Jialu Yan, Tingting Gao, Yilin Wei
Advised by Dr. German Creamer, PhD, CFA
Dec. 11th, 2015

Forecasting Rossmann Store Sales Prediction

Problem Understanding

It is very important for retail stores to save money on their inventory and increase profit by meeting demands. Thus It will help a lot on stores' earnings if their sales are predictable. In our project, we chose Rossmann as our client, which is a quite big European drug store with 3000 locations. The company collected sales data for 1,115 Rossmann stores, including holiday, promotion, competitors and so on. The store manager is required to predict the 6 weeks of daily sales for these stores.

We chose Machine Learning algorithms to help find the precise forecast result, and to find out the importance of factors which may affect the store sales. Besides it will also help the company to understand their data and relationships between sales and other factors.

How precisely will a data mining solution help us to address this problem? First of all, the scale and range of the dataset is more than important. Second, the relationship between variable and target need to be understood. Third, the type of variables should be clear for further application.

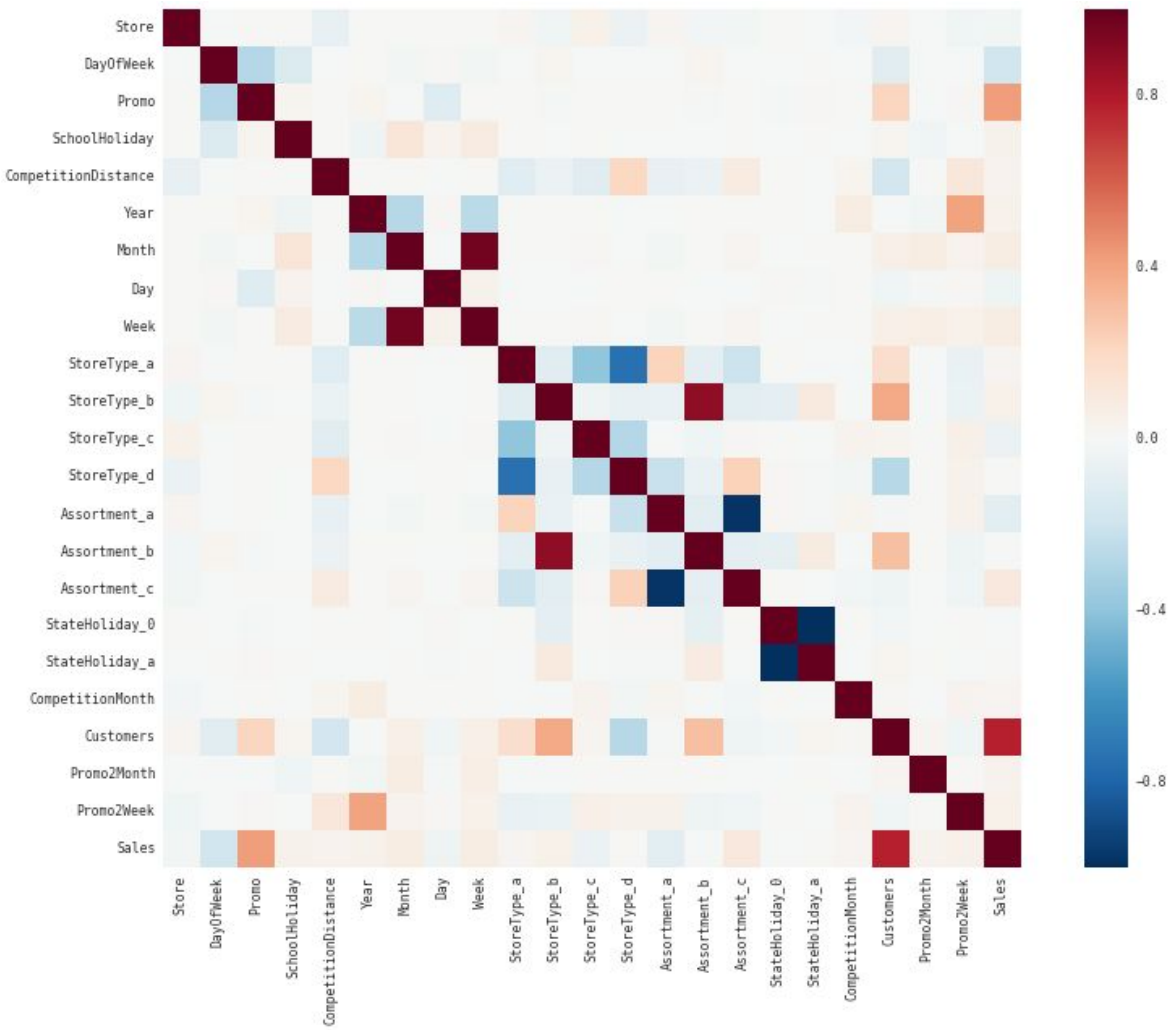
Scource: https://commons.wikimedia.org/wiki/File:Rossmann_Schriftzug_mit_Centaur.jpg

Data Understanding

The datasets we used in our project came from an on-going Kaggle competition. The original dataset contains two table, one is for each store's information (competitions and promotions), while the other one is for daily sales information based on each date (feature of the day and number of customers). Columns are described in following part (Source: Kaggle.com)

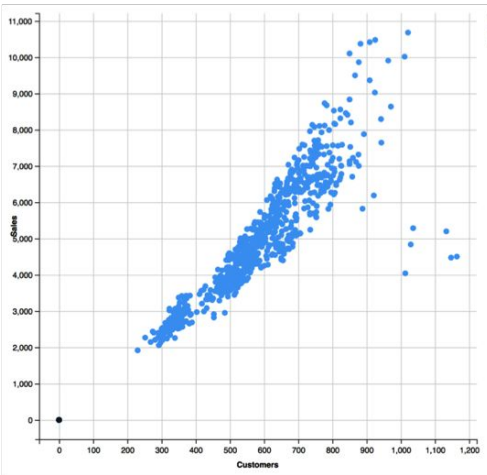
Correlation Matrix

The correlation matrix of data with Promo2 equal to 0 shows that Customers, Promo and StoreType_b positively relate to Sales, but DayofWeek and Assortment_a highly negatively relate to Sales.

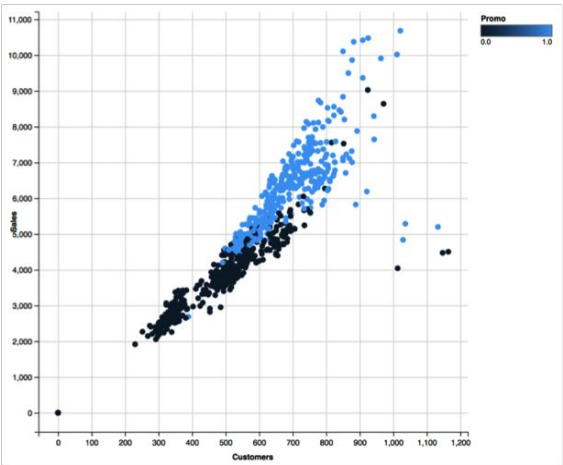


Relationship between Sales and Customers

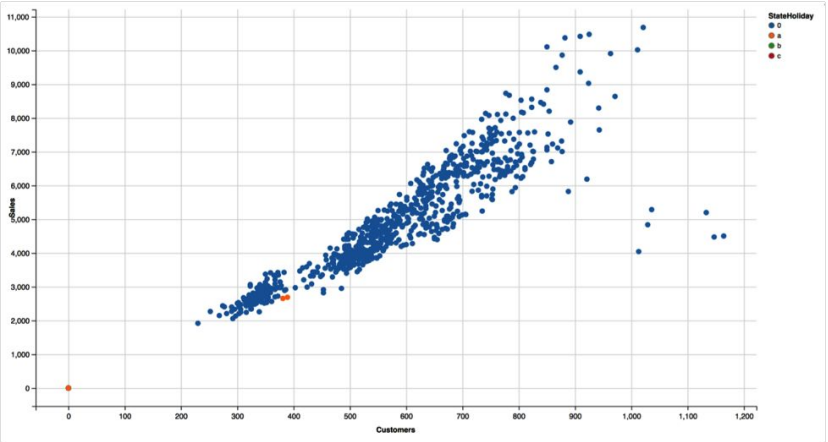
The scatter plots below gave us an overview on the relationship between Sales and Customers based on different situation (affected by different variables). Figure 1 showed that under general situation, Sales increase exponentially as the increase of Customer numbers. It indicates that customers have vital influence on the number of Sales. Figure 2 shows when there is promotion, it always attracts more customer, and get more profits even the customer numbers are the same. Figure 3 and Figure 4 here indicate that the relationships under StateHoliday and SchoolHoliday are not that obvious and needs further analysis.



