Project Report

Logistics & Supply Chain Industry

Group 3

**Submitted To:** Somayeh Alizadeh

**Course Number:** BAN240

**Submitted By:** Group 3

Priyanka Patil (155305196)

Niyati Yogesh kumar Shah (138845193)

Pranshu Verma (138048194)

Shikhir Dodeja (139738199)

Munish Kumar (131412199)

Contents

[Business Understanding 3](#_Toc48317547)

[Business Objectives 3](#_Toc48317548)

[Project Goal 3](#_Toc48317549)

[Project Plan 3](#_Toc48317550)

[Data Understanding 4](#_Toc48317551)

[Collect Initial Data 4](#_Toc48317552)

[Data Description 4](#_Toc48317553)

[Data Explore 4](#_Toc48317554)

[Data Preparation Introduction 5](#_Toc48317555)

[Before Processing Data: 5](#_Toc48317556)

[After Processing Data: 8](#_Toc48317557)

[Steps: 9](#_Toc48317558)

[After Processing Data 27](#_Toc48317559)

[Correlation Matrix 28](#_Toc48317560)

[Modeling 29](#_Toc48317561)

[Select Modeling Technique 29](#_Toc48317562)

[Modeling Technique 1: Multiple Linear Regression 29](#_Toc48317563)

[Modeling Technique 2: Random Forest Regressor 29](#_Toc48317564)

[Modeling Technique 3: Decision tree 30](#_Toc48317565)

[Modeling Technique 4 Gradient boosting 30](#_Toc48317566)

[Modeling Technique 5: Neural Networks 30](#_Toc48317567)

[Test Design 31](#_Toc48317568)

[Build model 31](#_Toc48317569)

[Models: Liner Regression 31](#_Toc48317570)

[Models: Gradient Boosting Regression 32](#_Toc48317571)

[Models: Random Forest Regression 34](#_Toc48317572)

[Models: Decision Tree Regression 35](#_Toc48317573)

[Models: Neural Network 37](#_Toc48317574)

[Assess Model 39](#_Toc48317575)

[Model assessment 39](#_Toc48317576)

[Evaluation 40](#_Toc48317577)

[Evaluate your results: 40](#_Toc48317578)

[Review process: 40](#_Toc48317579)

[Determine next steps: 41](#_Toc48317580)

[Business Impact 41](#_Toc48317581)

# Business Understanding

## Business Objectives

Our topic is related to how the supply chain industry is affected during pandemic situations such as COCID-19. We are trying to understand how a pandemic situation impacts supply chain, what are the consequences and challenges faced by the industry players. The supply chain industry is a very complex industry consisting of lots of uncertainties. However, at the same time it is booming at a tremendous rate as the consumption is increasing, with that the technology which is helping in easing the process of the supply chain. Our aim in this project was to understand the different factors involved in the shipment industry and which are the important ones that are having a bigger impact on the industry. Furthermore, the motive is to find ways in which supply chain industry is well prepared and help out the government and the rest world in pandemic situations such as COVID-19.

## Project Goal

The project goal for our topic is to predict the shipment value and what are the factors affecting the shipment. Furthermore, the goal is to find out which factors have the highest impact when it comes to predicting the shipment value. The aim is also to find out which factors does not have any impact on the target variable. The target is to find out the best possible model with highest accuracy and minimum errors. On a personal level, the goal is to learn the data related processes- be it cleaning, manipulation, exploring the dataset, running different machine learning models, evaluating the results, researching about different topics related to the industry and data science, etc.

## Project Plan

The projected started months ago when we chose the industry – Logistics and Supply Chain. Then going into further details about the industry especially during a pandemic situation. For this project, we searched for a numerous number of datasets on different websites such as Kaggle and finally selected this one as it involves medical supply goods and is a huge dataset. After selecting the dataset, we went into details about the dataset by exploring the dictionary well, finding out the meaning of certain terms present there. Then we explored all the variables one by one, looked for missing values and outliers, and made best adjustments possible for each error present. We performed statistics to look for skewness, correlation among variables and same way made few adjustments here and there. Finally, we moved onto the machine learning models and here we tested our dataset by running on different machine models and then finally selected the one with minimal errors and highest accuracy. We used different tools and techniques such as Python, SAS enterprise miner, MS-Excel to accomplish this project. It was a difficult and challenging task to complete this task as the dataset was huge with large number of outlier’s present, so we had to make certain adjustments. As the dataset was huge, it took a lot of time to run models due to limited commuting power that we possess as of now. However, it was an interesting project which helped us to gain a lot of knowledge and experience in data science field.

Data Understanding

## Collect Initial Data

We collected the data from Kaggle website by searching extensively there as well on other similar websites. Format: Datasets are available in the .txt and .csv file format

## Data Description

The dataset gives away supply chain information of goods including medical supplies in US during a pandemic situation such as COVID-19.

The name of the dataset is cfs-2012-pumf-csv and it has the following fields:

Fields: "'SHIPMT\_ID', 'ORIG\_STATE', 'ORIG\_MA', 'ORIG\_CFS\_AREA', 'DEST\_STATE',

'DEST\_MA', 'DEST\_CFS\_AREA', 'NAICS', 'QUARTER', 'SCTG', 'MODE',

'SHIPMT\_VALUE', 'SHIPMT\_WGHT', 'SHIPMT\_DIST\_GC', 'SHIPMT\_DIST\_ROUTED',

'TEMP\_CNTL\_YN', 'EXPORT\_YN', 'EXPORT\_CNTRY', 'HAZMAT', 'WGT\_FACTOR'

# Records: 4547661

## Data Explore

The main transactional dataset consists of 20 columns and 4.5million records in total.

Out of 20 columns, 13 are quantitative types and rest 7 are character types.

There is a variable ‘SHIPMT\_VALUE’ which depicts about the value of a shipment i.e. its price

The first variable SHIPMT\_ID does not have any duplicate records so it can be treated as an index and each id can be related to a shipment  
Data Quality

There are no missing values in any of the columns in the dataset

There are extreme values in the some of the quantitative variables such as SHIPMT\_VALUE

All the data combined and analyzed well tell us regarding the shipment value of a good or a commodity. We will create dummy variables for certain columns

Yes, it can be considered as a big dataset as the total records are around 4.5 million.

Our data set is both huge and capable enough to give us insights for the supply chain information of critical goods.

In the real world, we get unstructured data more than the structured data. We cannot expect good quality readymade dataset, we must work on it-clean and transform it into a structured dataset. As we all know, more the cleaned data, better our output result is for the model. Data quality is highly important to get the right prediction from our model. Better the dataset, better our model would be. In this scenario as well, we received a dataset with some issues which needed to be fixed. Starting from understanding the data dictionary to treating the outliers and coping with

# Data Preparation Introduction

In this, our team worked on importing, analyzing, and correcting the dataset using various techniques such data cleaning, data accuracy, data transformation, and data reduction.

