**Project Report**

**On**

**Freight Management**

**(Shipment Cost Prediction)**

**Rajesh Kumar (#800)**

**INSOFE Student (19th Batch)**

**Purpose**

*The purpose of the project is to provide a method/tool that can support the transportation cost variances analysis and provide the best shipment cost to the customer of “****Quick Freight****” company. This method or tool will help to predict future shipment costs and also will help to identify the ideal mix of each cost element needed to hit the targeted costs/profit.*

*Provide an innovative and useful tool for “****Quick Freight****” to report transportation cost to customer, and at the same time provides information regarding cost saving opportunities for their customers and the company itself.*

***Who is Freight Forwarder/Broker?***

*A freight broker is an individual or company that serves as a liaison between another individual or company that needs shipping services and an authorized motor carrier.*

*Though a freight broker plays an important role in the movement of cargo, the broker doesn't function as a shipper or a carrier. Instead, a freight broker works to determine the needs of a shipper and connects that shipper with a carrier willing to transport the items at an acceptable price.*

**State-of-the-Art**

***Domain Understanding:***

***Supply chain management (SCM):*** *is the management of the flow of goods and services, involves the movement and storage of raw materials, of work-in-process inventory, and of finished goods from point of origin to point of consumption.*

***Logistics management:*** *is the part of supply chain management that plans, implements, and controls the efficient, effective forward, and reverse flow and storage of goods, services, and related information between the point of origin and the point of consumption in order to meet customer's requirements.*

**

***Logistics understanding:*** *A freight forwarder should be familiar with the customs rules and regulations of every country through which freight will pass, in addition to understanding the associated service parameters and costs. Equally important is the forwarder’s knowledge of Incoterms (the standard international shipping terms) and demonstrated ability to manage transportation, risk, and costs, and to establish advantageous transport and delivery terms.*

***Customer service emphasis:*** *From a door-to-door service backed by customs clearance, storage, and distribution to a straight one-off transaction, a freight forwarder should understand and adapt to meet shippers’ needs. The objective is cost and time efficiency, with minimum downtime and as few obstacles to delivery as possible, complemented by maximum uptime for freight delivery.*

***State-of-the-art technology:*** *Sophisticated online tracking technology is a must for a competitive freight forwarder. Shippers should receive protected system passwords for complete confidentiality, enabling them to access accurate real-time data about shipment location. Simply by entering a purchase order number, shippers should be able to see the whole picture.*

**Method**

*In order to provide a better visibility on methodology and processes, I have created the following chart to show the steps of the data analysis:*

***Step 1) Pre- Processing:***

*Data provided by the company was pre-processed (e.g. handling the missing & outliers).*

***Step 2) Uni-Variate and Bi-Variate Analysis:***

*Uni-Variate and Bi-Variate analysis (by using plots) was done on multiples variables to find the correlations and importance of variables.*

***Step 3) Feature Engineering:***

*Featuring engineering was done to create the new variables like DELIVERY\_DAY from existing variables.*

***Step 4) Modelling***

1. ***Regression analysis****:*

*Regression analysis was used to individually test the correlation of each of the 10 variables with ORDER\_COST. Regression analysis has been applied to each set of data in order to obtain coefficients and indicators that statistically prove significant impact on ORDER\_COST.*

*After running the regression analysis and obtaining the results, some of the variables ended up not having any significance or correlation with the target value, in this case the freight cost. Some variables were not included in the analysis due to a lack of significance or impact on ORDER\_COST.*

1. ***GLM (Generalized Linear Models):***

***LASSO*** *creates a regression model that is penalized with the L1-norm which is the sum of the absolute coefficients. This has the effect of shrinking coefficient values (and the complexity of the model), allowing some with a minor effect to the response to become zero.*

***Ridge Regression*** *creates a linear regression model that is penalized with the L2-norm which is the sum of the squared coefficients. This has the effect of shrinking the coefficient values (and the complexity of the model) allowing some coefficients with minor contribution to the response to get close to zero.*

