Predicting On-time Delivery in the Trucking Industry

by

Rafael Duarte Alcoba
Bachelor of Science, Industrial Engineering, Federal University of Rio Grande do Sul, 2011

and

Kenneth W. Ohlund
Bachelor of Science, Marine Engineering, Massachusetts Maritime Academy, 2009

SUBMITTED TO THE PROGRAM IN SUPPLY CHAIN MANAGEMENT
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF ENGINEERING IN SUPPLY CHAIN MANAGEMENT

AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2017

© 2017 Rafael Duarte Alcoba and Kenneth W. Ohlund. All rights reserved.

The authors hereby grant to MIT permission to reproduce and to distribute publicly paper and
electronic copies of this thesis document in whole or in part in any medium now known or
hereafter created.

Signature redacted

Master of Engineering in Supply Chain Management Program
May 12, 2017

Signature redacted

Master of Engineering in Supply Chain Management Program
May 12, 2017

Signature redacted

Dr. Matthias Winkenbach
Director, MIT Megacity Logistics Lab
Thesis Supervisor

Signature redacted

Dr. Yossi Sheffi
Director, Center for Transportation and Logistics
Elisha Gray II Professor of Engineering Systems
Professor, Civil and Environmental Engineering
Predicting On-time Delivery in the Trucking Industry

by

Rafael Duarte Alcoba

and

Kenneth W. Ohlund

Submitted to the Program in Supply Chain Management
on May 12, 2017 in Partial Fulfillment of the
Requirements for the Degree of Master of Engineering in Supply Chain Management

ABSTRACT

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 six explanatory variables with statistical significance, results in a 76.4% resource reduction while incurring an impactful error of 2.4%. 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.

Thesis Supervisor: Dr. Matthias Winkenbach
Title: Director, MIT Megacity Logistics Lab
ACKNOWLEDGEMENTS

First, we would like to thank our thesis supervisor, Dr. Matthias Winkenbach, for his outstanding guidance and value added to this thesis. His ability to quickly develop insights and articulate the right questions to obtain valuable information is deeply appreciated in the research process.

Next, we would like to thank Pamela Siska, our writing instructor, for her dedication to providing comprehensive and constructive comments throughout the writing process.

We would also like to thank Coyote Logistics for their tremendous support in this thesis. Specifically, we would like to thank John Kochanek, Amit Prasad, Rohan Boricha, and Fateme Fotuhiardakani. Their expertise and prompt delivery of data enabled this thesis to have the depth we sought.

Lastly, we would like to thank the SCM Class of 2017 for their support throughout the year.

- Rafael & Ken

I would like to thank my wonderful family, particularly my parents, Syllas and Sandra, and my brother, Diego for their unconditional support in all stages of my life. I would like to especially thank my wife, Natalia for her love, and for being such a supportive life partner.

- Rafael Duarte Alcoba
# Table of Contents

List of Figures .................................................................................................................................. 6

1. Introduction ...................................................................................................................................... 7

2. Literature Review ............................................................................................................................. 9
   2.1 The relevance of predictive analytics in transportation.............................................................. 9
   2.2 Model selection in predictive analytics ......................................................................................... 10
   2.3 Evaluating Prediction Performance ............................................................................................ 14
   2.4 Section Summary ......................................................................................................................... 15

3. Methodology ..................................................................................................................................... 17
   3.1 Data Profiling ............................................................................................................................ 17
      3.1.1 Gathering Data ...................................................................................................................... 17
      3.1.2 Data Handling and Preparation ............................................................................................ 19
      3.1.3 Data Joins ............................................................................................................................ 20
   3.2 Data Analysis ............................................................................................................................. 21
      3.2.1 Multi-collinearity .................................................................................................................. 21
      3.2.2 Sampling and Partitioning Data ............................................................................................ 23
      3.2.3 Stepwise and Logistic Regression ......................................................................................... 24
      3.2.4 Model Validation - Neural Nets and Bootstrap Forest ......................................................... 27
      3.2.5 Performance Evaluation ....................................................................................................... 30

4. Results ............................................................................................................................................. 33
   4.1 Variable Selection & Logistic Regression ..................................................................................... 33
   4.2 Model Validation ........................................................................................................................... 37
      4.2.1 Neural Nets Model ................................................................................................................. 37
List of Figures

Figure 1 - Fishbone Diagram for Variables Potentially Affecting On-time Delivery ..................18
Figure 2 - NOAA’s Severe Weather Data Inventory Web Service ...........................................20
Figure 3 - PCA results reflecting no dimensionality reduction requirement.............................23
Figure 4 - Stratified Sampling for the Coyote Dataset ............................................................24
Figure 5 - Illustrative Neural Network example ......................................................................28
Figure 6 - Confusion Matrix: Meaning of each cell ..................................................................31
Figure 7 - Confusion Matrix, Reweighted ..............................................................................32
Figure 8 - Stepwise Regression for Variable Selection in the Coyote Dataset .........................34
Figure 9 - Confusion Matrix Results from the Logistic Regression Model ...............................35
Figure 10 - Adjusted Confusion Matrix Results from the Logistic Regression Model ............36
Figure 11 - Adjusted Confusion Matrix Results from the Neural Network Model ....................38
Figure 12 - Adjusted Confusion Matrix Results from the Bootstrap Forest Model ..................39
Figure 13 - Correlation Matrix for six explanatory variables ....................................................40
Figure 14 - US map used for geographical zones .....................................................................41
Figure 15 - Statistical Significance of six chosen variables and geographical zone variables ...42
Figure 16 - Time buckets for variable creation from timestamps ............................................43
Figure 17 - Statistical Significance of six chosen variables and time bucket variables ............44
Figure 18 - Confusion Matrix Results Comparison for Validation and Customer Test Data ......46
1. 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. Previous barriers to applying these tools have included data availability and processing power. As technology and innovation drive forward, methods facilitating the use of mathematical tools for predictions improve. Data collection and computing power are two key areas where improvements are changing the landscape of predictive applications. The Internet of Things (IoT) and Cloud storage have transformed data collection and storage capacity. Further, statistical software packages crunch large datasets enabling insights. These packages are supported by significant improvements in computing power and availability. Ultimately, advancement in these areas leads researchers to the cusp of making accurate predictions.

In the transportation industry, namely trucking, 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 deliveries, or loads, that require resources. Currently, firms in trucking 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, an exhaustive list of variables potentially affecting on-time delivery is compiled. Through an extensive variable selection process, six variables with high statistical significance are chosen. The six variables are on-time pickup, duration at the pickup facility, historical volume of the carrier,
incidents per volume of the carrier, contract versus spot tendering of the load and the facility type of the delivery facility (appointment versus open). A predictive model selection process is followed, ultimately leading to the selection of the logistic regression model.