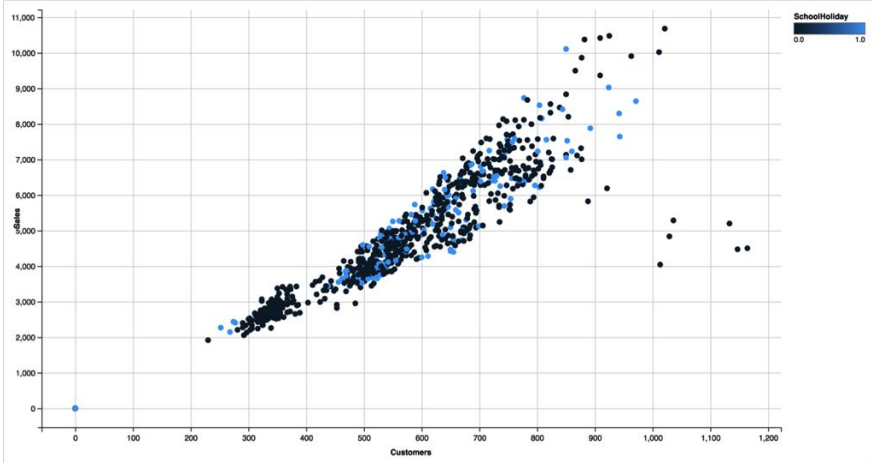
(1) Open



(2) With Promotion



(3) With StateHoliday



(4) With SchoolHoliday

Data Preparation

The goal of this step is to merge train and test containing time series information with store table and create features for prediction.

Create Features

We firstly consider what data should take into consideration to build model and make prediction. Since Open variable of some Kaggle test is missing, we fill all Nan of Open with 1. We drop records with open == 0 or Sales <=0. Because StateHoliday of Kaggle test is equal to 0 or a, we only keep train with StateHoliday == 0 or a. Then we replace Date with Year, Month, Day and Week. We create dummy variables for StoreType, Assortment and StateHoliday. We use HaveCompetitor to record whether certain store has competitors or not. We replace CompetitionOpenSinceYear with CompetitionMonth by calculating how many month from competitor opening to the day of sales record. Null value of this feature and CompetitionDistance is filled with mean. Promo2Week means how many weeks certain store have conducted long-term promotion. PromoInterval records the first month of each email marketing. We replace it with Promo2Month, which means how many month has the store conducted the most recent email promotion. For example, if a store send coupons via email in March, and a sales data is recorded in May, Promo2Month of this record is 2. We use this to build features for both train and test.

Split into Two Trains and Two Tests

Promo2 variable distinguish stores having Promo2 from stores not having Promo2. Stores conducting long-term promotion have Promo2SinceWeek and Promo2Interval. And this dataset is relatively large. Therefore, after creating features, we separately divide Kaggle train and Kaggle test into two based on whether Promo2 is equal to 0 or not. In this case, we need to build model for both trains to predict both tests. Then we combine two predicted results as our final predictions.

Predict Customers

Although our target variable is Sales, since we cannot know how many customers will come in future, Customers is also inaccessible in Kaggle test. Based on correlation matrix, Customers highly relate to Sales. We build models by two approached: not use Customers feature and prediction Sales based on Customers prediction. We use decision tree, KNN and random forest to predict Customers. We try to predict Customers with different periods of data: only use data of similar months (July, August and September) as test, use data of recent 3 month (May, June and July), use combination of previous two dataset.

The RMSPE is as follows:

Data	Similar months + recent 3 month			Similar months			Recent 3 months		
RMSPE	Promo2==0	Promo2==1	All	Promo2==0	Promo2==1	All	Promo2==0	Promo2==1	All
DT	0.1321	0.1370	0.1345	0.1151	0.1189	0.1170	0.1727	0.1650	0.1688
KNN	0.2679	0.2502	0.2594	0.2452	0.2339	0.2398	0.3026	0.2550	0.2793
RF	0.0897	0.0967	0.0932	0.0860	0.0893	0.0876	0.1253	0.1299	0.1277

Random forest has the lowest RMSPE, and using data of same month has the best result, so use data of same month with random forest makes a better prediction. We merge Customer prediction with original Kaggle test as our test.

Modeling

Models Comparison

We used cart (decision tree) first to predict. Because cart algorithm supports numerical target variables. It also supports continuous and discrete attributes. The target variable-"sales" is continuous. The attributes in the data are continuous and discrete. Cart is not very sensitive to outliers. It's resistant to irrelevant attributes. However, cart will generate too many leaf nodes and the model will be too complex. It will lead to overfit. Cart may have unstable trees. To avoid the problem of cart algorithm, we tried to use random forest algorithm. Based on our experience, the prediction accuracy of random forest is very high. It overcome the overfitting problem and handles large dataset very well. Random forest grows many decision trees. When a new input comes, each tree gives a classification and random forest chooses the classification having the most votes. Besides cart and random forest, we also tried to use K-nearest neighbors algorithm. K-nearest neighbors algorithm is easy understand. However, because the classification is based on similar attributes, irrelevant attributes may affect the prediction result a lot. So it's very important to remove irrelevant attributes first or weight them differently. If there are many attributes, it's hard to find the nearest neighbors for a new input. K-nearest neighbors is not very efficiently to run for a large data. The disadvantage of K-nearest neighbors explains the reason why the prediction accuracy is very low for our data. Because our data has so many attributes and the number of target variable - "Sales" is large, too.