***Elastic Net*** *creates a regression model that is penalized with both the L1-norm and L2-norm. This has the effect of effectively shrinking coefficients (as in ridge regression) and setting some coefficients to zero (as in LASSO).*

1. ***Clustering:***

*Applied the “****K-means clustering****” (k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster).*

1. ***Random Forest:***

*Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees.)*

***Bottlenecks:***

*To manage the Zip codes variables for Regions and Markets.*

*Around 10k orders where duplicate in dataset provided by customer.*

**Data**

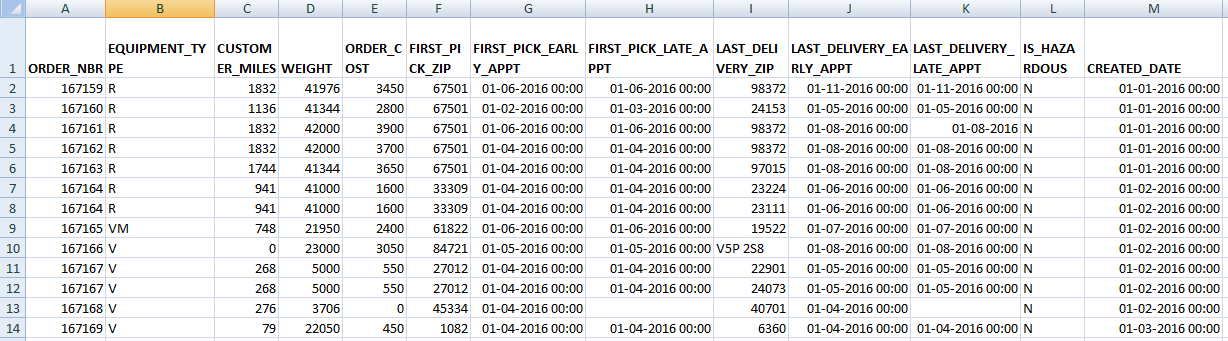
*A telephonic meeting with the “****Quick Freight****” representative (In our case* ***INSOFE*** *rep (Jagan)) was made in order to define expectations of the project. In this meeting, many hypotheses to their problems were treated from different angles in order to identify as many cost drivers and variables as possible that might be having some sort of impact on their total shipment costs. After identifying the variables, the company provided a data in excel sheet.*

*The above is industry specific process, but as I am the student of “INSOFE”, all the above communication was done by “****INSOFE****” employee and the raw data was provided to us in* ***Microsoft Excel*** *Sheet.*

*The company provided the mentioned data (Orders Created) with a time frame from* ***Jan 2016 until May 2016****. This timeframe would enable the analysis to find a big enough pattern to identify cost variances and its impact on shipment costs.*

*Around* ***37699*** *data records were contained in the excel sheet sent from the “****Quick Freight****” for the analysis.*

***Data Snapshot:***



***Explanation of all the Attributes:***

***#ORDER\_NBR***

*An unique number/ID for each new orders booked by customers of “Quick Freight”*

***#CUSTOMER\_MILES (DISTANCE)***

*Distance is defined as the total miles travelled to deliver a product.*

***#EQUIPMENT\_TYPE***

*Type of equipments e.g. "R" Refrigerated.*

***#WEIGHT***

*It is the actual weight of all items contained in an order number.*

***#ORDER\_COST***

*Cost of the order/shipment.*

***#FIRST\_PICK\_ZIP***

*Zip code (Location) from where the order was shipped.*

***#FIRST\_PICK\_EARLY\_APPT***

*Date and time for early pick up.*

***#FIRST\_PICK\_LATE\_APPT***

*Date and time for Late pick up.*

***#LAST\_DELIVERY\_ZIP***

*Zip code (Location) from where the order has to deliver.*

***#LAST\_DELIVERY\_EARLY\_APPT***

*Date and time for early delivery.*

***#LAST\_DELIVERY\_LATE\_APPT***

*Date and time for late delivery.*

***#IS\_HAZARDOUS***

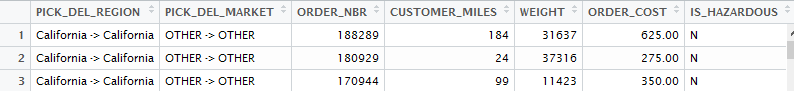
*Is the shipment/order is hazardous e.g. "Y" (HAZARDOUS)*

