QMBU 450

SPRING 2020

FINAL PROJECT REPORT

Walmart Recruiting- Store Sales Forecasting

Anıl Kul

Muhammet Said Kırılmaz

1. Introduction

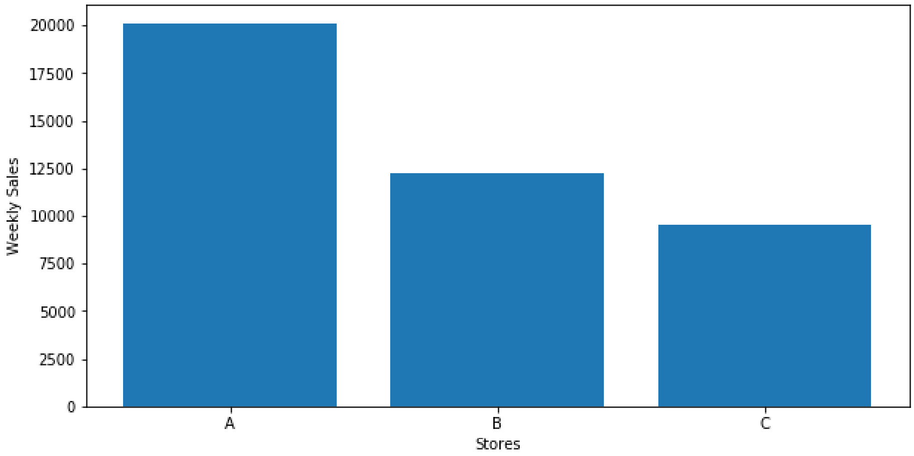
For this project we chose to work on the Walmart's store sales forecasting data. This data was used for a recruiting process. It is one of the well-known challenges from Kaggle. This project contains data visualization, pre-process of the data, and implementation of three different models.

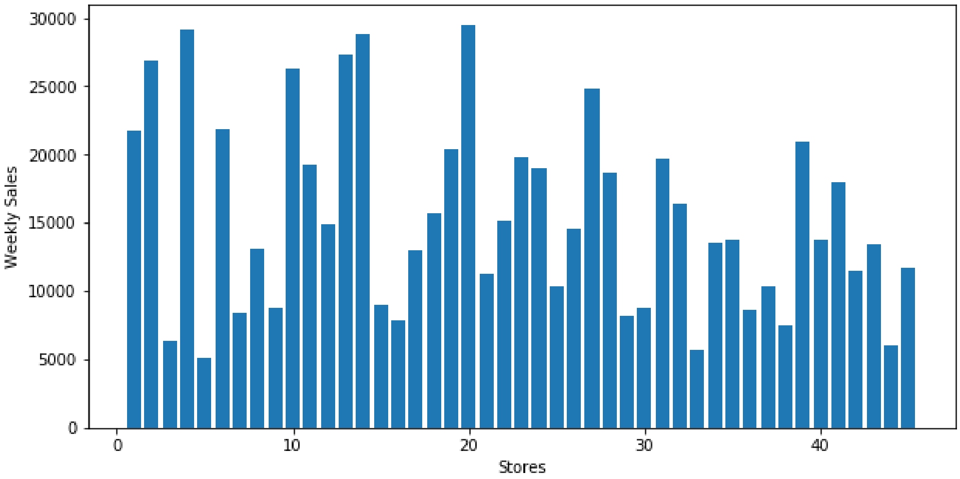
2. Data Set

Data set had historical sales of 45 Walmart Stores located in different region. Each store has different departments. Data set consist of 3 different csv files. First file is stores.csv. In This file had information about all 45 stores with their type and size. 22 of the stores type A, 17 of the stores type B, and remaining 6 of the stores are type C stores. Second file is train.csv. This file had historical data from 05-02-2010 to 2012-11-01. Also, this file had information about store number, department number, date, sales for given department in the given store, and whether the week is a special holiday week. The last file is features.csv. This file contains information about store number, date, average temperature in the region, cost of the fuel in the region, promotional markdowns, the consumer price index, the unemployment rate, and whether the week is a special holiday week. Our target is forecast the sales for each store.

