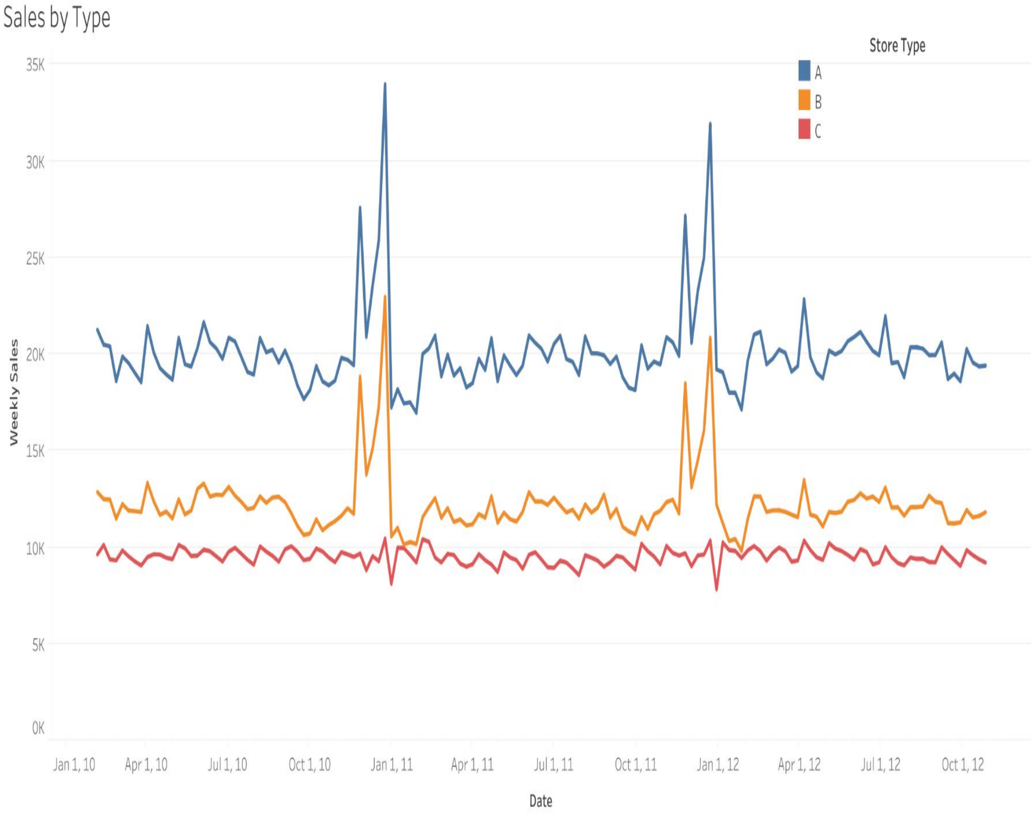
Weekly Sales at-a-glance

- Thanksgiving and Christmas peaks dominate the annual weekly sales.



Store 'Type'

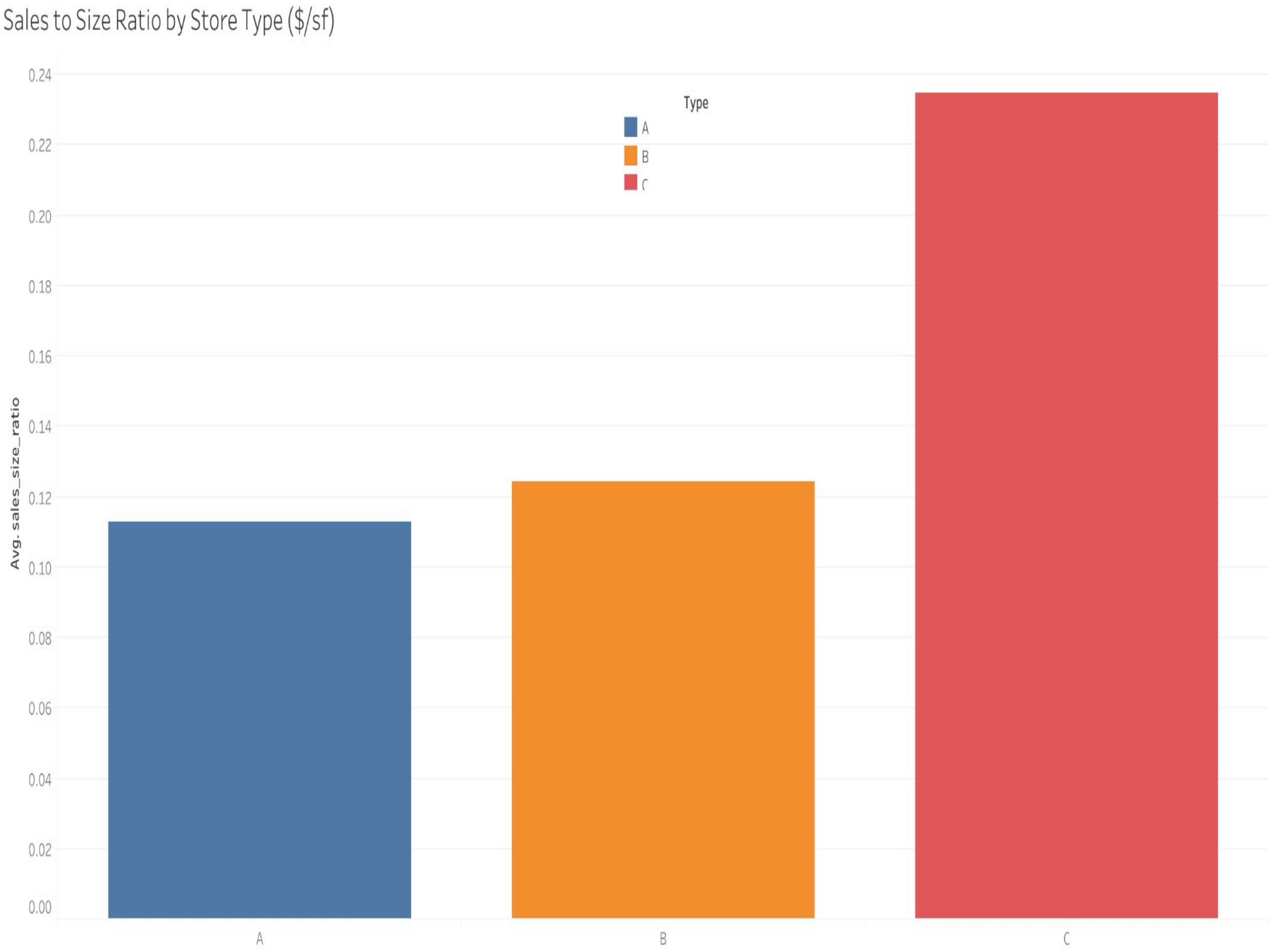
- 45 Walmart Stores are divided in three types based on Weekly Sales