Models Implementation

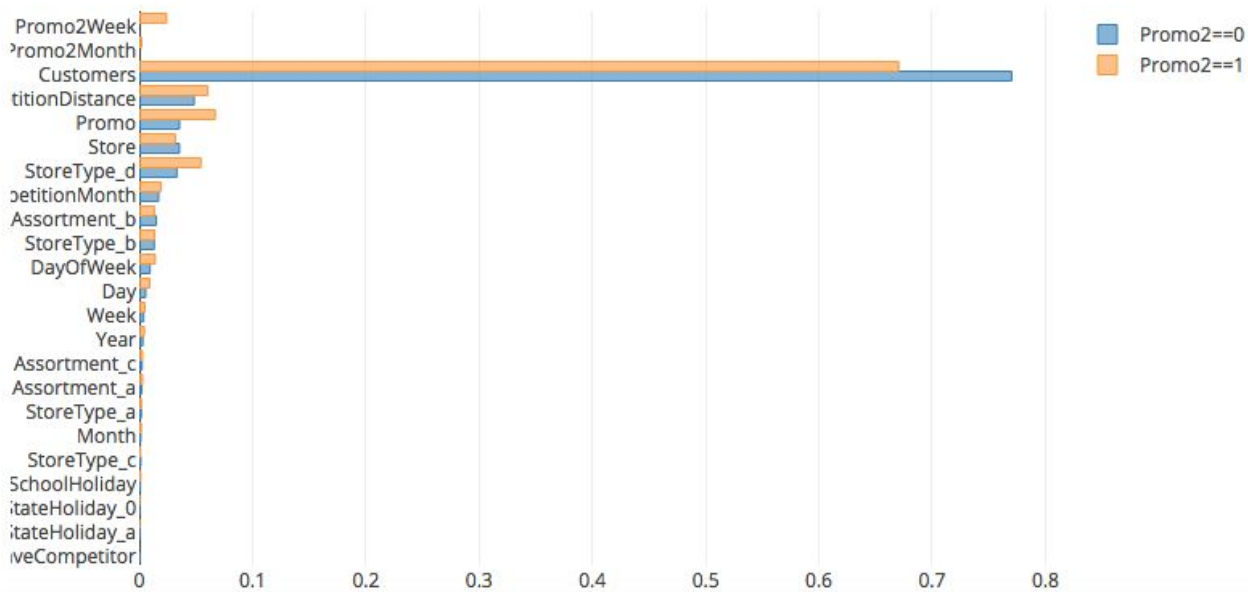
Predict Sales is similar to predict Customers. We also apply Decision Tree, KNN and Random Forest to three different periods of data: only use data of similar months (July, August and September) as test, use data of recent 3 month (May, June and July), use combination of previous two dataset. We do not use July 2015 data when predicting Sales with similar months data, which is different from predicting Customers,. The reason is that we get better result when only use July, August and September of 2013 and 2014.

The following table shows when we use similar months and recent 3 months data as train and apply random forest, we get the most accurate prediction.

Data	Similar months + recent 3 months			Similar months			Recent 3 months		
RMSPE	Promo2==0	Promo2==1	All	Promo2==0	Promo2==1	All	Promo2==0	Promo2==1	All
DT	0.1018	0.1266	0.1146	0.1036	0.1302	0.1171	0.1234	0.1345	0.1292
KNN	0.0820	0.0890	0.0855	0.0827	0.0901	0.0863	0.0871	0.0968	0.0922
RF	0.0702	0.0785	0.0744	0.0704	0.0800	0.0751	0.0862	0.0911	0.0887

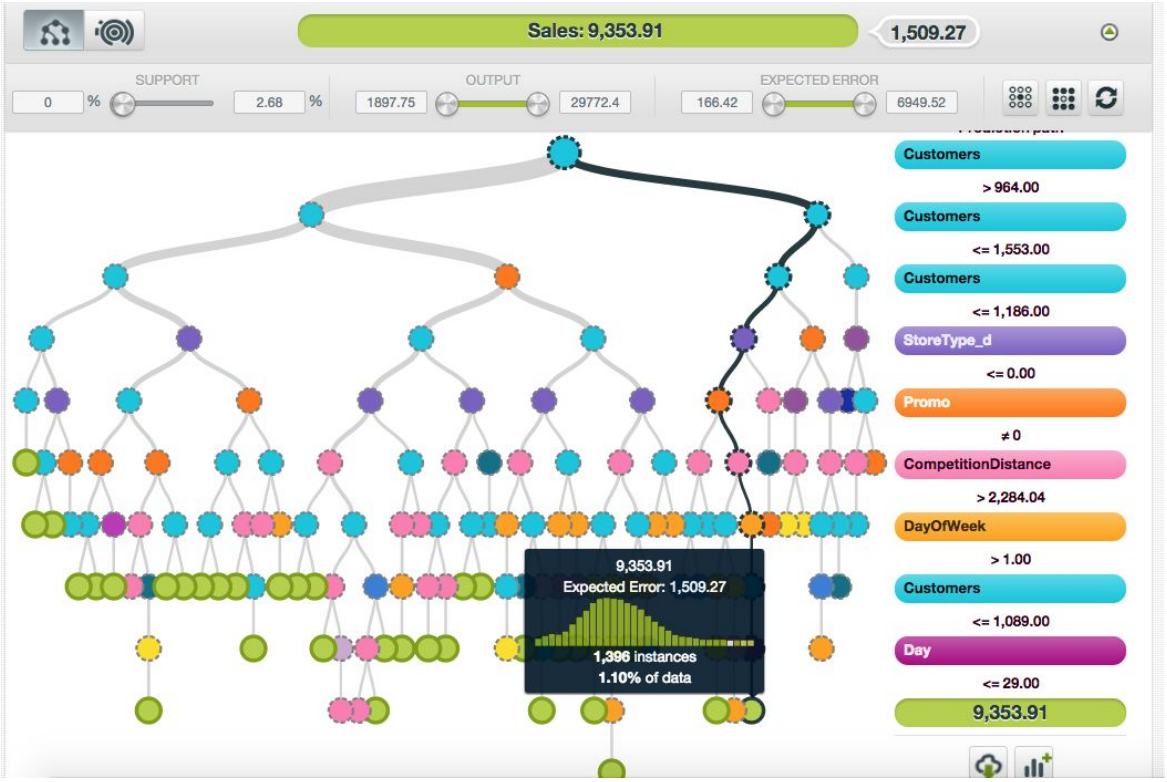
Feature importance of Random Forest

According to feature importance of random forest model, we find that Customers is highly important than other features. Therefore, the accuracy of Customers predicted by test data is very important. Top 5 important features to predict Sales with Promo2 equal to 0 is Customers, CompetitionDistance, Promo, Store and StoreType_d. Top 5 important features to predict Sales with Promo2 equal to 1 is Customers, Promo, CompetitonDistance, StoreType_b and Store.



Graph of Decision Tree

The following graph is a prediction for data with Promo2 equal to 0. Detailed rules is attached in appendix. The relative larger percentage of blue and pink nodes show that Customers and CompetitionDistance are important in this decision tree, which is similar to the conclusion of random forest.



The following two branches of decision tree explains when store is more far away from competitor, the sales will be higher.

Evaluation

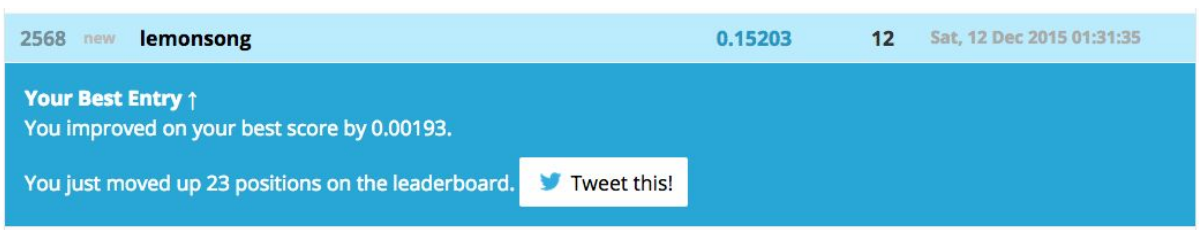
Model Evaluation

We evaluated the prediction result by comparing the prediction sales to the real sales. RMSPE is the evaluation standard. The lower the score, the better the prediction.

$$RMSPE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\hat{y}_i}{y_i}\right)^2}$$

The best prediction performance algorithm is random forest. When we use similar months and recent three months to predict the sales, the error rate of using random forest algorithm is only 0.0744. We think we can accept the error rate. The model is effective to predict the sales. We are confident that the model is valid and reliable.

The Kaggle rank floats frequently. Our Score of submission is 0.15203.



