Predicting Carrier Load Cancellation

By: Ali Al-Habib and Nicolas Favier
Advisor: Dr. Christopher Mejia
Topic Area(s): Transportation, Predictive Analytics

Summary: This research explores the key drivers of carriers' cancellations of truckloads using historical cancellation patterns. Truckload cancellations by carriers cause disruptions in the trucking operations. If these cancellations can be predicted, shippers and transportation brokers can avoid loss of money and resources. The research evaluates the applicability of different predictive including logistic regression, random forest, neural networks and k-nearest neighbors. The resulting models were capable of correctly predicting only 16% of the cancelled loads. Accordingly, it is recommended that business solutions be implemented in order to reduce the probability of cancellations.

Before coming to MIT, Ali graduated with a B.S. in Computer Science from the King Fahd University of Petroleum and Minerals. He then worked for Saudi Aramco for 12 years taking several roles mainly in the supply chain analytics and applications. Upon graduation, Ali will return to Saudi Aramco in Dhahran, Saudi Arabia.

Before coming to MIT, Nicolas graduated with a B.S. in Industrial Engineering from University of Cuyo and in Mechanical Engineering from ENISE. He then worked for Group Peñaflor for 2 years as a supply analyst. Upon graduation, Nicolas will return to Argentina continuing his professional career.

KEY INSIGHTS

1. Load cancellation by carriers represents a significant ratio of booked loads. Based on three-year data, 17% of booked loads were cancelled.
2. Average cost difference between cancelled load and recovery load is $145.
3. Load, shipper, and carrier characteristics are not highly correlated with cancellations; hence, models based on this data did not provide sufficient prediction accuracy.
4. More details pertaining to actual cancellation reasons need to be captured in order to build more accurate predictive models.

Introduction

Based on an initial analysis of three-year dataset from a third-party logistics (3PL) provider, approximately 17% of confirmed loads get cancelled (also known as bounced) by carriers. On average, each cancellation is estimated to result in $145 of extra cost to rebook the load with another carrier. In most cases, the cost of the rebooked load is higher than the original cost from the original carrier.

Our project focuses on building a predictive model using historical data to identify the main drivers for cancellations. The project started by analyzing three-year data and preparing it for building the model. Then, several models were evaluated using available and enriched datasets to assess the predictability of load cancellations. Finally, sensitivity analysis was developed to measure the tradeoff between prediction power and model error.

Data Analysis and Preparation

Before building a predictive model, a full understanding of the operation and available data was required. A site visit to the 3PL company resulted in a list of possible characteristics that might impact load cancellations (**Figure 1**). Three-year data of truckloads were used to build a predictive model that provides the company with cancellation probability for future loads.