Before Processing Data**:**

* Variables:

Categorical: 14

Quantitative: 6

* Observations: 4547661
* Data Dictionary:

|  |  |  |  |
| --- | --- | --- | --- |
| **CFS PUM File Data Dictionary** |  |  |  |
| **Field** | **Description** | **Type** | **Length** |
| **SHIPMT\_ID** | Shipment identifier | NUM | 7 |
| **ORIG\_STATE** | FIPS state code of shipment origin | CHAR | 2 |
| **ORIG\_MA** | Metro area of shipment origin | CHAR | 5 |
| **ORIG\_CFS\_AREA** | CFS Area of shipment origin | CHAR | 8 |
| **DEST\_STATE** | FIPS state code of shipment destination | CHAR | 2 |
| **DEST\_MA** | Metro area of shipment destination | CHAR | 5 |
| **DEST\_CFS\_AREA** | CFS Area of shipment destination | CHAR | 8 |
| **NAICS** | Industry classification of shipper | CHAR | 6 |
| **QUARTER** | Quarter in which the shipment occurred | CHAR | 1 |
| **SCTG** | 2-digit SCTG Commodity Code of the shipment | CHAR | 5 |
| **MODE** | Mode of transportation of the shipment | CHAR | 2 |
| **SHIPMT\_VALUE** | Value of the shipment in dollars | NUM | 8 |
| **SHIPMT\_WGHT** | Weight of the shipment in pounds | NUM | 8 |
| **SHIPMT\_DIST\_GC** | Great circle distance between ship-ment origin and destination (in miles) | NUM | 8 |
| **SHIPMT\_DIST\_ROUTED** | Routed distance between shipment origin and destination (in miles) | NUM | 8 |
| **TEMP\_CNTL\_YN** | Temperature controlled shipment - Yes or No | CHAR | 1 |
| **EXPORT\_YN** | Export shipment - Yes or No | CHAR | 1 |
| **EXPORT\_CNTRY** | Export final destination | CHAR | 1 |
|  |  |  |  |
|  |  |  |  |
| **HAZMAT** | Hazardous material (HAZMAT) code | CHAR | 1 |
|  |  |  |  |
|  |  |  |  |
| **WGT\_FACTOR** | Shipment tabulation weighting factor | NUM | 8 |

**After Processing Data:**

* Variables:
* Index: 1

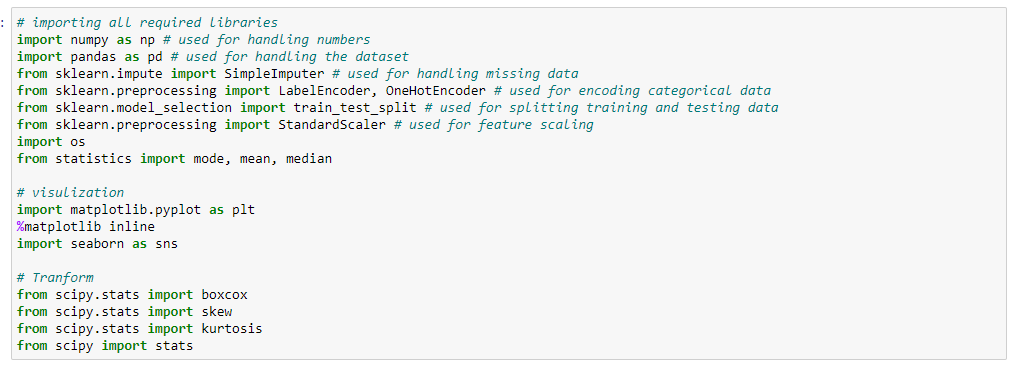
Numerical: 276

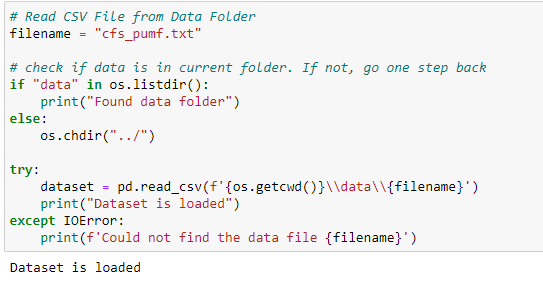
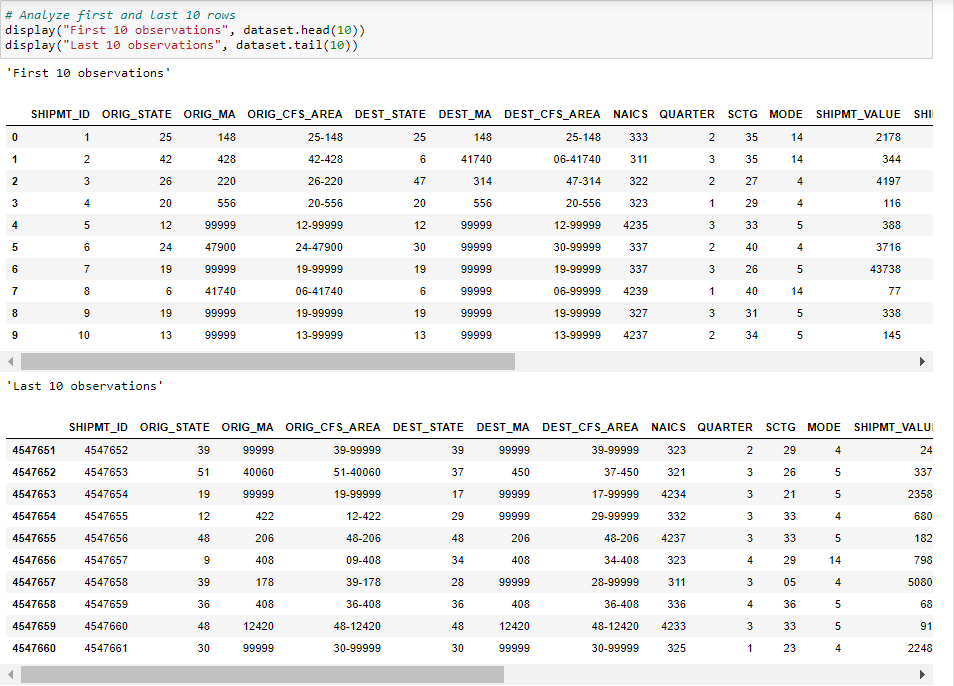
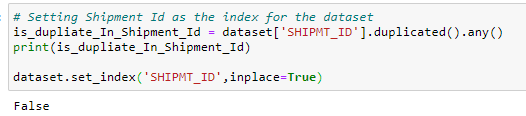
Continuous: 5