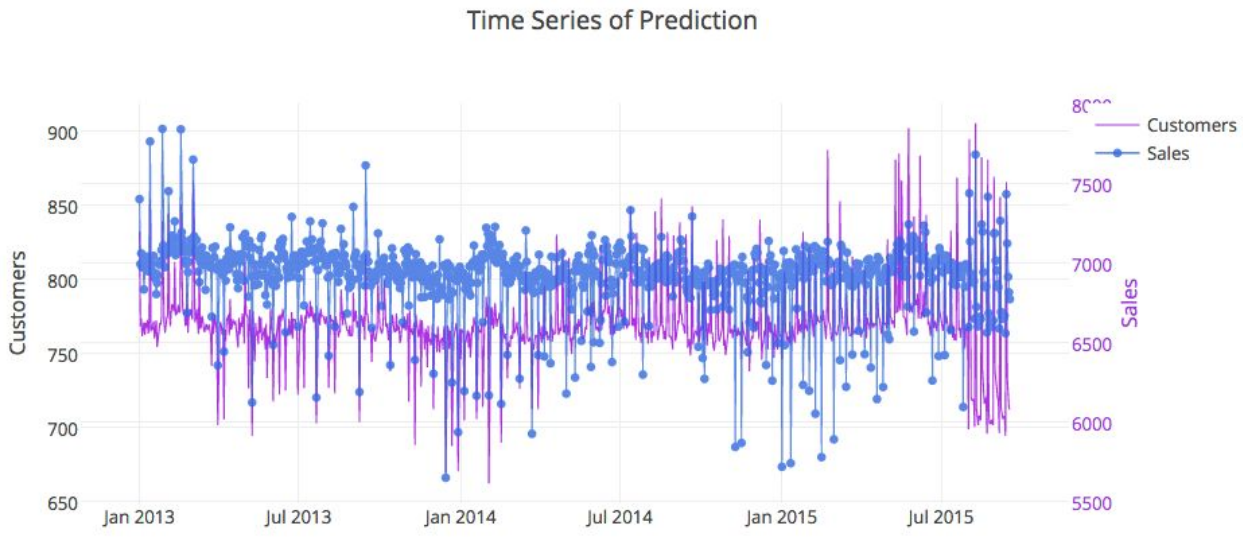
When we use similar months and recent three months to predict the sales, the result is better than using only recent months and only similar months. The result is also better than using all of dates. So time, especially similar month and recent time affects the prediction accuracy. Sales are also affected by time.

According to the predictive data, we find that the number of customers, short term promotions and store type B are positively related to the sales. So when the store type is B, and it has short term promotions and the number of customers is large, it's more likely that the store has good sales. This result is the same with the original data. Based on our knowledge and logic, more customers and short term promotions help increase the sales. So it's unlikely that the model makes catastrophic mistakes.

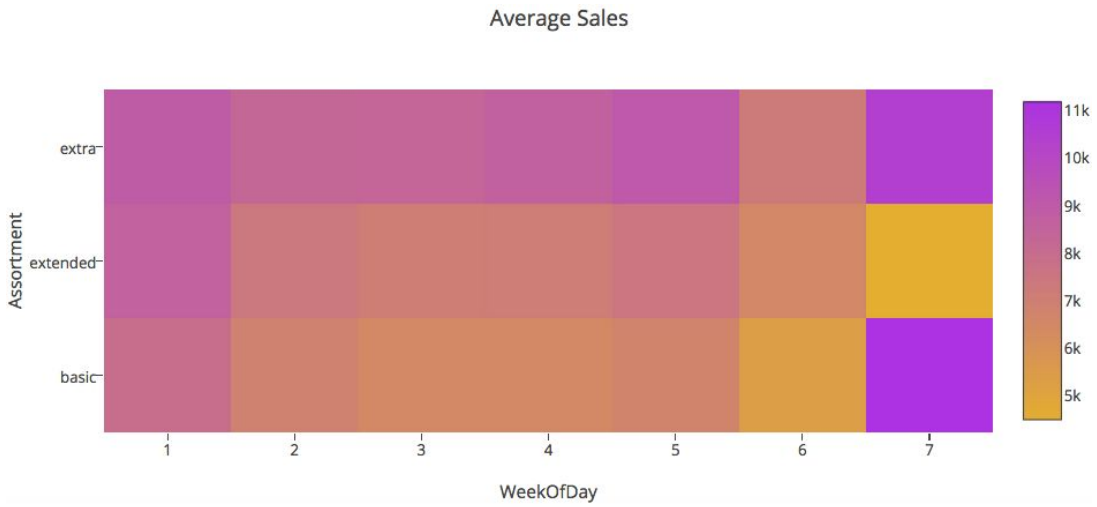
The difference between RMSPE of our calculation and that of Kaggle submission shows that usually the model has lower RMSPE for randomly splitted train and test, but it has higher RMSPE for Kaggle test dataset. In the read world, the model may become less effective because of the change of customers behaviors. We also need to do qualitative research to see which error rate can be accepted by the industry.

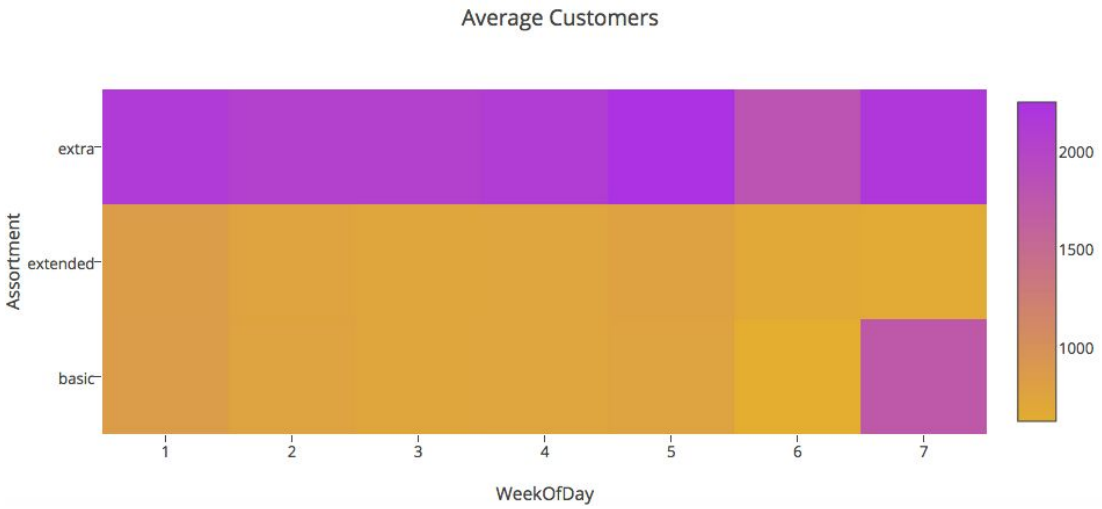
Business Meaning

The following time series graph of Customers and Sales reveals that the range of our prediction is larger than historical data. This may be caused by the lack of time series model.

