**CAPSTONE DATA SCIENCE PROJECT PROPOSAL**

**MOTOR CARRIER PERFORMANCE ANALISYS**

In being almost 10 years in the Trucking Industry I was fortunate enough to gather experience in almost every department of a Motor Carrier Company. With mainly having the role as a Director of Safety very often I was put in a position to overlook other departments activities as well. And finally, this experience can be utilized in performing analysis on a Motor Carrier performance.

Each Trucking company is mainly consisted of three departments: Accounting/Billing, Safety and Dispatch. Although there are already a lot of softwares that can perform analytics and the management can get an idea of what they should be expecting and should be corrected in order to maximize profit I am confident that I can apply what I learned from my data science classes and get good analysis on the datasets from company where I currently work at.

The Motor Carrier revenue comes from hauling loads for customers. The customers can be either direct or the load can be booked thru brokers. The dispatchers are booking the load, then are dispatching it to the driver they decided Is the best fir to pick up that load. And usually, it is dispatched to which driver with the proper trailer (for this Company it is either flatbed or reefer) would be nearby. After delivering the load the driver receives a bill of lading. That bill of lading together with the rate confirmation is sent to the factoring company and is getting billed by the factoring company. If all paperwork is correct than the carrier gets paid by the factoring company usually within a day since the factoring company received the paperwork. The Factoring Company is keeping records on which load when it was billed and when it was paid, for each I plan to do analysis and try to get useful results.

Each booked load is entered in the Company system with information which broker/customer was booked thru, where it was picked up, where it was delivered, the dates of pick up and delivery, which truck, trailer and driver hauled the load, when it was billed, how much the load was and many more information. This dataset will be part of my analysis as well.

There is quarterly International Fuel Tax Agreement that is filled on quarterly basis. In that filling the Company submits how many miles were driven in each state and how many gallons were fueled. In order for the Company to get to the final report, it is first gathered for each truck gallons and miles per state and the grand total is filled to the state. This dataset will be also used in my analysis to try go get useful information.

The Safety Department of the Company is keeping a track of how many accidents and incidents occurred. It keeps a record of was the driver at fault, was it incident, accident or cargo claim. Incident means a property damage was done on private property (it can either on a truck stop hitting or being hit by another vehicle, at a customer a damage to an object or another vehicle, or our equipment got damaged). Was it filled as a claim to the Company’s appropriate insurance and how much was paid. Also there is information in which state it occurred, what kind of trailer was hauled, which driver, the date and etc.

I have 3 datasets that I can use. One of the datasets is downloaded from the company’s database and consists information of 7000+ loads hauled by the Carrier in the past 2 years. From that dataset I can get predictive analytics and apply classification or clustering to get which broker the Company has the most loads with, then get information like in each quarter which were top brokers by number, then get an idea which were top brokers by revenue and see if there is a pattern that would give us an idea of what we should be expecting in the quarters to come. In that same dataset there is information if the load was hauled by a flatbed trailer and reefer trailer. A comparison can be done quarterly again of how was the performance per type of trailer and what should we expect in the quarters to come. In that dataset there is also how many miles were driven per each load, so what I want to do is classify the loads in 3 classes under 300 miles, between 300-500 miles and more than 500 miles and then check which were the top brokers used in each class and get an idea of which broker will be mostly used in the next year for each class.

The second dataset I have is from the factoring company thru which the loads are billed. The dataset has information of when the load was billed to the customer and how long it took the customer to pay the load. The goal from this dataset is to get information who pays the loads in the most short time and which broker is taking more time to pay the invoice. If a load is billed and not paid in 90 days then the factoring company is taking the money they paid for the load from other loads. Based on which brokers/customers are top customers I want to check if the company decides to remove the factoring company and starts billing the invoices directly what the management should be expecting in the terms of receiving payment from the customers.

The third dataset is the international fuel tax agreement report that each carrier files quarterly reporting how many miles were driven and how much fuel was bought in that quarter per state. The main analysis that I will be getting from this dataset is which are the top 10 states that have most percentage of the miles driven and how was the fuel ratio. If there are less gallons bought in the states the company pays tax, so with my analysis I want to check and get the idea of how to minimize paying fuel taxes each quarter by fueling enough in each state.

I plan mostly to use Orange for applying data mining techniques on the datasets. In some cases I can also try to use Gephi or code. I can use the Association Rule Mining to get which brokers are mostly used in particular states. That way we can get an idea if the truck is in a state that has high percentage a particular broker to have a load that will be acceptable for the Company.

The Association Rule Mining can also be applied to get an idea from the top 10 states where the loads were picked up to get an idea which class the load would be, under 300 miles, 300-500 miles or over 500 miles. That will give a better idea to the dispatcher what kind of mileage the driver is most likely to get. And the mileage is how the driver gets paid by, so this analysis would be helpful for the dispatcher to plan ahead of how many miles the driver will get, so they paycheck would be higher.

With applying the Association Rule on the dataset I got from the factoring I will get information of which broker usually how many days is needed to pay the load after it being billed. This can be used again to get an idea when brokers are expected to pay the Company and also if the Company removes the Factoring Company how long it will take to receive the money. I do want to see if I can apply clustering and see how the data will be classified and if I can get some meaningful information.

Clustering can be also applied on the International Fuel Tax Agreement dataset to check and see how the algorithm will group the states depending of which state has similar mileage with other states.

I am planning on using Bayesian optimization in Python as this will give me an idea about the brokers with most revenue or the broker with most loads.

There will be also included descriptive statistic on the dataset. These descriptive analytics will help to get meaningful conclusions about the Company operation as well. One of the most helpful descriptive statistics would the frequency distribution and ratio for all of the datasets.