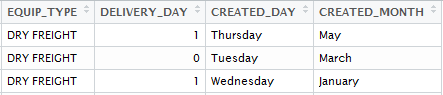
***#CREATED\_DATE***

*It is the actual day, month and year when an order has been created/booked.*

**Results**

*Various Pre- Processing techniques were applied to the raw data provided by customer.* ***Snap shot of the data after pre processing.***

**

**

*The results analysis criterion was based on the statistical theory where the P-value obtained from the calculations is compared to a standard deviation, deciding whether or not the values obtained are significant or not to the target which in this case has been set as the "Allocated Freight Cost". The standard deviation used as default for this analysis was 0.05. Any value greater than the standard deviation is not causing a significant impact on freight cost; but, any value below this standard is considered as a focus of distortion and a main cause for freight cost value fluctuation. The analysis was performed using the exact same variables for each origin in order to find a pattern and to be able to identify possible saving opportunities and identify main distortions through the analysis. Such variables or cost drivers that were analyzed in order to obtain the following results are:*

1. ***Regression Equation.***

*This particular equation shows the formula based on the coefficients obtained in the regression analysis. The formula indicates the variables that after the analysis were considered to be causing an impact on the target, which is the freight cost. This formula will be tested by substituting the total values for each variable and multiply them for the respective coefficient, and then add them all together in order to obtain the present value for the freight cost.*

*Some alterations might be encountered as this analysis will show on the formula test. Such alterations are higher or lower depending on the R-square value obtained during the regression analysis. The R-square will indicate the accuracy of the regression equation to predict future values; it is represented in the tables as a percentage. This formula is the key for this particular project since it will work as a prediction tool for freight costs.*

***Lm (formula = ORDER\_COST ~ CUSTOMER\_MILES + WEIGHT + IS\_HAZARDOUS +DELIVERY\_DAY, data = Train)***

***Estimate Std. Error t value Pr(>|t|)***

***# (Intercept)*** *2.138e+02 5.841e+00 36.605 < 2e-16 \*\*\**

***# CUSTOMER\_MILES*** *1.417e+00 7.777e-03 182.168 < 2e-16 \*\*\**

***# WEIGHT*** *1.767e-03 1.659e-04 10.647 < 2e-16 \*\*\**

***# IS\_HAZARDOUSY*** *1.461e+02 2.435e+01 6.001 2.02e-09 \*\*\**

***# DELIVERY\_DAY*** *-3.455e+01 3.209e+00 -10.765 < 2e-16 \*\*\**

1. ***P-value***

*This indicator was previously defined and indicates whether or not a variable has a significant impact on the target, which in this case is the freight cost. If the P-value of a particular variable is greater or equal to 0.05, it is considered as not having a significant impact on freight cost; however, if it is lower than 0.05 is considered as having a significant impact on freight cost variances. Even though it might be obvious that a particular variable or cost driver is causing cost variances, linear regression analysis will provide a better visibility in order to be able to measure it.*

1. ***R-Sq (R-square value)***

*This particular value indicates the accuracy of the regression equation. It can be found either in percentages or on a scale from 0.0 to 1.0. This indicator will tell the user how accurate the regression equation will be regarding prediction for future values.*