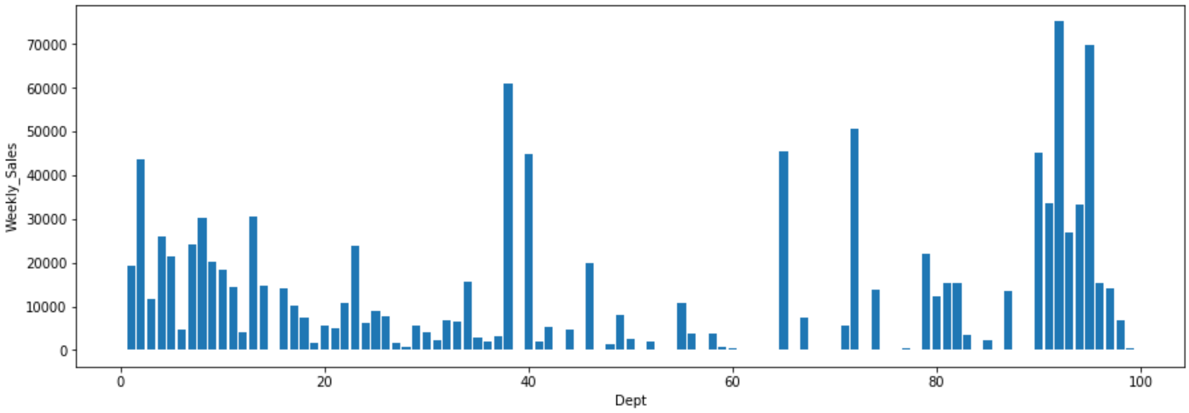
3. Pre-process and Visualization

After reading all three csv files, a merging operation implemented. First, we started to examine the store numbers and types. There is a link between store types and the weekly sales.

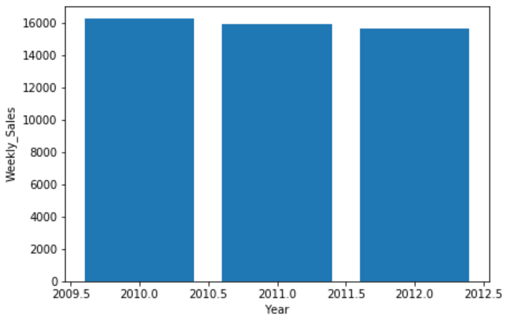


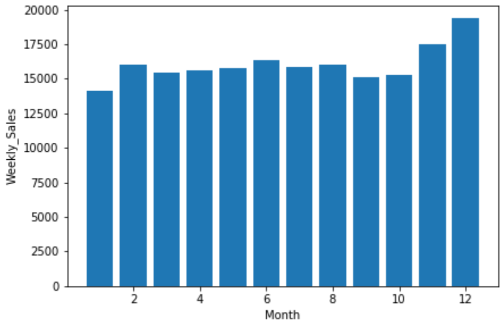


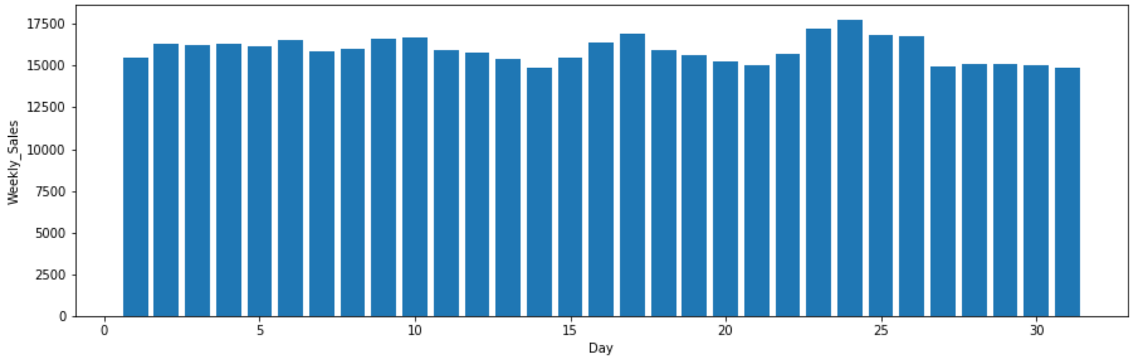
There is no direct correlation between the store number with the sales, but we categorized the store numbers into 4 different values. Number 1 assigned to stores had sales lower than 10000, 2 to stores had sales higher than 10000, but lower than 15000. Number 3 to stores had sales higher than 15000, but lower than 20000. We assigned 4 to stores had sales higher than 20000. After store numbers, we analyzed the department number feature. Also, there is no correlation between department number and sales.



We divided the departments into 4 groups by the sales numbers. First group is the departments had lower sales than 10000, second one sale between 10000 and 30000, and the third group had sales between 30000 and 50000. Last group had sales more than 50000. After department number, we investigated the date. For the day and year, we could not find clear trend, but in the November and December the sales slightly increase. We separated those months from the others as “Lucky Months.”

