A Machine Learning Approach to Delivery Time Estimation for Industrial Equipment

Douglas Halim, Alexander Hartman, Aditya Jariwala, Rishabh Mohan, Matthew A. Lanham

Purdue University Krannert School of Management

halimd@purdue.edu; hartma82@purdue.edu; ajariwa@purdue.edu; mohan35@purdue.edu; lanhamm@purdue.edu

Abstract

Our research focuses on obtaining better predictions for lead-time of made-to-order equipment for a large multinational corporation. In collaboration with this corporate partner, our team was tasked to create a deployable solution that could provide reliable delivery predictions. The motivation for this work is that when customers place orders for pieces of equipment, and they are provided an expectation that their product will be delivered in a timely manner. Without a delivery estimation system currently in place, the company cannot provide customers an expected time window, which is an inconvenience for the customers that have their own operational planning to use this equipment. To predict this lead-time, our team was provided access to tens of thousands of entries of equipment order data. We experimented with many models considering the unique aspects of the features and were able to obtain predictions of delivery time for each product line. Our predictive approach provides a solution to this business dilemma, by providing a highly accurate, cross-validated predictions of delivery time as well as a corresponding prediction interval. We believe our approach could be easily extended to other similar type supply-chain problems.

Introduction

The primary purpose of this study is to understand key factors in predicting lead-time for industrial grade equipment, and developing a model to accurately make these predictions. Compared to consumer, which tend to be stocked in warehouses and can be easily tracked, large made-to-order industrial equipment doesn't have this flexibility. The equipment not only has to account for customer specifications, but must also account for difficulties associated with large scale, international transportation. These predictive models could be used by companies to develop a better understanding regarding their supply chain, customer response time, and any issues that would slow down the delivery of equipment.

The motivation behind this collaborative project with an industry partner is two-fold:

1) Predictive Modeling for Supply Chain Applications
Predictive applications in supply chain management are becoming prevalent issues facing large companies today. The ability to forecast various stages of a product's life-cycle has translated easily for consumer goods, but not for large industrial equipment.

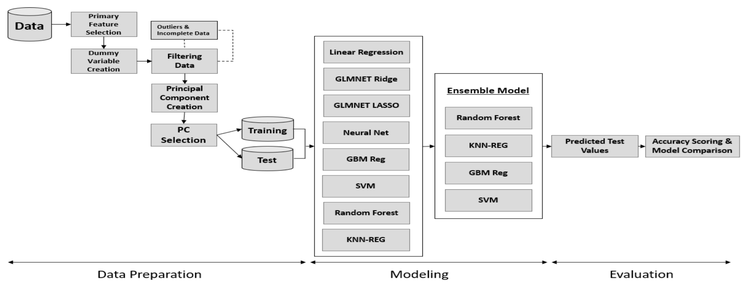
2) Lead time transparency
Industrial companies manufacturing specialized / customized goods have problems measuring all these combinations of equipment leaving their facilities. Predicting lead times allows companies to reduce operating costs, optimize capital, increase revenues and improve their competitive advantage. Clients will be able to better allocate resources efficiently and reduce the risk through certainty in their product purchases.

Methodology

Data Sources

The data that was used for the creation of our model was a dataset provided by the company partner that contained 1,200,986 rows and 51 columns, consisting of a combination of multiple data sources that the corporate partner had at its disposal. This data encompassed various information such as the type of product of each piece of machinery, manufacturer where the product was created, the buyer of the product, and various dates and coordinates related to the order and delivery. Each entry also had a unique serial ID and a measurement of order lead time in days, which became our dependent variable.

Figure 1: Predictive Model Identification



Exploratory Data Analysis

The first step of our process involved data cleaning and deciding which variables to retain for the final modeling. A large portion of the data was categorical, and many of the variables were completely correlated with others (e.g. product line and product division) meaning that we had to be diligent in ensuring that nothing in our final dataset was redundant. Furthermore, the large amount of categorical variables also meant that many dummy variables would have to be created. In addition, many of the columns, particularly involving dates, had up to 90% of the data missing, which had to be accounted for as well. Furthermore, clear outliers existed as a result of unclean data, which led to lead time entries of as little as zero days to multiple years, with either not possible in this scenario.

In addition, the coordinate variables were given but distance between the buyer and manufacturer had not been calculated. Since this represented an important facet of lead-time predicted, this was a variable that had to be added.

Data Preparation

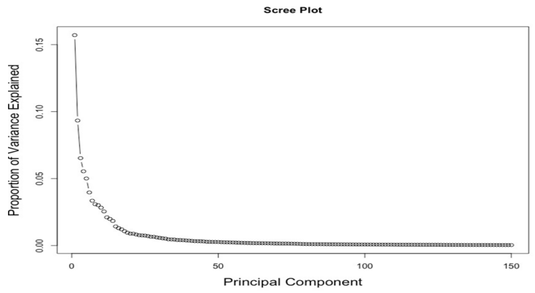
Before altering the data any further, we wanted to add the necessary features that were not present, such as the distance variable. Once these were calculated and added to each entry, certain columns were removed that were either missing a large majority of the entries and could not be calculated, or were not pertinent to our predictive model given the other variables present (e.g. dates, division, coordinates, etc.). In addition, the outlier rows were removed and the lead times were contained between 10 and 300 days, representing a reasonable range of expected lead times.