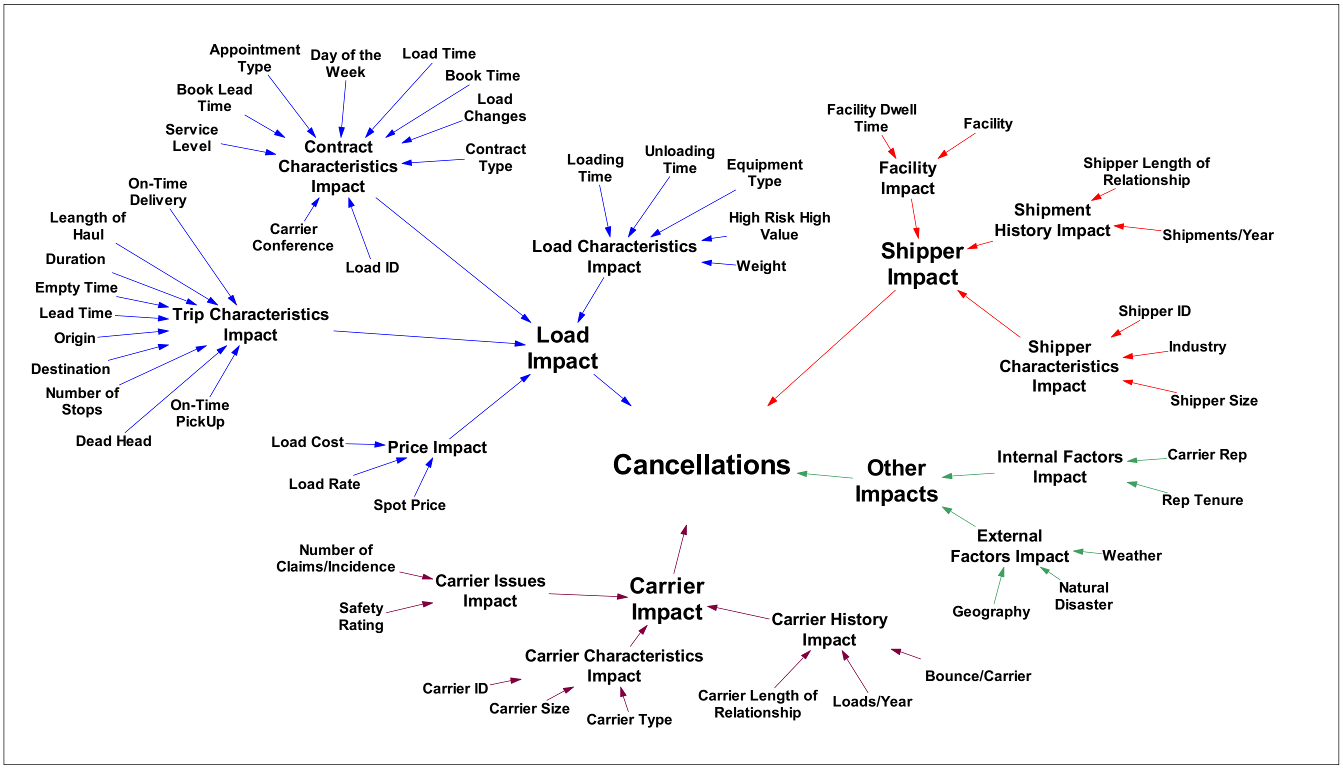

Figure 1: *Diagram of potential variables impacting the cancellation probability*

Provided data was analyzed and prepared to fit the expected outcomes. As the data was at stop-level (i.e. each load is represented by more than one record), the first step was to transform to load-level (one record per load). Details available at stop-level were aggregated and presented in the load-level data to capture information that might impact load cancellations. Outliers in the data were identified and cleaned to avoid any undesired impact on the model.

Initial data analysis showed that cancellation patterns differ with different characteristics of loads, carriers, and shippers. As an example, Figure 2 shows that cancellation rates for loads that pass through some cities are higher than other cities.

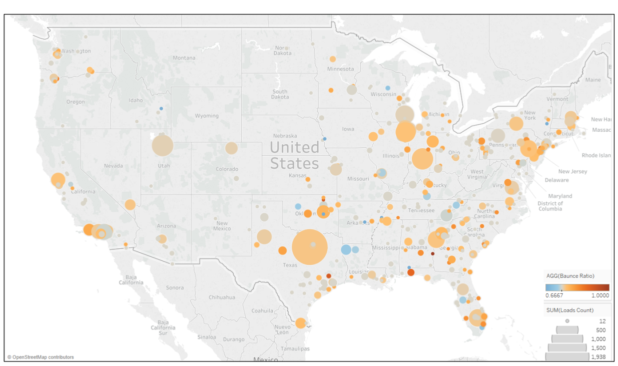

Figure 2: *Load volume and cancellation ratios by city*

Moreover, the analysis showed that the probability of cancellation is lower for loads that are booked within 24 hours of the pickup time. A few other characteristics also showed some variations in the cancellation rate such as shipper's industry and carrier's length of relationship with the company.

The final step in data preparation was to develop correlation and multi-collinearity analysis. These analyses were done to eliminate correlated features that might lead to model overfitting.

Model Building and Results

Multiple predictive models are commonly used to solve similar problems. After reviewing the pros and cons for different models, logistics regression was selected as the main model due to its self-descriptive nature. Machine learning models (neural networks, random forest, k-nearest neighbor) were used to validate model performance.

The initial dataset contained all basic information pertaining to the load, carrier, and shipper. After identifying the most significant features through predictor screening, first model was built using these basic features. The model results reflected that these features are not enough to predict carriers load cancellation (**Table 1**).

		Predictions		
		No	Yes	
Actual	No	652,501	2,956	**655,457**
	Yes	129,727	1,971	**131,698**
		782,228	**4,927**	**787,155**
Error				16.86%
Missed Bounces				98.50%

Table 1: *Confusion matrix for initial dataset model*

To improve prediction capability, the dataset was enriched with extra features that were extrapolated from the data or gathered from external sources. Weather alerts data were obtained from National Centers for Environmental Information (NCEI) and added to the dataset to test the impact of weather severity on cancellation decisions. Moreover, cancellation ratios were calculated for several features (like carriers, shippers, cities, zip codes, states) and added to the dataset to reflect cancellation patterns. Ratios for combined features were also calculated and added to the data to capture specific cancellation behaviors (like carrier-city and carrier-equipment type combinations). Although the new features improved the prediction power, the achieved accuracy was not good enough. The model correctly predicted only less than 17% of the cancellations when tested on a new dataset.

