Vishal Srinivasan

PGP DSBA | G3 JULY 2021

BUSINESS REPORT

Supply Chain Project CApstone

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| SNO | TITLE | PAGE NO |
| A | List of figures | 2 |
| B | List of Tables | 3 |
| 1 | Introduction | 4 |
| 2 | EDA and Business Implications | 5 |
| 2.1 | Visual Inspection of the Data | 5 |
| 2.2 | Univariate Analysis | 7 |
| 2.3 | Bivariate Analysis | 10 |
| 3 | **Data Cleaning and Pre-processing** | 13 |
| 3.1 | Removal of unwanted variables | 13 |
| 3.2 | Missing Value Treatment | 14 |
| 3.3 | Outlier Treatment | 15 |
| 4 | Model Building | 15 |
| 4.1 | Linear Regression Model | 16 |
| 4.2 | CART Regressor Model | 18 |
| 4.3 | Random Forest Regressor Model | 19 |
| 4.4 | Adaptive Boosting Regressor Model | 19 |
| 4.5 | Lasso Regressor Model | 19 |
| 4.6 | Ridge Regressor Model | 20 |
| 4.7 | Hyper Parameter Tuning | 20 |
|  |  |  |
| 5 | Model Validation | 21 |
| 6 | Final Interpretation or Recommendations | 22 |

**List of Figures**

|  |  |  |
| --- | --- | --- |
| SNO | TITLE | PAGE NO |
| 1 | Box Plot | 7 |
| 2 | Bar Plot | 8 |
| 3 | Count plot for the variable wh\_est\_year | 9 |
| 4 | Count Plot for the variable transport\_issue\_l1y | 9 |
| 5 | Heat Map | 10 |
| 6 | Histogram | 11 |
| 7 | Transport\_issue\_l1y vs Zone | 12 |
| 8 | Pair plot | 13 |
| 9 | Scatter plot for actual and predicted value | 18 |
| 10 | Model Comparison | 22 |
| 11 | Feature Importance | 22 |

**List of Tables**

|  |  |  |
| --- | --- | --- |
| SNO | TITLE | PAGE NO |
| 1 | Evaluation Metrics of Linear Regression Model | 16 |
| 2 | Evaluation Metrics of CART Regressor Model | 18 |
| 3 | Evaluation Metrics of Random Forest Regressor Model | 19 |
| 4 | Evaluation Metrics of AdaBoost Regressor Model | 19 |
| 5 | Evaluation Metrics of Lasso Regressor Model | 19 |
| 6 | Evaluation Metrics of Ridge Regressor Model | 20 |
| 7 | Hyper Parameter Tuning | 20 |

**INTRODUCTION**

A FMCG company has entered the instant noodles business two years back. Their higher management has notices that there is a miss match in the demand and supply. Where the demand is high, supply is low and where the demand is low, supply is high. In both the ways it is an inventory cost loss to the company; hence, the higher management wants to optimize the supply quantity in each warehouse in entire country.

**PROBLEM STATEMENT**

Prediction of the shipment of the product each time at the warehouse

**Goal & Objective**: The objective of this exercise is to build a model, using historical data that will determine an optimum weight of the product to be shipped each time to the warehouse. Also try to analysis the demand pattern in different pockets of the country so management can drive the advertisement campaign particular in those pockets. This is the first phase of the agreement; hence, company has shared very limited information. Once you can showcase a tangible impact with this much of information then company will open the 360-degree data lake for your consulting company to build a more robust model.

**Data dictionary:**

**Table

Description automatically generated**

**2 Exploratory Data Analysis (EDA) and Business Implications**

**2.1 Visual inspection of data**

The quick glimpse of the data is shown below

Graphical user interface, text, application, chat or text message

Description automatically generated

The dataset has 25000 rows and 24 columns

**Types of data present in the dataset**

Table

Description automatically generated

A quick glimpse on all the different types of data present in the dataset. The dataset has 8 object type variables, 14 integer type variables and 2 float type variables

**Five-point summary of the dataset**

Graphical user interface, application

Description automatically generated

The above figure shows the 5-point summary of all the different variables present in the data set

**Columns in the dataset**

Text, letter

Description automatically generated

The above figure shows the various columns present in the dataset

**Head of the Dataset**

Graphical user interface, text, application, chat or text message

Description automatically generated

The above figure shows the first 5 rows and columns of the dataset

**Tail of the dataset**

**Graphical user interface, application

Description automatically generated**

The above Figure Shows the last 5 rows and columns of the Dataset

**2.2 Univariate Analysis**

The term **univariate analysis**refers to the analysis of one variable. The purpose of univariate analysis is to understand the distribution of values for a single variable.