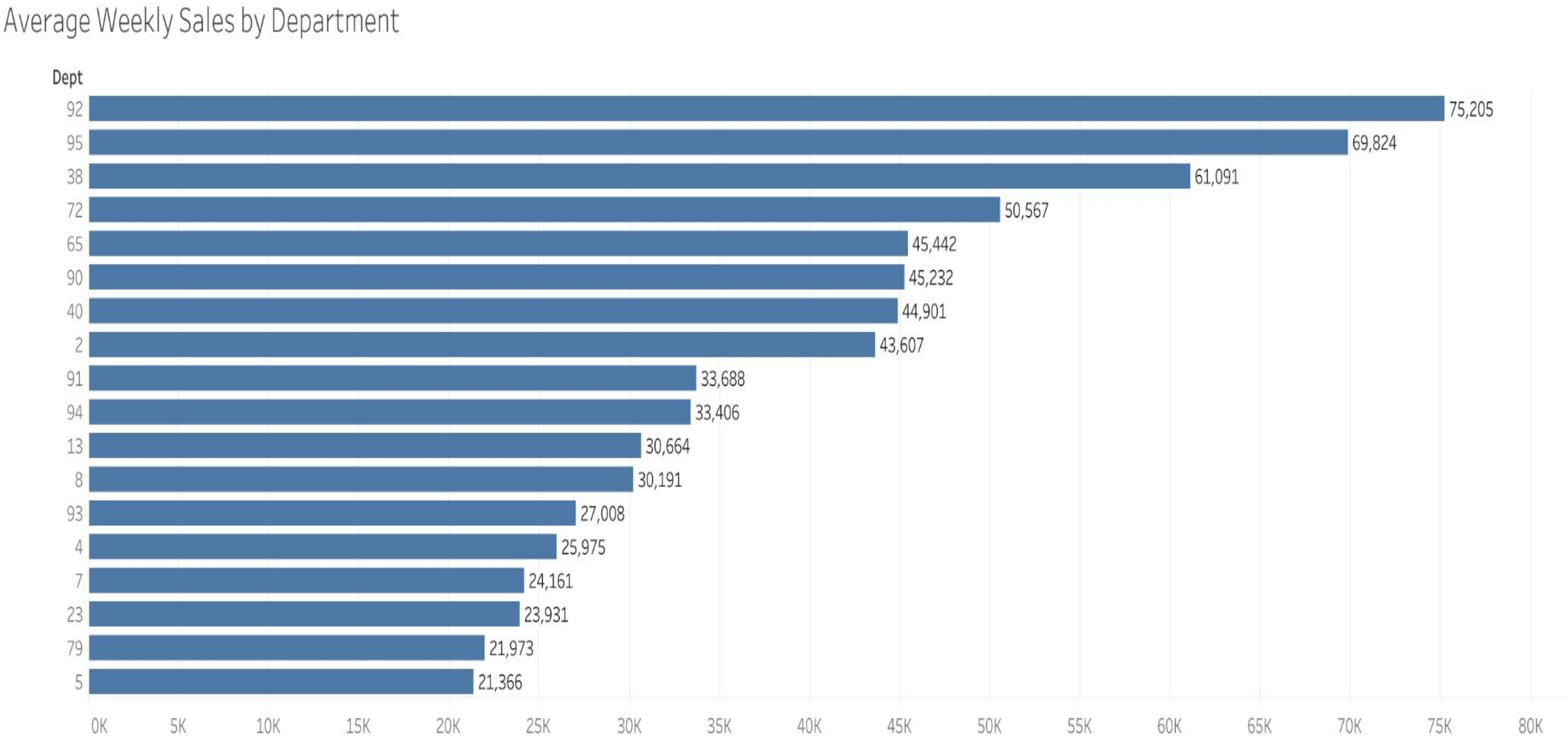
Average Weekly Sales to Size ratio:

Type ‘C’ is 2x

That of ‘A’



Seven departments are responsible for approx. 37% of all Weekly Sales aggregated over date range

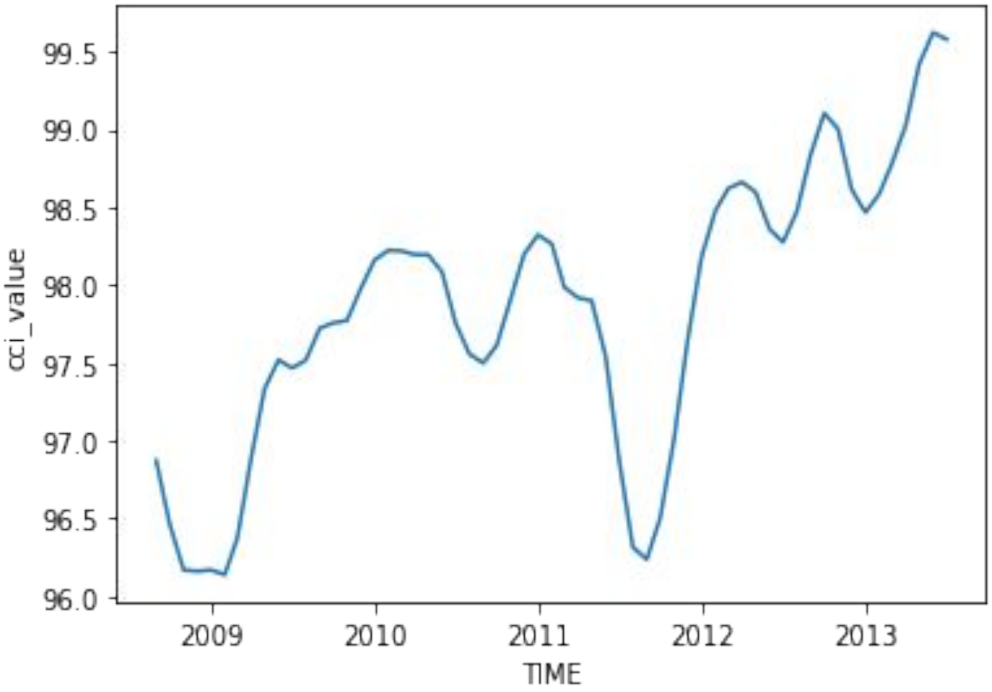


Consumer Confidence Index

CCI varies within a narrow range over data set date range.

