Capstone Project Submission

Name: Priyanka Shinde

Email: shindepriya7709@gmail.com

Contributor: Individual Project

Github Link: https://github.com/priyankashinde-DS/Rossmann_sales_prediction

Rossmann is a Germany drug store chain with over 3790 stores in Europe. In this challenge, we are given records of sales of each store on different dates, from 01/01/2013 to 31/07/2015. Sales can be affected by many factors such as holidays, promotions and competitions. Our goal is to explore the correlations between these features using Python and utilize the results of this exploratory data analysis to build a linear regression model that predicts the store sales from 08/01/2015 to 09/17/2015, 6 weeks in advance. Our project consist of 3 parts: data cleaning, exploratory data analysis and data modeling.

In DATA CLEANSING Some data fields should be converted to a more suitable data type for the convenience of the explanatory process.

We start by seeing what our data consits of. We want to see which variables are continuous vs which are categorical. After exploring some of the data, we see that we can create a feature. Number of sales divided by customers could give us a good metric to measure average sales per customer. We can also make an assumption that if we have missing values in this column that we have 0 customers.

For achieving the goal, training data will be divided into two parts, first half will contain 70% of the data and the other half that will to test the accuracy of predictions made by the model designed and trained over the first set. For the creation of model, several regression algorithms will be used such as XGBoost,LightGBM,KNeighbors gradientboost etc. The model having the highest accuracy(R2score) and the lowest (RMSE) will be choosen and will be trained over the entire dataset and then that model will be used to make predictions over the test data. we use root mean squared prediction error to find out the accuracy of our predictions. The root-mean-square error (RMSE), or RMSE is a frequently used measure of the differences between values (test and train values) predicted by a model or an estimator and the values actually observed. The predictions made by model i.e(XGBRgressor) is pretty close to the real value for sales.

- Following are some important collaborations we get from our project is:

1) **We selected three types of linear regression models, namely Linear Regression , Ridge Regression , and Lasso Regression. These models have R2SCORE around 0.675 for both train and test. They perform well, but not as well as other models.**

2) **Also , all three linear regression models are perfectly fitted ,as they have low bias and low vaiance.**

3) **Based on tree regressor ,we developed baseline models for further evaluation , namely Gradient Boosting Regressor, XGB Regressor , KNearestNeighbors Regressor,LightGBM Regressor.**

4) **According to RMSE and R2SCORE xgbr,knn and lgbm the most reliable models.**

5) **The KNearestNeighbors Regressor had a very high R2SCORE, however, it required much computational time, so we dropped it.**

6) **For the hyper parameter tuning ,we choose XGBR and LGBM based on the RMSE and R2SCORE obtained from the baseline models.**

7) **We observed that the LGBM also have high R2 SCORE and we can improve it using a specific parameter which is the number of estimators. However, it takes a lot of time to compute.**

8) **XGB has perform very good on performing hyperparameter tuning which has R2SCORE 0.96 (train data) & 0.96 (test data) and RMSE. with no underfitting and overfitting.The predictions is pretty close to the real value for sales.**

9) **It is also observed that the out all the features Customers is most important. this implies that customers are directly impacting the sales.**

10) **The most selling store type is A.**

11) **The store which is run promo leads to increase in sales and customers as compared to store which are not runing the promo.**

12) The number of stores opened during School Holidays were more than that were opened during State Holidays.Another important thing to note is that the stores which were opened during School holidays had more sales than normal.

13) Sales are increased during Chirstmas week.

Google Drive Link:
https://drive.google.com/drive/u/0/folders/1xEynnj0tOwFVPZCmffn5TViW-SxbV4cc