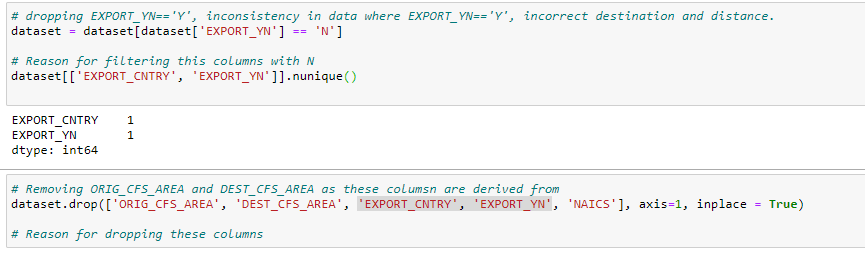
* Dummy: 271
* Target Variable: SHIPMT\_VALUE
* Dropped: 5
* # Observations: 3,792,057

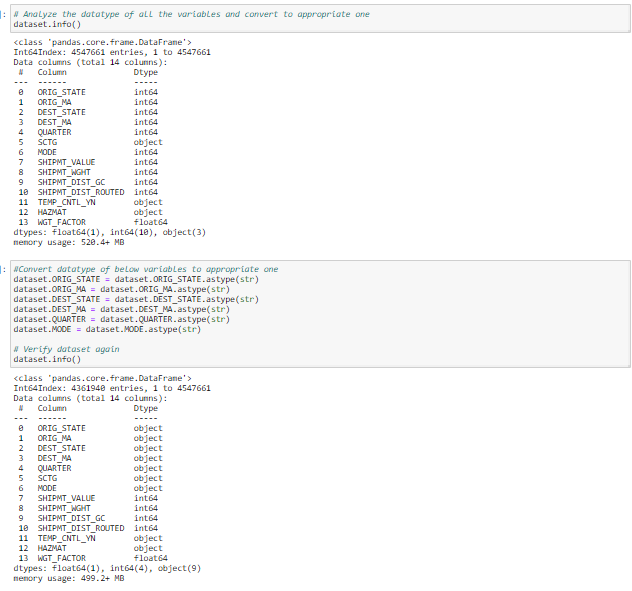
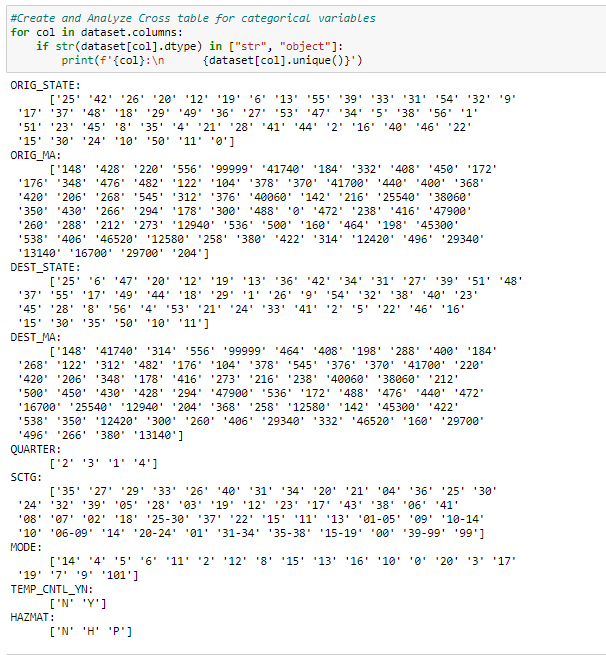
**Steps**:

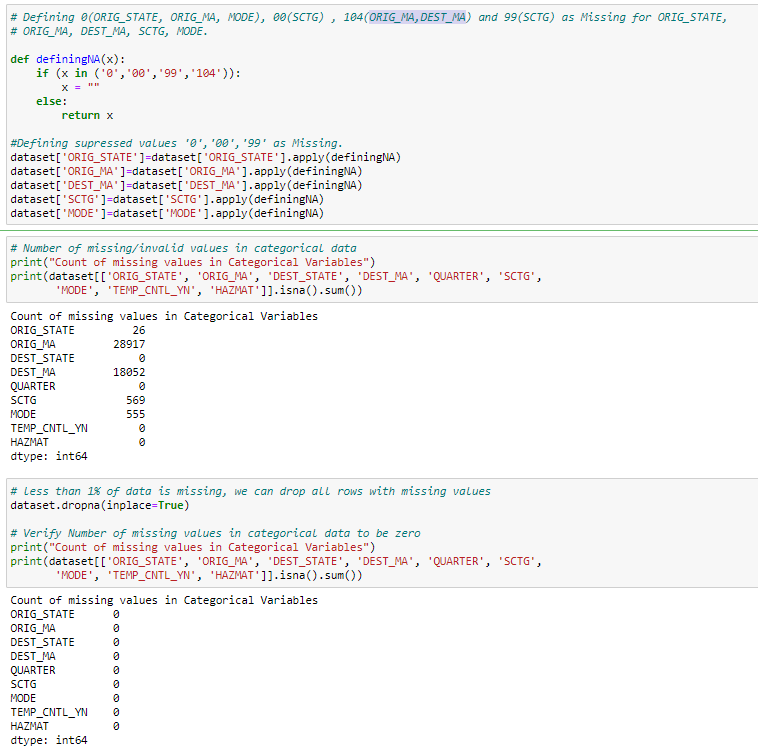
1. Libraries imported for data pre-processing:



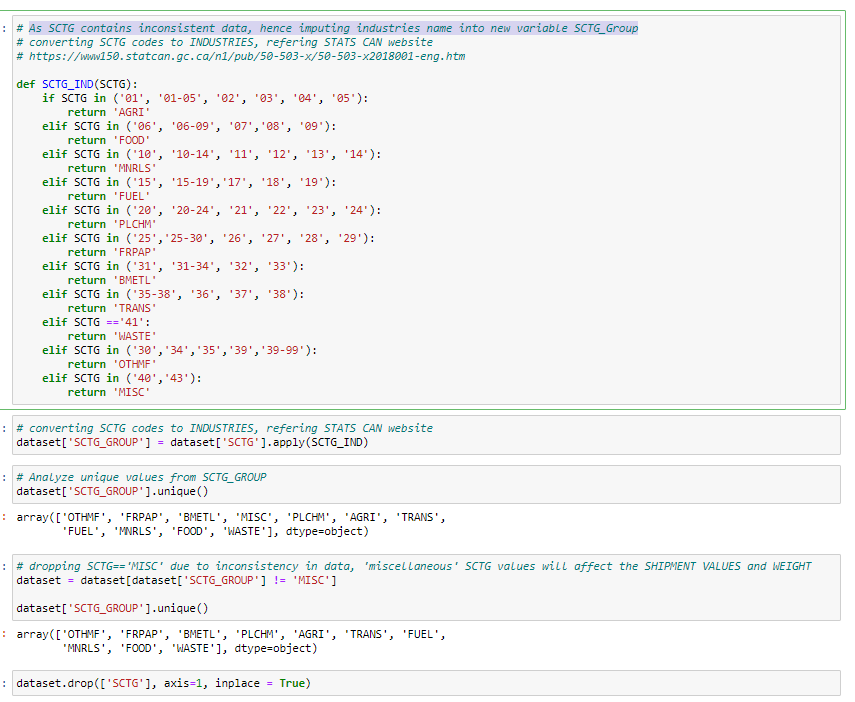
1. Importing the dataset:  
   