High - 99.5

Low - 96.0



Exploratory Data Analysis - Weekly Sales - Numerical Featues Correlation Heatmap

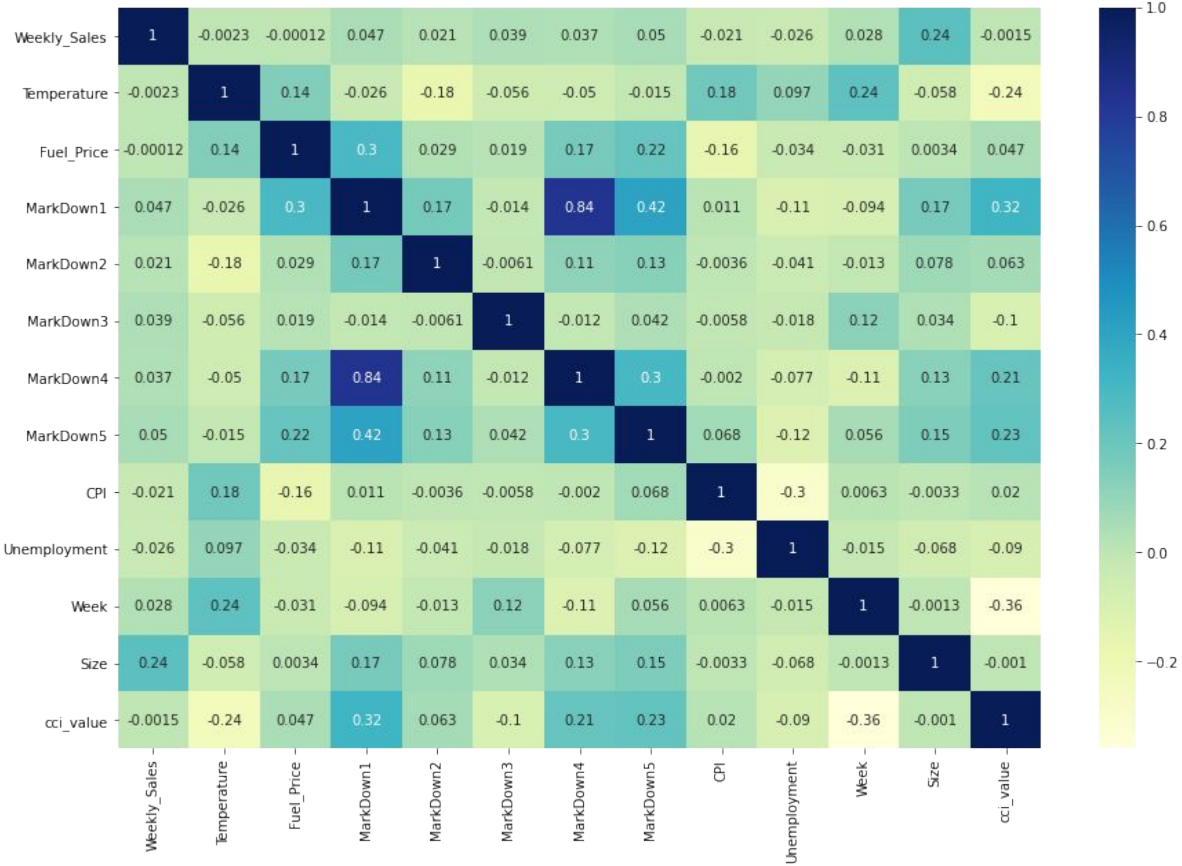
Largest: ‘Size’ +0.24

Smallest:

‘Fuel Price’ -0.00012

Practically Zero:

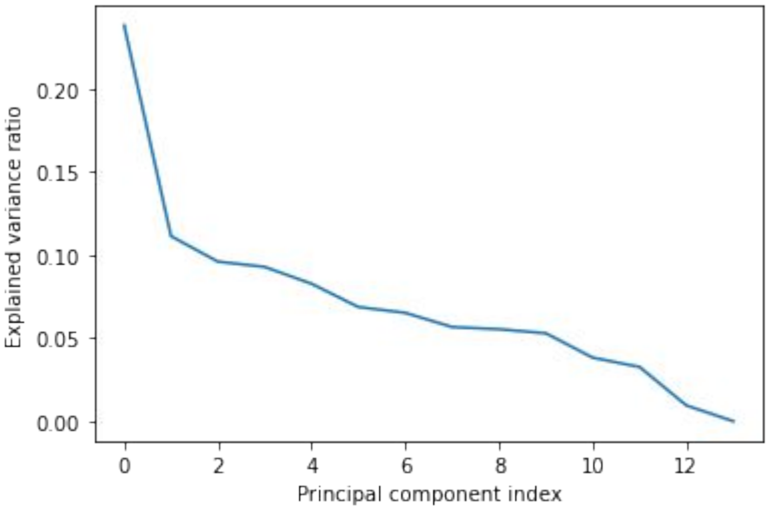
‘CCI’ -0.0015

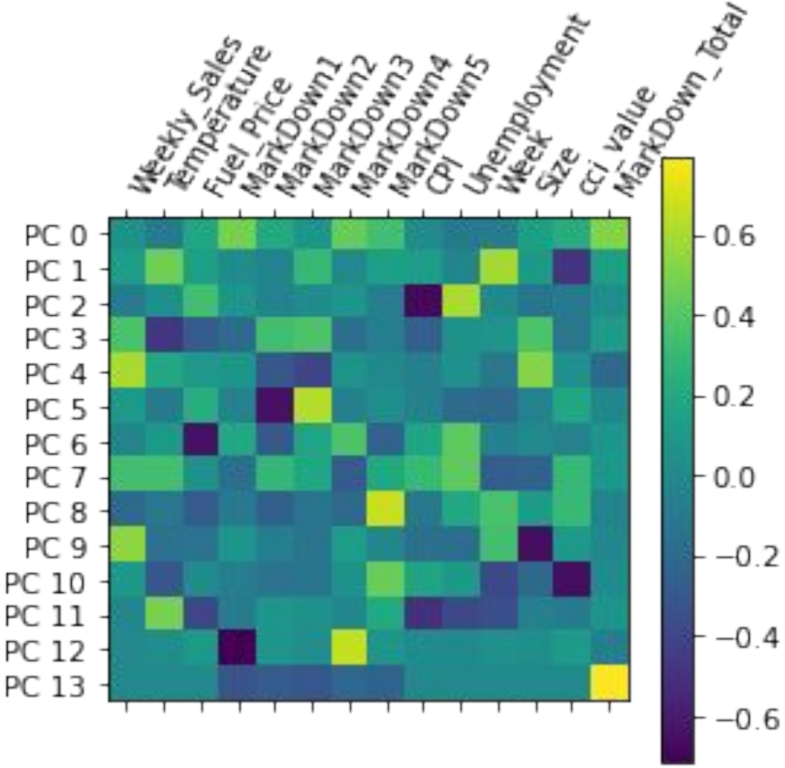


	Weekly_Sales	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	Week	Size	cci_value
Weekly_Sales	1	-0.0023	-0.00012	0.047	0.021	0.039	0.037	0.05	-0.021	-0.026	0.028	0.24	-0.0015
Temperature	-0.0023	1	0.14	-0.026	-0.18	-0.056	-0.05	-0.015	0.18	0.097	0.24	-0.058	-0.24
Fuel_Price	-0.00012	0.14	1	0.3	0.029	0.019	0.17	0.22	-0.16	-0.034	-0.031	0.0034	0.047
MarkDown1	0.047	-0.026	0.3	1	0.17	-0.014	0.84	0.42	0.011	-0.11	-0.094	0.17	0.32
MarkDown2	0.021	-0.18	0.029	0.17	1	-0.0061	0.11	0.13	-0.0036	-0.041	-0.013	0.078	0.063
MarkDown3	0.039	-0.056	0.019	-0.014	-0.0061	1	-0.012	0.042	-0.0058	-0.018	0.12	0.034	-0.1
MarkDown4	0.037	-0.05	0.17	0.84	0.11	-0.012	1	0.3	-0.002	-0.077	-0.11	0.13	0.21
MarkDown5	0.05	-0.015	0.22	0.42	0.13	0.042	0.3	1	0.068	-0.12	0.056	0.15	0.23
CPI	-0.021	0.18	-0.16	0.011	-0.0036	-0.0058	-0.002	0.068	1	-0.3	0.0063	-0.0033	0.02
Unemployment	-0.026	0.097	-0.034	-0.11	-0.041	-0.018	-0.077	-0.12	-0.3	1	-0.015	-0.068	-0.09
Week	0.028	0.24	-0.031	-0.094	-0.013	0.12	-0.11	0.056	0.0063	-0.015	1	-0.0013	-0.36
Size	0.24	-0.058	0.0034	0.17	0.078	0.034	0.13	0.15	-0.0033	-0.068	-0.0013	1	-0.001
cci_value	-0.0015	-0.24	0.047	0.32	0.063	-0.1	0.21	0.23	0.02	-0.09	-0.36	-0.001	1

Principal Component Analysis

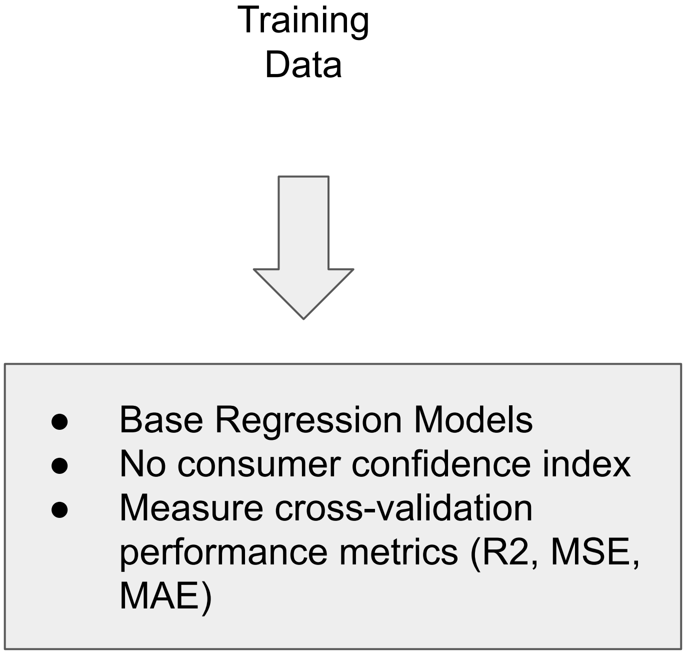
PC 0 and PC 1 responsible 34% of Weekly Sales Variation





Experiment Procedure:

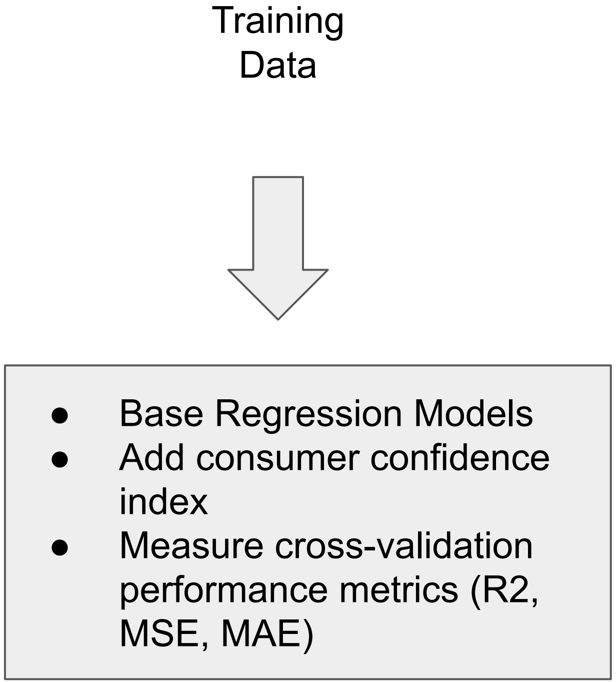
Part I.



Training Data

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

Part II.

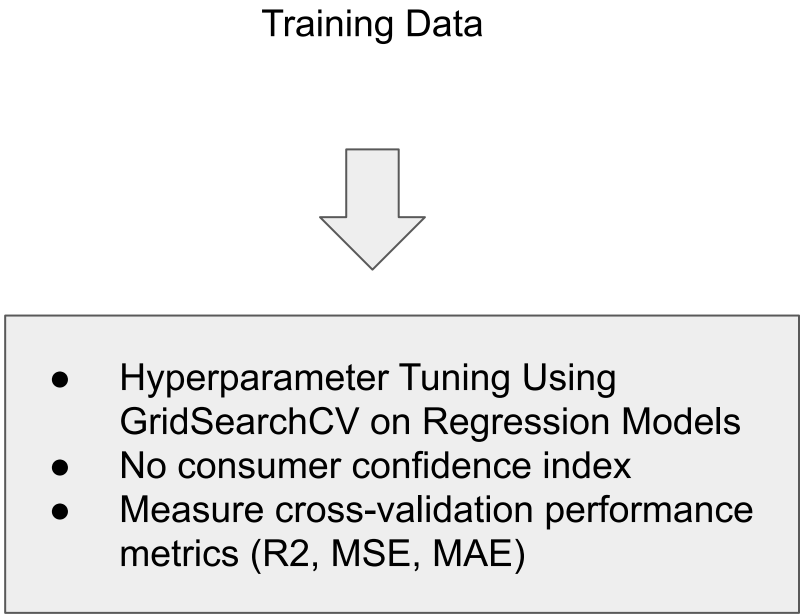


Training Data

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

Experiment Procedure:

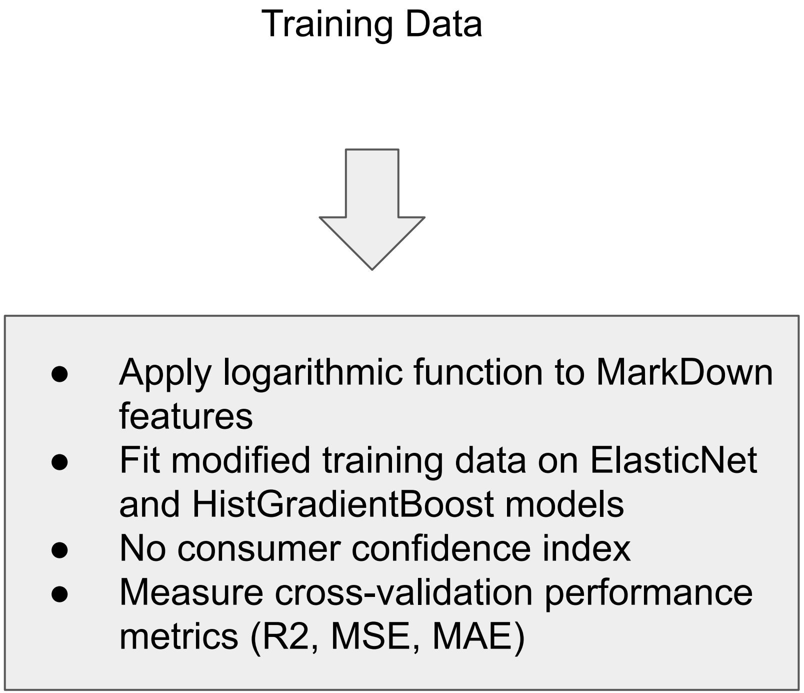
Part III:



Training Data

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

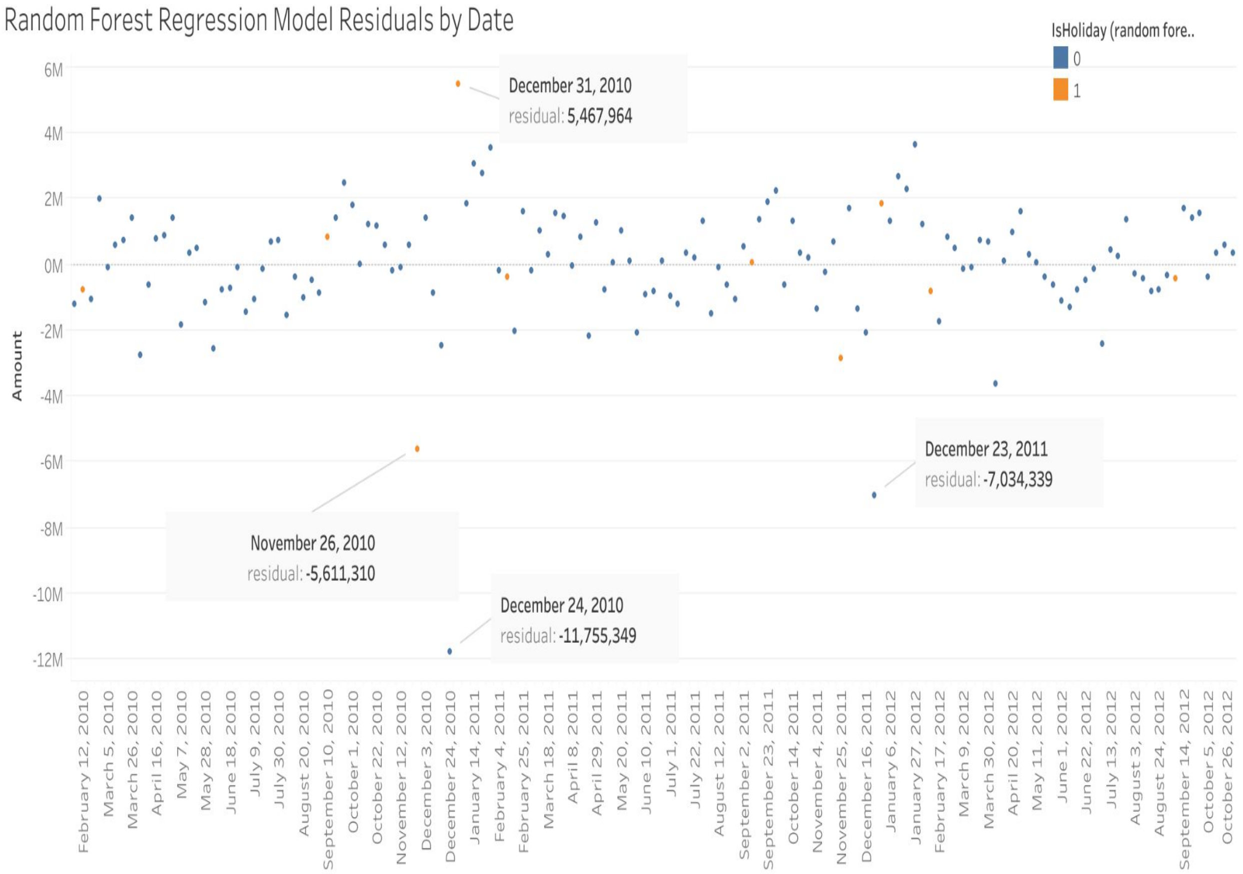
Part IV:



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)