***#Multiple R-squared: 0.8896, Adjusted R-squared: 0.8888***

1. ***F – Test:***

***#F-statistic: 1051 on 89 and 11607 DF, p-value: < 2.2e-16***

*The p value is less than 0.05. That means it’s significant.*

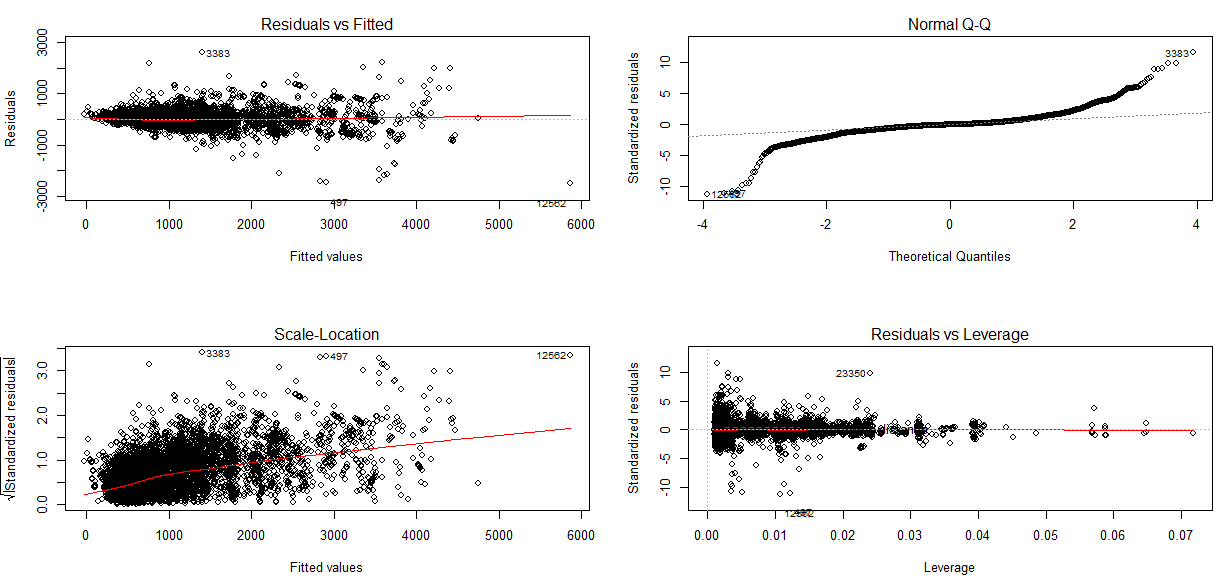
***The below plots shows that the linear regression follows all the fours assumptions:***

*1) Linearity*

*2) Independence of errors*

*3) Homoscedasticity (constant Variance)*

*4) Normality of errors*

**

1. ***Different errors were calculated between original and predicted values for comparison.***

***Error Comparison Chart:***

***# mae mse rmse mape***

***#train\_Lm*** *131.6923 50200.59 224.0549 0.1471939*

***#test\_Lm*** *137.0007 55537.91 235.6648 0.1491166*

***#train\_Lasso*** *131.8873 50234.55 224.1307 0.1474764*

***#test\_Lasso*** *137.1158 55534.85 235.6583 0.1492800*

***#train\_Ridge*** *141.8346 56537.08 237.7753 0.1616798*

***#test\_Ridge*** *147.2684 62199.73 249.3987 0.1640374*

***#train\_Elastic*** *131.8814 50232.89 224.1270 0.1474699*

***#test\_Elastic*** *137.1099 55534.99 235.6586 0.1492724*

***Comparison of the original shipment cost vs predicted shipment cost after applying the linear regression model.***

***Snap Shot:***

***V1 -> Original shipment cost***

***V2 -> Predicted shipment cost***

**

1. ***Also applied the “K-means clustering”*** *(k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster).*

*And then applied the Linear Regression on each clusters. Below is the comparison chart.*