Initial results from the logistic regression model yield significant negative predictive value. Negatives in the model represent on-time observations and predictions. Positives represent delays. Specifically, the model is robust in correctly predicting those loads that were on-time. This delivers value enabling the more effective allocation of resources desired from the predictions. This model can deliver value to firms across the trucking segment of the transportation industry. The methodology can be followed to create a predictive model for use across many industries. The following sections present the detailed process followed to arrive at the results.
2. Literature Review

This section presents an overview of the research related to using predictive analytics to drive decision-making. The primary goal is to identify an appropriate model to predict on-time performance in the trucking industry. First, it recognizes the relevance of predictive analytics in freight transportation. Next, it summarizes possible predictive models for binary response outcomes, which is the main goal of this thesis.

2.1 The relevance of predictive analytics in transportation

The use of summary statistics in analyzing performance metrics like on-time delivery is not new to the transportation industry. Predictive analytics, however, is a much newer and more powerful tool that can deliver quantifiable results to businesses. In his 2015 article “Predicting On-Time Deliveries: It’s not Rocket Science; It’s Data Science!”, Bill Kimler explains this idea. Kimler mentions the increasing availability of statistical software packages that enables users to make predictions through depth of understanding achieved by computing power. Those predictions enable business leaders to improve their performance in metrics like on-time delivery (Kimler, 2015). In trucking, a segment of the transportation industry, 3PLs fight for every opportunity to gain competitive advantage.

A recent study by Capgemini (2016) indicates that 3PL companies have been using predictive analytics to create a competitive advantage for their customers and to optimize the supply chain. In this study, the authors conducted a survey where 98% of all 3PLs identified that predictive analytics is extremely important to the future of their business. Companies mention that the use of predictive analytics will become a core competency of their supply chain, transcending traditional competitive advantage over competitors.
Trucking continues to be the most common form of domestic transportation within the US. An increase in outsourcing and supply chain complexity supports the growth trajectory of trucking. One reason for this growth is the point-to-point capability of trucks. Simply put, trucks can go places that freight trains and airplanes cannot. Mode forecasts by weight and value indicate that the basis for modality choices will not change. The Bureau of Transportation Statistics estimates that 2.36 billion tons of goods worth nearly $8 billion USD will be shipped annually via trucks by 2040. As the value-to-weight ratio for freight increases, the metrics by which carriers are measured will become even more important (USDOT, 2015). The desired outcome of this thesis is to provide a model that will enhance Coyote Logistics’ ability to perform better with fewer resources.

2.2 Model selection in predictive analytics

Predictive models are widely used in different industries and their implementation has been expanding consistently. The advances in computers’ ability to process large amounts of data and the overflow of available data have enhanced companies’ capabilities through data science. Predictive modeling uses actual observations from the past to predict the probability of an outcome, also called response, dependent variable or output variable. Shmueli, Bruce, Stephens, and Patel (2017) suggest two widely used models to predict categorical response using continuous and categorical predictors: logistic regression and neural nets. Random forests also prove to be an effective approach to predictive modeling (Shmueli et al., 2017).

Recent research sheds light on the desire to predict delays and methods for prediction. In his 2014 MIT thesis, Gold Truong seeks to provide Amazon with a model to forecast linehaul transit times and on-time delivery. He selects the quantile regression forest machine learning technique. In his model selection process, Truong explains his reasoning for choosing this method (Truong, 2014). Our objective of exploring this thesis is to provide strategic direction, justification,
and validation for the selection of the multiple logistic regression model for predicting on-time delivery in trucking.

At the core of his thesis, Truong (2014) shares the same desire to predict delays. He differs in his model selection and end goal. The accurate prediction of delays is a midpoint in his research. Ultimately, he shares a well-defined process to arrive at a valid model given some data. A simple first step in predicting delays is to identify what variables could cause them. In Amazon’s transportation problem, a preceding calculation was used that ignored variables such as departure time, destination type, and weather (Truong, 2014). Truong’s main goal is to design an optimal linehaul schedule, one derived from including appropriate variables. Predicting on-time performance for transportation is an intermediate step in his research. He breaks the variables into groups based on the controllability of the events. With this step, he lays the groundwork for a model focused on shipper process optimization. Optimizing factors within Amazon’s control makes sense in this case (Truong, 2014).

Amazon’s problem exhibits a data component that Truong (2014) addresses. The component is outliers. Some inputs with significant variability pose a risk to distorting the results of some types of models. As a result, Truong (2014) selects a random forests model for three main reasons. First, random forests can manage a large number of variables, some of which may have no explanatory value. Second, random forests models are not prone to overfitting. Third, random forests generally are highly accurate classifiers (Truong, 2014). One main factor in selecting a model type is the way in which the firm will use the model. A firm that is highly curious about the variables and their explanatory value may not desire a random forests model. A random forests model is a “black box” approach that makes results very difficult to link with intuition. Truong’s thesis provides excellent information on variable selection and parsing based on the sponsor
company’s end goal. These crucial preliminary steps offer insights and add value in the variable selection process for the predictive model for Coyote. The random forests model can also be used to validate the results of the multiple logistic regression model upon completion.

Popular alternatives to random forests models delivering the same categorical response include logistic regression and neural nets models. In exploring the predictive model selection process, Schumacher, Roßner, and Vach (1996) state that there is no consensus about when and under which circumstances logistic regression models should be used or preferred over neural nets. Discussions surrounding model selection are attributed to the fact that the method should be customized based on the characteristics of the data available and the model objective (Schumacher, Roßner, & Vach 1996). The method that presents a better predictive performance in one set of data may not lead to better results for another.

Following the reasoning of Schumacher et al. (1996), Dreiseitl and Ohno-Machado (2002) state that there is no single algorithm able to outperform the other in a given dataset and application area. Their study presents a methodology review confronting logistic regression with neural nets on biomedical data classification. The authors argue that besides the technical performance of the model and its predictive power, one crucial characteristic to take into account is the interpretability of the results. They suggest that interpretability is responsible for the wide acceptance of logistic regression in cases where model interpretation is more important than predictive performance. In those cases, neural nets’ “black box” approach does not lead to satisfactory results.

