Springboard Introduction to Data Science Capstone Project

Maritime Data Intelligence System - Predicting the revenue of Ship Chartering Business

Now as the data is cleaned up, wrangled into shape and explored. Now it's time to perform some in-depth data analysis using machine learning.

Shipping industry is very competitive, Recently these conscious shippers are getting more aware of all the costs by analyzing their history data to make more precise decisions. Prediction of the revenue on the particular charter by creating a machine learning model based on the previous year's available data will help the shipper to choose best trades to get the maximum profit.

The goal of this project is build a model for earning prediction to help shipowners to choose the charter wisely to maximize their profits by identifying the cost affecting factors on the voyage.

The problem is supervised regression problem.

After doing preliminary exploration, earnings of these voyages are dependent on mainly below mentioned factors: - 1. Vessel type - there are different types of vessels; most profitable are MIDRANGE, HANDYMAX, AFRAMAX, PANAMAX 2. Cargo type- there are mainly two categories of cargo - dry cargo (coal, iron etc.) and liquid cargo (crude oil, gasoline, vegetable oil etc.). Most profitable are FO, CPP, ULSD, Gasoline, Naphtha and Gas Oil etc. 3. Dwt - this is the capacity of the vessel. 4. Actual Cargo lift - How much cargo the particular vessel is carrying affects the earnings. 5. Month - The period of the year in which this voyage is happening also affects the total earnings. 6. Region - the region in which the voyage is starting, loading and discharging also affect its earnings. Most profitable region is Transatlantic area. 7. Daily expense - It includes their daily fuel usage, port fees, jetty charges, food for crew etc. Their 60% of expense is on fuel.

We think these are the main predictors which will affect the earnings of the voyage.

We decided to apply Linear regression to the cleaned data to build the model and see the results.

R-Squared tells us is the proportion of variation in the dependent (response) variable that has been explained by the model.

STATISTIC CRITERION to choose best model are - R-Squared - Higher the better (> 0.70)
Adj R-Squared - Higher the better
F-Statistic - Higher the better
Std. Error - Closer to zero the better
t-statistic - Should be greater 1.96 for p-value to be less than 0.05
AIC - Lower the better
BIC - Lower the better
Mallows cp - Should be close to the number of predictors in model
MAPE (Mean absolute percentage error) - Lower the better
MSE - (Mean squared error) Lower the better
Min_Max Accuracy => mean(min(actual, predicted)/max(actual, predicted)) Higher the better

Evaluation metrics explain the performance of a model. We will use Confusion matrix. Confusion matrix is an N X N matrix, where N is the number of classes being predicted. For the problem which has N=2 hence we get a 2 X 2 matrix.

Here are a few definitions for a confusion matrix -
Accuracy : the proportion of the total number of predictions that were correct.
Positive Predictive Value or Precision : the proportion of positive cases that were correctly identified.
Negative Predictive Value : the proportion of negative cases that were correctly identified.
Sensitivity or Recall : the proportion of actual positive cases which are correctly identified.
Specificity : the proportion of actual negative cases which are correctly identified.