Machine Learning Foundation – Batch 03

Capstone Project

Bawani A. Guruge

Dialog Axiata

Registration Number- 252

Market Sales Prediction Regression Model

1.0 Introduction

This project focuses on sale prediction of a logistic company. The dataset was collected for 60 days, this is a real database of a Brazilian logistics company. The dataset has twelve predictive attributes and a target that is the total of orders for daily treatment. The problem was considered as a regression type machine learning model.

2.0 Data

The data was taken from 'Daily Demand Forecasting Orders Data Set' in the UCI Machine Learning Repository.

Data Set Characteristics:	Time-Series	**Number of Instances:**	60	**Area:**	Business
Attribute Characteristics:	Integer	**Number of Attributes:**	13	**Date Donated**	2017-11-21
Associated Tasks:	Regression	**Missing Values?**	N/A	**Number of Web Hits:**	114917

The data was represented in 13 columns and 60 rows. The column names of the data were long so changed column names which is easier to use in the model.

```
[124] #renaming the columns
      data.columns = ['Week of the month','Day of the week ', 'Non-urgent order',
             'Urgent order', 'Order type A', 'Order type B', 'Order type C',
             'Fiscal sector orders', 'traffic controller sector',
             'Banking orders (1)', 'Banking orders (2)', 'Banking orders (3)',
             'Total_orders']
      data.head()
```

There were no missing values or duplicates in the dataset. All columns were either int or float data type.

```
[129] data.info()

      <class 'pandas.core.frame.DataFrame'>
      RangeIndex: 60 entries, 0 to 59
      Data columns (total 13 columns):
       #   Column                    Non-Null Count  Dtype
      ---  ------                    --------------  -----
       0   Week of the month         60 non-null     int64
       1   Day of the week           60 non-null     int64
       2   Non-urgent order          60 non-null     float64
       3   Urgent order              60 non-null     float64
       4   Order type A              60 non-null     float64
       5   Order type B              60 non-null     float64
       6   Order type C              60 non-null     float64
       7   Fiscal sector orders      60 non-null     float64
       8   traffic controller sector 60 non-null     int64
       9   Banking orders (1)        60 non-null     int64
       10  Banking orders (2)        60 non-null     int64
       11  Banking orders (3)        60 non-null     int64
       12  Total_orders              60 non-null     float64
      dtypes: float64(7), int64(6)
      memory usage: 6.2 KB
```

Box plot shows multiple outliers for ‘Total order’. These were removed from dataset.

3.0 Methodology

As mentioned earlier, data pre-processing was conducted on the raw data which was taken from the repository. After conducting the data cleaning correlation matrix was drawn to identify any relationship between the variables.

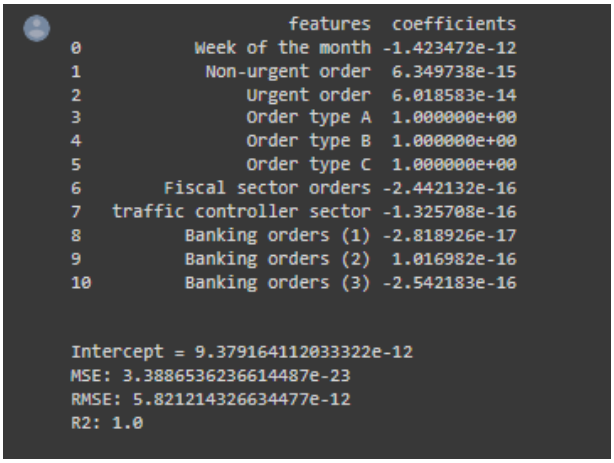
X and Y variables were identified as below. ‘Days of the week’ was neglected from the X features due to the less relationship it showed to the other variables.

```
[ ]  #'Day of the week ' has been removed from features as result of correlation matrix
     X = [ 'Week of the month',  'Non-urgent order',
           'Urgent order', 'Order type A', 'Order type B', 'Order type C',
           'Fiscal sector orders', 'traffic controller sector',
           'Banking orders (1)', 'Banking orders (2)', 'Banking orders (3)' ]

     y = ['Total_orders']

     X_train, X_test, y_train, y_test = train_test_split(data[X], data[y], test_size=0.3, random_state=42)
```

Linear Regression model was used with while using 70% for training data and 30% for testing data. Multiple iterations were conducted to see which features can make a difference to the metrices. MSE, RMSE and R2 values were obtained to evaluate the performance of the regression model.

4.0 Results

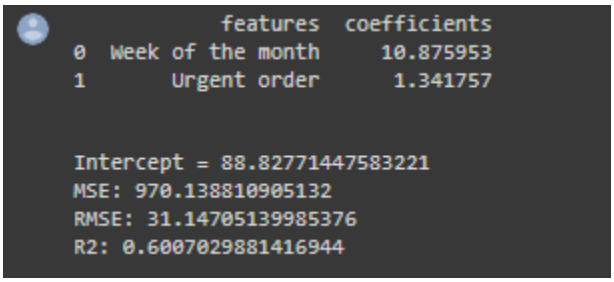
Using all the Variables for the X features.



	features	coefficients
0	Week of the month	-1.423472e-12
1	Non-urgent order	6.349738e-15
2	Urgent order	6.018583e-14
3	Order type A	1.000000e+00
4	Order type B	1.000000e+00
5	Order type C	1.000000e+00
6	Fiscal sector orders	-2.442132e-16
7	traffic controller sector	-1.325708e-16
8	Banking orders (1)	-2.818926e-17
9	Banking orders (2)	1.016982e-16
10	Banking orders (3)	-2.542183e-16

```
Intercept = 9.379164112033322e-12
MSE: 3.3886536236614487e-23
RMSE: 5.821214326634477e-12
R2: 1.0
```

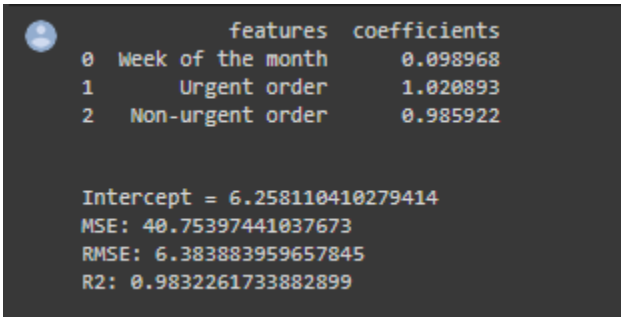
2nd iteration with two features for X.



	features	coefficients
0	Week of the month	10.875953
1	Urgent order	1.341757

```
Intercept = 88.82771447583221
MSE: 970.138810905132
RMSE: 31.14705139985376
R2: 0.6007029881416944
```

3rd iteration with three variables for X



	features	coefficients
0	Week of the month	0.098968
1	Urgent order	1.020893
2	Non-urgent order	0.985922

```
Intercept = 6.258110410279414
MSE: 40.75397441037673
RMSE: 6.383883959657845
R2: 0.9832261733882899
```

5.0 Conclusion and Discussion

The skill or performance of a regression model must be reported as an error in those predictions. We cannot calculate accuracy for a regression model. Hence MSE, RMSE and R2 values were obtained.

While using all the variables for X, the model was able to get R2 score of 1. And in-order to check the accuracy, 2nd and 3rd iterations were conducted. Model was able to get R2 score of 0.98 with the variables urgent order, non-urgent order and week of the month. This can be considered important features for the X variable while we consider the target orders per week.