FUTURE SALE PREDICTION

I. Project Statement

For retail stores, if you want to keep a customer coming back with your business, you need to have enough product to satisfy customer needs. The problem is how many products do you need to order to provide a good customer but not keep the product in the inventory so long. This project will predict the total sales of each product each month so the business can maximize the sales of each month and reduce the money for inventory. Our goal in this project is to predict total sales for each item (as provided in the test file) next month.

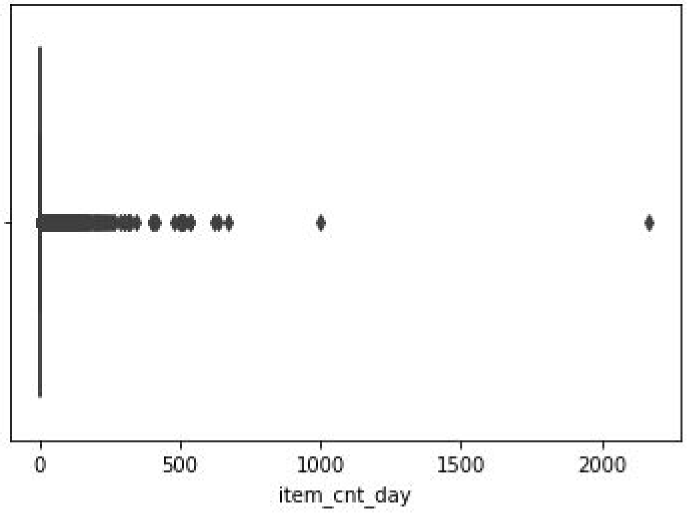
II. Data Wrangling

a. Gather and Loading the data

- Data was provided from the kaggle competitor

b. Cleaning the data

- Our data was cleaned and has no missing value, our goal is to predict the total sale of each item for next month. Column item_cnt_day is used for our training model.
- We found that item_cnt_day has some outliers, so we use a filter to remove the outlier. Make sure that our data is in good shape for our training and exploration.

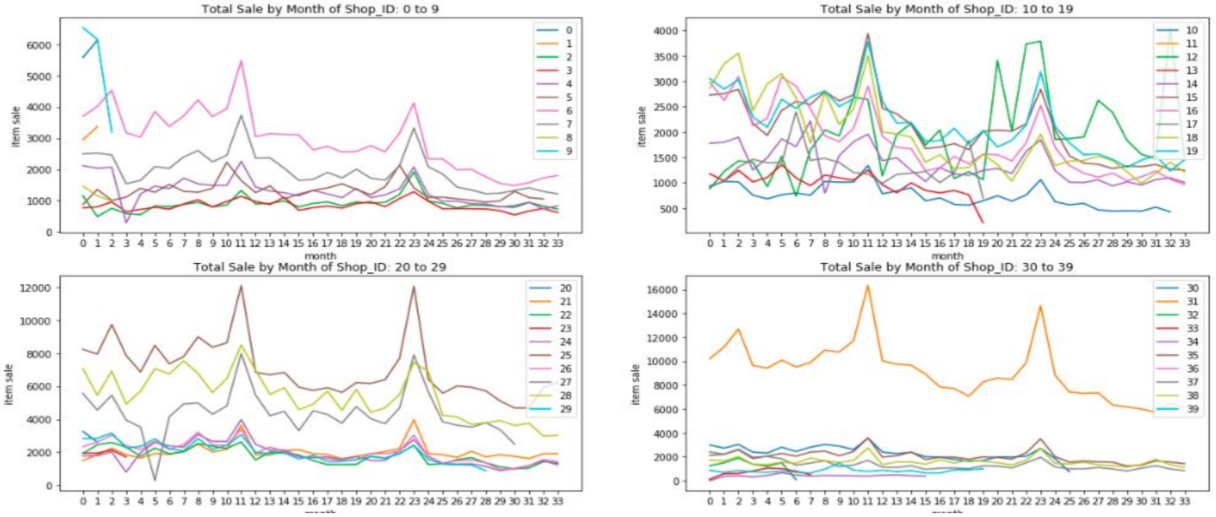


- Also we found that Item_cnt_day and item_price was have some negative data, then I just removed so it not make any noise later