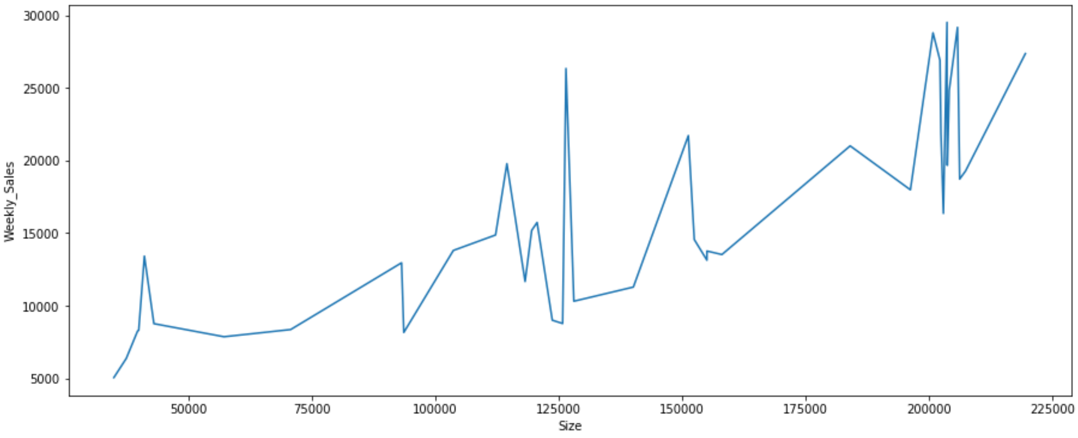




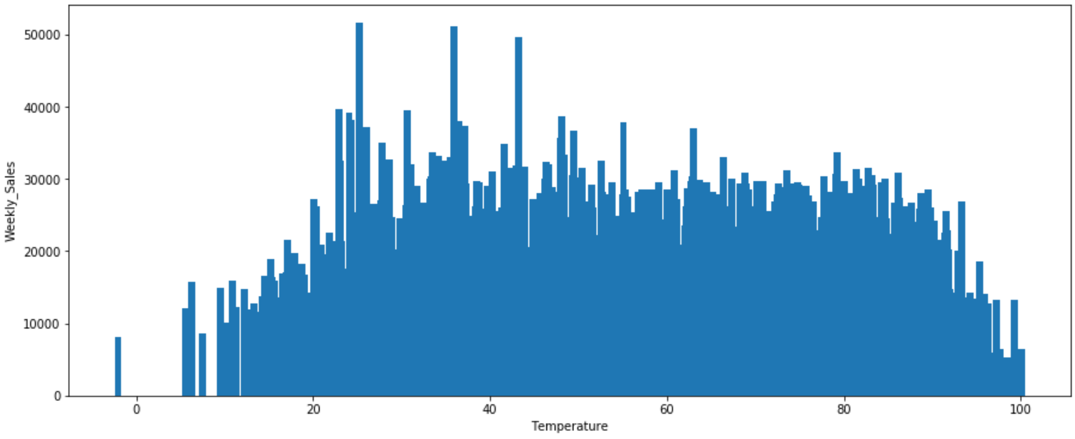


Also, there are special days that Walmart offers promotions during the year. Some of those days are Super Bowl, Labor Day, Thanksgiving, and Christmas. We assigned a new feature that separates those days from the normal days called special days.

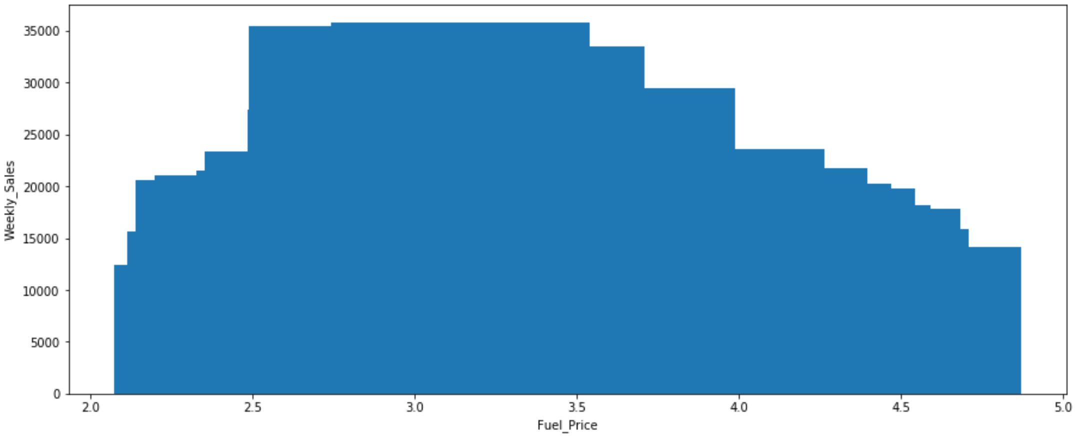
Size is another feature that affects the sales. Although there are some outliers, we can see a linear trend. When the size is increasing the sales numbers increases as well.