Any remaining incomplete cases were removed as well. To break down the data further and allow us to test models on a smaller dataset, the data was subset by type of product, for which there were 43 in total. From there, the remaining categorical variables were transformed into dummy variables to be used in the final modeling process, resulting in a dataset of around 1000 columns depending on the type of product line the data was submitted on. After the data was normalized to 90% confidence on lead time, the data was then ready for the model building process.

Data Partitioning and PCA

In order to ensure that the models were not only trained correctly but also produced accurate prediction in the test set, on the data was divided into two subsets using a 80/20 split Given the large size of the product line data, this partition provided ample data for both training and testing However, due to the hundreds of columns, the dataset was still impossible to build models on, so Principal Component Analysis (PCA) was conducted to compile the key variables and ultimately reduce the dimensions of the dataset before modeling. From both the scree plot and eigenvalues, we determined that around 13 principal components accounted for approximately 75% of the variance in the data, so 13 PCs were chosen to feed into models going forward.

Figure 2: Scree Plot



Results

Decision Model
For our analysis, we decided to train and implement multiple machine learning models, consisting of Linear Regression, Ridge Regression, Least Absolute Shrinkage and Selection Operator (LASSO) Regression, Artificial Neural Net, Random Forest, Gradient Boosting Model (GBM), Support Vector Machine (SVM), and a final ensemble model. To score the model, we not only looked at RMSE (Root-mean-square deviation), but also calculated the percentages of instances the predicted lead time fell between +/- 7, 10, and 14 days of the actual lead time.

This method of scoring was requested by the company partner, and would help determine if the models could properly measure within an industry-acceptable threshold of days if implemented for real-world problems. For consistency, each of the models were performed on the product line with the largest number of entries, with the results corresponding to only this product.

Model Comparison Models

Figure 3: RMSE Comparison

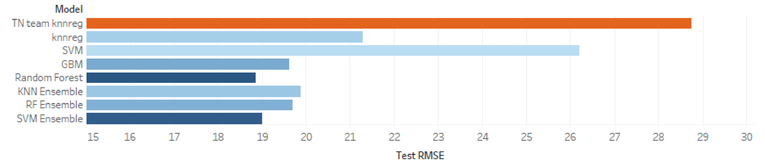
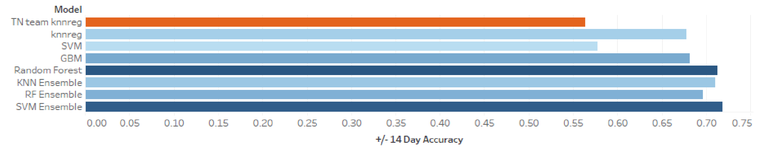


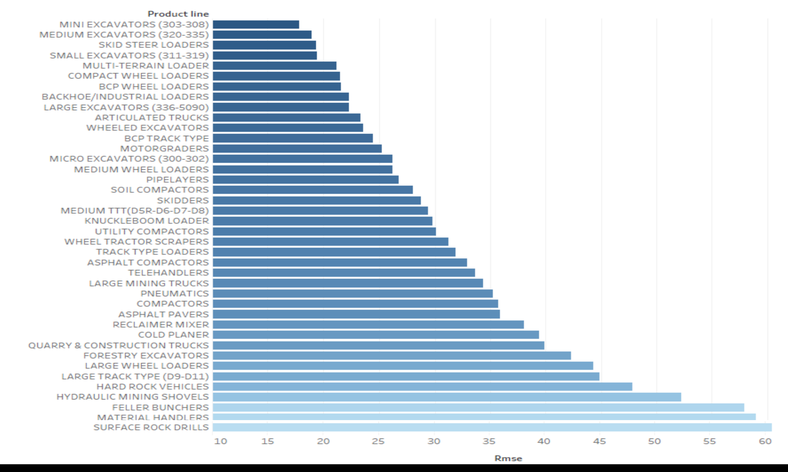
Figure 4: Accuracy Comparison



Given the measurements of RMSE and the accuracy of predicting lead time within certain thresholds, the Random Forest model is the superior model for predicting lead time in our problem. Even with all of the machinery made-to-order, and much of machinery taking months to build and deliver, this model is able to predict lead time within 14 days of the actual time around 71.2% of the time, and predict within 10 days and 7 days 61.8% and 52.4% of the time, respectively. As the products are complex and countless factors and outlying conditions affect the overall lead-time of them, even a model with this kind of accuracy can be extensively helpful when applied to a real-world setting.

In addition, each of the individual products lines were assessed to see if any possesses additional outstanding variances leading to increased model error.

Figure 5: Specific Product-Line Comparison (RMSE)



Conclusions

In order to improve our modeling in the future, additional data on smaller product lines and parts, as well as more factory level features would need to be collected. Furthermore, additional factor-level variance test would be helpful in identifying potential areas in the business to improve efficiencies.
Manufacturers are able to utilize predictive modeling approaches to provide customers with accurate information regarding shipment information for their products. Predicting lead time allows manufacturers to utilize just-in-time manufacturing alleviating high holding costs, allocate resources efficiently during peak seasons, and effectively manage suppliers.

Acknowledgements

Special Thanks to the Purdue Business Information and Analytics Center for its support.