III. Exploratory Data Analysis

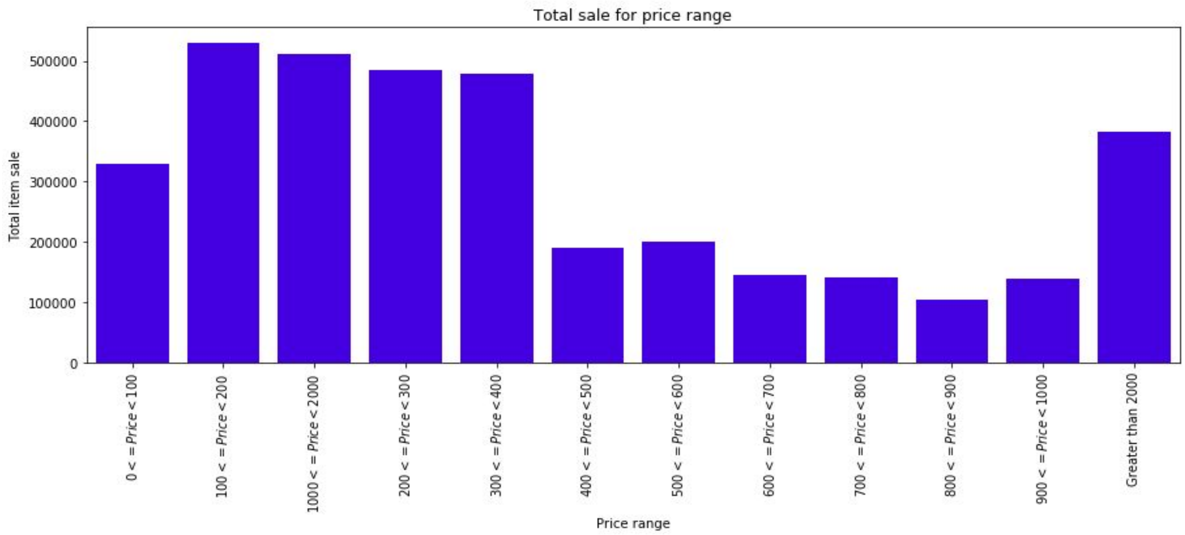
a. Visualize the total sale each month of each shop

First we gotta see how many items of each shop are sold during the year 2013 and 2015. As your graphs are shown below each graph contains the total item sale of 10 shops. All the shops have peak sales in December of each year and total sales are decreasing.



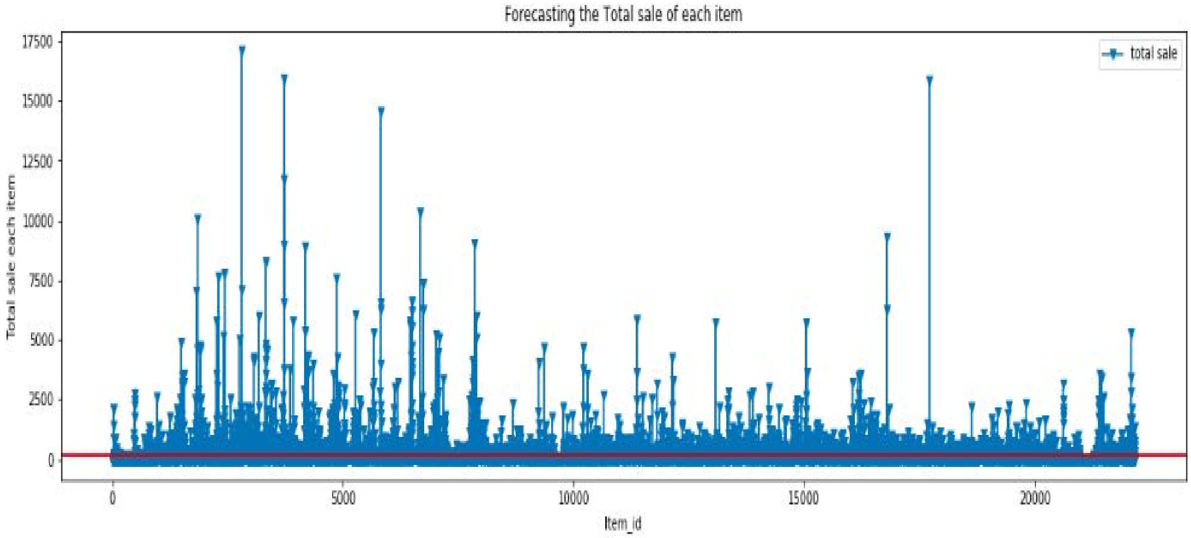
b. Total sale for price range

Now, we are going to check which range of prices most people are going to buy the most. The graph below are you can see that we are divided in to 12 intervals range prices, we have more items was sale if the item price are less than $400.