Besides its wide applicability to medical research, logistic regression is extensively used in other fields. Through a study of logistic regression in mechanical engineering, Phillips, Cripps, Lau, and Hodkiewicz (2015) reinforce the influence of the results’ interpretability in order to decide which model to use in industry applications. In the financial sector, logistic regression is
the leading predictive tool to detect fraud (Albashrawi, 2016). Another study that illustrates the broad application of logistic regression is introduced by Tian (2010) in the procurement area. Logistic regression is used to predict the likelihood of suppliers winning a request for quote, matching buyers and suppliers in the E-Procurement environment. Application of logistic model in the transportation industry can be found in a study conducted by Gebeyehu and Takano (2007). In this paper, the authors show the applicability of the model for evaluating citizens’ perceptions of the bus conditions and analyzing travelers’ choice behavior.

Due to Coyote’s desire for an explanatory model with high interpretability of model results, this thesis focuses on the logistic regression model. This type of predictive model can be used to deliver powerful, robust results with high visibility of the predictors. Logistic regression works with the same principles as linear regression. Logistic regression, however, is used when the output variable is categorical, while linear regression is used for continuous variables. Its main goal is to predict the probability of a new observation belonging to a specified class. In predicting a binary output variable (0 or 1), based on probability, it predicts the likelihood that a new observation would have to be classified as 0 or 1. Logistic regression models follow two steps:

1. Estimate probabilities of a new observation to belong to each class (0 or 1).
2. Classify each observation to one of the classes based on a specified cutoff value.

In this model, the predictors are related to the output variable through a nonlinear function called the logit. One particular advantage of this method is that it can also be extended to more than two classes - ordinal or multinomial (Shmueli et al., 2017).

The multiple logistic regression model is widely used in predictive analytics applications for many reasons. Its methodology of transforming the probabilities of variables taking certain values into a linear equation results in reliability. Unlike classification methods such as random
forests, naïve-bayes, and SVMs, multiple logistic regression is not a “black box” approach. Users can see how explanatory variables impact the model. Shmueli et al. (2017) outline the pros and cons of the logit model. Though less robust than linear regression, if stable data can be provided, logistic regression can provide great predictive power. That predictive power can be delivered so long as certain problems are addressed. Collinearity can prove to be a problem. It occurs when highly correlated variables distort the results of the model. This problem can be dealt with through principal component analysis (PCA) or using a correlation matrix. (Shmueli et al., 2017)

2.3 Evaluating Prediction Performance

Finally, predictive performance is an important step in determining accuracy of the model. Accuracy can be shown using a confusion matrix. A confusion matrix is an intuitive visual and mathematical representation of the robustness and performance of a model. While the confusion matrix delivers the model’s results, many times the logit model uncovers useful information in the process. This useful information is most satisfying to business leaders who seek to link their intuition to the model. In the model for predicting on-time delivery in trucking, Coyote can benefit from knowing which explanatory variables influence results the most. (Shmueli et al., 2017)

To avoid overfitting and to ensure that the chosen model is able to generalize beyond the dataset on hand, the concept of data partitioning is recommended by Shmueli et al. (2017). Usually the entire dataset is divided into three groups: a training set, a validation set and a test set. After developing a model with the training set, the validation set is used to check how the model performs. The validation partition is used to compare models and select the best performing approach. By using the validation set to define the best model, overfitting is still possible as the validation data might fit the model better than other datasets would. To avoid that, the test partition
data is used to ensure new data is used to assess the performance of the chosen model (Shmueli et al., 2017).

Several methods to evaluate the predictive model can be used in each of the three partitioned groups. As the main goal of the model is to predict whether a new observation belongs to a specified class, James, Witten, Hastie, and Tibshirani (2014) suggest the results of the confusion matrix be used to assess the model’s performance. More specifically, for a predictive model that aims to be directly implemented in the industry, performance and robustness are key characteristics to be considered.

2.4 Section Summary

Accurate predictions for core metrics like on-time performance in transportation are pivotal for firms to succeed in the future. Firms that allocate an exorbitant amount of resources on manually tracking loads will struggle to compete. A number of classification methods can be used to predict delays, thereby more effectively allocating resources. The aforementioned research examines some similar goals to ours with a number of approaches. It is critical to identify how the end user of the model will interact with it. If the user does not care why the predictive model works, a “black box” approach will suffice. Those business leaders with interest in linking their intuition with the model will desire methods that enable interaction. The multiple logistic regression model is an appropriate solution for such leaders. As a result, we recommend the application of the logistic regression model to Coyote for more effective resource allocation. Among the many benefits, simplicity, robustness, and transparency support the predictive power of the logistic regression model. The research explored offers insights into the key steps of variable selection and data preparation. Though model selection differs, these different approaches offer value by
providing a means to validate the multiple logistic regression model for predicting on-time delivery in the trucking industry.
3. Methodology

This section identifies the steps taken to select the appropriate variables and approach for building a model to predict on-time delivery in trucking. The data profiling and data analysis subsections explain the steps in depth. First, a site visit facilitated the observation of the operation, agreement on scope and deliverables, and brainstorming of variables. Once the data were received, a handling process and checks for outliers were performed. An appropriate software selection was made to achieve the agreed upon output. External data needed for variables not included in the Coyote’s database were obtained. Next, the data were restructured in order to effectively extract its information and adapt it for the purpose of a predictive model. Finally, following multicollinearity checks, partitioning, and sampling of the dataset, different models were analyzed for performance. The logistic regression model, the top performing model, was selected and validated.

3.1 Data Profiling

3.1.1 Gathering Data

A site visit was made to Coyote’s headquarters in Chicago, IL. The Senior Director of Supply Chain showed us the operation and explained the functions of personnel in different roles. We spoke with individuals in those roles. Through these interviews, we developed an understanding of the processes and functions at Coyote Logistics. In essence, Coyote matches shippers with carriers using a proprietary software.

The company made a recommendation to create a list of variables that could impact on-time delivery in trucking. We agreed that bucketing the variables through a fishbone diagram enabled the creation of an exhaustive list in the brainstorming process. The buckets chosen were load, lane, carrier, process, facility, and Coyote. Attributes of these six categories were added to the diagram shown in Figure 1.
Central to this thesis is the notion that the model will enable 3PLs to allocate their resources more appropriately. Resource allocation, for this thesis, is defined as the resources used to track loads. Coyote, a 3PL that differentiates themselves by high service levels, tracks every load.

In addition to the high-level discussion of the thesis outcome, we agreed on 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 from October 1, 2014 to September 30, 2016. We determine this time horizon to be robust due to recency of the data and the inclusion of two calendar cycles. The geographical scope of this thesis is restricted to the loads within the United States. Additionally, to narrow the impact of possible variables, the company restricted the study to FTL (full truckloads).
3.1.2 Data Handling and Preparation

The data handling and preparation process is critical to developing an accurate model. Through weekly calls and an open line of communication with Coyote, we classified outliers among the data. Sanity checks verified by Coyote enabled us to clean the data appropriately. Our objective at this stage was to have an exhaustive and accurate dataset.