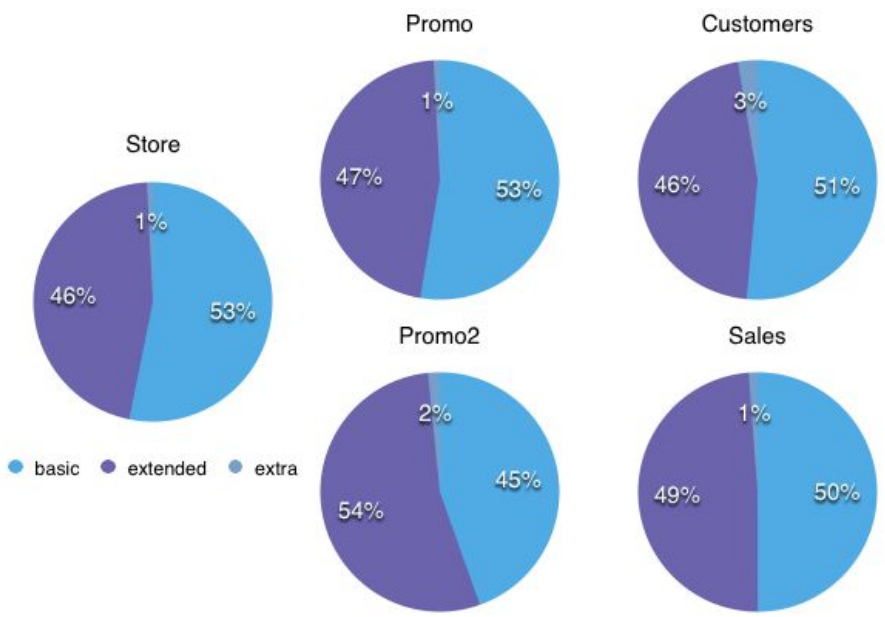


The following heatmaps present average Sales and Customers of each store on each weekday grouped by assortment of store. It shows that extra assortment stores have higher sales than other stores. All assortment type of stores have higher sales in Monday. In Sunday, while extra and basic stores have higher sales, extended stores have the lowest sales.





The following pie graphs reveals that basic stores conduct more short-term promotion, while extended stores conduct more email promotion. Comparing customers and sales graphs, we find that basic stores have more customers than extended stores do, while extended stores have higher sales. It demonstrates that the customers of extended stores are more profitable than that of basic stores. Extra stores take a very small percentage of all stores.



Deployment

Prediction of Sales

The stores can organize supply chain management and investment based on the sales prediction. When the number of customers prediction goes up, the sales will increase, which means the demand will also increase . The store needs to equip with enough products for the upcoming increasing demands. When the stores want to do short time promotion, they should prepare more supplies.

Deployment of human resources is similar. Based on heatmap of average customers, stores can arrange sellers appropriately.

New Store Opening

The result can also be used by people who are considering to open a new branch. The type of stores and the location can be considered based on the result. The store of type B has relatively good sales. There's a need to find the optimal location. The firm can choose some important factors like competition distance, store type, assortment type and competition time to do the experiment design and find which combination is the best. If the stores want to increase the

sales by promotion, basic stores should consider short time daily or weekly promotion instead of long term email promotion. Extended stores should consider long term email promotion.

Risk

If there are some errors when collecting the data, the whole model will be affected a lot. Thus, besides using the model prediction result, the firm should do some qualitative research as well. The firm can search for the successful drug store chain sales and analyze which attributes are important to sales and how they will affect the sales.

Customer Performance Understanding

In addition to the store sale forecasting and promotion arrangement, it is also recommended to understand customer performance based on the result. From dataset we may find a regular cycle of sales going up and down, neglecting holidays and sales. Thus it will also be very helpful to provide new event on the high peak, and avoid over-supply in lower valley of the cycle.

Appendix

Contribution

Jialu Yan	• Group meetings participation • Code: Data Visualization.R, Jialu_656project_2.py • Poster: Organize overall structure • Report: Problem Understanding; Data Understanding (brief introduction, Relationship between Sales and Customers); Deployment (Customer Performance Understanding); Double check with format.
Tingting Gao	• Group meetings participation • Code: Prediction with similar month • Poster: Prediction with similar month • Report: Modeling (models comparison), Evaluation (model evaluation), Deployment (prediction of sales, new store opening, risk)
Yilin Wei	• Group meetings participation • Code: visualizatio.ipynb, getdata.ipynb, predictcustomers.ipynb, predictionwtcustomers.ipynb • Poster • Report: Data Understanding (correlation matrix), Data Preparation, Modeling (models implementation, feature importance of random forest, graph of decision tree), Evaluation (business meaning)