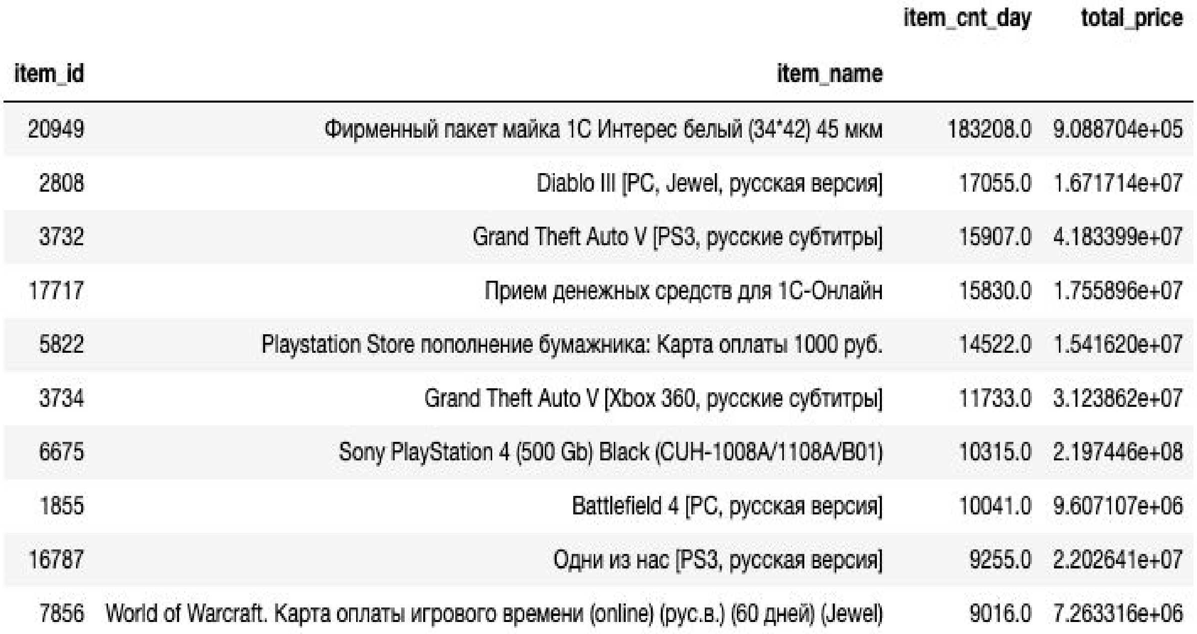
c. Total sale for each item

As you notice in the graph, the red line is the average of the total sale of items. We have around 3000 unique items and most of them have total sales above the average. And some items have a lot more than another, as you see the peak in the graph.