- Several dates produce unusual residuals



Random Forest Regression Model Residuals by Date

December 31, 2010
residual: 5,467,964

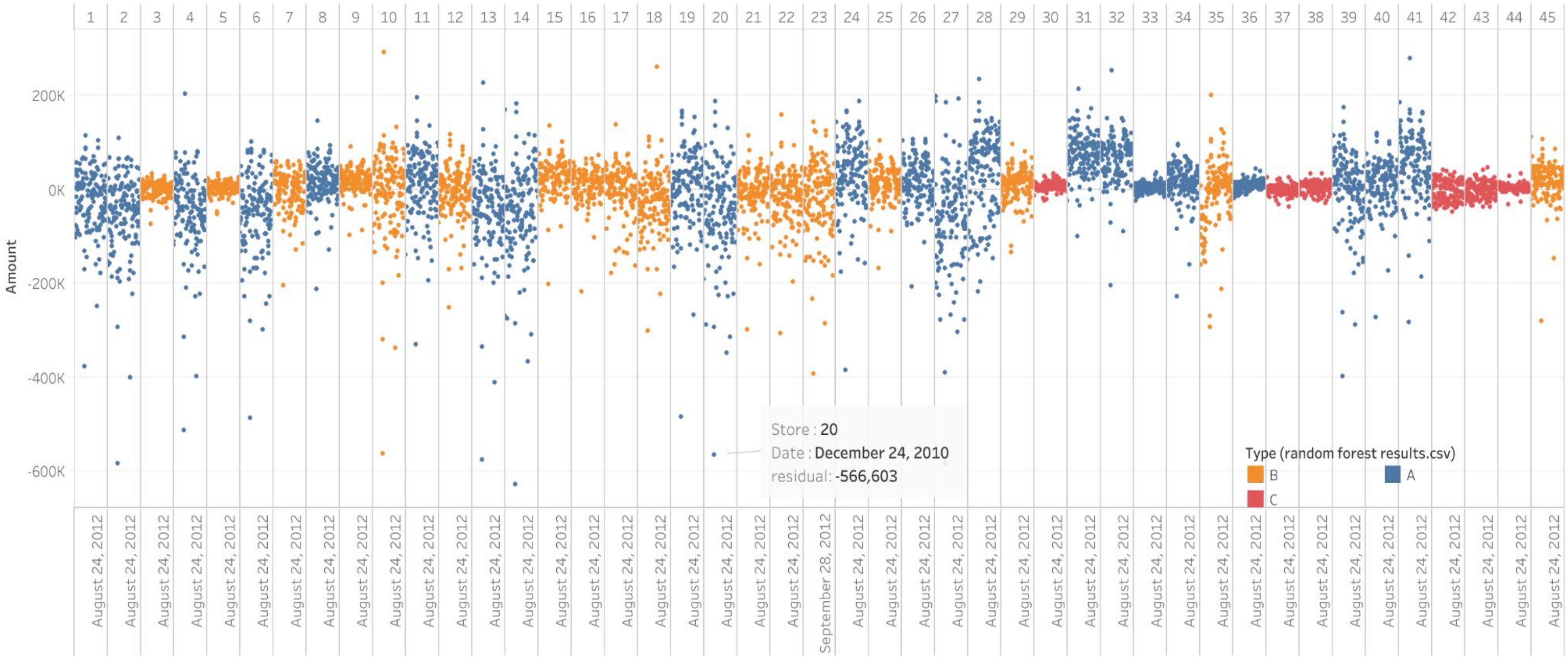
November 26, 2010
residual: -5,611,310

December 24, 2010
residual: -11,755,349

December 23, 2011
residual: -7,034,339

Type A Stores produce model residuals outliers

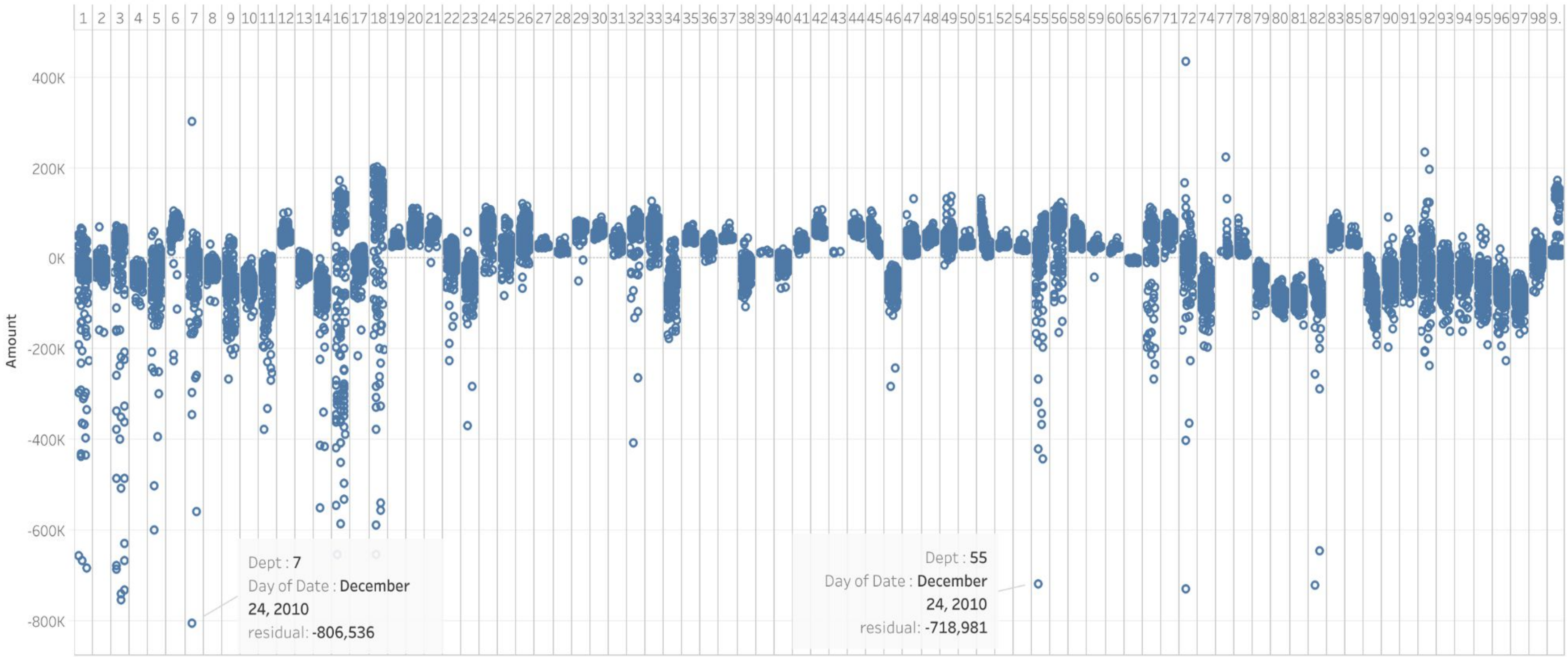
Random Forest Regression Model Residuals by Store and Date



