Predicting On-time Delivery in the Trucking Industry

By: Rafael Duarte Alcoba and Kenneth W. Ohlund
Thesis Advisor: Matthias Winkenbach
Topic Area: Transportation, Predictive Analytics

Summary: On-time delivery is a key metric in the trucking segment of the transportation industry. If on-time delivery can be predicted, more effective resource allocation can be achieved. This research focuses on building a predictive analytics model, specifically logistic regression, given a historical dataset. The model, developed using explanatory variables with statistical significance, results in a significant resource reduction while incurring a relatively small impactful error. Interpretability and application of the logistic regression model can deliver value in predictive power across many industries. Resulting cost reductions lead to strategic competitive positioning among firms employing predictive analytics techniques.

Before coming to MIT, Rafael graduated with a B.S in Industrial Engineering from the Federal University of Rio Grande do Sul, Brazil. He then worked for Anheuser-Busch Inbev, having several roles in the Supply Planning & Performance Management team. Upon graduation, Rafael will join Bayer in Whippany, NJ.

Before coming to MIT, Ken graduated with a B.S. in Marine Engineering from Massachusetts Maritime Academy and then worked as an engineer aboard LNG carriers engaged in worldwide trade. He also worked for Transocean in the offshore Oil and Gas industry. Upon graduation, Ken will join GE Aviation in Lynn, MA.

Introduction

If firms could accurately predict the future with certainty, profits could be maximized and shareholders would prosper. Despite the complex nature of predictions, many mathematical tools exist that enable firms to do just that. As technology and innovation drive forward, methods facilitating the use of mathematical tools for predictions improve.

In the transportation industry, third-party logistic firms (3PLs) have a stake in making accurate predictions. A key metric by which 3PLs are measured on is on-time delivery. If on-time delivery could be predicted with some degree of certainty, then efforts could be focused on those loads that require resources. Currently, trucking firms commonly allocate resources to tracking and supporting each load tendered. This inefficiency and its associated costs represent a great opportunity for firms to gain a competitive edge, if corrected.

This research focuses on predicting on-time delivery in the trucking industry. Coyote Logistics, a Chicago-based 3PL, sponsored this thesis. With a high degree of data availability, we complied an exhaustive list of variables potentially affecting on-time delivery. Through an extensive selection process, variables with high statistical significance were chosen. A predictive model selection process led to the choice of the logistic regression model.

Variable selection and data preparation

Selecting the right variables and predictive model requires a number of steps. A comprehensive understanding of the business operation enables the creation of an exhaustive list of potential variables. A brainstorming session with industry experts yielded the list of variables shown in Figure 1.

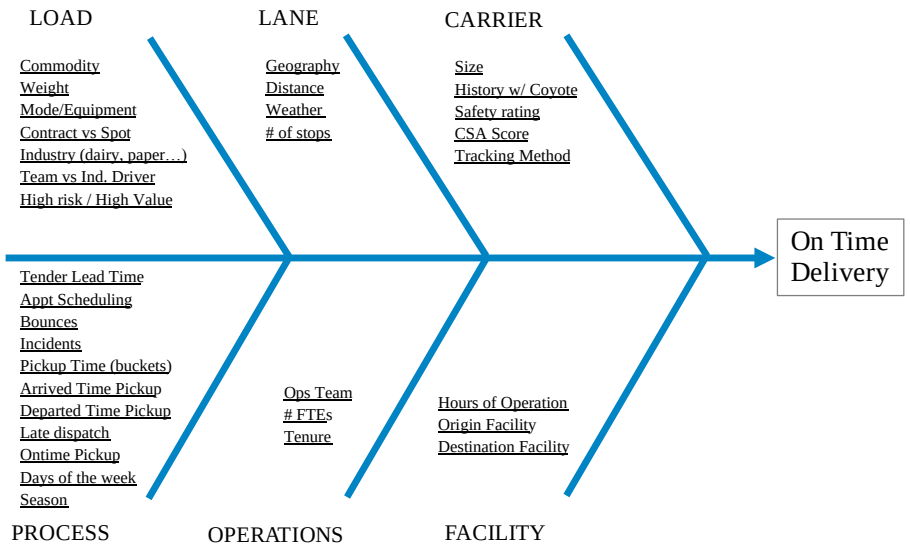

Figure 1. Fishbone Diagram for Variables Potentially Affecting On-time Delivery

In addition to the list of variables, we also chose to use a binary decision variable for on-time performance. On-time for 3PLs is usually defined as within one hour of the appointment time. The time horizon used for the analysis in this thesis is two years. We determine this time horizon to be robust due to recency of the data and the inclusion of two calendar cycles. Data handling and preparation are critical to developing an accurate model. Through open lines of communication with our sponsor company, we classified data outliers.

