Final Presentation- Philip Morris International

Case Study #3 Supervised Learning Problem(Regression task)

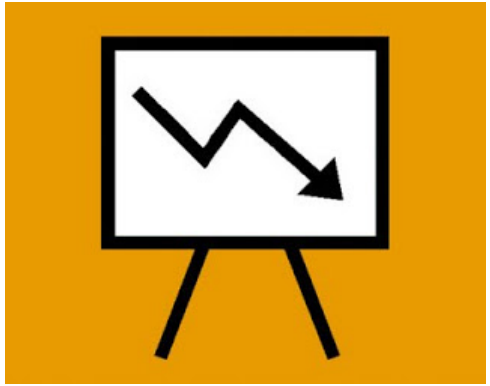
Name :Prashant Bansod

Github : ***https://github.com/prashant-bansod/usecase_3***

Problem Statement: Identifying the attributes in the surrounding that affect the sale

Situation

Every company wants to succeed and gain an edge on the competition. Achieving the revenue goals translates into maximizing sales. Many companies distribute their goods at a physical Point Of Sales (POSs). For all of them the challenge is to devise a strategy that will drive the sales at POSs.



Solution

Designing a Machine learning model which will identity the important attributes in the surrounding that impact the sales

Steps to develop a Machine Learning model

1) Exploratory Data Analysis
2) Deciding the Target Variable
3) Data Pre-processing
4) Preparing Data for the Machine Learning Model
5) Applying the Machine Learning Model
6) Deciding the most important features

Deciding the Target Variable

1. As we want to know the impact of the surroundings on the total sales, the target feature is going to be the total sale.

2. We are going to decide the effect of the Surrounding attributes on the total sale

3. As the target variable is continuous, we are going to apply regression algorithmm for this problem

Data Preparation Method

1 .The dataset was prepared using the both sales_granular.csv and Surroundings.json file.

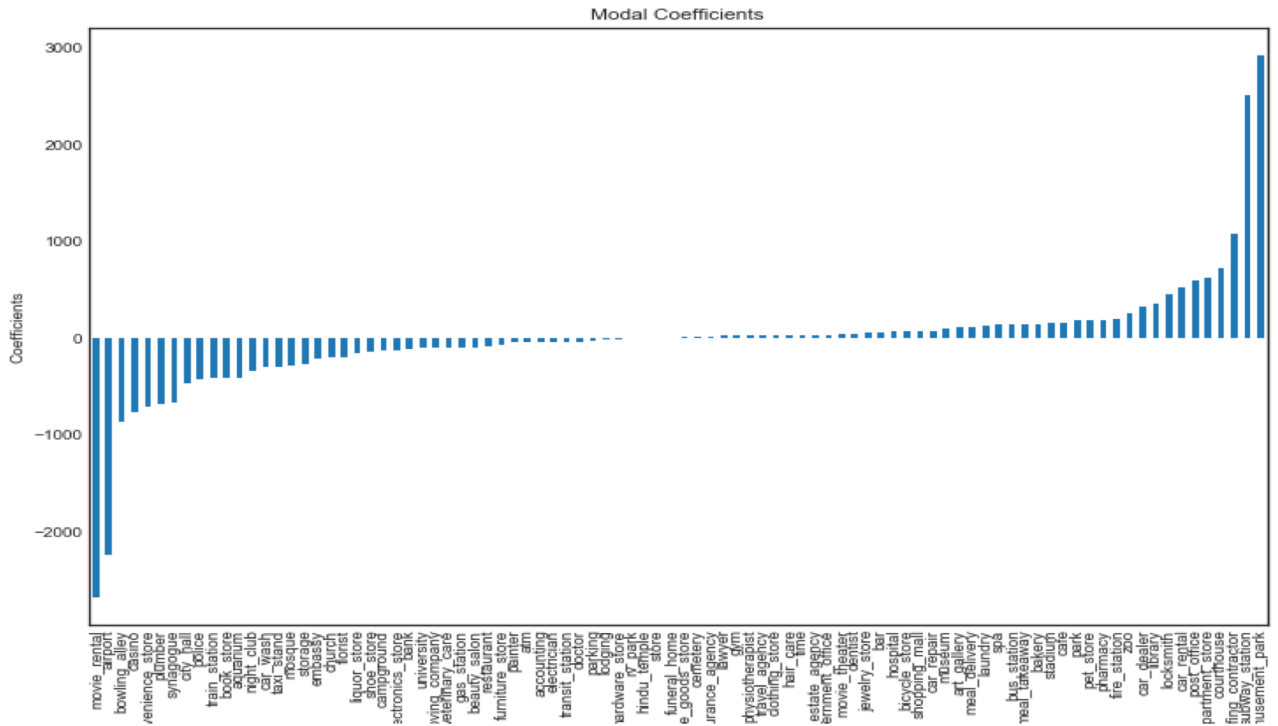
2. Extraction of the features is done using Surroudings.json where we have all the attributes related to the surroundings