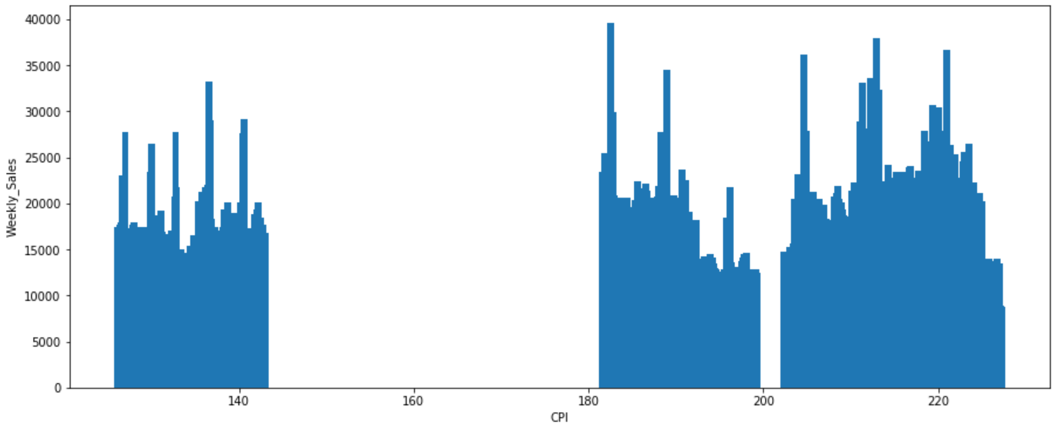
Then we moved to temperature in the region. We can observe lower sales at the numbers lower than 20 degree and higher than 80 degree. We divided the temperature into cold, warm, and hot groups.



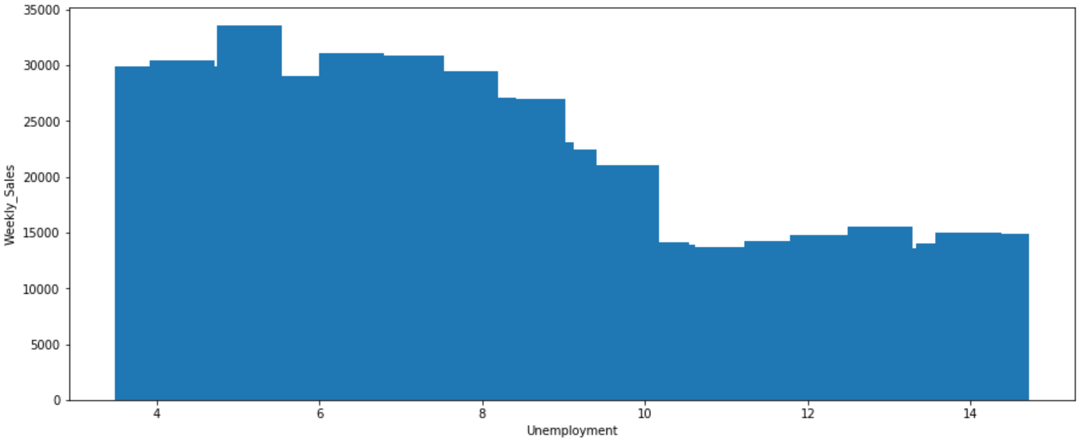
When we observed the fuel prices, we saw that there are some relatively higher sales numbers, between 2.5 and 4. We called the region between 2.5 and 4 as "Good Price." So we separated that part from others into two different groups.



When we looked to CPI numbers we could not see a lineer correlation with the sales, so we dropped it from the dataset.



Analyzing the unemployement rates showed us a clear correlation between the sales and unemployement rate. When the unemployement rate is higher the sales are lower.



Data set gives information about the promotional markdowns. If there is a promotion data set gives the exact promotion rate, but if there is no promotion the value is n/a. So, we dropped the n/a's and put zeroes to that blank cells. After all of those analyzes and categorizations, we used one hot encoding to implement categories to our model. Data splitted into two different groups as train and test with the rate of 80% to 20%.

4. Model Implementation

Our models trying to find the weekly sales of each store with our pre-processed data features. Store category, department category, date, fuel price, temperature, store type, and store size, promotions are the features that we used to forecast the sales. We believe that there are relations between those features and the weekly sales numbers. We used three different machine learning algorithms.

