Predicting Weekly Revenue of Walmart

Sophia Song
Data Science Career Track, 2019

Introduction:

Walmart is an American retail operation that operates a chain of grocery stores by the same name. Since its opening in 1962, it has expanded its ventures to many other countries, including Mexico, China, and Germany.

As of 2019, Walmart is the world's largest company by both revenue and number of employees, at $514.405 billion per year and 2.2 million people, respectively.

Problem

We're trying to predict a department - store combination's weekly revenue using previous information about revenue as well as a couple of geographical factors, like temperature, gas price, and unemployment rate.

Data

The columns included are:

- Store ID - *integer*
- Department ID - *integer*
- Date - *string*
- IsHoliday - *boolean*
- Temperature - *float*
- Gas-Price - *float*
- MarkDown1, MarkDown2, Markdown3, MarkDown4, MarkDown5 - *float*
- CPI - *float*
- Unemployment - *float*
- Type - *string*
- Size - *integer*
- Weekly_Sales - *float*

The data comes in two .csv files, one for testing and one for training. The training dataset contains 282,000 rows and 16 columns, one of which is the weekly revenue column that’s the object of our prediction, and the testing test is identical in format with the training, but doesn’t have the weekly revenue.

Data Cleaning

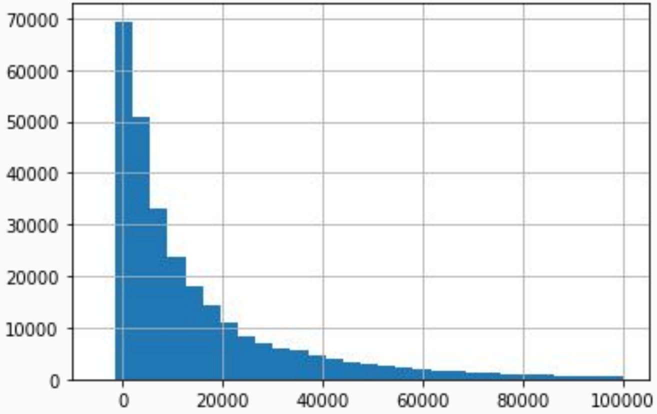
There were missing values in the markdown columns, which I assumed meant no promotional events.

There were also an outlier in the weekly sales column that interfered with our analysis, so I removed it.

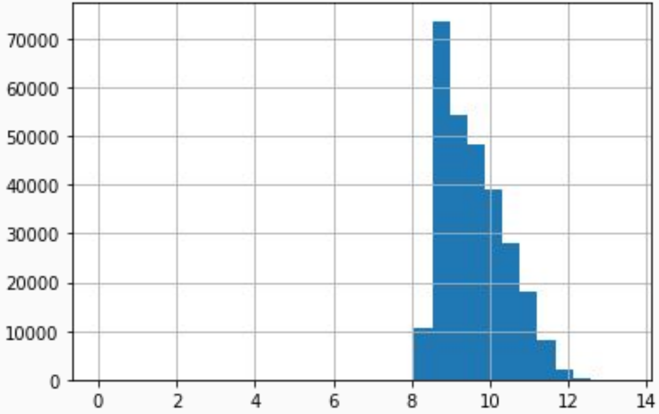
Exploratory Data Analysis

Weekly Sales

The weekly sales column have an exponential distribution. For easier analysis, I normalized the weekly sales using a logarithmic scale.