Store : 20
Date : December 24, 2010
residual: -566,603

Residual outliers associated with particular Departments

Random Forest Regression Model Residuals for Type A Stores by Dept, and Date

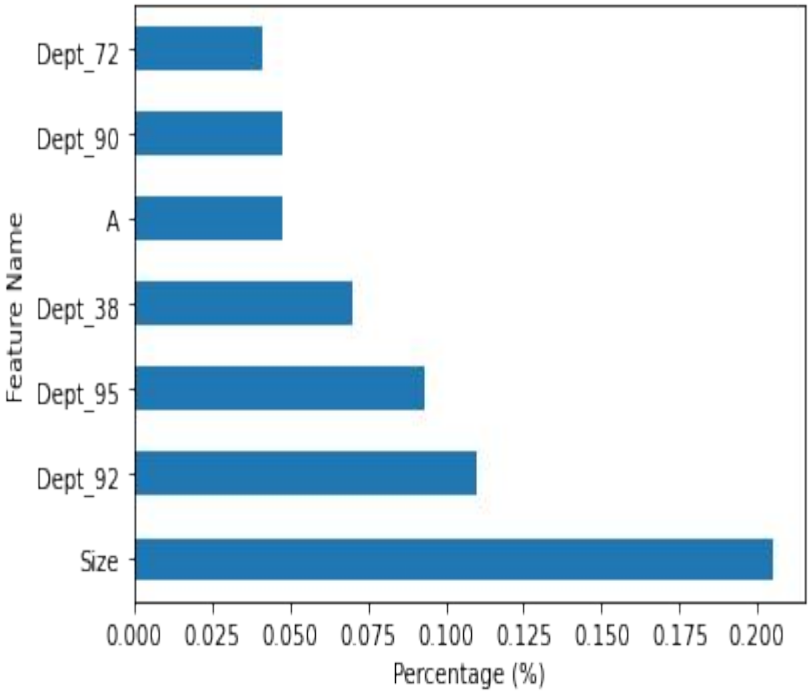


Dept : 7
Day of Date : December 24, 2010
residual: -806,536

Dept : 55
Day of Date : December 24, 2010
residual: -718,981

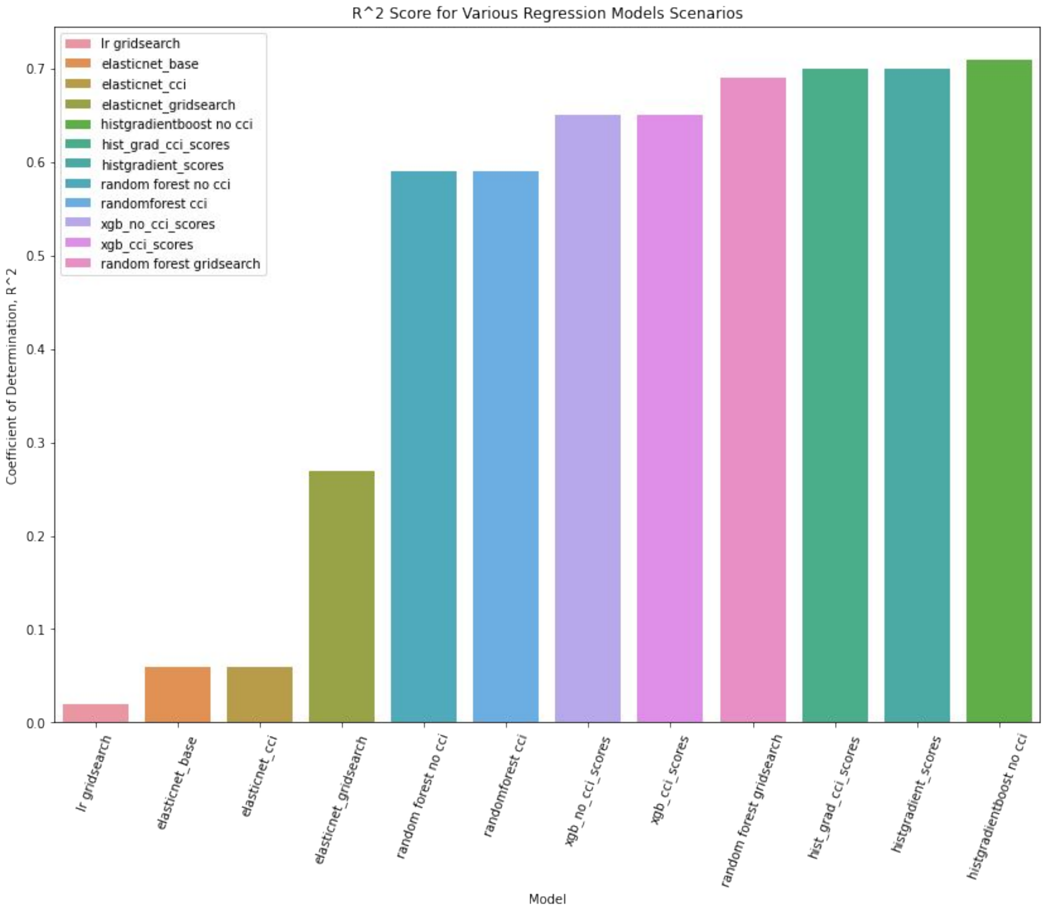
Results- XGBoost Regressor Feature Importance

- Store Size has the largest importance on weekly sales.
- Departments 92, 95, 38, 90, 72 are within top size of largest average weekly sales



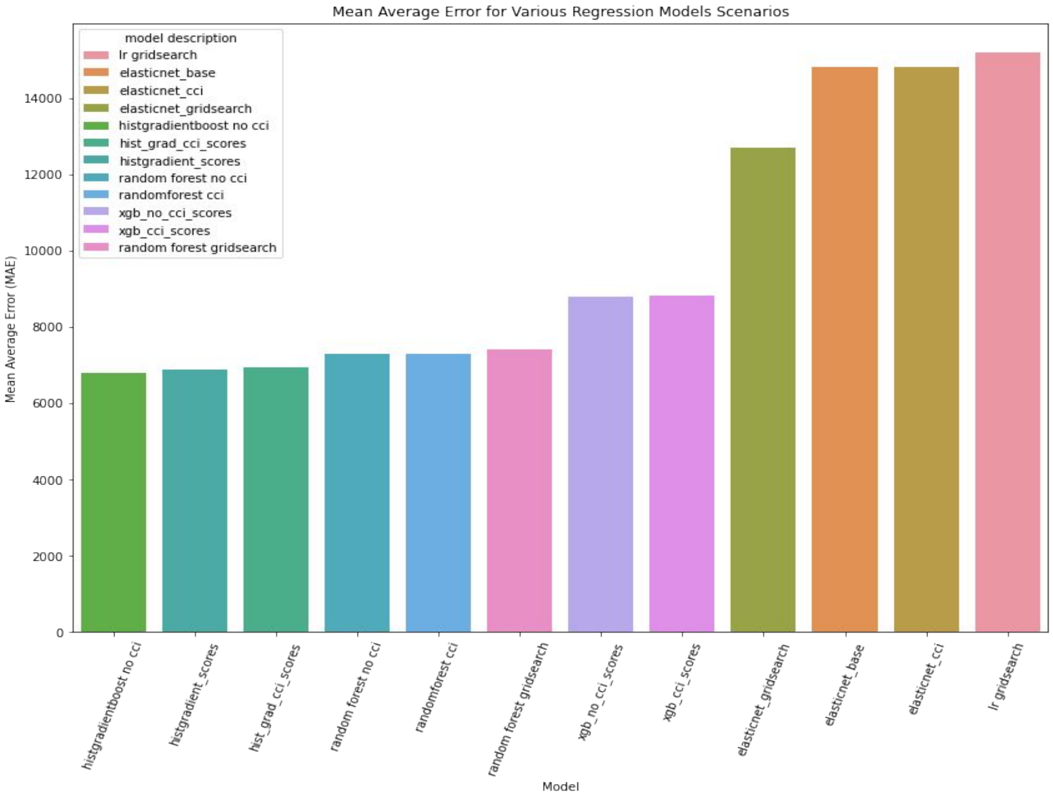
Results:

- R2 worsens slightly with cci data inclusion for HistGradientBoosting, Random Forest, XGBoost Regressors
- Minimal improvement for already low ElasticNet R2 score
- When compared to basic model, GridSearchCV produces larger improvement in R2 scores than CCI data