d. Top 10 item sale

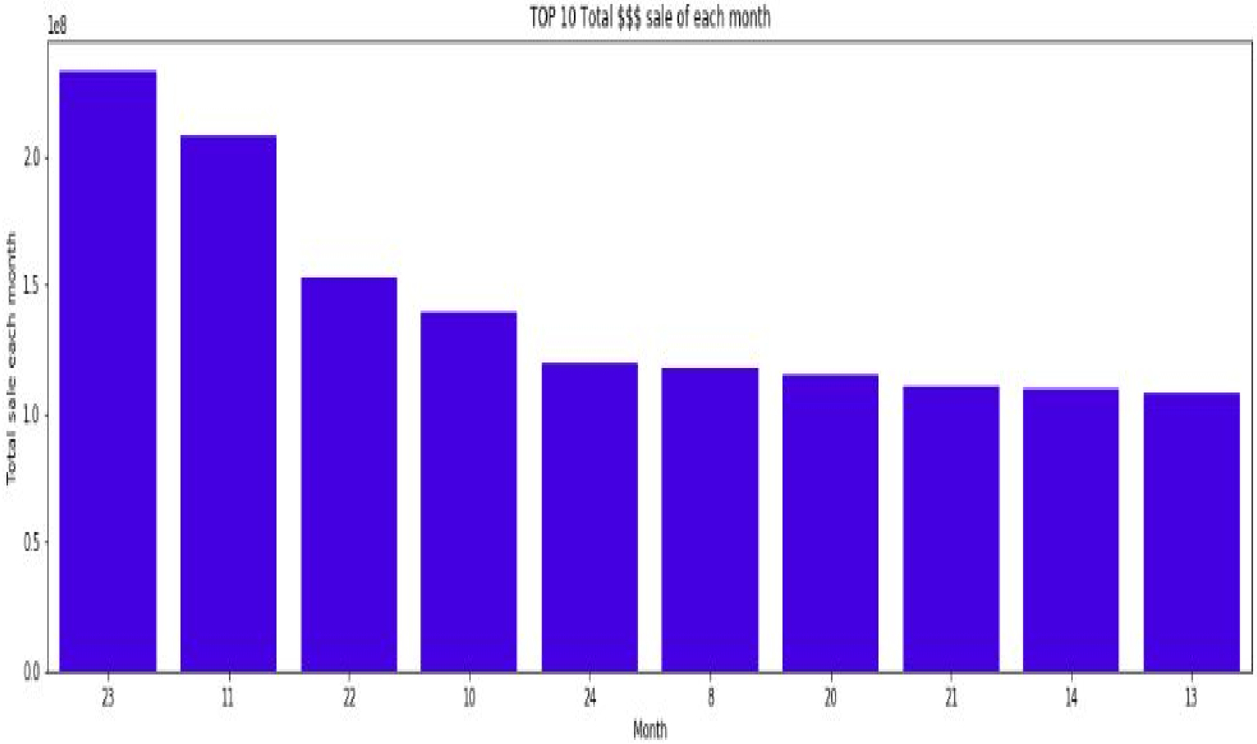
We will get more detail about which top 10 items are sold the most. You can tell that people are buying a lot of video games.



		item_cnt_day	total_price
item_id	item_name		
20949	Фирменный пакет майка 1С Интерес белый (34*42) 45 мкм	183208.0	9.088704e+05
2808	Diablo III [PC, Jewel, русская версия]	17055.0	1.671714e+07
3732	Grand Theft Auto V [PS3, русские субтитры]	15907.0	4.183399e+07
17717	Прием денежных средств для 1С-Онлайн	15830.0	1.755896e+07
5822	Playstation Store пополнение бумажника: Карта оплаты 1000 руб.	14522.0	1.541620e+07
3734	Grand Theft Auto V [Xbox 360, русские субтитры]	11733.0	3.123862e+07
6675	Sony PlayStation 4 (500 Gb) Black (CUH-1008A/1108A/B01)	10315.0	2.197446e+08
1855	Battlefield 4 [PC, русская версия]	10041.0	9.607107e+06
16787	Одни из нас [PS3, русская версия]	9255.0	2.202641e+07
7856	World of Warcraft. Карта оплаты игрового времени (online) (рус.в.) (60 дней) (Jewel)	9016.0	7.263316e+06

e. Top 10 month for total sale

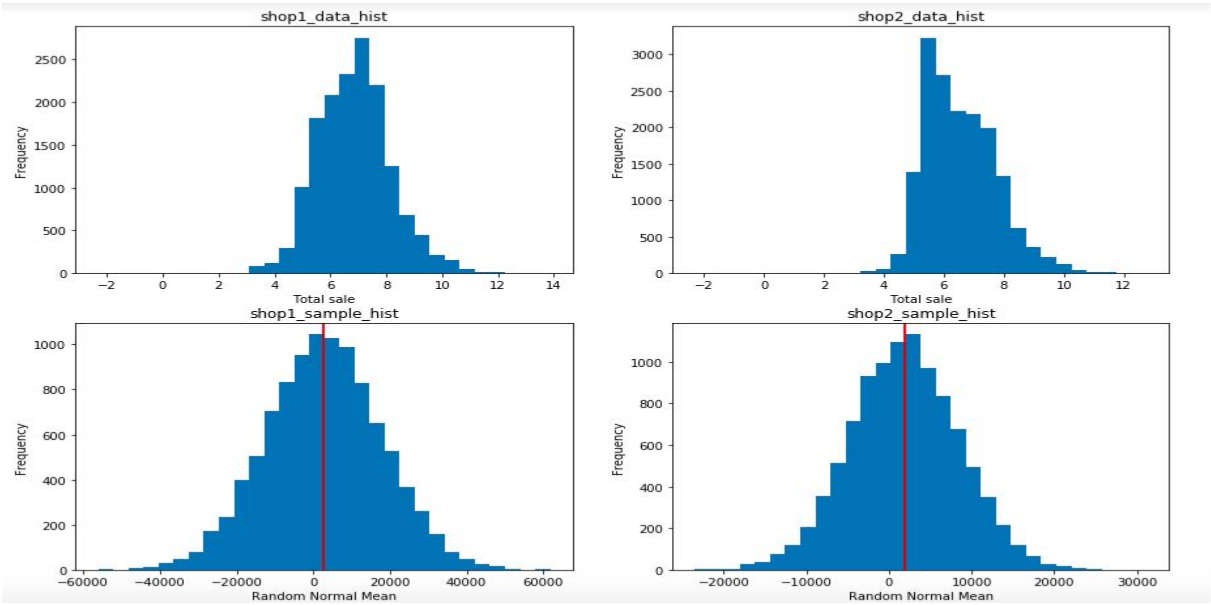
To better prepare for the stock of the shop, you can see how many items are sold each month. As the graphs are showing, the store is mostly sold in November and December of each year.