Our dataset had more than fifty input variables, nominal and continuous, and one binary output variable. We looked to identify, from the input variables, which combination would allow Coyote to predict on-time delivery. In the interest of reducing the number of variables that might yield better performance in the validation set, we used the stepwise regression approach. The stepwise approach enabled us to explore different combination of variables with a quick and flexible interface. Figure 2 presents the chosen six variables with high statistical significance for our model.

Effect Summary

Source	LogWorth	PValue
OnTime-1Hrs of StartSegment	368.011	0.00E+00
Incidents per Volume	96.607	2.47E-97
Contract-Spot	75.867	1.36E-76
Duration at StartSegment	75.860	1.38E-76
FacilityType Appt of EndSegment	59.558	2.77E-60
HistoricalVolume	58.502	3.15E-59

Figure 2. Stepwise Regression for Variable Selection

Multi-collinearity can play a role in distorting the results of a model by duplicating the statistical value through corresponding variables. We explored approaches aimed at identifying multi-collinearity and mitigating it. Of those approaches, principal component analysis, multiple correspondence analysis, and a correlation matrix were performed.

Through a systematic literature review, we identified three widely used models to predict categorical response using continuous and categorical predictors. The three models are logistic regression, neural nets and bootstrap forest. Due to Coyote's desire for an explanatory model with high interpretability of model results, this thesis focuses on the logistic regression model. The neural nets and bootstrap forest models validate the selected model.

The dataset used presented a very imbalanced proportion of on-time and delayed loads. The small representation of delayed observations reflects the high service level that Coyote provides its customers. To develop a model capable of capturing the useful information that distinguishes the underrepresented class from the dominant class, we used stratified sampling. Stratified sampling is a method of sampling data used when classes are presented in a very unequal proportion in the original dataset.

Besides certifying a correct representation of both classes (0 and 1), it is also extremely important to avoid overfitting. This ensures that the chosen model is able to generalize beyond the dataset at hand. To

mitigate this risk, we use the concept of data partitioning- dividing the stratified dataset into two groups: training and validation. The model is developed using the training set and evaluated using the validation set. The performance on the validation set provides insight into the model's predictive power.

Performance evaluation

Once we developed and ran different models on our dataset, we determined how to measure the predictive performance of each. Different methods to evaluate a model's performance can be used. Even though adjusted R squared is widely used, as the main goal of our model was to predict a binary outcome, we used a confusion matrix. The confusion matrix is a two-by-two table that classifies the actual response levels and the predicted response levels. The diagonal elements of the confusion matrix indicate correct predictions, while the off-diagonals represent the incorrect predictions.

Model results

Though misclassification rate is a widely used indicator of fit for the models, application of the model is a crucial consideration. Specifically, we considered the tradeoff between resource reductions and missed delayed loads. Missed delayed loads pose a serious problem for implementing the model. If the model predicts a load will be on-time and it is late without tracking, Coyote's service level suffers. These missed delays are represented in the top right quadrant in the matrix. They represent the error for predicting on-time when the load is actually delayed. Because of the severity of consequences, emphasis is placed on minimizing this error.

While minimizing this error is critical, it must be balanced with a reasonable resource reduction. We quantify this reduction by dividing the number of predicted delays (0's) by the total number of observations. In practice, this assumes that Coyote will track only those loads that the model predicts to be delayed. In the validation results presented in Figure 3, 23.6% of the loads are predicted to be delayed and should be tracked. We are not as concerned with predicting a delay that in fact is on-time. Since Coyote currently tracks all loads, tracking roughly 21% wastefully is acceptable in light of the significant resource reduction. Tracking a load that will be on-time also does not penalize Coyote's service level in any way.

Although the initial process of brainstorming variables through the fishbone diagram was exhaustive, we continued to search for additional data to improve the model. Despite some variables having a small enough p-value to be included in the model, we omitted them. The decision to exclude the new variables was substantiated by only a relatively minor improvement observed in the confusion matrix performance. Including variables with such small explanatory power under fabricated boundaries is forced. This leads to a less robust model. Since robustness is important in our model, we leave out variables with small explanatory power.