Results -

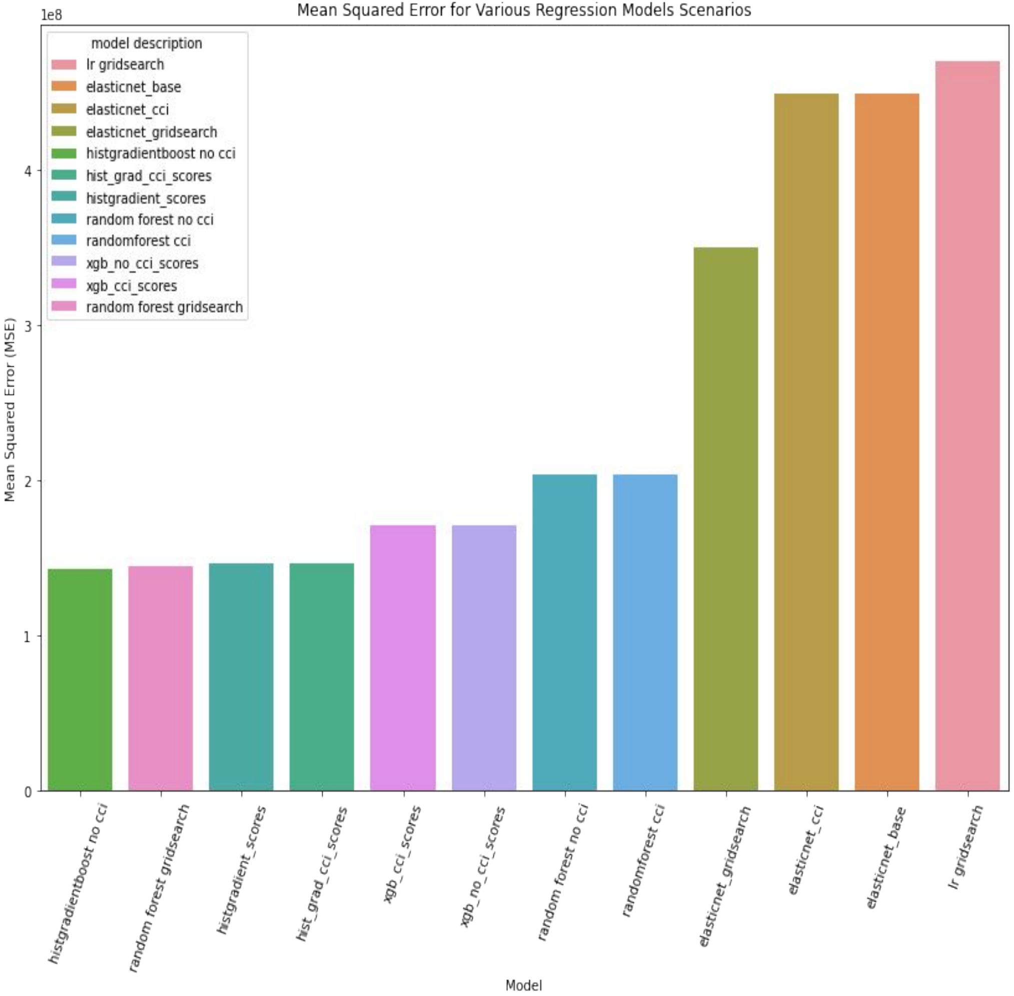
Mean Average Error (MAE) increases minimally with inclusion of CCI data for both ensemble and linear regression models



Results:

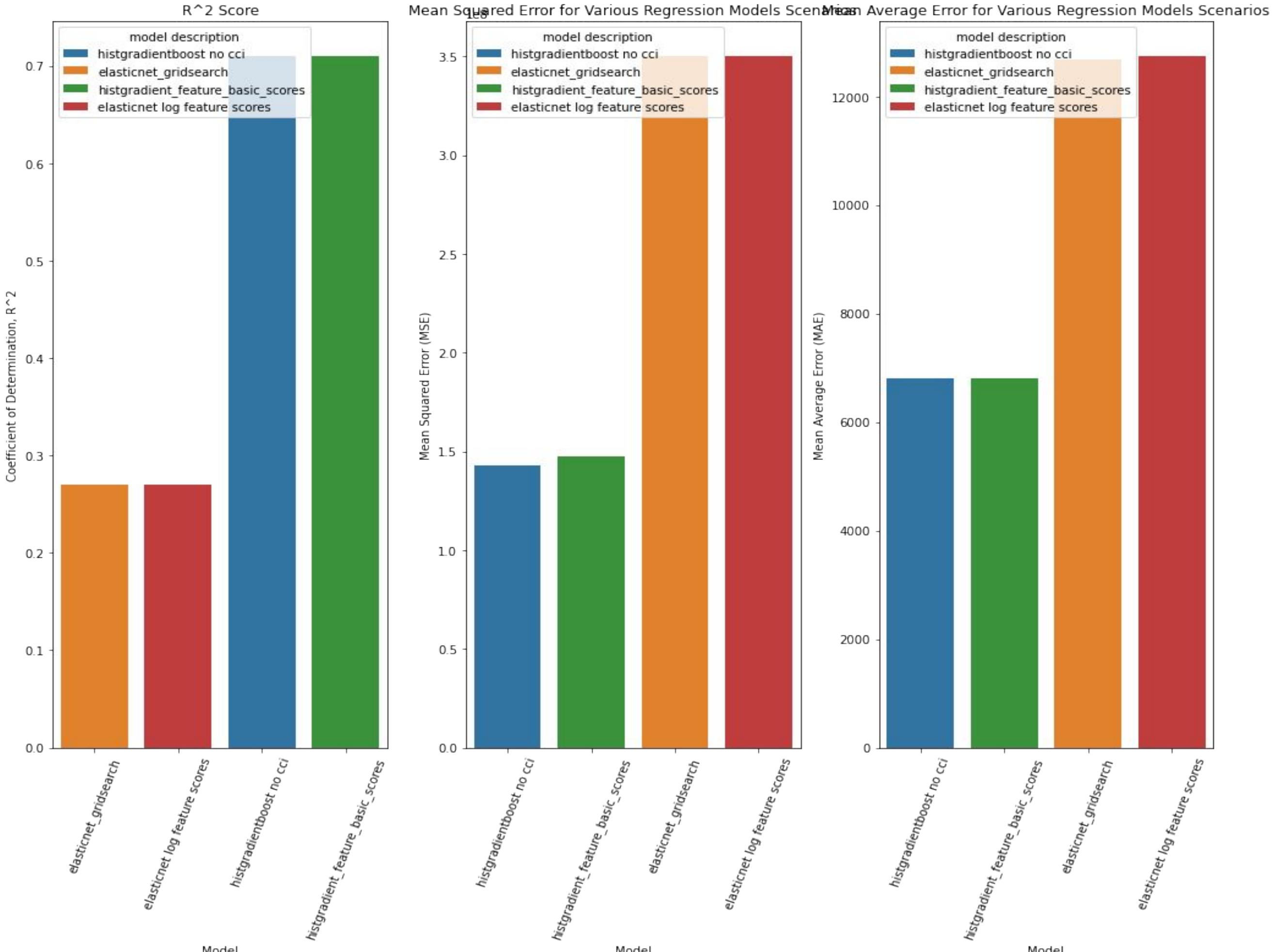
- CCI data increases Mean Squared Error For HistGradientBoost, XGBoost models

- CCI data decreases Mean Squared Error For RandomForest, ElasticNet models



Feature Engineering Results

Logrithmic transformation produces minimal improvement in R2, MSE, and MAE metrics



Conclusions:

- Based on provided data and regression models, 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.

Conclusions:

Consumer Confidence Index Data does not improve model performance

- R2 worsens with cci data inclusion for HistGradientBoosting, Random Forest, XGBoost Regressors
- When compared to basic model, GridSearchCV produces larger improvement in R2 scores than CCI data
- Mean Average Error (MAE) increases minimally with inclusion of CCI data for both ensemble and linear regression models

Conclusions:

Feature Engineering involving logarithmic transformation of MarkDown features does not improve model performance