2. Snapshot for first and last 10 records of the dataset  
   
3. We found that SHIPMT\_ID does not contain any missing or duplicate values, so we marked it as column index.  
   
4. From the EXPORT\_YN, we removed the observations with Export as Y because of incorrect destination and distance.
5. Team decided to remove **ORIG\_CFS\_AREA and DEST\_CFS\_AREA** becausethese columns are derived from **ORIG\_STATE, ORIG\_MA and DEST\_STATE, DEST\_MA** respectivelyby concatenation
6. As we already filtered the records from EXPORT\_YN so EXPORT\_YN and EXPORT\_CNTRY are of no use because there is only one single value. i.e. “N”

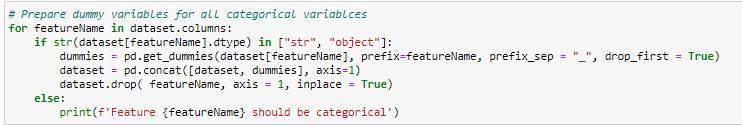


1. For categorical features, datatype was imputed as “int” whereas it should be either “str” or “object”. So, we manually updated the datatype.  
   
2. Preprocessing for Categorical Variables
   1. Analyzing unique values:  
      
   2. We found few missing/invalid values such as (0, 00, 99, 104) in ORIG\_STATE, ORIG\_MA, MODE, SCTG, DEST\_MA as per data dictionary. So, we removed those observations which stands for less than 1% of overall dataset.