		Predictions		
		No	Yes	
Actual	No	59,883	3,735	**63,618**
	Yes	8,903	1,722	**10,625**
		68,786	**5,457**	**74,243**
Error				17.02%
Missed Bounces				83.79%

Table 2: *Confusion matrix for enriched dataset model*

Results obtained from the logistic regression models indicated that load cancellations are not predictable using the information currently captured by the company. These results were validated using machine learning models, which gave similar results.

Unpredictability Testing and Sensitivity Analysis

Further tests were conducted to confirm the conclusion reached from the models' results. Two main hypotheses were made and tested in order to confirm this conclusion. The first hypothesis was that prediction accuracy will improve if the model is applied only on loads with enough historical data. The second hypothesis was that prediction accuracy will improve if the model is applied only on a very short time horizon. Both hypotheses were tested by running the model on smaller datasets with enough historical records and with short time horizon. The results confirmed that prediction power could not be improved any further. Accordingly, the unpredictability conclusion was confirmed.

The final test was a sensitivity analysis over the threshold used in the logistic regression model. The threshold is the cutoff value used to classify the result from the regression model as cancelled or not cancelled. For the base analysis and all tests, a threshold of 0.5 was used. This indicates that if the model output is a value between 0.5 and 1 the load is predicted as cancelled; while if the output value is between 0 and 0.5 the load is predicted as not cancelled.

Lowering the threshold enables predicting more cancelled loads accurately. However, the impact of this action on the wrongly classified cancellations and the overall model accuracy is huge. Figure 3 illustrates the tradeoff provided by selecting lower thresholds for the logistics regression model.

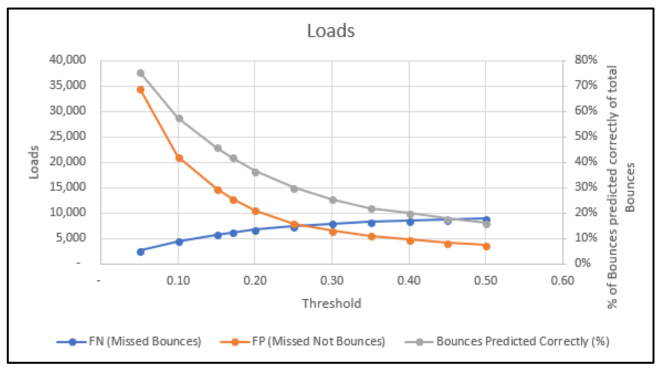

Figure 3: *Logistics regression threshold tradeoff*

For a threshold value of 0.17 (which represents the average cancellation ratio), false negatives decreased 31% compared to the base value (0.5); However, false positives increased 244%. In terms of number of loads, the threshold reduction allowed accurate prediction of additional 2,743 cancellations. However, it also increased the number of loads wrongly predicted as cancelled by 9,130 loads.

Conclusion

As load cancellations cannot be predicted accurately using the available data, other business measures need to be taken to reduce the cancellations' probability or impact. Business practices such as imposing penalties on carriers who cancel loads within 24 hours might discourage cancellations. Educating carriers on the impact of cancellation and encouraging them to cancel with a longer timeframe, when cancellation is inevitable, might also reduce their impact.

Prediction accuracy of the cancelled loads can be improved using lower thresholds for the logistics regression model. However, the impact of lower threshold on the misclassification of the uncancelled loads and the overall model accuracy must be considered. As highlighted in the previous section, lowering the threshold to the average cancellation ratio will classify three uncancelled loads as cancelled for every correctly predicted cancellation.

This research exploited the available dataset and concluded that the available features are not enough to correctly predict cancellations. However, the research identified some features that provide insights into cancellations. Future research can focus on surveying carriers about the actual cancellation causes and capturing data that can be incorporated into future models.

Finally, it is known that the transportation industry is very complex with many stakeholders that are not necessarily interconnected. This complexity might confound the possibility of building a good predictive model using the company data solely. As carriers work with many shippers and brokers at the same time, cancellations might be a consequence of delays or cancellations in other loads that are not managed by the same company. This fact implies that companies will always have a limited view of all the factors that might impact the load cancellations and consequently hinder the ability to build a sound predictive model.