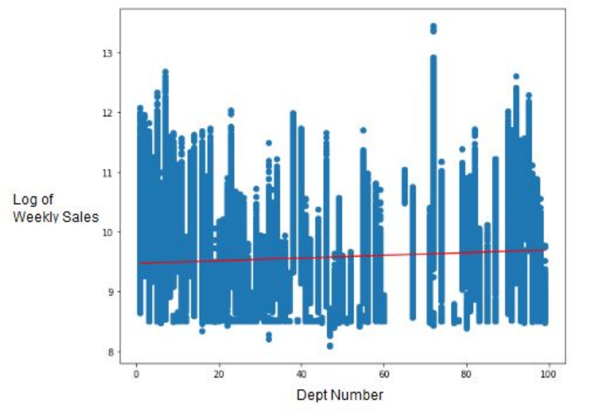


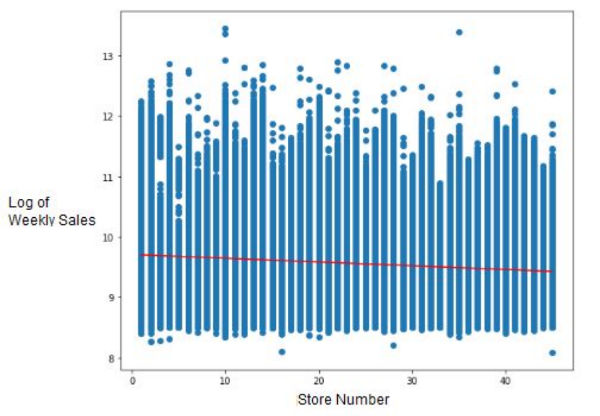
Before

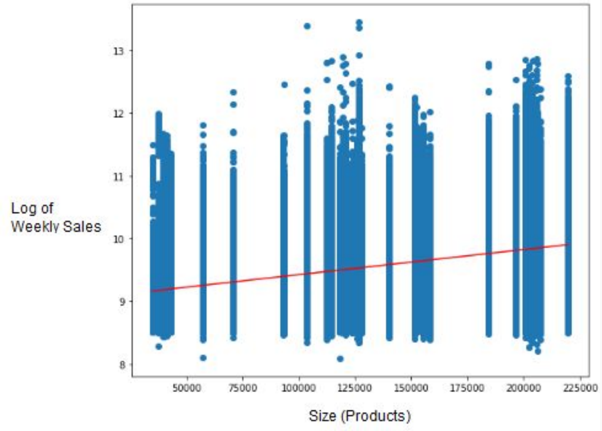


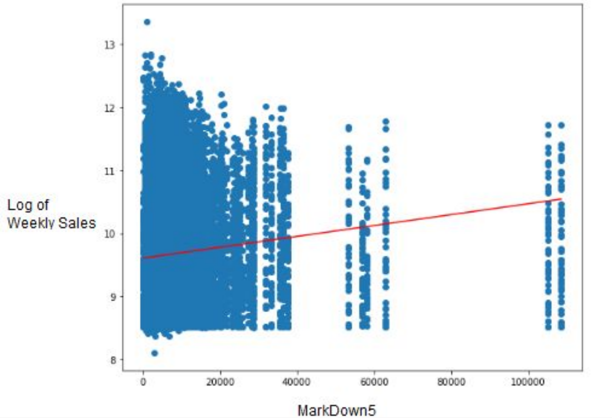
After

Target's relationship with other variables



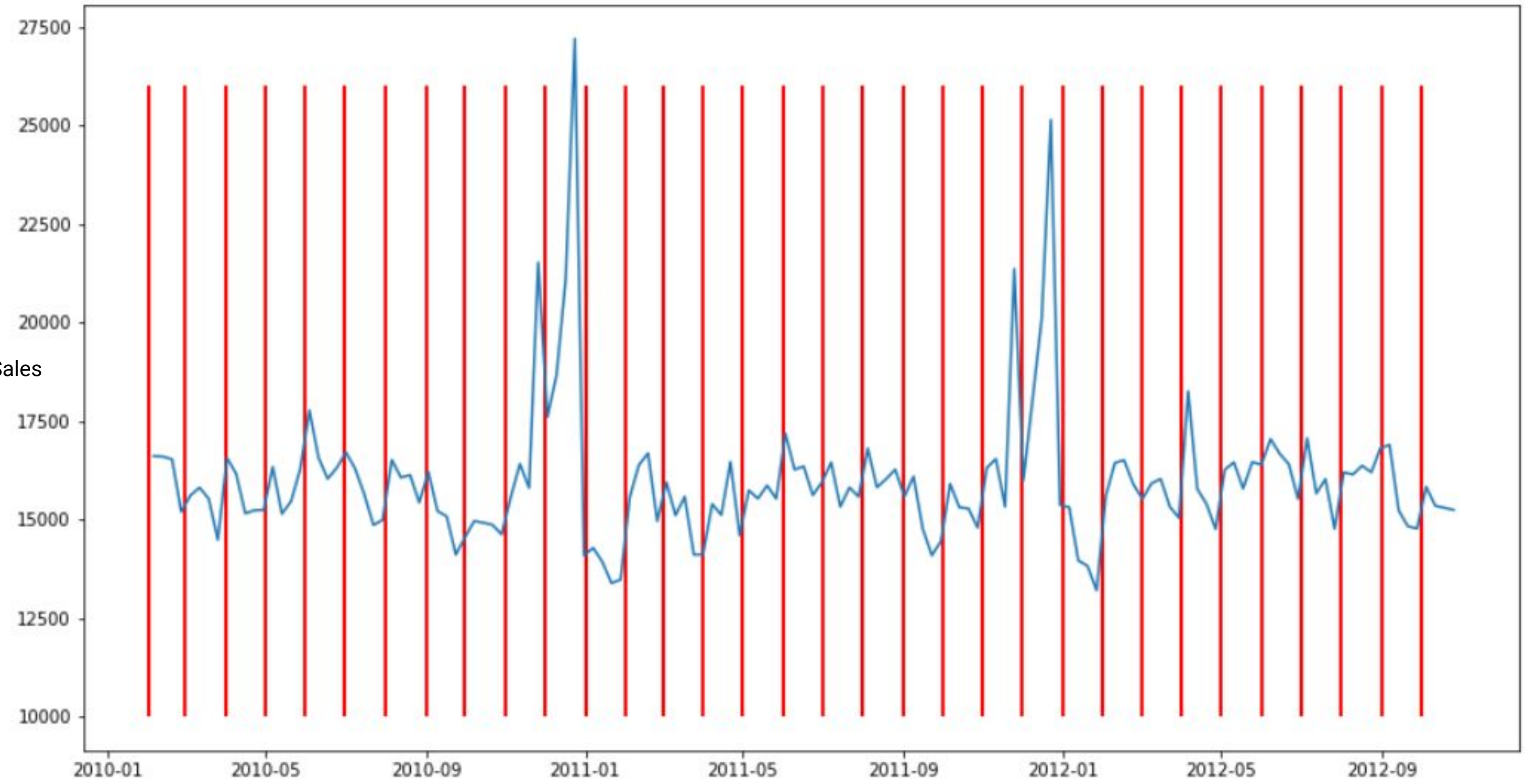






Date

Average Weekly Sales

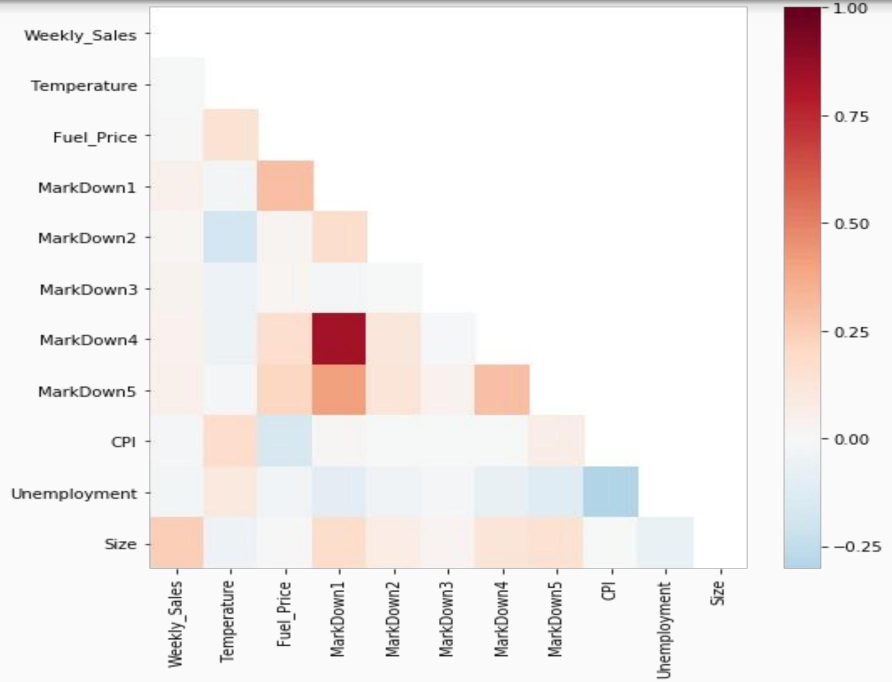


Date

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Weekly_Sales   R-squared:                       0.007
Model:                            OLS   Adj. R-squared:                  0.007
Method:                 Least Squares   F-statistic:                     226.1
Date:                Mon, 16 Sep 2019   Prob (F-statistic):               0.00
Time:                        18:28:53   Log-Likelihood:            -3.2323e+06
No. Observations:              282451   AIC:                         6.465e+06