**Box Plot**

**Graphical user interface, application, table, Excel

Description automatically generated**

The above figure shows the 5-point summary of the numerical variables present in the datatype in a pictorial manner

**BAR PLOT**

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

The above figure shows the bar plot for the categorical data present in the Dataset

Chart

Description automatically generated

Count Plot for the variable wh\_est\_year

Chart, bar chart, waterfall chart

Description automatically generated

Count Plot for the variable transport\_issue\_l1y

**Inference**

* There are a greater number of Warehouses in the rural area than in the urban area
* The capacity of warehouse of the company is more in the large category followed by the mid category and the small category
* There are greater number of warehouses present in the North Zone followed by the West zone, South zone, and the Least number of Warehouses in the East Zone
* There are a greater number of company owned warehouses than the rented warehouses
* From the Certificate Bar Plot, we find that the warehouses belonging the company has been certified in the “C” certification the greatest number of times

**2.3 Bi-variate Analysis:**

The Bivariate Analysis is the simplest form of quantitative analysis. It involves the analysis of two variables, and it is used find the empirical relationship between the two variables.

Bivariate analysis can help determine to what extent it becomes easier to know and predict a value for one variable, if we know the value of the other variable

Here we have used Heat map, correlation matrix, Histograms and pair plot for the bivariate analysis

**Heat Map**

**Timeline

Description automatically generated**

Heat map tells the correlation and collinearity between variables from the above heat map we can observe that there is a heavy correlation and presence of multicollinearity. Multicollinearity is not acceptable in regression

**Correlation Matrix**

Table

Description automatically generated

**Histogram**

**Chart, waterfall chart

Description automatically generated**

The histogram of all the numerical variables in the dataset


Description automatically generated

transport\_issue\_l1y vs Zone

**Pair plot**

**Diagram

Description automatically generated**

From the above pair plot, we can observe how the data is distributed and relation and patterns between each variable.

**3 Data Cleaning and Pre-processing**

**3.1 Removal of unwanted variables**

**Table

Description automatically generated**

The WH\_Manager\_ID is removed as we already have the Ware\_house\_ID as the index

The Variable storage\_issue\_reported\_l3m is also removed from the dataset as the target variable is what causing the storage issue

**3.2 Missing Value treatment**

**Table

Description automatically generated with medium confidence**

The above figure shows us the total number of missing values present in the dataset

The variable workers\_num has 990 missing variables and it can be replaced by the mean imputer function as the distribution is normally distributed

The variable wh\_est\_year has 11881 missing variables, and all the missing values are removed. From the Pair plot we found out that this is one of the variable which has the most relation with the target variable, so instead of dropping the whole variable we have just dropped the missing variables from the variable

Graphical user interface, text, application

Description automatically generated

The above figure shows the new data frame after the wh\_est\_year variable has been dropped

The approved\_wh\_govt\_certificate variable has 908 NA variables, and all these variables are off the years 2021,2022, 2023.From this we find that newly established warehouses have not been certified by the government

**3.3 Outlier treatment**

**Text, table

Description automatically generated with medium confidence**

From the above figure we find the skewness is between -5 to+5 so no outlier treatment is required

**4 Model building**

**STEP 1:** All the object variables are converted to Categorical codes

Graphical user interface, text, application

Description automatically generated

The Data set after all the object variables are converted to categorical codes

**STEP 2**: Constructed a X and y matrix for modelling

Text

Description automatically generated

The X and y matrix was constructed for model building and the Target Variable product\_wg\_ton is taken in the y matrix

**STEP 3**: The dataset is divided into test and train set in a 70:30 ratio. This is done by invoking the sklearn. model\_selection import train\_test\_split.