A new dataset was provided to test the robustness of the model. Figure 3 shows the comparison of the results from the Customer Test Data and the validation data.

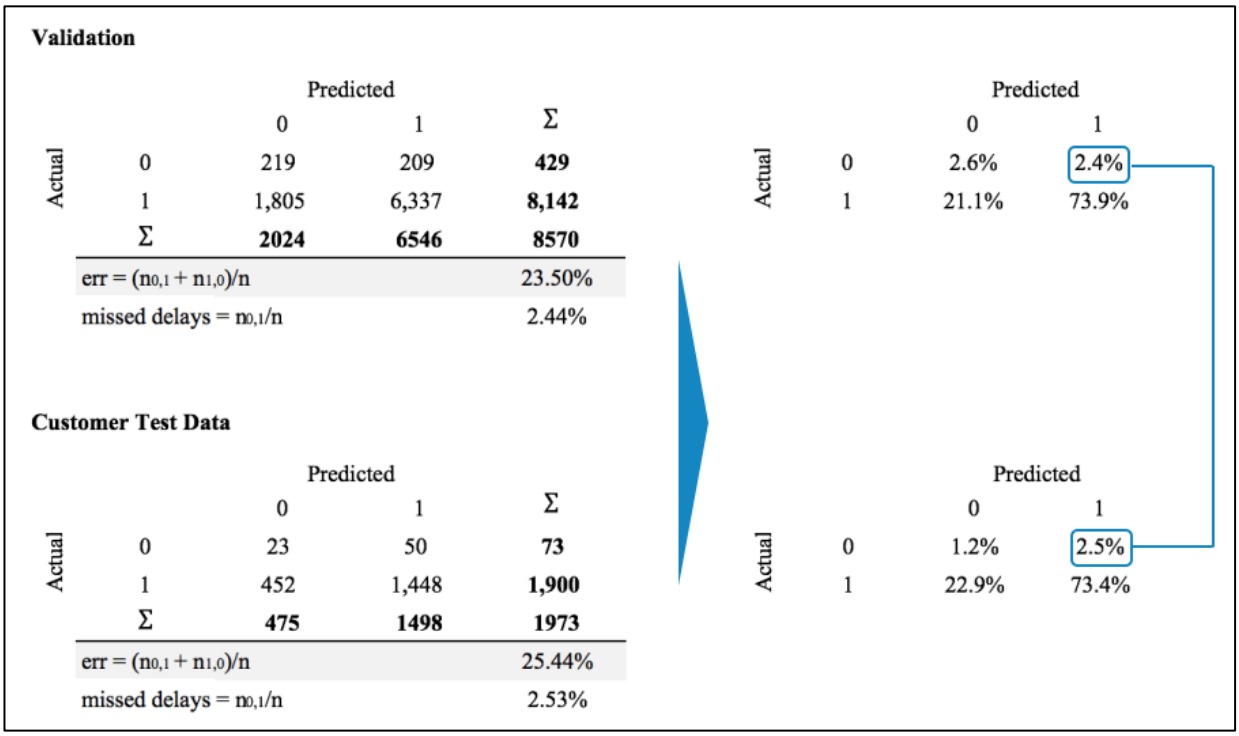

Figure 3. Confusion Matrix Results Comparison for Validation and Customer Test Data

Conclusion

In predictive modeling, where a binary response is required, misclassification is the overall metric of choice. Predictive models can be tailored to optimize other, more specific metrics. For a company that focuses on achieving a high service level, the minimization of missed delays is critical. As the rate of missed delays decreases, the sensitivity of the model increases. The reaction of the model to increased sensitivity is a higher misclassification rate and lower overall usefulness of the model. These tradeoffs are key drivers in the thesis.

Just as important as understanding tradeoffs and model performance metrics is the comprehension of the implications of adding new variables. Throughout the thesis, our desire to improve the model and deliver better results tempts us to include data that marginally improve performance. Although it is possible to improve overall misclassification by adding some extra variables, we avoid this. We find that adding variables without very high explanatory power adds complexity, reduces robustness, and can lead to overfitting.

Despite all the possible extensions of this research, our findings present valuable insights to Coyote Logistics. The thorough modeling process validates much of the intuition from the experts. Data-backed decisions enable firms to have greater success and gain competitive advantage. This research represents a step in the right direction for Coyote in investigating predictive analytics for their operations. The model, developed using six explanatory variables with statistical significance, results in a 76.4% resource reduction while incurring an impactful error of 2.4%.