Df Residuals:                  282441   BIC:                         6.465e+06
Df Model:                           9
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept     2.606e+04    494.488     52.711      0.000    2.51e+04     2.7e+04
Temperature     23.9436      2.471      9.691      0.000      19.101      28.786
Fuel_Price   -1483.2012    102.857    -14.420      0.000   -1684.798   -1281.604
MarkDown1        0.1904      0.014     13.386      0.000       0.163       0.218
MarkDown2        0.0513      0.009      5.919      0.000       0.034       0.068
MarkDown3        0.1500      0.008     19.263      0.000       0.135       0.165
MarkDown4       -0.0727      0.021     -3.545      0.000      -0.113      -0.033
MarkDown5        0.1940      0.011     17.401      0.000       0.172       0.216
CPI            -25.0676      1.203    -20.834      0.000     -27.426     -22.709
Unemployment  -388.1754     24.458    -15.871      0.000    -436.113    -340.238
==============================================================================
Omnibus:                   205761.418   Durbin-Watson:                   2.001
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          5733316.107
Skew:                           3.215   Prob(JB):                         0.00
Kurtosis:                      24.114   Cond. No.                     9.45e+04
==============================================================================
```

Intercorrelations

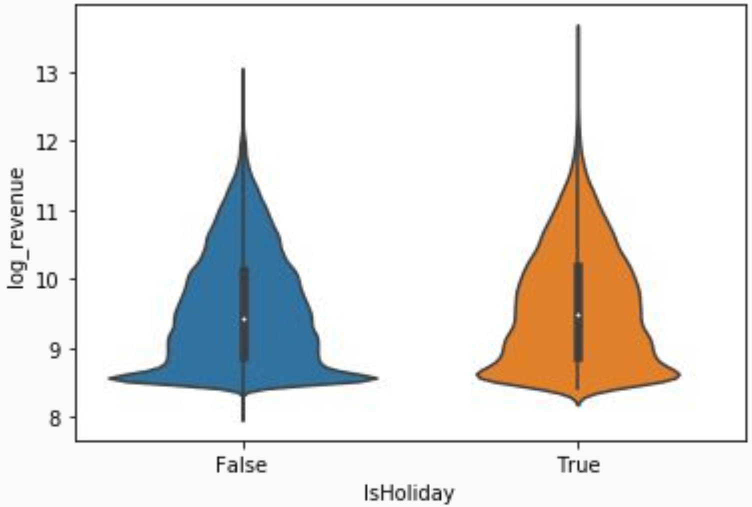
- Positive between MarkDown Columns
- Negative between Unemployment and CPI



Inferential Statistics

Holiday