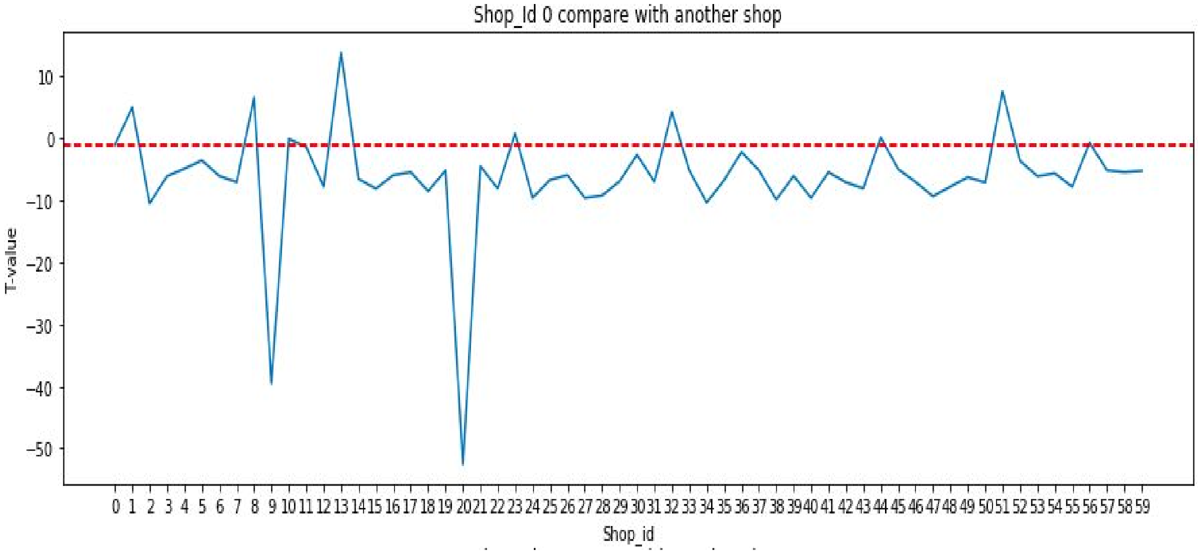
IV. Statistics Analyst

a. Frequency statistics test

Above we see that item sale in company, now we are using the Frequency test to see the compare of each store perform with another. We are resampling the data of each store over the mean and standard deviation. The graphs below are showing how our data are presenting before and after the sampling. We can get the T-value to see how our two shops are different.