***1st:***

***# With variable "WEIGHT","CUSTOMER\_MILES","DELIVERY\_DAY"***

***# with 4 Clusters as suggested by Elbow graph***

***#*** ***mae mse rmse mape***

***#train\_lm\_clust1*** *216.39198 88314.92 297.1783 0.1558501*

***#test\_lm\_clust1*** *221.18765 88765.61 297.9356 0.1507899*

***#train\_lm\_clust2*** *94.45837 20541.89 143.3244 0.1352266*

***#test\_lm\_clust2*** *104.03637 26431.69 162.5782 0.1448428*

***#train\_lm\_clust3*** *89.35460 23647.86 153.7786 0.1220900*

***#test\_lm\_clust3*** *88.66386 20385.13 142.7765 0.1231248*

***#train\_lm\_clust4*** *401.45561 310004.74 556.7807 0.1818439*

***#test\_lm\_clust4*** *401.14189 287912.33 536.5746 0.1864107*

***2nd:***

***# With only two variables "CUSTOMER\_MILES","DELIVERY\_DAY"***

***# with 4 Clusters as suggested by Elbow graph***

***# mae mse rmse mape***

***#train\_lm\_clust1*** *113.28248 29461.239 171.64277 0.1259150*

***#test\_lm\_clust1*** *125.74788 38631.845 196.54985 0.1372295*

***#train\_lm\_clust2*** *56.60306 9322.906 96.55520 0.1140783*

***#test\_lm\_clust2*** *57.64711 9662.817 98.29963 0.1163318*

***#train\_lm\_clust3*** *180.09371 64457.928 253.88566 0.1631142*

***#test\_lm\_clust3*** *200.28459 86147.093 293.50825 0.1644825*

***#train\_lm\_clust4*** *402.35789 296921.145 544.90471 0.1887735*

***#test\_lm\_clust4*** *384.07134 261404.677 511.27750 0.1437333*

***3rd:***

***# With only two variable "CUSTOMER\_MILES","DELIVERY\_DAY"***

***# with 3 Clusters only***

***# mae mse rmse mape***

***#train\_lm\_clust1*** *185.26193 69103.31 262.8751 0.1600499*

***#test\_lm\_clust1*** *196.69800 80805.25 284.2626 0.1794879*

***#train\_lm\_clust2*** *87.02783 19234.24 138.6876 0.1300547*

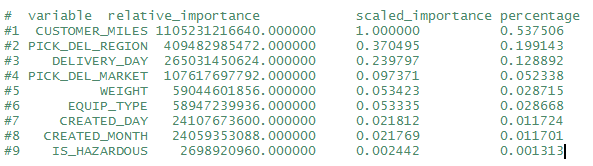
***#test\_lm\_clust2*** *90.19242 20580.20 143.4580 0.1329716*

***#train\_lm\_clust3*** *380.04458 262528.12 512.3750 0.1647014*

***#test\_lm\_clust3*** *454.15093 377268.79 614.2221 0.2187172*

1. ***Random Forest:*** *(Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees.)*

*Applied the Random Forest to know the important Variables.*

**

***Analysis***

*The purpose of this project was to create an effective tool or method to identify transportation cost variances for company “****Quick Freight****” which has been completely proven. The use of linear regression analysis throughout this entire project shows that this particular task can be easily performed. It is also considered a powerful tool either to predict future values for freight costs or to identify cost savings opportunities and trigger actions plans. Company “****Quick Freight****” (in this case* ***INSOFE****) was completely satisfied with the results obtained in this project and with the benefits that it will bring to the company processes.*

*This analysis has proven that statistical methods and calculations can be used to identify improvement opportunities within the distribution and logistics chain. Today's businesses are forced to cut operational costs in order to keep up with the competition. The above model/tool will provide the opportunity to “****Quick Freight****” company to stay competitive in market.*

***Limitations***

*When performing this project and its respective analysis, the following limitations were kept in the mind.*

* *This study was limited to company “****Quick Freight****” which has provided the information*
* *No comparisons were made to the other companies due to a lack of information and research.*
* *Time constraint to complete the project by 9th Dec 2016.*

***Appendices***

***Project Code in “R”:***

******

***HTML File Created by “R Mark Down”***

******