The next step was to select an appropriate software package to analyze the data and build the model. As presented by Shmueli et al. (2017), some widely used models to predict categorical response using continuous and categorical predictors are logistic regression, neural nets and random forests. Thus, we required a statistical software package with capabilities of performing those models. JMP met the requirements for the thesis and presented important features that we were seeking, such as:

- Powerful statistics with dynamic graphics that enabled a deep understanding of the data
- Ease of connection with different datasets, from Microsoft Excel to SQL
- Automation of data cleaning tasks, which accelerated the task
- Quick and interactive table joins
- The profiler feature enables the simulation of the effects of variables and easily communicates those effects with business leaders

Despite the appropriateness of this software for this thesis, JMP is not as flexible as a coding software like R or Python, which allow the user to explore more customized functions. Ease of use and lack of need for coding in this thesis drove us to support our decision to select JMP.

Though Coyote provided us with an extensive amount of data, some data were not included. Data for the weather observed during the shipments were not recorded by Coyote. We
retrieved weather data from NOAA’s Severe Weather Data Inventory. Initially, we matched the loads to the weather data based on zip code. Due to concerns regarding the high degree of granularity for the zip code, we iterated and matched the weather data at the county level. The county level data provides a better representation of severe weather within a region.

![NOAA's Severe Weather Data Inventory Web Service](image.png)

Figure 2. NOAA's Severe Weather Data Inventory Web Service

At this stage, we performed the relevant checks and eliminated outliers, which were validated by the experts. Finally, external data were retrieved and added at the level of granularity to capture regional severe weather events.

3.1.3 Data Joins

The data presented to us by Coyote have information pertaining to a load at each point, or node, in the system. A number of challenges associated with analyzing the data with information solely from the nodes are evident. A connection must be made to create a segment representing the leg of a trip. As a result, we merged each sequential node to create segments. The segments allow us to observe each leg of the trip as one record. Each observation consists of data from the
origin point and the destination point. Through this approach, we built a predictive model that independently assesses the attributes of each trip leg for statistical significance.

Just over 95% of the data consists of loads with one origin and one destination. These loads are point-to-point shipments with only one segment. The remaining data have more than one segment in a one-to-many or many-to-one configuration.

3.2 Data Analysis

3.2.1 Multi-collinearity

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

The correlation matrix is the most useful tool for us in analyzing the continuous variables. Correlation matrices can be used to ascertain the trends between any two continuous variables in a dataset. A cutoff point can be selected to determine the level of correlation as a decimal between 0 and 1. For our analysis, we use the default cutoff point of 0.5. The only two variables with correlation are the number of trucks and the number of drivers in a company. This link is intuitive and we agreed with Coyote to eliminate one of the variables and use just one in the analysis.

The correlation matrix presents the coefficients of correlation \( (r) \) that summarizes the strength of the linear relationships between each pair of variables \( (X \) and \( Y) \). There are several methods to calculate correlation matrix. In this thesis, we use the pearson product-moment correlation. The coefficient of correlation is calculated as follows:

\[
r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}
\]
If there is an exact linear relationship between $X$ and $Y$, the correlation between the pair is 1 (positive) or -1 (negative). If the correlation is 0, it means that there is no linear relationship between the variables. For more detailed mathematical information regarding correlation matrix, please refer to Rodgers and Nicewander (1988).

Principal component analysis (PCA) and multiple correspondence analysis (MCA), though powerful in exposing correlation, are not intuitive. New vectors $(Z_i)$ are formed with weights $(a_{i,p})$ for each variable and the spread of the data points indicates a lack of correlation. These vectors are termed eigenvectors, and the magnitude of variation explained by each principal component is called eigenvalue. PCA can identify the need for dimensionality reduction in the dataset. The principal components for the Coyote dataset dictate that no dimensionality reduction is required. PCA is an effective tool for identifying correlation in continuous data. The PCA results can be seen in Figure 3. For finding correlation between categorical variables, MCA is the appropriate method. A similar lack of correlation is observed in the MCA for the categorical variables in the dataset.

In PCA, we aim to find a new set of variables $Z_1, Z_2, Z_3, \ldots Z_p$, that are weighted averages of the original variables $X_1, X_2, X_3, \ldots X_p$. In this case, all pairs of $Z$ present 0 correlation:

$$Z_i = a_{i,1} (X_1 - \bar{X}_1) + a_{i,2} (X_2 - \bar{X}_2) + \cdots + a_{i,p} (X_p - \bar{X}_p), \quad i = 1, \ldots, p$$

The new variables $Z's$ are ordered by variance, from the largest variance ($Z_1$) to the smallest ($Z_p$). The weights $(a_{i,p})$ are used in computing the principal component scores. Figure 3 shows the relationship between the number of components and the corresponding percentage of total variation it accounts for (in the eigenvalues table). For more detailed mathematical information regarding principal components analysis, please refer to Jolliffe (2002).
3.2.2 Sampling and Partitioning Data

The dataset used for this thesis presents a very imbalanced proportion of on-time and delayed loads. The small representation of delayed observations reflects the high service level that Coyote provides to its customers. In order to develop a model capable of capturing the useful information that distinguishes the underrepresented class from the dominant class, Shmueli et al. (2017) recommend a technique called stratified sampling.

Stratified sampling is a method of sampling data used when classes are presented in a very unequal proportion in the original dataset. As suggested by Shmueli et al. (2017), by performing stratified sampling we force our model to work with a dataset with an equal number of observations (50% on-time and 50% delayed). In our case, there are 26,146 minor class (delayed) observations in the dataset. By defining an even distribution for the sample, the method maintains the total
number of the minor class 26,146 observations and randomly selects 26,146 majority on-time observations from the original dataset, representing an undersampling technique. Through this technique, the model would be able to explore more information from the minority class and increase its predictive power. Figure 4 shows the number of observations for each class resulting from the stratified sampling. There are 26,146 delayed observations and 26,146 on-time observations in the stratified dataset, divided in two groups: training and validation.

![Distributions Validation Stratified by OnTime of EndSegment=Training](image)

![Distributions Validation Stratified by OnTime of EndSegment=Validation](image)

**Figure 4. Stratified Sampling for the Coyote Dataset**

Besides certifying a correct representation of both classes (0 and 1), it is also extremely important to avoid overfitting and ensure that the chosen model is able to generalize beyond the dataset on 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. Following rule of thumb, our dataset is partitioned as follows: 75% training set, 25% validation set.