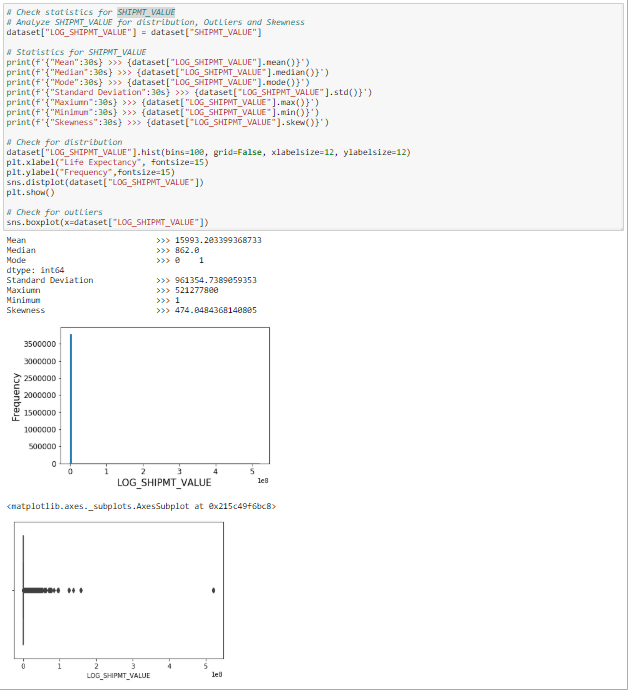
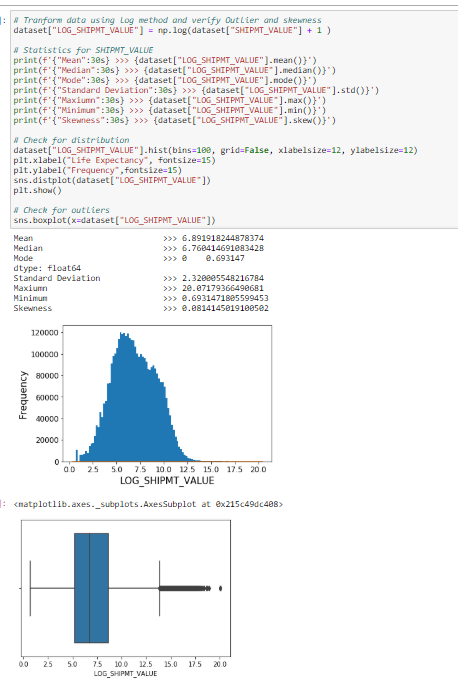
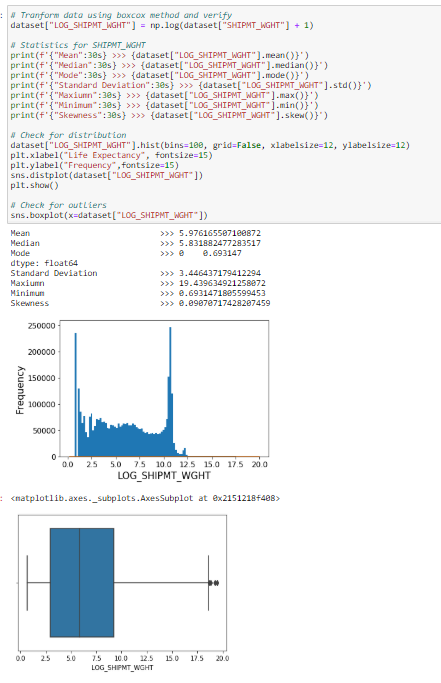
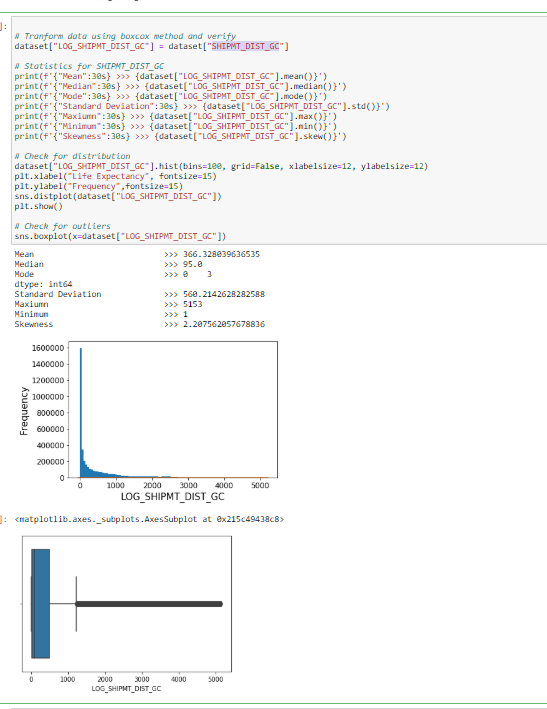
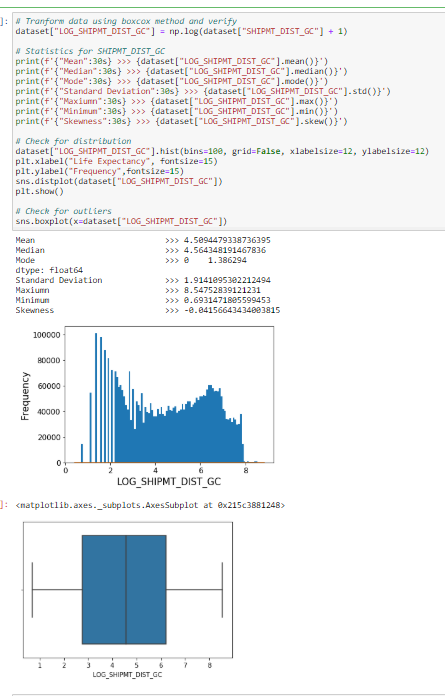
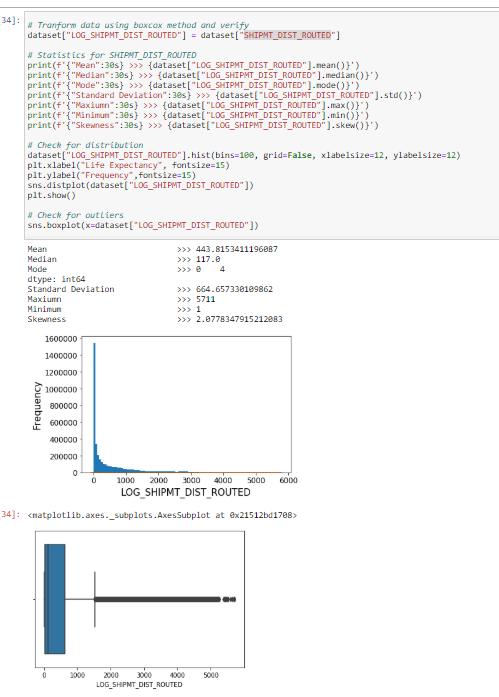
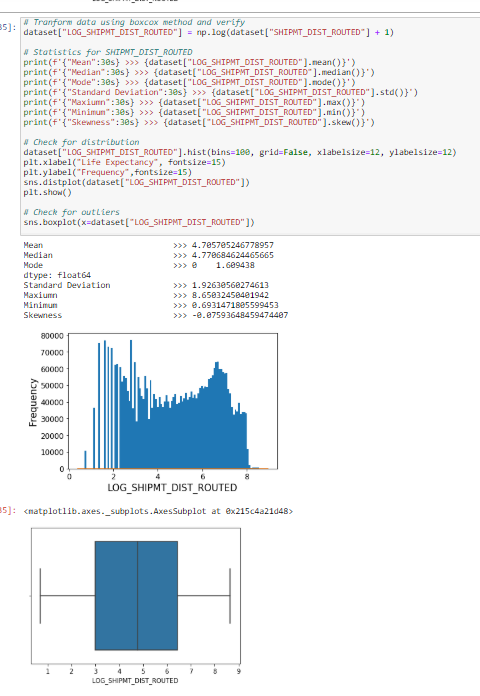
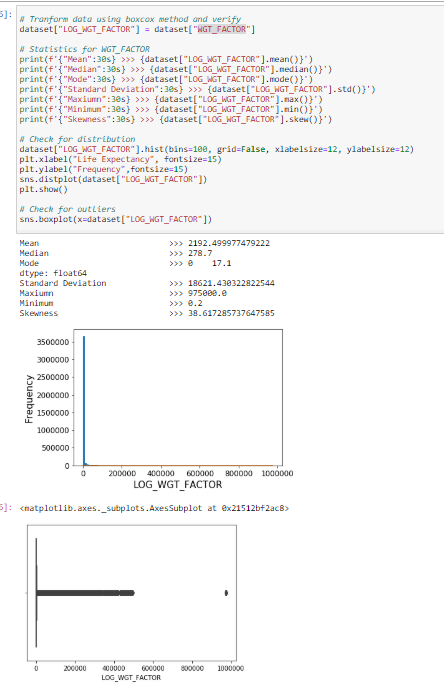
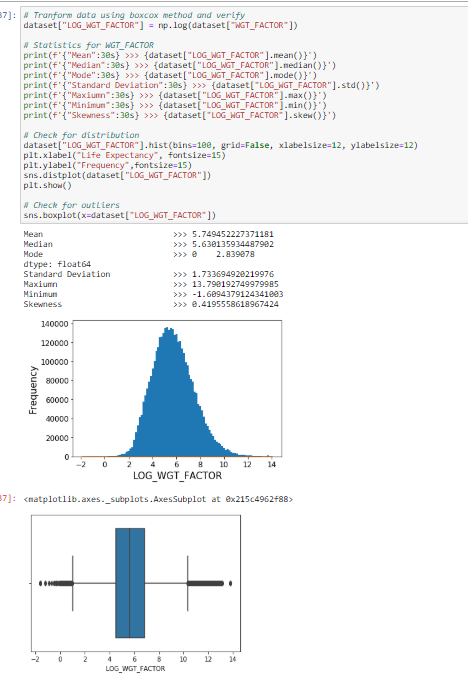


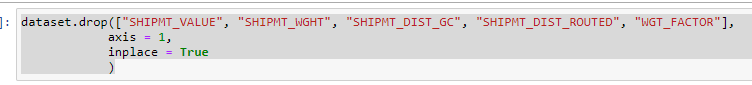
* 1. As SCTG contains inconsistent data such as few observations contain group number such as 01-05, 06-09 whereas other contains individual industry codes. Hence, we decided to gather industries name from “**STATS Canada”** andderived a new variable named as SCTG\_ Group. Also, there were few codes such as 40 and 43 where industry name was not mentioned. So, we grouped them into one and removed them.



