**Second Capstone Milestone Report:**

* An introduction to the problem: What is the problem? Who is the Client?

My company is a global shipping, logistics / supply chain and currently solving an optimization problem to minimize the cost of moving an empty container within North America and to Asia. I will use the same model to predict the demand/supply situation for this week up to the next two weeks.

The problem is to forecast the return date when a customer will return the empty container. This prediction/model will help to have a consistent flow of equipment inventory up to two weeks in advance. This will keep the customer satisfied by providing them enough equipment when they need it.

Client will be my current company which is intensively looking ways to save money by minimizing moving empty containers within and going out of North America. This prediction/model will be an important part of an end to end prediction from the time the shipment arrives via ship at the port terminal to be loaded into rail ramp until empty container is returned empty.

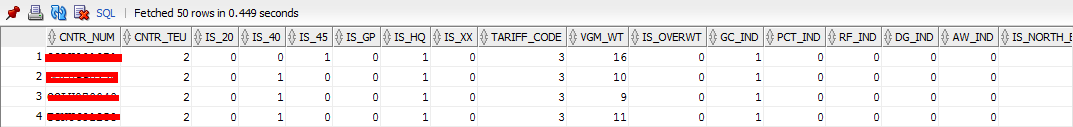
* A deeper dive into the data set:
  + What important fields and information does the data set have?
  + What are its limitations i.e. what are some questions that you cannot answer with this data set?
  + What kind of cleaning and wrangling did you need to do?
  + Are there other datasets you can find, use and combine with, to answer the questions that matter?

Features

Important features on the data includes shipment related information such as container size, container type, cargo weight, cargo description, number of containers in the shipment and arrival datetime in port terminal. Some limitations on this data set includes traffic conditions on the road, weather and delays due to administrative reasons such as customs.

Data Wrangling and other datasets

Cleaning the data includes missing data on cargo weight, travel time (target variable) and customer data. Any missing data on those variables will be filtered out in the dataset. Other datasets that can be combined with shipment data are related to customer information such as allowable free time to return the empty container and geographical location of customer site. These information will help to gauge the distance of the customer facilities to the intended return facility and the allowed days that customer can keep the empty container.



Continued…

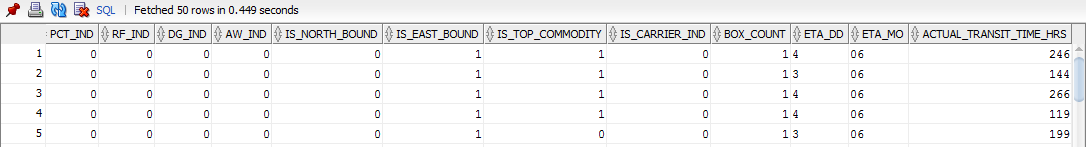


Figure: Sample data from the dataset

* Any preliminary exploration you’ve performed and your initial findings. Test the hypotheses one at a time. Often, the data story emerges as a result of a sequence of testing hypothesis e.g. You first tested if X was true, and because it wasn't, you tried Y, which turned out to be true.

Preliminary exploration

Initially, I created a model based on XGB linear and Random Forest models based on the features such as container size, type, cargo weight, arrival datetime at port terminal and trade. However, the mean absolute percentage error (MAPE) is high at 60% although much better than the current company MAPE= 90%. However, after discovering a pattern about different return days from different customer per return facility, I tested and updated the model to include customer pattern on commodity/cargo description, number of containers on the shipment (customers can request more than 1 container for bigger quantity of goods) and if the cargo is overweight. An overweight cargo can slow down travel time via truck.

The current MAPE is at 25% which is an improvement from initial MAPE=60%.

* Based on these findings, what approach are you going to take? How has your approach changed from what you initially proposed, if applicable?

Approach

I will create a dataset for each top/major customers per return facility then use XGB linear model to predict the number of days a customer will return the empty container. The training/test dataset will be 80/20 percent, learning rate will be tested from 0.01 to .5 with L1/L2 regularization and boosting iteration from 5, 10 and 50. Codes are written in R and mainly using caret library which is easy to use.

Figure: Cargo weight and arrival day are most importance features in this model

Importance Variable

1 100.000000 Cargo\_Weight

2 55.250652 Arrival\_Day

3 34.505150 40FT

4 16.771749 BOX\_COUNT

5 15.916074 COMMODITY

6 14.013976 Arrival\_Month

7 5.386412 General Cargo

8 0.000000 45FT

9 0.000000 High Cube

10 0.000000 Overweight Cargo