### 3.2.3 Stepwise and Logistic Regression

The predictive analytics model we create shows the combinations of variables that lead to delays. Moreover, based on those variables, importance is placed on the likelihood of a particular
load being delivered on-time. To align our model selection with the desire for an explanatory model with high interpretability of model results, we started our analysis using the logistic regression model. Its main goal is to predict the probability of a new observation belonging to a specified class. In predicting a binary output variable (0 or 1), based on probability, it predicts the likelihood that a new observation would have to be classified as 0 or 1.

Our dataset had more than fifty input variables, nominal and continuous, and one binary output variable. We look 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 present better performance in the validation set, we use the stepwise regression approach. The stepwise approach enables us to explore different combination of variables with a quick and flexible interface.

The logistic stepwise regression function is used to select a subset of variables for a regression model that deals with a nominal or categorical response. It is also known as standard forward search, because it starts from an empty model, and at each step the model selects a variable that increases maximum likelihood fit. As we partitioned the data and created a validation set, as described in section 3.2.2, the stepwise function tries to maximize the Validation $R^2$, which means that the function attempts to find a model that maximizes the $R^2$ statistic for the validation set. The $R^2$ statistic can be calculated as follows:

$$R^2 = 1 - \frac{RSS}{TSS}$$

TSS represents the sum of squares of the differences between the actual responses in the validation set and their mean, while RSS represents the sum of squared prediction errors:

$$TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$
For more detailed mathematical information regarding $R^2$ statistic, please refer to James et al. (2014).

Once the set of variables that presents higher information regarding on-time performance is defined, we run the logistic regression model. Logistic regression's main goal is to predict the probability of a new observation to belong a specified class. In predicting a binary output variable (0 or 1), based on probability, it predicts the likelihood that a new observation would have to be classified as 0 or 1. Logistic regression models follow two steps:

1. Estimate probabilities of a new observation to belong to each class (0 or 1).
2. Classify each observation to one of the classes based on a specified cutoff value.

The model works on the same principles as linear regression, but instead of using $Y$ as the dependent variable, it introduces the nonlinear function logit.

To calculate the logit, some intermediate steps are required. First, we calculate the logistic response function. Assume that $p$ represents the probability of belonging to group 1, where $p \in [0,1]$, and that $x_1, x_2, \ldots, x_q$ represent the input variables. The logistic response function is calculated as follows:

$$ p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q)}} $$

After calculating the logistic response function, we should find the ratio of probability of belonging to group 1 to the probability of belonging to group 0, known as odds:

$$ odds (Y = 1) = \frac{p}{1 - p} $$

Combining the two equations above we find that:
odds \((Y = 1) = e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_q X_q)}\)

Taking the natural logarithm on both sides we find the log(odds), also known as logit:

\[
\log (\text{odds}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_q X_q
\]

The logit function is the dependent variable for the logistic regression model and it is a linear function of the \(q\) predictors. A logit = 0 represents an even odds of 1, which means the probability of being 1 and 0 are the same (probability = 0.5).

For more detailed mathematical explanation regarding logistic regression model, please refer to Shmueli et al. (2017).

### 3.2.4 Model Validation - Neural Nets and Bootstrap Forest

In addition to logistic regression, Shmueli et al. (2017) suggest neural nets (neural networks) and bootstrap forest as two commonly used models to predict categorical response using continuous and categorical predictors. To validate our results with other models’ output, we run both models on the same dataset for results comparison.

Initially, we ran a neural network, which is a model based on the functioning of the human brain. It mimics the biological activity in the brain where information processing is performed by neurons. The model combines input information in a complex and flexible way to capture relationships between predictors and the response. A neural network can be seen as a set of hidden nodes, which are nonlinear functions of the original inputs. With neural nets in JMP, the user can specify up to two layers of hidden nodes, with each layer containing as many hidden nodes as desired.

Figure 5 supports the understanding of the principles that underline the neural network method presenting an example of a neural network design.
The figure shows two predictors (nodes 1 and 2) and two output variables (nodes 6 and 7). In this system, there are 3 hidden layers (nodes 3, 4 and 5). Assume that $W_{i,j}$ represents the weights connecting node $i$ to node $j$. The bias values $\theta_j$, is a coefficient not subject to iterative adjustment, that defines the level of contribution of node $j$. Each node takes the inputs $x_i$ and computes the output by taking the weighted sum:

$$S_j = \theta_j + \sum_{i=1}^{p} W_{ij} x_i$$

In the next step, the function $g$, known as transfer function, is applied to the result of the calculation above. In this thesis, we used the sigmoidal function (TanH), as follows:

$$g(s) = -1 + \frac{2}{1 + e^{-s}}$$

In the case illustrated in figure 5, we aim to calculate the output value of nodes 6 and 7. We can write the output of node $j$ as:
Output \( j = g \left( \theta_j + \sum_{i=1}^{p} w_{ij} x_i \right) = -1 + \frac{2}{1 + e^{-(\theta_j + \sum_{i=1}^{p} w_{ij} x_i)}} \)

For more detailed mathematical explanation regarding neural network, please refer to Shmueli et al. (2017).

Secondly, we implement the bootstrap forest model, which is a variant of random forests. In this model, random samples are generated from the original dataset, and this sampling process is termed bootstrapping. Then, predictors are randomly selected from the available variables, a classification tree fits each sample, and thus a forest is created. The process is repeated and the final model is the average of all of the trees, producing a “bootstrap aggregated” model.

Denote a set of independent observations \( Z_1, Z_2, Z_3, \ldots, Z_n \), with individual variance \( \sigma^2 \), the associated variance of the mean of the observations \( \bar{Z} \) is \( \sigma^2 / n \). The idea behind this concept is that we can bootstrap the data, by taking repeated samples from the training dataset, and thus reduce the variance and increase the prediction accuracy. In the bootstrap forest, the method generates \( B \) different bootstrapped training datasets and averages all predictions \( \hat{f}^*b(x) \) to obtain:

\[
\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^*b(x)
\]

The \( \hat{f}_{bag}(x) \) is called bagging or bootstrap forest. For further details on bootstrap forest, please refer to James et al. (2014).

The next step is to identify the pros and cons of each model, recommend a model to Coyote, and finally substantiate our recommendation based on the predictive performance of the models and their qualitative features.
3.2.5 Performance Evaluation

Once we develop and run three different models on our dataset, we determine how to measure the predictive performance of each. Different methods to evaluate a model’s performance can be used. The most common metric for regression models is the adjusted $R^2$:

$$R^2_{adj} = 1 - \frac{n - 1}{n - p - 1} (1 - R^2)$$

where $n$ is the number of observations, $p$ is the number of independent variables and $R^2$ measures the percentage of the total variation of the dependent variable explained by the model, according to formulation presented in section 3.2.3.