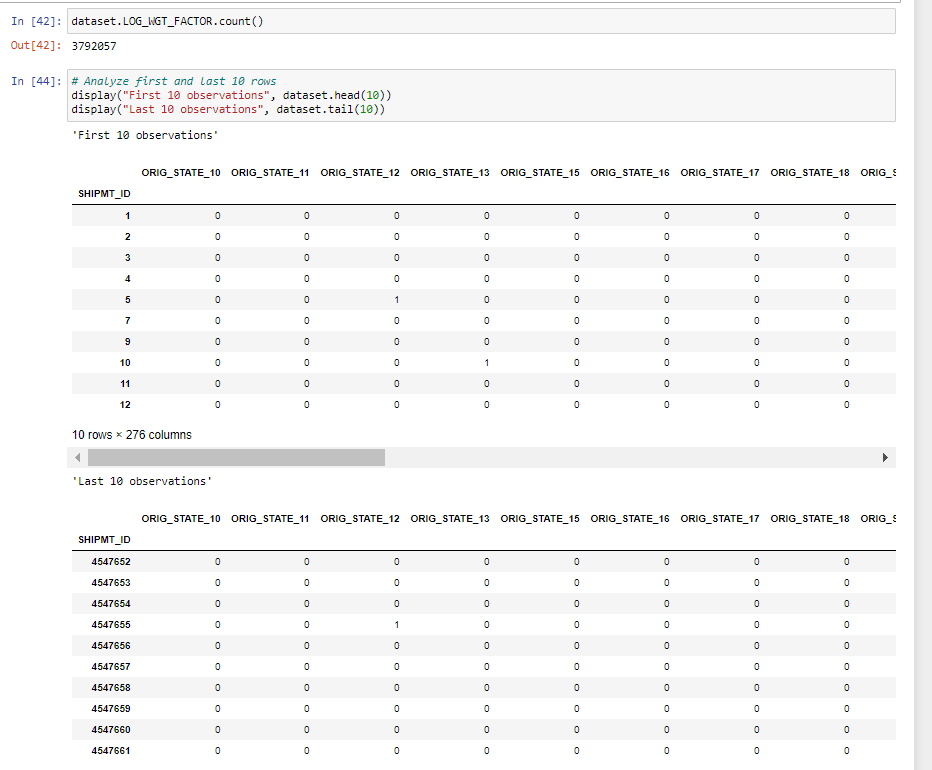
* 1. For all the categorical variables, we created dummy variables as we are planning to perform regression analysis.  
     

1. Preprocessing for Quantitative Variables
   1. Analyzed “**SHIPMT\_VALUE**”, “**SHIPMT\_WGHT** “, “**SHIPMT\_DIST\_GC**”, “**SHIPMT\_DIST\_** **ROUTED**” and “**WGT\_FACTOR**” features and found that there is too much noise all of these. We also found that data is highly skewed and there are many outliers. We also observe that the variance if quite high. So, we decided to normalize the data and transform it. We tried various techniques such as **box-cox, cube, exponential and log**. After analyzing all, we found that **log (x + 1)** is the best possible formula to transform the data. Though it removed the skewness, yet there are many outliers which cannot be removed as they are the important datapoints.  
        
      Before Transformation

  
After Transformation:  
  
Shipment Weight  
  
  
SHIPMT\_DIST\_GC  
  
  
SHIPMT\_DIST\_ROUTED  
  
  
WGT\_FACTOR  
  


* 1. After the log transformation, we removed the columns with raw data and kept the transformed features.  
     
  2. We did not find any missing/incorrect/inconsistent data in the data. For example, we check the value and weight for value which either 0 or less than 0 as these could be invalid.

After Processing Data**:**

Snapshot of data after pre-processing:  


The skewness of the dataset, we made certain decisions and actions to improve the quality of our dataset.

Correlation Matrix**:**

**We f**oundhigh correlation between SHIPMT\_DIST\_GC AND SHIPMT\_DIST\_ROUTED, hence team decided to remove SHIPMT\_DIST\_GC. ****

# Modeling

As per the supply chain capstone project, we are working on building a modelthat will predict the shipment price. From the obtained that we found labeled data with several categorical and continuous variable. As the label was continuous, we selected various regression algorithm and applied on the pre-processed data

## Select Modeling Technique

### Modeling Technique 1: Multiple Linear Regression

For the first model, team decided to go with MLR (Multiple Linear Regression) with all the variable selected after removing the multi-colinear variables.

**Modeling Assumptions:**

For MLR, there should be more than 2 independent variables. Linear relationship must exist between dependent and independent variables. We analyzed it by using correlation matrix. MLR requires residuals to be normally distributed. Independent variables should not be correlated with each other i.e. no multicollinearity. We found SHIPMT\_DIST\_GC and SHIPMT\_DIST\_ROUTED to be highly correlated. So, we removed one of them. Various of error term should be similar

Modeling Technique 2: Random Forest Regressor  
It is one of the most important algorithms in data science/machine learning. Random forest is a supervised learning algorithm that uses ensemble learning method to perform classification and regression. It is a bagging technique, not a boosting technique, wherein the trees here run in parallel. While building the trees in random forest, there is no interaction. It runs by developing a multitude of decision tree at training time and outputting the class which is the mean prediction (regression) of the individual trees.Basically, it is a meta-estimator which combines the result of multiple predictions.

There are no as such assumptions for random forest which is the special thing about it, however there are some modifications:

1. To ensure that the ensemble model does not rely majorly on any individual feature by limiting the number of features that can be split on each node to some percentage of the total.

2. To avoid overfitting, each tree draws a random sample from the original data set, adding some randomness.

### Modeling Technique 3: Decision tree

Decision Tree algorithm is a supervised learning algorithm. Decision tree algorithm can be used for solving regression and classification problems too. We have performed decision tree regressor for our dataset and used CART technique.

Assumptions while creating Decision Tree.

The below are the some of the assumptions we make while using Decision tree:

1. At the beginning, no roots are defined. The whole training set is considered as the root.
2. Attributes are preferred to be categorical. If the values are continuous then they should be converted to discrete values before starting to build the model.
3. Records are distributed recursively based on importance of variables. For example, in our case the most important variable to split is SHIPMT\_WGHT.

### Modeling Technique 4 Gradient boosting

For the second model, team decided to go with an ensemble model, and we wanted to implement boosting so, we decided to implement Gradient Boosting regressor with all the variables and overlapping the samples.

Assumptions while creating Gradient boosting model

The below are the some of the assumptions we make while using Gradient boosting:

1. Independence of observations
2. Assumptions related to the interaction depth. If set to 1, strictly additive model is assumed. As we increase the interaction depth, this assumption is relaxed.

### Modeling Technique 5: Neural Networks

We decided to implement Neural Network as one of our models. We applied the most basic model i.e. a sequential model with all the layers densely connected. It can be considered as a basic feed forward network where every neuron is connected every other neuron in the next layer.

Assumptions while creating a sequential neural network model:

1. Neurons within the same layer do not interact or communicate to each other.
2. Artificial Neurons are arranged in layers, which are sequentially arranged.
3. All inputs enter into the network through the input layer and passes through the output layer.
4. All hidden layers at same level should have same activation function.

Test Design**:**

For the testing purpose, we divided the dataset into train and test with 80:20 ratio. We planned to use R-square, Mean Absolute Error and Mean square error to check the performance of each model. We will use R-square of train and test to check if the model is overfitting or underfitting. Also, we are planning to run different sets of train and test using different random seed.

# Build model

## Models: Liner Regression

Model descriptions: Linear regression with 267 variables and 1 label variable. The model performed quite well. 70% of the variance can be explained by this model with mean absolute error is $9081. However, as the data is skewed, the highest error seen is $ 987594 which is huge.