**STEP 4**: Model Building

The Following Models were Built

1.Linear Regression Model

2.CART Regressor Model

3.Random Forest Regressor Model

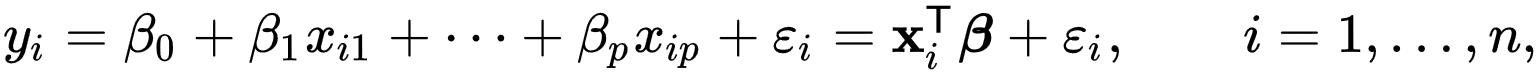
4.AdaBoost Regressor Model

5.Lasso Regressor Model

6.Ridge Regressor Model

**4.1 Linear Regression Model**

**Linear regression** is a Linear approach for modelling the relationship between a scalar response and one or more explanatory variables (Dependent and independent variables). The case of one explanatory variable is called as simple linear regression, for more than one, the process is called **multiple linear regression**

****

**Coefficient of Independent Variables**

Table

Description automatically generated

**INTERCEPT**

A picture containing shape

Description automatically generated

The Intercept of the Model was calculated

Now we have coefficient and intercept to form an equation but how to check the performance of the equation. We can check the performance of the equation by R Square or RMSE

**Evaluation Metrics of Linear Regression Model**

|  |  |  |
| --- | --- | --- |
|  | TEST | TRAIN |
| MAE | 5052.314 | 4988.67 |
| MSE | 43900552.56 | 42205437.39 |
| RMSE | 6625.75 | 6496.58 |
| R2 | 0.676 | 0.681 |

**Scatter Plot for Actual Value and Predicted Value**

Chart, scatter chart

Description automatically generated

**4.2 CART Regressor Model**

**Evaluation Metrics of CART Regressor Model**

|  |  |  |
| --- | --- | --- |
|  | TEST | TRAIN |
| MAE | 5095.162 | 5008.395 |
| MSE | 42579164.7 | 40322828.91 |
| RMSE | 6525.27 | 6350.02 |
| R2 | 0.6864 | 0.6958 |

After Doing the Grid Search CV the RMSE value came down to 6200.957

**4.3 Random Forest Regressor Model**

**Evaluation Metrics of Random Forest Regressor Model**

|  |  |  |
| --- | --- | --- |
|  | TEST | TRAIN |
| MAE | 4719.21 | 1709.5 |
| MSE | 3843201.07 | 4923235.76 |
| RMSE | 6199.356 | 2218.83 |
| R2 | 0.7169 | 0.9628 |

After Doing the Grid Search CV the RMSE value is 6201.297, as the RMSE value after Grid Search CV is higher, the model performs good even without tuning

**4.4 AdaBoost Regressor Model**

**Evaluation Metrics of AdaBoost Regressor Model**

|  |  |  |
| --- | --- | --- |
|  | TEST | TRAIN |
| MAE | 4705.18 | 85.756 |
| MSE | 40485522.3 | 115082.32 |
| RMSE | 6362.823 | 339.237 |
| R2 | 0.7018 | 0.999132 |

After Doing the Grid Search CV the RMSE value is 6408.270, as the RMSE value after Grid Search CV is higher, the model performs good even without tuning

**4.5 Lasso Regressor Model**

**Evaluation Metrics of Lasso Regressor Model**

|  |  |  |
| --- | --- | --- |
|  | TEST | TRAIN |
| MAE | 9102.75 | 8986.66 |
| MSE | 125328989.1 | 121216071.7 |
| RMSE | 11195.043 | 11009.81 |
| R2 | 0.07694 | 0.085 |

After Doing the Grid Search CV the RMSE value came down to 6623.298 and R2 value went up to 0.68

**4.6 Ridge Regressor Model**

**Evaluation Metrics of Ridge Regressor Model**

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | TEST | TRAIN |
|  | MAE | 5056.9 | 4994.42 |
|  | MSE | 43928981.79 | 42217135.75 |
|  | RMSE | 6627.894 | 6497.47 |
|  | R2 | 0.67646 | 0.6815 |

After Doing the Grid Search CV the RMSE value came down to 6625.885 and

R2 value went up to 0.679

**4.7 Hyper Parameter Tuning**

Hyper-parameter tuning refers to the process of find hyper-parameters that yield the best result. This, of course, sounds a lot easier than it actually is. Finding the best hyper-parameters can be an elusive art, especially given that it depends largely on your training and testing data.