Even though adjusted $R^2$ is widely used, as the main goal of our model was to predict a binary outcome, we use a confusion matrix. James et al. (2014) suggest the results of a confusion matrix to assess the performance of predictive models. For nominal or ordinal responses, statistical software packages commonly provide a confusion matrix report. 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. Shmueli et al. (2017) point out that the confusion matrix provides an intuitive visual and mathematical representation of the performance of a model. Figure 6 shows a confusion matrix, also known as classification matrix, with the description of each of the quadrants in the two-by-two table.
A main accuracy method used to evaluate model’s performance based on the confusion matrix is the estimated misclassification rate, also known as overall error rate. It is calculated as follows:

\[
err = \frac{n_{0,1} + n_{1,0}}{n}
\]

where, \( n = n_{0,0} + n_{0,1} + n_{1,0} + n_{1,1} \)

Other common performance measures are

\[
Sensitivity = \frac{n_{0,0}}{n_{0,0} + n_{0,1}}
\]

\[
Specificity = \frac{n_{1,1}}{n_{1,0} + n_{1,1}}
\]

It is crucial to note that all the three models (logistic regression, neural nets and bootstrap forest) are built using a well-balanced dataset. As previously stated, by performing a stratified sampling technique, we force our model to work with a dataset with an equal number of both types of observations (50% on-time and 50% delayed). Therefore, the direct results from the confusion matrix reflect the balanced data and do not represent the performance of the real data. To correct
this problem, we recalculate the confusion matrix through a reweighting technique. The revised confusion matrix counterbalances the observations in the matrix based on the real proportion of the observations in the dataset.

The original dataset presents 95% of on-time observations, and the remaining 5% delayed. The stratified sampling changes the proportion to an even balance, thus each delayed observation in the whole data is worth 10 in the sample data (50/5). In addition, each on-time observation in the whole data is worth 0.53 in the sample data (50/95). These values are called sampling weights and the reweighted matrix is calculated as follows:

<table>
<thead>
<tr>
<th>Actual Class</th>
<th>Predicted Class</th>
<th>C₀</th>
<th>C₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>C₀</td>
<td>n₀₀ /10</td>
<td>n₀₁ /10</td>
<td></td>
</tr>
<tr>
<td>C₁</td>
<td>n₁₀ /0.53</td>
<td>n₁₁ /0.53</td>
<td></td>
</tr>
</tbody>
</table>

*Figure 7. Confusion Matrix, Reweighted. Adapted from Data Mining for Business Analytics: Concepts, Techniques, and Applications in JMP Pro. (p. 248), by Shmueli et al., 2017, Hoboken, NJ: John Wiley & Sons. Copyright 2017 by John Wiley & Sons. Adapted with permission.*
4. Results

This section presents the results obtained through applying the different types of predictive models previously mentioned. The logistic regression model is the chosen model, whereas the neural nets and bootstrap forest models validate the selected model. We discuss the outcome from the variable selection process and the subsequent evaluation of predictive models created using those variables. In addition, we explore attempts to improve the performance of the models.

4.1 Variable Selection & Logistic Regression

The stepwise approach determines statistically significant variables in a given dataset by using iterative algorithms. Stepwise provides an adequate and efficient method of selecting variables with results that can match exhaustive techniques. This widely used method, termed stepwise regression, is popular in variable selection. Despite its popularity, stepwise iterates instead of performing the computationally demanding task of generating an exhaustive list of variable combinations (Shmueli et al., 2017). The results given by stepwise can provide insight into whether it will be a sufficient technique in variable selection. In this thesis, the stepwise regression approach was an appropriate tool for variable selection, namely due to the results.

The results from the stepwise regression reveal six explanatory variables with very high predictive power. The predictive power is quantified using the p-value and the log worth, which is a function of the p-value:

\[ \text{LogWorth} = -\log_{10}(p\text{-value}) \]

LogWorth transforms the p-value in order to provide an adequate scale for graphing. From the formula above, it is noted that a LogWorth equal to 5 represents a p-value of 0.00001:

\[ (-\log_{10} 0.00001 = 5) \]
The p-value can be found based on the F-statistic using this distribution.

\[ F = \frac{RSS_0 - RSS}{RSS \frac{n - p - 1}{n - p - 1}} \]

\[ RSS = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \]

where \( n \) is the number of observations, \( p \) is the number of independent variables and \( RSS_0 \) is the residual sum of squares for the whole model. For further details on F-statistic and p-value calculation, please refer to James et al. (2014).

Figure 8 shows the output from Stepwise platform. The p-values are so low that each variable exceeds the maximum limit for this effect summary tool.

Figure 8. Stepwise Regression for Variable Selection in the Coyote Dataset

The p-values for the top six explanatory variables are smaller than 0.00001. These values are several orders of magnitude smaller than the typical required minimum p-value of 0.05. A p-value equal to 0.05 represents a 5% chance of the results randomly happening. Our hypothesis, given the extremely small p-values, is that a model formed using these six explanatory variables yields good results.

While the Stepwise approach determines that these six variables offer the most predictive power, it was crucial to validate the selection with the experts at Coyote. Four of the six variables are known when the load is tendered. They include:
- The number of incidents per volume for the carrier
- Whether the load tendered on the spot market or contracted
- Whether the delivery facility operates with delivery appointments
- The historical volume of the carrier (with Coyote)

Two explanatory variables are known from the events unfolding at the pickup location. These include:

- Whether the pickup was on-time (within one hour of scheduled time)
- The duration at the pickup location

These six variables appear to make logical sense for affecting on-time delivery. When discussed with Coyote, their experts agreed that these factors indeed played a role in whether a load would be on-time. After this validation, we move forward with the process of running models with these six variables.

The direction from our literature review drives us to run the logistic regression model first. We run the logistic regression model with the six variables on the stratified dataset with no missing values. Specifically, we aim to avoid any distortion of the model by lacking values or having to infer values for some observations. The model yields the results shown in Figure 9.

![Confusion Matrix](image)

**Figure 9. Confusion Matrix Results from the Logistic Regression Model**

This confusion matrix reflects a misclassification rate of 35.51% on a sample with an equal number of observations on each class (50% on-time and 50% delayed). To estimate the performance of the model for the original imbalanced dataset (95% on-time and 5% delayed),
Shmueli et al. (2017) outline a procedure for adjusting the sampled confusion matrix. The actual number of delays should be divided by 10 (5%/50%) and the actual number of on-time observations should be divided by 0.53 (95%/50%). The adjusted confusion matrix is shown in Figure 10.