4.1 Random Forest Regressor

Random Forest is an ensemble method that consists of a large number of decision trees. It trains different parts of the same training set randomly and tries reducing the variance with averaging multiple deep decision trees. Algorithm also uses bagging. Bagging repeatedly selects a random sample with replacement of the training set and fit trees to these samples.

Averages of Random Forest, it can handle binary features, categorical features, and numerical features. It is also parallelizable, and prediction is faster than training because we can use generated forest for future uses. Random forest can handle outliers and unbalanced and n/a values. Disadvantages are size of trees can use a lot of memory in the large data sets and there would be an overfitting problem.

4.2 Ada Boost Regressor

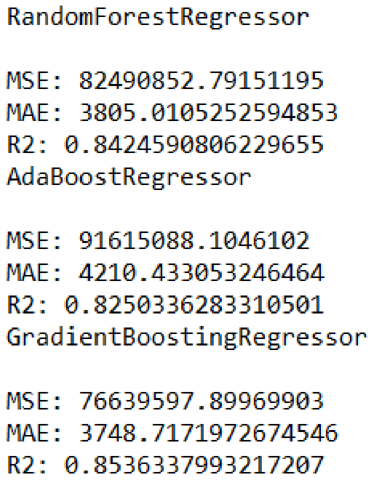
Ada Boost starts with a weak prediction model like decision tree. After evaluating the first tree, we increase the weights of those observations that are difficult to classify and lower the weights for those that are easy to classify. The second tree is therefore grown on this weighted data. So, we can improve our predictions. New model had tree one and tree two. Then the error computed from this ensemble model of both trees and grow a third tree to make prediction. This process repeated for a specified number of iterations. Predictions of the final ensemble model is therefore the weighted sum of the predictions made by the previous tree models. AdaBoost is sensitive to noisy data and outliers.

4.3 Gradient Boost Regressor

Gradient boost is another ensemble machine learning technique that produces a prediction model with weak prediction models. In Gradient Boost target is minimize the loss function. Gradient boost assumes Y and seeks approximation in the form of weighted sum of weak learners. In every iteration, it adds new tree and gives a weight according to loss function of whole model until minimizing the loss function. Choosing the best function to find the arbitrary loss is an optimization problem in general.

5. Outcome

We tried many models, but we decided to use these three models in our project. And the results are as follows:



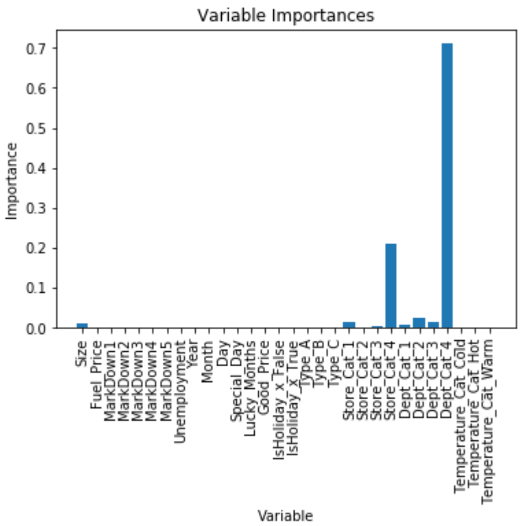
```
RandomForestRegressor

MSE: 82490852.79151195
MAE: 3805.0105252594853
R2: 0.8424590806229655
AdaBoostRegressor

MSE: 91615088.1046102
MAE: 4210.433053246464
R2: 0.8250336283310501
GradientBoostingRegressor

MSE: 76639597.89969903
MAE: 3748.7171972674546
R2: 0.8536337993217207
```

Our R2 score is very good. Almost 85 percent of the observed variation can be explained by the models input. We compared our results with the results in the Kaggle. Our MAE score with using Gradient Boosting Regressor would be the 339th score in the competition in 690 teams. We can say that our feature selection and categorization processes and our choice of the models are good. So, our predictions are better than most of the teams in the competition.



We also looked to importance of the features in the data set. Some features' importance values are higher than the others and some of them are zero. Temperature categories' importance numbers are zero, so we excluded them from the model and the results of the models did not change.

6. Conclusion

In this project, we analyzed data from 45 different Walmart Stores between the dates from 05-02-2010 to 2012-11-01. We first investigated the data and made some operations to increase the success of our model. With the categorizations and one hot encoding with 30 different features, we estimated the weekly sales with success. We learned about how to visualize and pre-process the data and chose the right features for the modeling. Also, we learned how to implement different models and chose the right models to use to forecast.