3. The total sale is being calculated by adding all the sales for a particular store.

4. As the scales of the features is not same, the feature scaling step is important

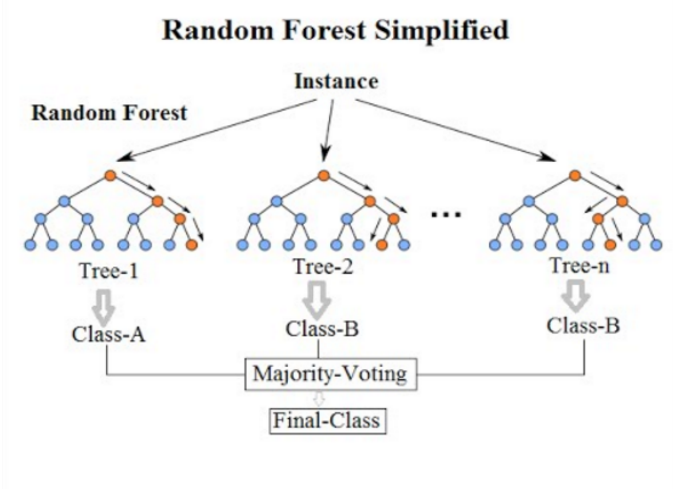
Applying Machine Learning Model

After preprocessing the data, Lasso Model was used to understand the features in the data



Why Random Forest Machine Learning Algorithm

It can handle categorical features, numerical features without any need for scaling and also provide a pretty good indicator of feature importance. It overcomes the problem of overfitting as well.



Random Forest Steps :
1) Take different subset of training samples
2) Build decision trees on each sample
3) For regression task, the process follows the same as described above except the final predicted value would be the average of all the values predicted by individual decision trees.

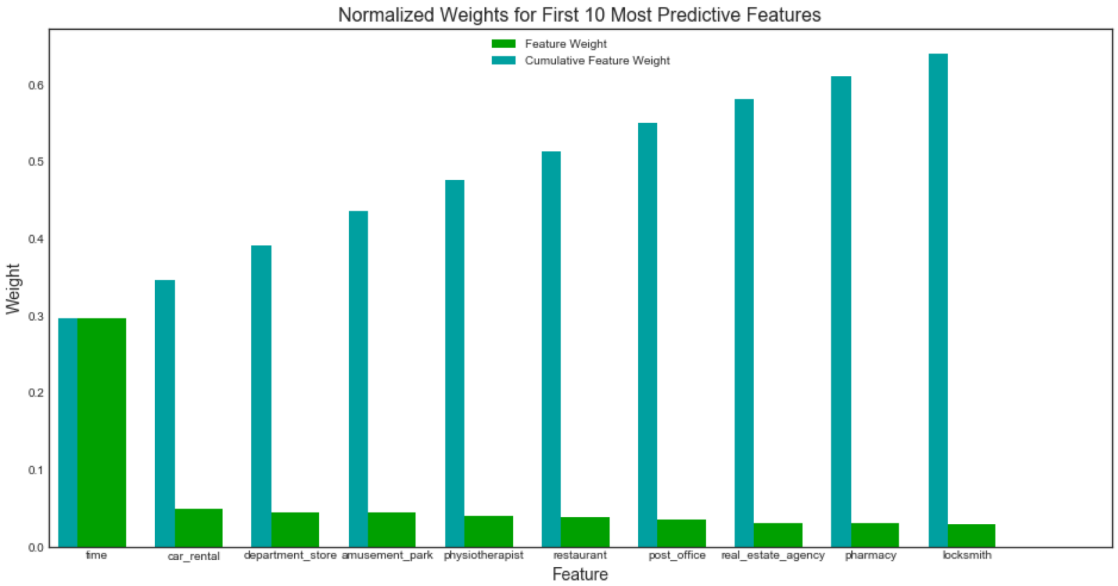
Final Accuracy of the model :

R^2 score=0.87

Feature Importance

By using using the Gradient Boosting Regressor we can get the understanding of the important features in the data. To drive the sales at the point of sale (POS), top 5 attributes are

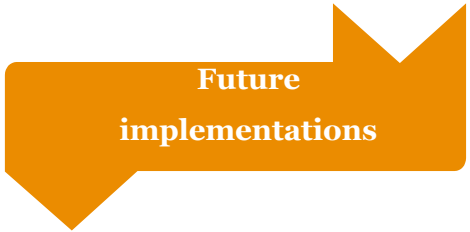
1. Time (Hourly sales)
2. Car rental service
3. Department store
4. Amusement park
5. Physiotherapist



Challenges & Lessons Learned

-Understanding and extracting the features
-Invested considerable time in data preparation

- Enhanced my knowledge of the data preparation process
- Got hands on a real life project with the complex data



-Application of XGBoost algorithm to improve the accuracy further
- Better Hyperparameter tuning