**LOGISTIC REGRESSION**

<table>
<thead>
<tr>
<th></th>
<th>Predicted</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Actual</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>219</td>
<td>209</td>
</tr>
<tr>
<td>1</td>
<td>1,805</td>
<td>6,337</td>
</tr>
<tr>
<td>(\Sigma)</td>
<td>2024</td>
<td>6546</td>
</tr>
</tbody>
</table>

err = \(\frac{m_{1,0} + n_{0,1}}{n}\) = 23.50%

missed delays = \(\frac{n_{0,1}}{n}\) = 2.44%

*Figure 10. Adjusted Confusion Matrix Results from the Logistic Regression Model*

The adjusted misclassification rate that can be expected from running the model on an ordinarily sampled dataset is 23.5%.

Though misclassification rate is a widely used indicator of fit for the models, application of how the model would be used in practice is a crucial consideration. Specifically, we consider 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 and 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. In the adjusted confusion matrix, the error for missed delays is 2.44% of the total observations.

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 results presented in Figure 10, 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.

As we gain depth of knowledge for each of the four components of the confusion matrix, we expand our sense of predictive power for the logistic regression model. With the model in Figure 10, it is possible to allocate 23.6% of the current tracking resources and miss only 2.44% of loads that should be tracked. The reason for this drastic reduction in tracking with a relatively small error is the high negative predictive value of the model. The negative predictive value offers the greatest certainty that when a load is predicted to be on-time, it will in fact be on-time.

4.2 Model Validation

4.2.1 Neural Nets Model

Despite our satisfaction with the initial logistic regression model results on a conceptual level, other models offer value by validating those results. With the same stratified dataset without any missing values, we run a neural network model. Figure 11 displays the results from the neural network model.
Figure 11. Adjusted Confusion Matrix Results from the Neural Network Model

With an overall misclassification rate within 3% difference from the logistic regression model and a missed delay error within 0.4%, it is clear that the neural network produces similar results. Additionally, a similar percentage of loads are predicted to be delayed. Therefore, the practical use of this model can be comparable to the logistic regression model.

The interpretability of the results for the neural network model is also a point for discussion. It appears that the neural network model produces a smaller error for missed delays while only marginally increasing the misclassification rate and resource allocation. While these facts are evident, we question the robustness of this model and its duplicability. We are interested in how the model performs given new data. Despite new questions and interests, the neural network model validates our results.

4.2.2 Bootstrap Forest Model

In addition to neural networks, bootstrap forest is an appropriate method for model validation. The process of aggregating many decision trees is a very different approach from the transformation that takes place in the logit function. The variance of this approach offered significant value for our validation attempts.

With the same stratified dataset, we run a bootstrap forest model using the six variables. Figure 12 displays the results from the model.
Figure 12. Adjusted Confusion Matrix Results from the Bootstrap Forest Model

The results reflect a very close misclassification rate to that of the logistic regression model. An improvement in the error for missed delays is observed. A similar tracking recommendation is offered by the model. Each of these metrics validates the logistic regression model results. We share similar thoughts from running neural networks regarding the improvements of the bootstrap forest.

4.3 Attempts to improve the model

4.3.1 Final Correlation Check

As we confirm that logistic regression is the focus model for the thesis, we shift gears towards further validation and improvement. Our first point of validation is to ensure that the six explanatory variables are not correlated in any way. We previously performed correlation checks on the entire dataset through covariance matrices, PCA and MCA. A correlation matrix verifies the level of correlation among those variables through our lens.

Figure 13 shows the relative correlation of variables in the dataset on a scale of -1 to 1. A value of zero indicates no correlation and a value of 1 or -1 indicates the strongest possible correlation.
The largest absolute value in the correlation matrix was 0.3359, which is well below our cutoff threshold of 0.5. Any value above 0.5 is deemed too correlated, thus distorting the results of the model.

4.3.2 Search for Additional Variables

Although the initial process of brainstorming variables through the fishbone diagram was exhaustive, we continue to search for additional data to improve the model. If a variable proves to have explanatory power, we perform the same correlation exercise to maintain model integrity. We also examine the possibility of a dual model where preliminary decisions can be made and then improved with new information.

The first new variable we examine is geographical zones in the US. Using a US map divided into six regions, we assign zones for the origins and destinations. We selected these six zones to emulate Coyote’s use of teams for different geographical zones. The zones can be seen in Figure 14.
The zones we develop from the map are as follows:

- Pacific
- Mountain
- Central
- South Central
- Southeast
- Northeast

There is statistical significance for the loads with origins and destinations in the Pacific region. In addition, there is explanatory power for loads that were picked up in the Mountain region. The Log Worth and p-values can be seen in Figure 15.
**Figure 15. Statistical Significance of six chosen variables and geographical zone variables**

Despite the three aforementioned geographical zone variables having a small enough p-value to be included in the model, we omit 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 geographical zones as variables.

Next, we explore the transformation of the timestamps provided in the data to time buckets. The time buckets were given by Coyote's experts based on their best estimates for common facility times. The time buckets used for the analysis are shown in Figure 16.
<table>
<thead>
<tr>
<th>Time bucket</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early Morning</td>
<td>4:01 AM</td>
<td>8:00 AM</td>
</tr>
<tr>
<td>Morning</td>
<td>8:01 AM</td>
<td>12:00 PM</td>
</tr>
<tr>
<td>Afternoon</td>
<td>12:01 PM</td>
<td>2:00 PM</td>
</tr>
<tr>
<td>Late Afternoon</td>
<td>2:01 PM</td>
<td>5:00 PM</td>
</tr>
<tr>
<td>Rush Hour</td>
<td>5:01 PM</td>
<td>7:00 PM</td>
</tr>
<tr>
<td>Evening</td>
<td>7:01 PM</td>
<td>10:00 PM</td>
</tr>
<tr>
<td>Night</td>
<td>10:01 PM</td>
<td>4:00 AM</td>
</tr>
</tbody>
</table>

*Figure 16. Time buckets for variable creation from timestamps*

We transform the timestamps for six variables into the buckets shown in Figure 16. The variables include facility open and close times as well as appointment times. When running the logistic regression model including the time bucket variables, we notice a trend similar to including the geographical zones. Figure 17 shows the statistical significance of the variables including the time bucket variables.
Figure 17. Statistical Significance of six chosen variables and time bucket variables

The same conclusion is drawn regarding the applicability of the information from the time bucket variables. Potential weakening of the model in its robustness trumped the marginal improvement of adding these new time bucket variables.

The act of exploring new variables adds value to the thesis. In addition to continuing to add to the exhaustive list of variables, it adds to our comprehension of the confusion matrix results. Ultimately, we believe that bucketed or zoned variables deliver less explanatory power than continuous and less rigid variables.