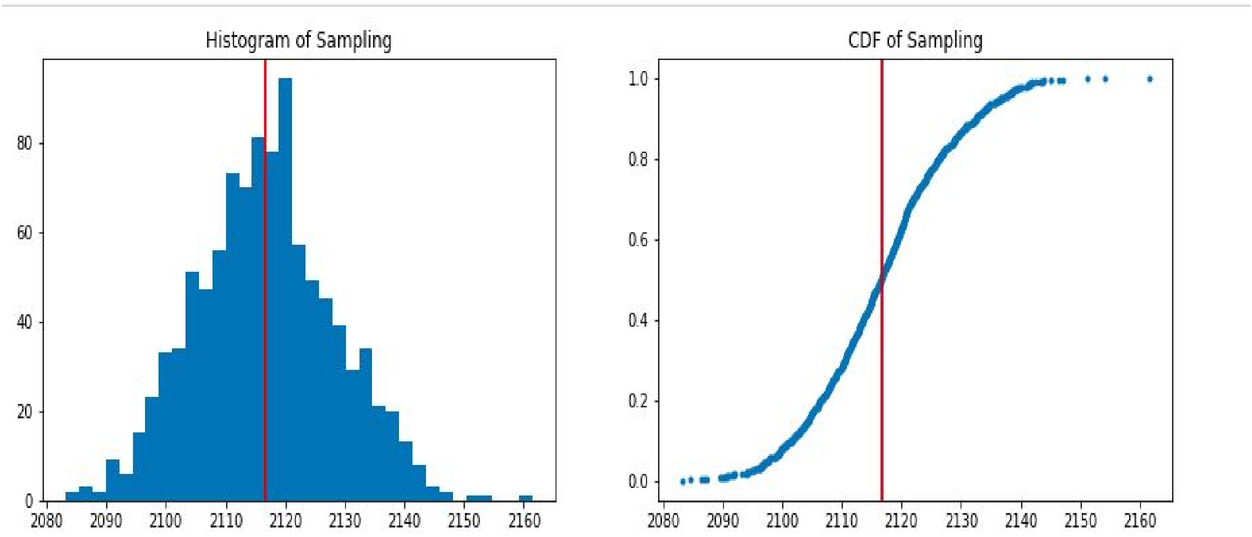
Here, this is the graph showing that the T-value of 1 shop compares to the rest of the shop. The red line is the baseline for example graph below, you can see that we compare shop_id 0 with the other shop. If the points are above the red line, it means shop_id 0 performs better or has more sales. If below the red line, it means the shop_id 0 has less sales than others.



b. Hypothesis test

We used the hypothesis test to tell our shop that profit is greater average all the time, or that will be different. First we set our null hypothesis is greater than the average and alternative hypotheses are less than or equal to the average. We choose a 95% confidence interval for our test. So we will have a significance levels is 5%.

After the hypothesis test, we get the p-value equal 0.54. So we can not rejected the null hypothesis.



V. Building predicting machine learning model

a. Prepare data for our model

Our data now has less features, so we need to do some feature engineering to add a few more features to help our model have better training.

Our goals are to predict the total sale for each item of each shop. So we will use a statistics method to create the new feature over the item_price and item_cnt_day. After we are done with feature engineering, we end up with our data and have 10 features to train our model.

We use the train_test_split in scipy to split our data to the training set and test set.

You can see more detail I did on github.

b. KNN regression model

First we will try a simple machine learning model called the KNN regression model. KNN models can predict the total sale of an item based on the number neighborhood we are set. We use the hyperparameter to choose the best number neighbor, as a result give us that best number neighbors is 5 and best test score or R square of testing set is 0.55 and the RMSE is equal 0.55.