Rules of Decision Tree

```
IF customers > 964 AND
    IF customers > 1553 AND
        IF customers > 2856 AND
            IF assortment_b > 0 AND
                IF customers > 3424 THEN
                    sales = 15579.66216
                IF customers <= 3424 THEN
                    sales = 13981.37234
            IF assortment_b <= 0 AND
                IF customers > 3530 AND
                    IF customers > 4216 THEN
```

```
            sales = 29772.39535
        IF customers <= 4216 AND
            IF promo = 0 THEN
                sales = 21540.85714
            IF promo != 0 AND
                IF store > 689 THEN
                    sales = 25586.51563
                IF store <= 689 THEN
                    sales = 22116.57692
    IF customers <= 3530 AND
        IF storetype_b > 0 AND
            IF customers > 3113 AND
                IF promo = 0 THEN
                    sales = 18179.24762
                IF promo != 0 THEN
                    sales = 19969.18382
            IF customers <= 3113 THEN
                sales = 16962.59184
        IF storetype_b <= 0 AND
            IF promo = 0 THEN
                sales = 19493.6
            IF promo != 0 AND
                IF year > 2014 THEN
                    sales = 24473.07692
                IF year <= 2014 THEN
                    sales = 22082.73276
IF customers <= 2856 AND
    IF storetype_b > 0 AND
        IF assortment_b > 0 AND
            IF customers > 2183 AND
                IF store > 812 AND
                    IF dayofweek > 6 THEN
                        sales = 11498.59524
                    IF dayofweek <= 6 THEN
                        sales = 9403.99367
                IF store <= 812 AND
```

```
            IF customers > 2483 THEN
                sales = 12647.59048
            IF customers <= 2483 THEN
                sales = 10875.26957
        IF customers <= 2183 AND
            IF customers > 1941 THEN
                sales = 8593.72081
            IF customers <= 1941 THEN
                sales = 7209.07692
    IF assortment_b <= 0 AND
        IF customers > 2000 AND
            IF customers > 2682 THEN
                sales = 15317.67925
            IF customers <= 2682 AND
                IF competitiondistance > 680 AND
                    IF dayofweek > 6 THEN
                        sales = 14329.54839
                    IF dayofweek <= 6 THEN
                        sales = 12109.07563
                IF competitiondistance <= 680 THEN
                    sales = 14565.66038
        IF customers <= 2000 AND
            IF promo = 0 AND
                IF competitiondistance > 1570 THEN
                    sales = 12402
                IF competitiondistance <= 1570 THEN
                    sales = 9748.61017
            IF promo != 0 THEN
                sales = 11736.91304
IF storetype_b <= 0 AND
    IF customers > 1964 AND
        IF promo = 0 AND
            IF competitiondistance > 295 AND
                IF day > 29 THEN
                    sales = 19168.65217
                IF day <= 29 THEN
```

```
            sales = 16311.91971
      IF competitiondistance <= 295 AND
         IF customers > 2254 AND
            IF competitiondistance > 60 THEN
               sales = 15905.21348
            IF competitiondistance <= 60 THEN
               sales = 19342.23077
         IF customers <= 2254 AND
            IF competitionmonth > -1 THEN
               sales = 13338.28571
            IF competitionmonth <= -1 THEN
               sales = 17599.05882
   IF promo != 0 AND
      IF competitiondistance > 295 AND
         IF customers > 2431 THEN
            sales = 21928.90099
         IF customers <= 2431 AND
            IF store > 733 AND
               IF competitionmonth > 50.5 THEN
                  sales = 18849.32
               IF competitionmonth <= 50.5 THEN
                  sales = 22847.63636
            IF store <= 733 AND
               IF customers > 2124 AND
                  IF store > 468 THEN
                     sales = 20402.72619
                  IF store <= 468 THEN
                     sales = 18525.65812
               IF customers <= 2124 THEN
                  sales = 17931.14103
      IF competitiondistance <= 295 AND
         IF customers > 2289 AND
            IF dayofweek > 1 THEN
               sales = 17723.68889
            IF dayofweek <= 1 THEN
               sales = 20013.08
```

```
            IF customers <= 2289 AND
                IF storetype_a > 0 AND
                    IF dayofweek > 2 THEN
                        sales = 15206.5
                    IF dayofweek <= 2 THEN
                        sales = 16747.06742
                IF storetype_a <= 0 THEN
                    sales = 18437.7
IF customers <= 1964 AND
    IF competitiondistance > 952.71951 AND
        IF store > 712 AND
            IF storetype_d > 0 THEN
                sales = 29158.33333
            IF storetype_d <= 0 AND
                IF customers > 1725 THEN
                    sales = 19325.96907
                IF customers <= 1725 AND
                    IF competitionmonth > -7.66667 AND
                        IF promo = 0 THEN
                            sales = 15635.25
                        IF promo != 0 THEN
                            sales = 17683.86747
                    IF competitionmonth <= -7.66667 THEN
                        sales = 11355.66667
        IF store <= 712 AND
            IF promo = 0 AND
                IF customers > 1719 THEN
                    sales = 14401.4
                IF customers <= 1719 THEN
                    sales = 12817.61702
            IF promo != 0 AND
                IF customers > 1708 AND
                    IF store > 386 AND
                        IF customers > 1839 THEN
                            sales = 17024.01299
                        IF customers <= 1839 THEN
```

```
                    sales = 15385.14173
            IF store <= 386 THEN
                sales = 20286.87097
        IF customers <= 1708 AND
            IF store > 406 THEN
                sales = 13890.91398
            IF store <= 406 THEN
                sales = 16052.85366
IF competitiondistance <= 952.71951 AND
    IF customers > 1766 AND
        IF competitiondistance > 345 AND
            IF competitiondistance > 805 THEN
                sales = 13583.42553
            IF competitiondistance <= 805 THEN
                sales = 15519.08014
        IF competitiondistance <= 345 AND
            IF storetype_a > 0 AND
                IF promo = 0 THEN
                    sales = 12198.54595
                IF promo != 0 THEN
                    sales = 13966.74863
            IF storetype_a <= 0 THEN
                sales = 15892.725
    IF customers <= 1766 AND
        IF promo = 0 AND
            IF competitiondistance > 245 AND
                IF competitiondistance > 805 THEN
                    sales = 10591.4898
                IF competitiondistance <= 805 THEN
                    sales = 12536.02674
            IF competitiondistance <= 245 AND
                IF storetype_c > 0 THEN
                    sales = 13421.89189
                IF storetype_c <= 0 THEN
                    sales = 10528.53712
        IF promo != 0 AND
```

```
                IF dayofweek > 1 AND
                    IF competitionmonth > 78.60294 THEN
                        sales = 13522.49407
                    IF competitionmonth <= 78.60294 AND
                        IF storetype_c > 0 THEN
                            sales = 14546.66667
                        IF storetype_c <= 0 THEN
                            sales = 11951.51923
                IF dayofweek <= 1 THEN
                    sales = 14343.28378
IF customers <= 1553 AND
    IF customers > 1186 AND
        IF promo = 0 AND
            IF storetype_b > 0 AND
                IF assortment_a > 0 THEN
                    sales = 7740.58716
                IF assortment_a <= 0 THEN
                    sales = 5436.79808
            IF storetype_b <= 0 AND
                IF competitiondistance > 687.62689 AND
                    IF store > 687 AND
                        IF storetype_d > 0 AND
                            IF competitiondistance > 1555 THEN
                                sales = 11677.61111
                            IF competitiondistance <= 1555 THEN
                                sales = 20863.82353
                        IF storetype_d <= 0 AND
                            IF store > 809 AND
                                IF store > 953 THEN
                                    sales = 11608.40816
                                IF store <= 953 AND
                                    IF storetype_c > 0 THEN
                                        sales = 7679.41379
                                    IF storetype_c <= 0 THEN
                                        sales = 10007.43885
                            IF store <= 809 AND
```

```
                IF store > 787 THEN
                    sales = 14529.92857
                IF store <= 787 THEN
                    sales = 11391.24706
            IF store <= 687 AND
                IF customers > 1411 THEN
                    sales = 11091.78704
                IF customers <= 1411 AND
                    IF store > 104 AND
                        IF customers > 1278 THEN
                            sales = 10015.41732
                        IF customers <= 1278 THEN
                            sales = 9230.45985
                    IF store <= 104 THEN
                        sales = 10914.31183
        IF competitiondistance <= 687.62689 AND
            IF store > 563 AND
                IF store > 947 THEN
                    sales = 10091.9901
                IF store <= 947 AND
                    IF customers > 1389 THEN
                        sales = 9192.16456
                    IF customers <= 1389 THEN
                        sales = 8024.98045
            IF store <= 563 AND
                IF customers > 1296 AND
                    IF store > 491 THEN
                        sales = 11613.725
                    IF store <= 491 THEN
                        sales = 10251.22286