After exploring new variables, we test a hypothesis that a dual model can be used to give a planning advantage to Coyote. Essentially, a dual, or two-part, model could allow Coyote to track loads using information available at two separate times. First, the model would be run once the load is tendered using the four variables known at that time. Then, once the load has cleared the pickup location, whether it was on-time and the duration at the pickup facility would be known.
A second model could be run, potentially enhancing Coyote’s ability to predict delays. The main caveat with this hypothesis would be whether the same loads would remain in the same confusion matrix quadrants. Inconsistency among the confusion matrices could prove the dual model hypothesis invalid. The reason is that it is not feasible to suddenly track a load that has previously been predicted to be on-time.

We assigned binary values to the predictions of the models with four and six variables. We summed the binary values to determine the consistency among the predictions. Unfortunately, this simple arithmetic analysis indicated unwanted changes in confusion matrix section movement. Specifically, 5.7% of observations change from on-time predictions to delay predictions. This is an unsatisfactory figure in terms of magnitude. Therefore, the hypothesis for a dual model was disproved.

Our attempts to improve the model and its use, though unsuccessful, are not fruitless. We make a conscious decision to omit variables that do not contain a high degree of statistical significance. That process drives our desire to obtain new data unseen by the model to ascertain its robustness.

4.4 Test data

Coyote graciously provided us with data not used to build the model to gauge its accuracy and robustness. This provides a significant opportunity to test the model. The new dataset consists of 1973 observations. We run the new “Customer Test Data” through the model. Figure 18 shows the comparison of the results from the Customer Test Data and the validation data.
### Validation

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted</th>
<th>0</th>
<th>1</th>
<th>Σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>219</td>
<td></td>
<td>209</td>
<td>429</td>
</tr>
<tr>
<td>1</td>
<td>1,805</td>
<td></td>
<td>6,337</td>
<td>8,142</td>
</tr>
<tr>
<td>Σ</td>
<td>2024</td>
<td></td>
<td>6,546</td>
<td>8,570</td>
</tr>
</tbody>
</table>

\[
\text{err} = \frac{(n_{0,1} + n_{1,0})}{n} \\
\text{missed delays} = \frac{n_{0,1}}{n}
\]

### Customer Test Data

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted</th>
<th>0</th>
<th>1</th>
<th>Σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>23</td>
<td></td>
<td>50</td>
<td>73</td>
</tr>
<tr>
<td>1</td>
<td>452</td>
<td></td>
<td>1,448</td>
<td>1,900</td>
</tr>
<tr>
<td>Σ</td>
<td>475</td>
<td></td>
<td>1,498</td>
<td>1,973</td>
</tr>
</tbody>
</table>

\[
\text{err} = \frac{(n_{0,1} + n_{1,0})}{n} \\
\text{missed delays} = \frac{n_{0,1}}{n}
\]

**Figure 18. Confusion Matrix Results Comparison for Validation and Customer Test Data**

The results from Customer Test Data reflect the very robust nature of the logistic regression model. The largest difference observed is 1.8%, which results from misclassifying on-time loads. An important consideration is that the missed delays error changed very little. A massive variance in this error would be grounds for disqualifying the model. Fortunately, we do not observe any significant variance.

### 4.5 Section Summary

The analysis of the comprehensive dataset yields six variables with statistical significance. The six variables are on-time pickup, duration at the pickup facility, historical volume of the carrier, incidents per volume of the carrier, contract versus spot tendering of the load and the facility type of the delivery facility (appointment versus open). With those variables, we ran several models. The model with the most robust and applicable performance was the logistic regression model. Subsequently, we selected this model for the thesis.
5. Discussion

5.1 Findings from Final Model

Many metrics can be used to measure models. In predictive modeling, where a binary response is required, misclassification is the overall metric of choice. Misclassification, however, is a very general metric. 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 this 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.

We discuss the idea of a dual model. This type of model that compiles a wide range of loads to track which would subsequently be reduced with new information is interactive in a business sense. Unfortunately, the dual model does not align with our logistic regression method.

5.2 Suggestion for future investigation

Although our model delivers potentially huge resource reductions (76.4%) while minimizing the missed delays, Coyote expressed the desire to improve the missed delay error. The company suggested comparing the missed delay error with the total actual delays, instead of using the total number of observations (actual delays + actual on-time). Our efforts to reduce the missed
delay error gives marginal results at the expense of skyrocketing the misclassification rate and rendering the model useless. It would be interesting to see if another modeling approach could deliver better results related to the dual model. Also, a method with the ability to optimize certain model metrics by adjusting parameters would be of interest.

We are interested in the loads that fall into the missed delays category. Specifically, what is the controllability of the loads that are predicted to be delayed? Can Coyote representatives do anything about it? Are the tracking techniques effective? We would like to see the results from the implementation of the model to gather more data and answer these questions.

Additionally, we have discussed the different sampling methods with Coyote and analyzed the sensitivity analysis associated with pulling certain levers. Undersampling the more plentiful class was critical for building our model. If additional research was done on optimizing stratified sampling techniques, we feel our analysis could be improved.

It is important to note that the dataset used in this study presented information from the pick-up of a load (start segment) and its delivery (end segment). What occurred in between these two events was not used in this thesis. The increased availability of online information through new technologies, such as telemetry, and the readiness to store those records on remote servers using cloud servers could contribute to future research. The development of a predictive model able to capture information from online records could bring new insights and complement the analysis presented in this study.
6. Conclusion

This research focuses on predicting on-time delivery in the trucking industry. If firms can accurately predict the future with certainty, decisions can be made to be more profitable. For 3PLs, predicting on-time delivery for more effective resource allocation is key to remaining competitive in the future.

Working with a firm like Coyote Logistics that fully understands the power and importance of data is advantageous and leads to valuable findings in this thesis. A high degree of data availability enabled the creation of many variables. The variable selection process resulted in the selection of six explanatory variables with statistical significance. Using those six variables, the logistic regression model was run, presenting predictive value to Coyote.

Though access to data was relatively high in this thesis, there are infinite possibilities of data that could improve the model. If more data could be obtained from the carrier, it could add value to the model. Also, as real time traffic data collection and availability improves, it could be integrated into the model.

A subject of interest for future investigation is the controllability of the loads that are delayed. Does predicting a delayed load correctly help get the load there any faster? What is the impact of tracking and resources on a load predicted to be delayed? Loads with attributes that make them likely to be delayed could be given to certain carriers to reduce the likelihood of that delay occurring.

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.
Reference List


51