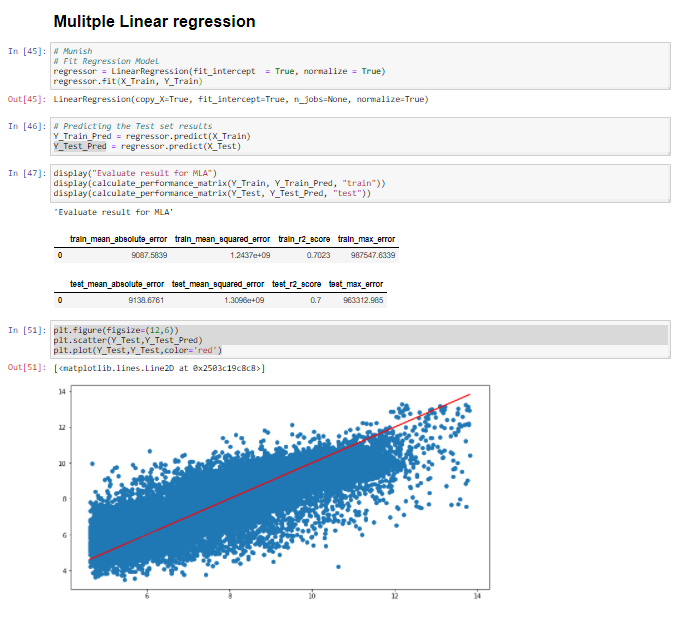
Parameter settings:

fit\_intercept: True

Set to true to calculate the intercept.

Normalize: True

Normalize all the input variables before using them in regressor.



## Models: Gradient Boosting Regression

Model descriptions: Gradient Boosting is an ensemble-based machine learning algorithm which combines (additive model) many weak predictive trees sequentially to provide improved predictive accuracy. GBM implemented by us tries to optimize the differentiable mean squared loss function through multiple iterations. We tuned the model by running multiple iterations of it using different sets of hyperparameters. In the end, the model with hyperparameters mentioned in in the parameter setting section below is selected because of high R-square value which indicates the power of model in terms of explaining the variability of the target variable and the stability of the model across training and validation samples. The model provided satisfactory R-square of 79% on training dataset and 77% on validation dataset.

Parameter settings:

1. Loss Function: Least squares regression

Classic loss function for regression implementation.

2. learning\_rate:1

Learning rate/shrinkage indicates the scale by which to shrink the contribution of each tree. Tried different values of learning rate (e.g. 0.01, 0.1, 0.5, 1,2) and finally tuned it to the value of 1 as it resulted in best model.

3. n\_estimators: 100

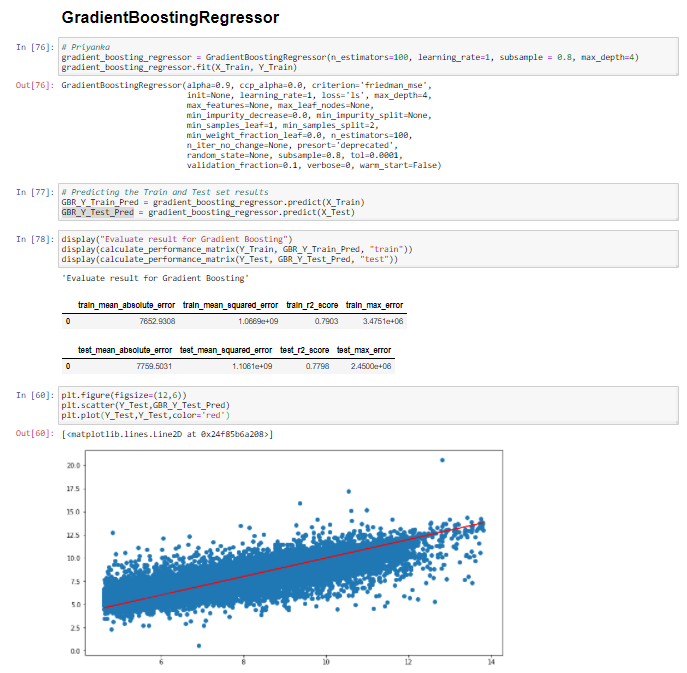
n\_estimators represents the number of trees to use in GBM. Tried values of 50, 100 and 500. Selected 100 as it was most feasible to train the model considering the time and resources required and it resulted in better results compared to 50 and 500.

4. subsample: 0.8

to implement stochastic gradient boosting wherein, 80% of data was sampled out for each iteration. this is generally done to reduce the variance and increase the bias

5. max\_depth: 4

maximum depth of the individual predictor tree. Tuned it to 4 as it gave better results compared to default value of 2 and values higher than 4.



## Models: Random Forest Regression

Model descriptions: Random forest is a supervised learning algorithm that uses ensemble learning method to perform classification and regression. It is a bagging technique, not a boosting technique, wherein the trees here run in parallel. As we saw, that the accuracy for multiple linear regression is not accurate, we decided to go with ensemble modelling for which we chose random forest regressor.

Parameter settings: max\_depth= it helps to reduce the overfitting to increase the accuracy

n-estimators= Ideal number of estimators to get the best accuracy is by keeping the number of estimtors between 64 and 128.



## Models: Decision Tree Regression

Model descriptions:

The target value, shipment value is continuous, so it is a decision tree regressor problem. In the cart model technique, interval target criterion is generally set to “variance” and nominal target criterion is set to “Gini”.

Decision tree with 267 variables and 1 label variable. The model performed quite well. Total of 70% of variables can be explained by the model with mean absolute error $9294. However, as the data is highly skewed, the highest error generated is $971166, which is huge.

Parameter settings:

max\_depth: The first parameter we tuned is max\_depth. We fit a decision tree with depths ranging from 6 to 16 and checking the R-square value and mean absolute error. The deeper the tree, the more splits it has and it captures more information about the data.

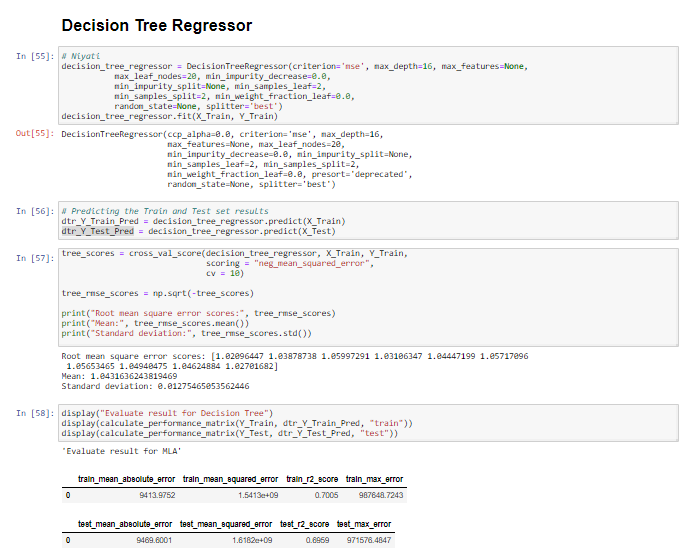
min\_samples\_leaf: The minimum number of samples we tried with values of 1 and 2 to be at a leaf node. It describes the minimum number of samples at the leaf’s.