                IF customers <= 1296 THEN
                    sales = 8974.71239
IF promo != 0 AND
    IF storetype_d > 0 AND
        IF store > 827 AND
            IF store > 860 THEN
```

```
            sales = 14382.44643
        IF store <= 860 AND
            IF customers > 1447 THEN
                sales = 26328.5
            IF customers <= 1447 THEN
                sales = 21289.42424
    IF store <= 827 AND
        IF store > 262 AND
            IF competitiondistance > 900 THEN
                sales = 13786.41818
            IF competitiondistance <= 900 THEN
                sales = 11309.13483
        IF store <= 262 AND
            IF customers > 1338 THEN
                sales = 17572.17241
            IF customers <= 1338 THEN
                sales = 14422.19481
IF storetype_d <= 0 AND
    IF competitiondistance > 2000.13058 AND
        IF customers > 1367 AND
            IF store > 636 AND
                IF customers > 1479 THEN
                    sales = 16571.95161
                IF customers <= 1479 THEN
                    sales = 14587.94857
            IF store <= 636 AND
                IF store > 406 THEN
                    sales = 11838.4031
                IF store <= 406 THEN
                    sales = 13491.95681
        IF customers <= 1367 AND
            IF competitionmonth > 83.77778 THEN
                sales = 13264.05911
            IF competitionmonth <= 83.77778 AND
                IF dayofweek > 1 AND
                    IF customers > 1301 THEN
```

```
                    sales = 11982.43038
                IF customers <= 1301 THEN
                    sales = 10899.75092
            IF dayofweek <= 1 THEN
                sales = 12434.48358
        IF competitiondistance <= 2000.13058 AND
            IF customers > 1348 AND
                IF assortment_b > 0 THEN
                    sales = 5909.9697
                IF assortment_b <= 0 AND
                    IF store > 1091 THEN
                        sales = 21122.75
                    IF store <= 1091 THEN
                        sales = 11839.30524
            IF customers <= 1348 AND
                IF assortment_c > 0 AND
                    IF competitiondistance > 1350 AND
                        IF competitiondistance > 1605 THEN
                            sales = 11050.85294
                        IF competitiondistance <= 1605 THEN
                            sales = 13900.05634
                    IF competitiondistance <= 1350 THEN
                        sales = 10788.79405
                IF assortment_c <= 0 AND
                    IF dayofweek > 1 AND
                        IF store > 479 AND
                            IF competitiondistance > 1680 THEN
                                sales = 11953.75
                            IF competitiondistance <= 1680 THEN
                                sales = 9539.06302
                        IF store <= 479 THEN
                            sales = 10419.28857
                    IF dayofweek <= 1 THEN
                        sales = 10898.05155
IF customers <= 1186 AND
    IF storetype_d > 0 AND
```

```
IF competitiondistance > 904.60747 AND
    IF competitionmonth > 66.34859 AND
        IF competitiondistance > 1555 AND
            IF customers > 1048 AND
                IF competitionmonth > 103.5 THEN
                    sales = 14045
                IF competitionmonth <= 103.5 THEN
                    sales = 11971
            IF customers <= 1048 THEN
                sales = 11378.85567
        IF competitiondistance <= 1555 AND
            IF customers > 1100 THEN
                sales = 19175.08108
            IF customers <= 1100 THEN
                sales = 15961.48889
    IF competitionmonth <= 66.34859 AND
        IF promo = 0 THEN
            sales = 9590.51701
        IF promo != 0 AND
            IF dayofweek > 1 AND
                IF customers > 1051 THEN
                    sales = 11620.77258
                IF customers <= 1051 THEN
                    sales = 10500.8907
            IF dayofweek <= 1 THEN
                sales = 12239.58191
IF competitiondistance <= 904.60747 AND
    IF competitionmonth > 13.5 THEN
        sales = 7958.01515
    IF competitionmonth <= 13.5 AND
        IF competitiondistance > 510 THEN
            sales = 11932.16327
        IF competitiondistance <= 510 AND
            IF competitiondistance > 430 THEN
                sales = 7336.19355
            IF competitiondistance <= 430 THEN
```

```
                sales = 11293.98148
IF storetype_d <= 0 AND
   IF promo = 0 AND
      IF competitiondistance > 2333.19073 AND
         IF customers > 1031 AND
            IF store > 611 THEN
               sales = 9411.25731
            IF store <= 611 AND
               IF store > 527 THEN
                  sales = 7415.76471
               IF store <= 527 THEN
                  sales = 8698.19291
         IF customers <= 1031 THEN
            sales = 8190.81211
      IF competitiondistance <= 2333.19073 AND
         IF customers > 1068 AND
            IF assortment_c > 0 AND
               IF store > 1072 THEN
                  sales = 11652.16667
               IF store <= 1072 THEN
                  sales = 8572.58654
            IF assortment_c <= 0 THEN
               sales = 7919.74368
         IF customers <= 1068 AND
            IF assortment_c > 0 AND
               IF store > 1069 THEN
                  sales = 10123.69841
               IF store <= 1069 AND
                  IF competitionmonth > 34.6614 THEN
                     sales = 8097.1134
                  IF competitionmonth <= 34.6614 THEN
                     sales = 6615.33735
            IF assortment_c <= 0 THEN
               sales = 7208.45698
   IF promo != 0 AND
      IF competitiondistance > 2284.03691 AND
```

```
    IF dayofweek > 1 AND
        IF customers > 1089 THEN
            sales = 10444.34091
        IF customers <= 1089 AND
            IF day > 29 THEN
                sales = 10347.81356
            IF day <= 29 THEN
                sales = 9353.91117
    IF dayofweek <= 1 THEN
        sales = 11111.45227
IF competitiondistance <= 2284.03691 AND
    IF customers > 1071 AND
        IF store > 1094 THEN
            sales = 15257.92857
        IF store <= 1094 AND
            IF competitiondistance > 1287.65447 AND
                IF competitiondistance > 1605 AND
                    IF competitionmonth > 58.47048 THEN
                        sales = 10299.74708
                    IF competitionmonth <= 58.47048 THEN
                        sales = 9037.64394
                IF competitiondistance <= 1605 AND
                    IF store > 358 THEN
                        sales = 11877.33884
                    IF store <= 358 THEN
                        sales = 10008.71765
            IF competitiondistance <= 1287.65447 AND
                IF dayofweek > 1 THEN
                    sales = 9021.97645
                IF dayofweek <= 1 THEN
                    sales = 9886.36286
    IF customers <= 1071 AND
        IF dayofweek > 2 AND
            IF store > 1095 THEN
                sales = 13035.88889
            IF store <= 1095 THEN
```

```
                    sales = 8241.13118
                IF dayofweek <= 2 THEN
                    sales = 8988.46642
IF customers <= 964 AND
    IF customers > 612 AND
        IF promo = 0 AND
            IF customers > 794 AND
                IF storetype_d > 0 AND
                    IF competitionmonth > 66.37727 AND
                        IF competitiondistance > 2405 THEN
                            sales = 8730.84566
                        IF competitiondistance <= 2405 AND
                            IF competitiondistance > 1165 AND
                                IF competitiondistance > 1350 THEN
                                    sales = 10202.57813
                                IF competitiondistance <= 1350 THEN
                                    sales = 14224.57143
                            IF competitiondistance <= 1165 THEN
                                sales = 6986.6
                    IF competitionmonth <= 66.37727 AND
                        IF customers > 867 THEN
                            sales = 8493.67526
                        IF customers <= 867 THEN
                            sales = 7760.62113
                IF storetype_d <= 0 AND
                    IF competitiondistance > 2328.93423 AND
                        IF customers > 863 THEN
                            sales = 7545.12966
                        IF customers <= 863 THEN
                            sales = 6951.70756
                    IF competitiondistance <= 2328.93423 AND
                        IF customers > 874 AND
                            IF competitiondistance > 1290.53585 AND
                                IF competitiondistance > 1605 AND
                                    IF competitionmonth > 58.32445 THEN
                                        sales = 7347.98294
```

```
                IF competitionmonth <= 58.32445 THEN
                    sales = 6387.69811
            IF competitiondistance <= 1605 THEN
                sales = 8306.75294
        IF competitiondistance <= 1290.53585 THEN
            sales = 6552.12801
    IF customers <= 874 AND
        IF competitiondistance > 910 AND
            IF store > 1105 THEN
                sales = 9383.28571
            IF store <= 1105 THEN
                sales = 6397.10543
        IF competitiondistance <= 910 THEN
            sales = 5984.96648
IF customers <= 794 AND
    IF storetype_d > 0 AND
        IF customers > 708 AND
            IF competitiondistance > 6927.87879 AND
                IF competitiondistance > 8170.08197 THEN
                    sales = 7484.62524
                IF competitiondistance <= 8170.08197 THEN
                    sales = 9510.11628
            IF competitiondistance <= 6927.87879 AND
                IF competitionmonth > 84.05769 THEN
                    sales = 8192.71028
                IF competitionmonth <= 84.05769 THEN
                    sales = 6952.06431
        IF customers <= 708 AND