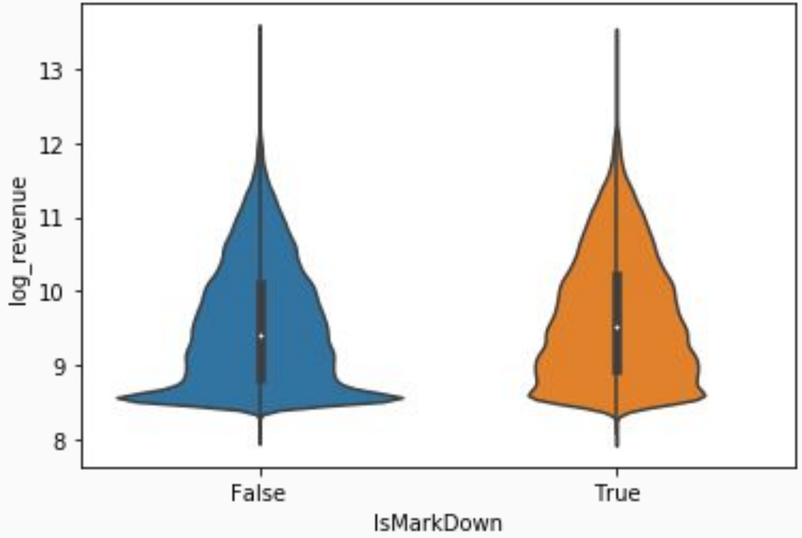
Distribution of the log of revenues for Holiday and Non-Holiday weeks



Holiday weeks see significantly more sales than non-holiday weeks

Markdown

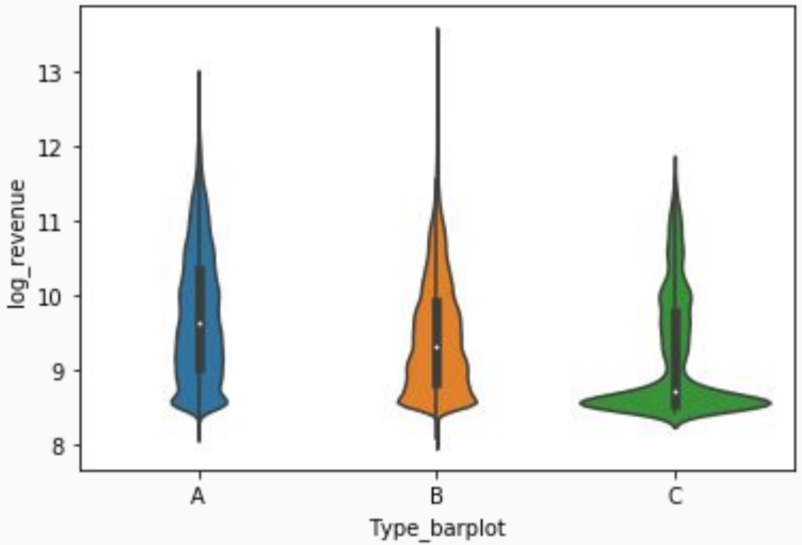
Distribution of the log of revenues for MarkDown and Non-Markdown weeks



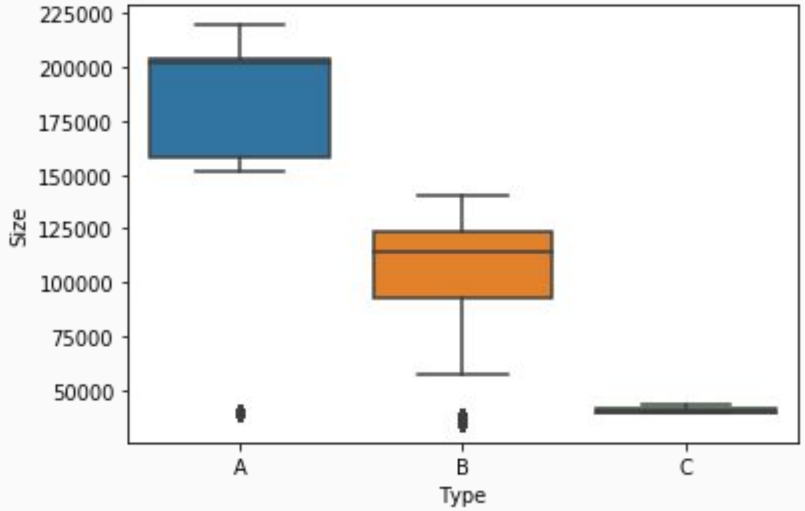
Markdown weeks generally see significantly more sales than non-markdown weeks, though the highest sale is made in a non-markdown week

Store Type

Distribution of the log of revenues for different store types



Distribution of size for each store



Feature Engineering

New Features

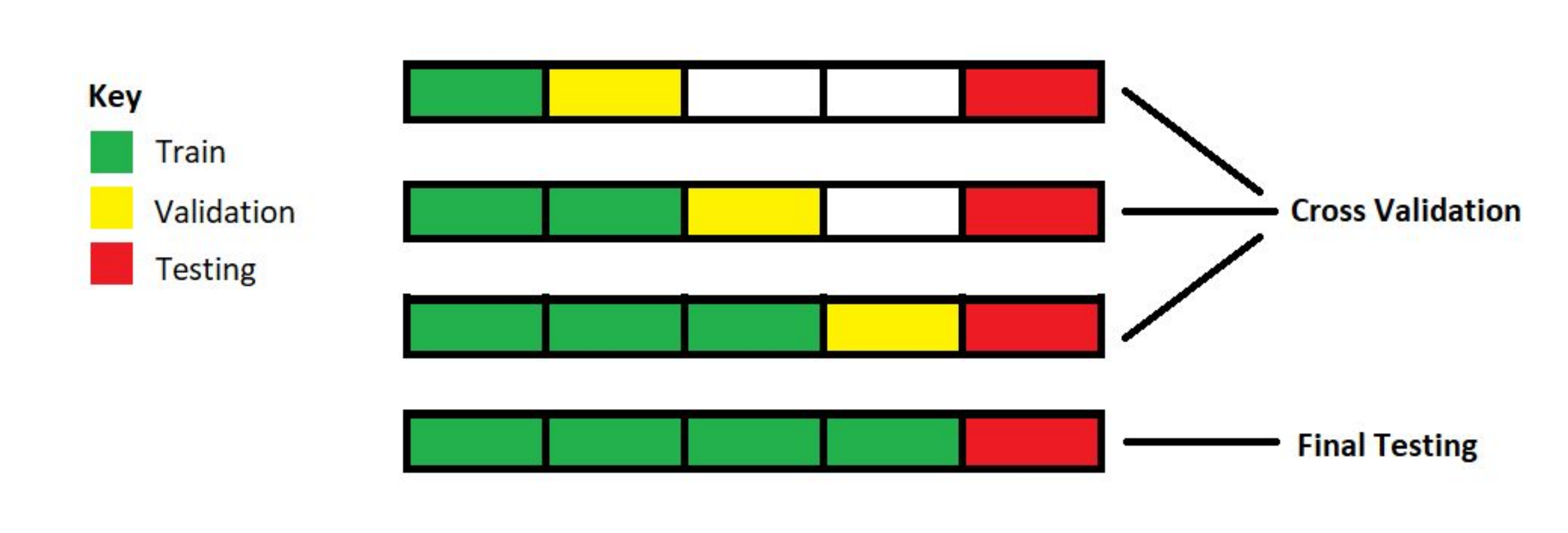
- Total Markdowns
- Type A, B, and C store dummy variables
- Year, Month, Week of Year, and Day
- Yearly Median value of each Store-Dept combination
- Lagged Revenue and its difference from Median
- Big Holidays
- Department 72 dummy variable

Machine Learning

Models I tried

- Linear Regression
- Ridge Regression
- Lasso Regression
- RandomForestRegressor
- XGBoost
- Support Vector Regressor

Training and Testing



Feature Selection

- For Linear Regression and all of its closely related algorithms (Lasso, Ridge), I used RFE (Recursive Feature Elimination) to select for the optimal number of features.

Results

In the models I tried both the unscaled, unmodified dataset and a standardized version of it

Model	MAE(Scaled)	MAE(unscaled)
Linear Regression	16030153634809.34	2641.02
Ridge Regression	14979.07	2641.02
Lasso Regression	15045.89	2631.40
RandomForestRegressor	11766.21	2573.75
XGBoost	13107.88	2224.51
SVR	13345.28	13431.92

1566.00

Final MAE using our test Dataset