Checking all the possibilities, decision tree works best with the hyper parameters of following:

Max\_depth: 16. After that it start to overfit.

min\_samples\_split: 10%

min\_samples\_leaf: 2



## Models: Neural Network

Model descriptions: Sequential Neural Network with densely connected layers and early stopping algorithm to prevent over fitting of the model. Loss function here in this model is ‘mse’ and optimizer as ‘adam’. We tuned the model by running multiple iterations of it using different sets of hyperparameters.

Model performs satisfactory with training and test R-sq. equal to 0.86 and 0.82 respectively. Mean Sq. error for training and test set equal to 6426 and 5987.

Parameter settings:

Input layer- 1

Hidden layer- 4

No. of nodes in hidden layer- 268

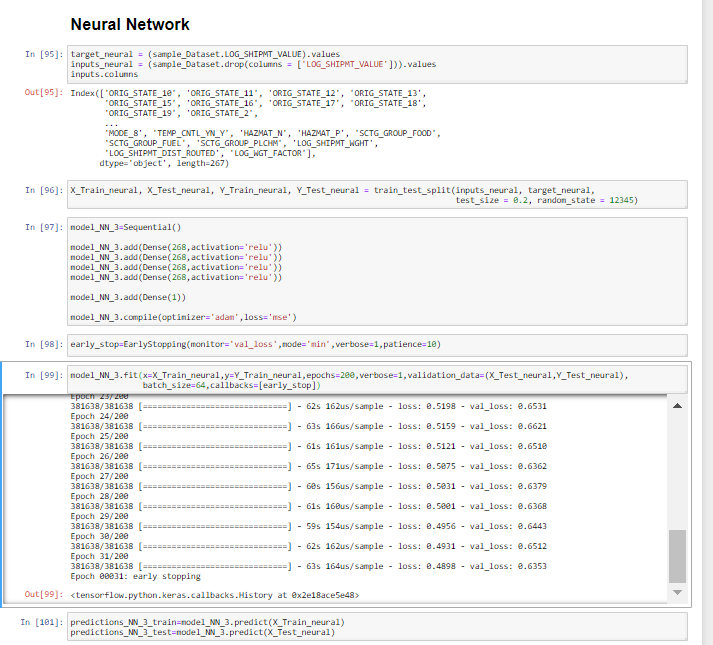
Output layer- 1

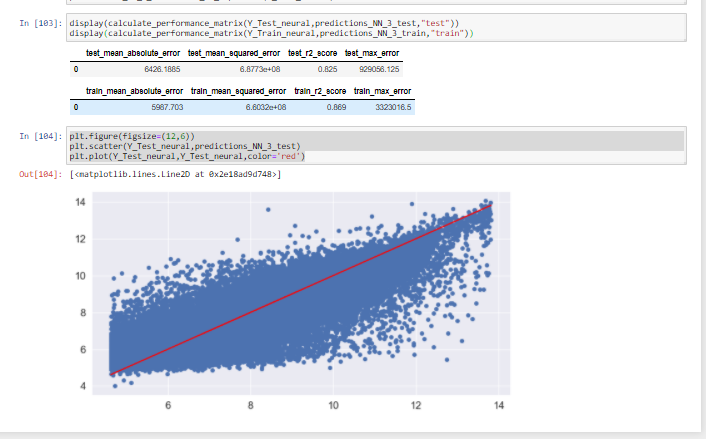
Epochs-100, early stopped at 31

Optimizer- adam

Loss- mse

early stopping with patience=10, val\_loss will be considered to stop the epochs.





# Assess Model

## Model assessment

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | Multiple Linear Regression | Decision Tree | Random Forest Regressor | Gradient Boosting | Neural Networks |
| R-square | Train | 0.70 | 0.70 | 0.9198 | 0.79 | 0.869 |
| Test | 0.70 | 0.70 | 0.8399 | 0.78 | 0.825 |
| Mean Absolute Error | Train | 9087.58 | 9413.98 | 4148.3454 | 7652.93 | 5987.703 |
| Test | 9138.68 | 9469.60 | 5463.2436 | 7759.50 | 6426.1885 |
| Mean Square Error | Train | 1.24E+09 | 1.54E+09 | 3.28E+08 | 1.07E+09 | 660320000 |
| Test | 1.31E+09 | 1.62E+09 | 5.40E+08 | 1.11E+09 | 687730000 |
| RMSE | Train | 35266.13 | 39259.39 | 18100.83 | 32663.44 | 25696.69 |
| Test | 36188.40 | 40226.86 | 23230.58 | 33258.08 | 26224.61 |

# Evaluation

Evaluate your results:  
For predicting the Shipment Value, we decide to keep the cut off value for R-square as 80%. We found that there are two models which are Random Forest Regressor and Neural Networks with train accuracy as 93% and 86% and test accuracy as 84% and 83%. As there is difference of more than 5% in the train and test R-square for Random Forest Regressor, this model is overfitting on the training set. So, we rejected the Random Forest Regressor and selected the “Neural Networks” model for now.

## Review process:

Thorough analysis of the dataset was done yet there is still skewed, and few features are missing from the dataset which should be captured to improve the results.

Most of the models required huge computational power which was lacking in current situation as we have more than 250 features along with 4 million observations.

Team thoroughly checked the hyperparameters and analyzed the use of them and their effects conceptually and discussed within the team.

Information and opinion related to different models was shared amount the team which further helped the team to build various models.

As the number of features are high, we tried to use PCA or clustering however which is again computationally expensive, so we avoided it.

As stated, many of the times, we mostly concentrated on data preprocessing and did not consider the algorithms in advance. Team suggested to do data preprocessing keeping the various algorithms in mind to be used.

## Determine next steps:

1. Using GridSearchCV to tune the hyperparameters for models build earlier.
2. Analyze if any other Predictive model can be implemented.
3. Implement PCA and/or clustering to reduce the number of variables.
4. Find missing features which will helps to explain the remaining variance.
5. Team found that the data contains both the shipment and delivery data because of which data is highly skewed so, we are planning to cluster the data in their subcategories and build different predictive models.

# Business Impact

As for this project, we selected Neural Network model because it met the specified conditions of R-square i.e. 80%. So, this model can be implemented in estimating the shipment cost in the logistics. By implementing this model, value of the shipment can be identified early, and it can help to avoid the possible deviation of around $5000 to $7000. It will help for retailer and buyers to know the amount being spend in the shipmen. Also, this model can then further be analyzed to identify the route with possible least amount of shipment cost and plan the shipment ahead.