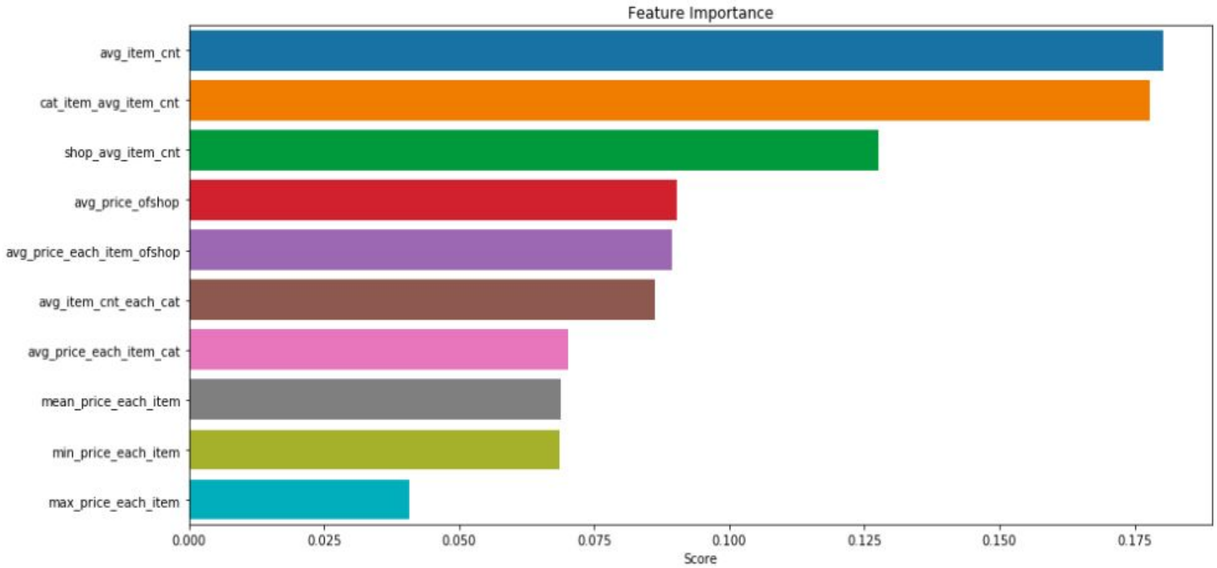
c. Random forest regression model

Next we tried the random forest regression model to predict our data, random forests return the mean of prediction of the decision tree, so it can give us better performance and more accuracy.

We first define the parameter for the random forest, then use the hyperparameter for tuning and choose the best performing. We use RandomizedSearchCV to choose the best parameter.

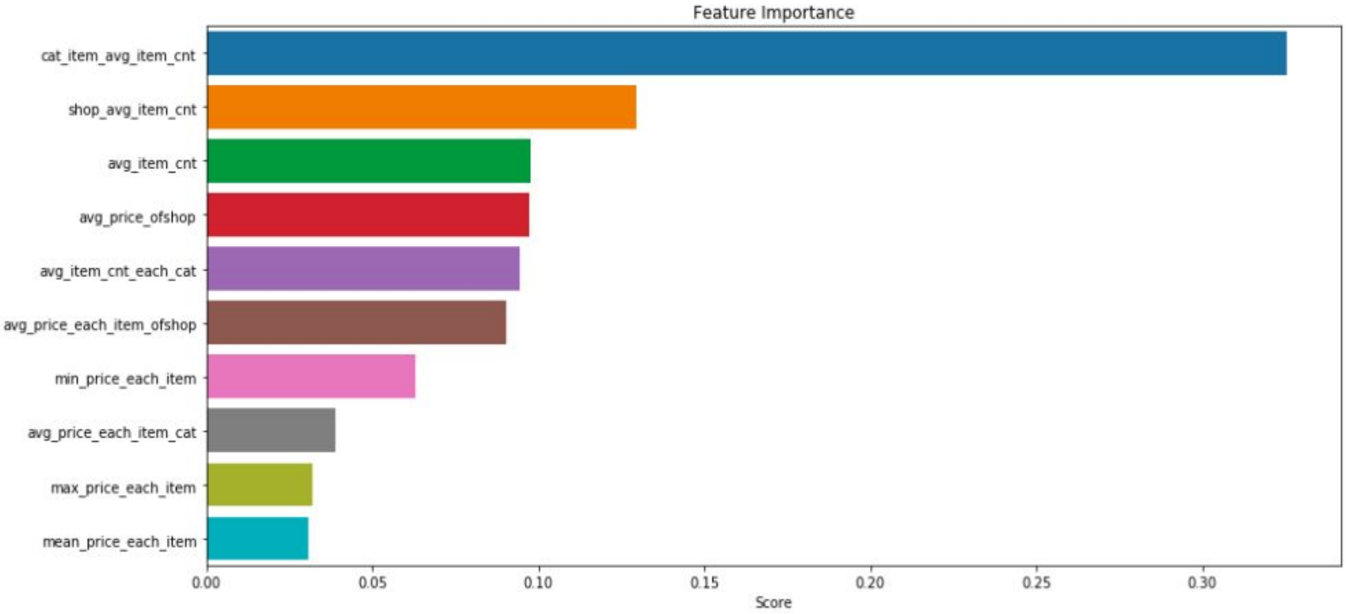
After we fit the data, and predict on our test set. We get the result for the model score or R square is 0.68 and the RMSE is 0.46.

And below are our features and the importance score of each feature.



d. Xgboost regression model

Next we are trying the Xgboost regression model, because Xgboost is one of the most popular machine learning algorithms these days. It can provide the best predicted model.



VI. SUMMARY

- The total sales for the 2 year are decreasing.
- The best month for sales is December.
- People buy items for less than $400.
- The best model for predicting the total sale is Xgboost.
- Checking our code at github.