Business Report

SMDM Project Business Report DSBA

*Capstone project – Supply Chain Management*

***Sanjay Srinivasan***

***PGP-DSBA Online***

***JULY’ 21 Batch***

***Date: 29-05-2022***

# *INDEX*

|  |  |  |
| --- | --- | --- |
| ***S. No*** | ***Contents*** | ***Page* No** |
| **1)** | **Introduction of the business problem** | 4 |
|  | a) Defining problem statement | 4 |
|  | b) Need of the study/project | 4 |
|  | c) Understanding business/social opportunity | 4 |
|  |  |  |
| **2)** | **Data Report** | 6 |
|  | a) Understanding how data was collected in terms of time, frequency and methodology | 6 |
|  | b) Visual inspection of data (rows, columns, descriptive details) | 7 |
|  | c) Understanding of attributes (variable info, renaming if required) | 8 |
|  |  |  |
| **3)** | **Exploratory data analysis** | 10 |
|  | a) Univariate analysis (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones) | 10 |
|  | b) Bivariate analysis (relationship between different variables , correlations) | 14 |
|  | c) Removal of unwanted variables (if applicable) | 16 |
|  | d) Missing Value treatment (if applicable) | 17 |
|  | e) Variable transformation (if applicable) | 18 |
|  | f) Addition of new variables (if required) | 18 |
|  | g) Addition of new variables (if required) | 20 |
|  |  |  |
| **4)** | **Business insights from EDA** | 21 |
|  | a) Is the data unbalanced? If so, what can be done? Please explain in the context of the business | 21 |
|  | b) Any business insights using clustering  (if applicable) | 21 |
|  | c) Any other business insights | 23 |

***List Of Figures***

|  |  |  |
| --- | --- | --- |
| ***S.No*** | ***Content*** | ***Page No*** |
| 1.1 | Sample data of product weight across years | 6 |
| 1.2 | Number of times frequently product refilled across zone | 6 |
| 1.3 | Sample dataset | 7 |
| 1.4 | Shape of the dataset | 7 |
| 1.5 | Description of the data. | 8 |
| 1.6 | Variable info of the dataset. | 8 |
| 1.7 | Missing value count in the dataset. | 9 |
| 1.8 | Missing value count in percentage. | 9 |
| 1.9 | Univaraite Analysis | 11 |
| 1.10 | Count plot for approved\_wh\_govt\_certificate | 12 |
| 1.11 | Count plot for Location\_type | 12 |
| 1.12 | Count plot for WH\_capacity\_size | 12 |
| 1.13 | Count plot for Zone | 13 |
| 1.14 | Count plot for Warehouse in regional zone | 13 |
| 1.15 | Count plot for Warehouse owner type | 13 |
| 1.16 | Boxplot for Product weight ton vs. zone across location type. | 14 |
| 1.17 | Stripplot for Product weight ton vs. flood\_impacted. | 14 |
| 1.18 | Violinplot for Product weight ton vs. electric\_supply with flood\_impacted. | 14 |
| 1.19 | Violinplot for Product weight ton vs. wh\_regional\_zone with wh\_capacity\_size. | 15 |
| 1.20 | Pairplot for the bivariate analysis. | 15 |
| 1.21 | Sample Correlation data for the bivariate analysis. | 16 |
| 1.22 | Plotting correlation in the heatmap for the bivariate analysis. | 16 |
| 1.23 | Shape of the dataframe after dropping unwanted variables | 17 |
| 1.24 | Info of the dataframe after dropping unwanted variables. | 17 |
| 1.25 | Missing values before treating. | 17 |
| 1.26 | Missing values after treating | 18 |
| 1.27 | Shape of the dataset after variable transformation. | 18 |
| 1.28 | Sample dataset after variable transformation. | 18 |
| 1.29 | Info of the dataset after variable transformation. | 19 |
| 1.30 | Transforming right skewed target variable. | 19 |
| 1.31 | Finding skewness of the data frame. | 20 |
| 1.32 | Boxplot before scaling | 20 |
| 1.33 | Boxplot after scaling | 21 |
| 1.34 | Dendogram of hierarchial clustering | 21 |
| 1.35 | Dendogram of hierarchial clustering after Truncating | 21 |
| 1.36 | Finding inertia from K-means clustering | 22 |
| 1.37 | Elbow curve for K – means | 22 |
| 1.38 | Silhouette score | 22 |
| 1.39 | Silhouette width | 22 |
| 1.40 | Silhouette Sample value | 23 |

# Model building and interpretation.

*a) Build various models (You can choose to build models for either or all of descriptive, predictive or prescriptive purposes).*

*b.) Test your predictive model against the test set using various appropriate performance metrics.*

*c.)Interpretation of the model(s).*

A machine learning model is **built by learning and generalizing from training data, and make predictions for the business problem.**

In regression analysis, model building **is the process of developing a probabilistic model that best describes the relationship between the dependent and independent variables**.