            IF competitiondistance > 1630.40157 AND
                IF dayofweek > 5 THEN
                    sales = 7088.78101
                IF dayofweek <= 5 THEN
                    sales = 6457.76176
            IF competitiondistance <= 1630.40157 THEN
                sales = 5864.70306
    IF storetype_d <= 0 AND
```

```
        IF competitiondistance > 973.4108 AND
            IF customers > 691 AND
                IF assortment_c > 0 THEN
                    sales = 6446.75894
                IF assortment_c <= 0 AND
                    IF competitiondistance > 2184.4355 THEN
                        sales = 6144.82927
                    IF competitiondistance <= 2184.4355 AND
                        IF competitiondistance > 2075 THEN
                            sales = 7214.32911
                        IF competitiondistance <= 2075 THEN
                            sales = 5581.14602
            IF customers <= 691 AND
                IF competitiondistance > 6527.16615 AND
                    IF storetype_c > 0 THEN
                        sales = 6405.81818
                    IF storetype_c <= 0 AND
                        IF competitiondistance > 17277.2619 THEN
                            sales = 5245.8526
                        IF competitiondistance <= 17277.2619 THEN
                            sales = 5900.64766
                IF competitiondistance <= 6527.16615 THEN
                    sales = 5347.93706
        IF competitiondistance <= 973.4108 AND
            IF customers > 701 THEN
                sales = 5361.76084
            IF customers <= 701 THEN
                sales = 4713.68188
IF promo != 0 AND
    IF customers > 771 AND
        IF storetype_d > 0 AND
            IF customers > 868 AND
                IF dayofweek > 1 AND
                    IF competitionmonth > 66.33889 THEN
                        sales = 10771.98404
                    IF competitionmonth <= 66.33889 AND
```

```
            IF competitiondistance > 6945.43478 AND
                IF competitiondistance > 8495 THEN
                    sales = 9304.67516
                IF competitiondistance <= 8495 THEN
                    sales = 12865.13333
            IF competitiondistance <= 6945.43478 AND
                IF competitiondistance > 6620 THEN
                    sales = 7613.43243
                IF competitiondistance <= 6620 THEN
                    sales = 9484.16614
        IF dayofweek <= 1 AND
            IF competitiondistance > 848.37302 THEN
                sales = 11191.953
            IF competitiondistance <= 848.37302 THEN
                sales = 8281.09524
    IF customers <= 868 AND
        IF dayofweek > 2 THEN
            sales = 8769.82852
        IF dayofweek <= 2 THEN
            sales = 9596.52935
IF storetype_d <= 0 AND
    IF competitiondistance > 2322.91485 AND
        IF customers > 857 AND
            IF dayofweek > 2 AND
                IF assortment_c > 0 THEN
                    sales = 8919.30907
                IF assortment_c <= 0 THEN
                    sales = 8309.19382
            IF dayofweek <= 2 THEN
                sales = 9469.17229
        IF customers <= 857 AND
            IF dayofweek > 2 THEN
                sales = 7843.11306
            IF dayofweek <= 2 THEN
                sales = 8462.60973
    IF competitiondistance <= 2322.91485 AND
```

```
        IF customers > 853 AND
            IF store > 1088 AND
                IF store > 1103 THEN
                    sales = 11444.19149
                IF store <= 1103 THEN
                    sales = 7327.80952
            IF store <= 1088 AND
                IF store > 169 AND
                    IF dayofweek > 2 THEN
                        sales = 7292.05623
                    IF dayofweek <= 2 THEN
                        sales = 7847.02792
                IF store <= 169 AND
                    IF assortment_c > 0 THEN
                        sales = 9829.8625
                    IF assortment_c <= 0 AND
                        IF store > 92 THEN
                            sales = 8761.1236
                        IF store <= 92 THEN
                            sales = 7478.09845
        IF customers <= 853 AND
            IF store > 404 AND
                IF store > 1103 THEN
                    sales = 9347.07143
                IF store <= 1103 THEN
                    sales = 6542.9645
            IF store <= 404 AND
                IF competitiondistance > 1269.86486 THEN
                    sales = 7912.83368
                IF competitiondistance <= 1269.86486 THEN
                    sales = 7055.07662
IF customers <= 771 AND
    IF storetype_d > 0 AND
        IF customers > 689 AND
            IF dayofweek > 1 AND
                IF competitiondistance > 3092.11745 THEN
```

```
            sales = 8383.84039
         IF competitiondistance <= 3092.11745 THEN
            sales = 7687.42336
      IF dayofweek <= 1 THEN
         sales = 9002.71903
   IF customers <= 689 AND
      IF dayofweek > 1 AND
         IF competitiondistance > 3944.63886 THEN
            sales = 7490.84263
         IF competitiondistance <= 3944.63886 AND
            IF store > 240 THEN
               sales = 6725.00173
            IF store <= 240 THEN
               sales = 7770.49223
      IF dayofweek <= 1 THEN
         sales = 8219.27708
IF storetype_d <= 0 AND
   IF competitiondistance > 1739.93022 AND
      IF customers > 681 AND
         IF year > 2014 THEN
            sales = 7535.83379
         IF year <= 2014 AND
            IF customers > 725 THEN
               sales = 7214.9635
            IF customers <= 725 THEN
               sales = 6759.45013
      IF customers <= 681 AND
         IF competitionmonth > 124.35413 THEN
            sales = 7131.30233
         IF competitionmonth <= 124.35413 THEN
            sales = 6240.8723
   IF competitiondistance <= 1739.93022 AND
      IF dayofweek > 1 AND
         IF customers > 693 AND
            IF store > 272 THEN
               sales = 5778.08259
```

```
                IF store <= 272 THEN
                    sales = 6562.98722
            IF customers <= 693 THEN
                sales = 5370.65315
        IF dayofweek <= 1 THEN
            sales = 6534.54673
IF customers <= 612 AND
    IF customers > 436 AND
        IF storetype_d > 0 AND
            IF promo = 0 AND
                IF customers > 520 AND
                    IF competitiondistance > 3911.73532 AND
                        IF customers > 558 THEN
                            sales = 6071.19739
                        IF customers <= 558 THEN
                            sales = 5594.12104
                    IF competitiondistance <= 3911.73532 THEN
                        sales = 5396.63384
                IF customers <= 520 AND
                    IF customers > 478 THEN
                        sales = 5091.70099
                    IF customers <= 478 THEN
                        sales = 4661.62917
            IF promo != 0 AND
                IF customers > 520 AND
                    IF dayofweek > 2 THEN
                        sales = 6498.5379
                    IF dayofweek <= 2 THEN
                        sales = 6962.59586
                IF customers <= 520 AND
                    IF competitiondistance > 3492.14286 THEN
                        sales = 5975.60412
                    IF competitiondistance <= 3492.14286 AND
                        IF competitiondistance > 1400 AND
                            IF competitiondistance > 1855 THEN
                                sales = 5597.81481
```

```
                IF competitiondistance <= 1855 THEN
                    sales = 3679.18803
            IF competitiondistance <= 1400 THEN
                sales = 5823.62222
IF storetype_d <= 0 AND
    IF customers > 526 AND
        IF promo = 0 AND
            IF competitiondistance > 1722.22425 AND
                IF competitionmonth > 112.70021 THEN
                    sales = 5473.69509
                IF competitionmonth <= 112.70021 AND
                    IF store > 1076 THEN
                        sales = 5788.50658
                    IF store <= 1076 THEN
                        sales = 4829.02004
            IF competitiondistance <= 1722.22425 AND
                IF competitiondistance > 657.29301 THEN
                    sales = 4603.1449
                IF competitiondistance <= 657.29301 THEN
                    sales = 4095.63514
        IF promo != 0 AND
            IF customers > 565 THEN
                sales = 5542.90269
            IF customers <= 565 THEN
                sales = 5068.32579
    IF customers <= 526 AND
        IF promo = 0 AND
            IF customers > 487 THEN
                sales = 4243.55196
            IF customers <= 487 THEN
                sales = 3913.35552
        IF promo != 0 AND
            IF storetype_a > 0 THEN
                sales = 4640.85406
            IF storetype_a <= 0 THEN
                sales = 3791.92537
```

```
IF customers <= 436 AND
    IF customers > 338 AND
        IF storetype_d > 0 AND
            IF promo = 0 AND
                IF customers > 395 THEN
                    sales = 4254.7401
                IF customers <= 395 THEN
                    sales = 3762.29647
            IF promo != 0 THEN
                sales = 4922.38889
        IF storetype_d <= 0 AND
            IF customers > 385 THEN
                sales = 3642.70115
            IF customers <= 385 THEN
                sales = 3148.20528
    IF customers <= 338 AND
        IF customers > 258 THEN
            sales = 2775.6074
        IF customers <= 258 THEN
            sales = 1897.75685
```

Reference

https://www.kaggle.com

http://scikit-learn.org/stable/supervised_learning.html#supervised-learning

http://edoc.hu-berlin.de/master/timofeev-roman-2004-12-20/PDF/timofeev.pdf

The notes of BIA 656

I pledge on my honor that I have not given or received any unauthorized assistance on this assignment/examination. I further pledge that I have not copied any material from a book, article, the Internet or any other source except where I have expressly cited the source.

Signature: Yilin Wei, Tingting Gao, Jialu Yan Date: 12/11/2015