As your data evolves, the hyper-parameters that were once high performing may no longer perform well. Keeping track of the success of your model is critical to ensure it grows with the data.

One way to tune your hyper-parameters is to use a grid search. This is probably the simplest method as well as the crudest. In a grid search, you try a grid of hyper-parameters and evaluate the performance of each combination of hyper-parameters.

|  |  |  |
| --- | --- | --- |
| Model Name | Before Tuning | After Tuning |
| Linear Regression | 6625.75 | 6625.75 |
| CART | 6525.27 | 6200.957 |
| Random Forest | 6199.356 | 6201.297 |
| Adaptive Boost | 6362.823 | 6408.27 |
| Lasso | 11195.043 | 6623.298 |
| Ridge | 6627.894 | 6625.885 |

The RMSE values of all the models used after and before tuning are listed above

**5. Model validation**

**Regression model evaluation metrics**  
  
    The MSE, MAE, RMSE, and R-Squared metrics are mainly used to evaluate the prediction error rates and model performance in regression analysis.

* **MAE** (Mean absolute error) represents the difference between the original and predicted values extracted by averaged the absolute difference over the data set.
* **MSE** (Mean Squared Error) represents the difference between the original and predicted values extracted by squared the average difference over the data set.
* **RMSE** (Root Mean Squared Error) is the error rate by the square root of MSE.
* **R-squared** (Coefficient of determination) represents the coefficient of how well the values fit compared to the original values. The value from 0 to 1 interpreted as percentages. The higher the value is, the better the model is.

Diagram

Description automatically generated with low confidence

Chart, bar chart

Description automatically generated

Model Comparison

The above plot shows the model performance based on the RMSE value

From the above Plot we come to know that the Random Forest Regressor Model is the optimum model to be used for Prediction as per the RMSE value

**6. Final interpretation / recommendation**

The Final interpretation is based on feature Importance

From the most optimum model.i.e., the Random Forest Model, we find the most important features which influences the target variable

Chart

Description automatically generated

Feature Importance

From the above plot we find that the wh\_est\_year, approved\_wh\_govt\_certificate and the transport\_issue\_1ly are the most important features

**Insights**

* It is observed that the greatest number of warehouses are located at Zone-6 and this further indicates that there is higher demand in that Zone
* The availability of generator in the warehouse doesn’t affect much
* About 54.3% of the warehouses are company owned and the rest 45.7% are rented
* There are a greater number of Warehouses in the rural area than in the urban area
* The capacity of warehouse of the company is more in the large category followed by the mid category and the small category
* There are greater number of warehouses present in the North Zone followed by the West zone, South zone, and the Least number of Warehouses in the East Zone
* From the Certificate Bar Plot, we find that the warehouses belonging the company has been certified in the “C” certification the greatest number of times
* It is clear from the observations that the variable “storage\_issue\_reported\_l3m” plays an important role in suggesting the company about the quantity to be shipped
* North Zone has the highest significance than the other zones

**Recommendations**

* The oldest warehouse which is established in the year 1996 and it needs to be maintained properly with all the prerequisites before the product is shipped to those warehouses
* Some frequent repairs to be done on older warehouses so that the storage issues faced by those warehouses can be prevented
* The transport issues faced by the warehouses also influences the quantity to be shipped to the warehouses, so those issues should be addressed by the company so that the quantity stored in the warehouses can be optimized and supply chain issues can be mitigated
* It is found that from the observations the warehouses with A+ certification from the government performs well in the business than the warehouses with other certifications and the company should strive to get A+ certifications so that the company has a greater number of warehouses that performs well
* The areas where there is more number of competitors should be addressed and more number of warehouses should be established in those areas
* The distance from the hub impacts the business as it induces the transport issue and this issue to be addressed by the company
* The temperature of machines in the warehouses should be monitored carefully as it may give rise to storage issues
* There are 5896 warehouses that are under performing and those warehouses should be reassessed and preventive efforts to be made by the company like increasing number of workers working in the warehouses
* The Location of the warehouse plays an important role in supply and demand of the product and those locations where there is low demands less quantity of product can be shipped and increasing the product quantity in areas of higher demand
* In areas of high demand, the retail shops should be increased so that there is enough product to be sold