**Steps involved in model building:**

1. Problem Statement
2. Data Collection.
3. Data Cleaning.
4. Exploratory Data Analysis.
5. Model Development
6. Train the Model.
7. Test the Models.
8. Applying Models
9. Inferences, Recommendations and business insights based on the model

The four main analytical models organisations can deploy are:

1. Descriptive
2. Diagnostic
3. Predictive
4. Prescriptive

***Descriptive* -** ***It generally uses historical data*** from a single internal source to pinpoint when an event occurred. ***Descriptive analytics are often displayed on dashboards and in reports.***

***Diagnostic* -** A diagnostic model is a framework for **identifying, analysing and interpreting data in a given context to identify possible needs.** An effective diagnostic model **allows identifying reliable data to help clients better understand their company's strengths, deficiencies, and opportunities for improvement, to later articulate a targeted intervention and measurement strategy.**

***Predictive -*** Predictive modelling**is a mathematical process used to predict future events or outcomes by analyzing patterns in a given set of input data.**

***Prescriptive* -** Prescriptive analytics **utilizes similar modelling structures to predict outcomes and then utilizes a combination of machine learning, business rules, artificial intelligence, and algorithms to simulate various approaches to these numerous outcomes.**

***Model building is performed with the following Model:***

* Linear Regression
* Lasso Regression (L1 Regularization Model)
* Ridge Regression (L2 Regularization Model)
* Support Vector Model Regression
* Huber Regression
* Random Forest Regressor model
* Artificial Neural Network Regressor Model
* Ada Boost Regressor Model

***To find Parameters for Regression model:***

***RMSE (Root Mean Squared Error)*** - The root-mean-square error (RMSE) is a **frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed.** The best model will have the value in the **range 0.2 - 0.5**

***MSE (Mean Squared Error)*** - In Statistics, Mean Square Error (MSE) is defined as **Mean or Average of the square of the difference between actual and estimated values.** The best model will have the value lesser value (0 is the best value for the model).

***MAE (Mean Absolute Error)*** - The **MAE measures the average magnitude of the errors in a set of forecasts, without considering their direction.**

***R-Squared value* -** R-squared is a **statistical measure that represents the goodness of fit of a regression model.** The ideal value for r-square is 1. The closer the value of r-square to 1, the better is the model fitted. The best model will have the value higher value(1 is the best value for the model).

***Linear Regression Model:***

Linear regression analysis is used to **predict the value of a variable based on the value of another variable**. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

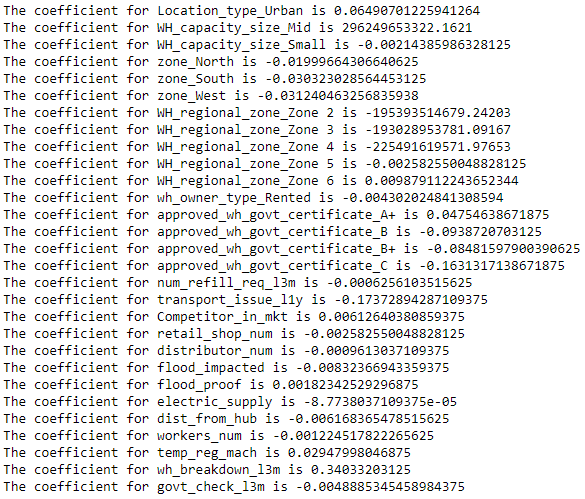


Fig 1.1 Coefficient for Linear Regression model



Fig 1.2 Intercept for Linear Regression model



Fig 1.3 R-Square for Linear Regression model

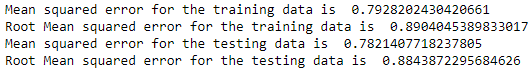


Fig 1.4 Mean Squared Error and Root Mean Squared Error for Linear Regression model



Fig 1.5 Mean Absolute Error for Linear Regression model



Fig 1.6 Mean Absolute Percentage Error for Linear Regression model

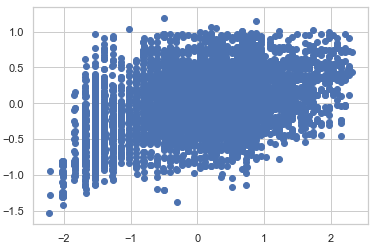


Fig 1.7 Scatter plot for Linear Regression model



Fig 1.8 Variance Inflation Factor for Linear Regression model

***Assumptions of Linear Regression:***

* Linear relationship.
* Multivariate normality.
* No or little multicollinearity.
* No auto-correlation.
* Homoscedasticity.

***Lasso Regression Model:***

In statistics and machine learning, lasso (least absolute shrinkage and selection operator; also Lasso or LASSO) is a **regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model.**

***Assumptions of Lasso Regression:***

* Linearity - linear regression needs the relationship between the predictor and target variables to be linear
* Independence
* No Heteroskedasticity
* No Multicolinearity.

The objective of this report is to find that, how the machine learning model supports the supply chain to overcome the demand and supply mismatch in every zone. A FMCG company has entered into the instant noodles business two years back. The data is gathered based on the FMCG Company’s demand and supply mismatch of the product instant noodles. The higher management has noticed that there is a mismatch in the demand and supply of instant noodles.

The demand and supply mismatch can be overcome by following these: first of all, finding the demand and supply mismatch. Secondly, find the optimum weight of the product been shipped to each warehouse at different zone and regions of the country.