Analysis of Inefficiencies in Shipment Data Handling

By Name(s): Rohini Prasad, Gerta Malaj
Thesis Advisor: Dr. Matthias Winkenbach

Topic Areas: Database Analytics, Tracking and Tracing, Risk Management

Summary: This thesis analyses the errors that occur in shipment data for a freight forwarder. We used descriptive and predictive analytics to identify the relationships between shipment attributes and errors, and to predict the likelihood of errors. We used data visualization, K-means, Naïve-Bayes classifiers, and Neural Networks in the course of our analysis. Results suggest that neural networks are the best predictive model for analyzing error in shipment data.

Before coming to MIT, Rohini Prasad graduated with an MBA from the Indian School of Business and then worked with KPMG Advisory for two years. Prior to that, she graduated from the University of Pune with a Bachelor's degree in Engineering (Information Technology) and worked with Sapient Global Markets for three years. Upon graduation, Rohini will join Deloitte Consulting's Strategy & Operations practice as a Senior Consultant in Boston, MA.

Before coming to MIT, Gerta worked at IBM and Salary.com as a Professional Services Consultant. Prior to that, she graduated with a Bachelor's degree in Mathematics from Wellesley College. Gerta is interested in the intersection of technology and social impact.

KEY INSIGHTS

1. Data entry errors in shipment tracking data poses a significant risk to supply chain visibility. Monitoring error occurrences and time taken to rectify them can reveal significant areas of risk.
2. Our thesis proposes a reusable approach for data entry error detection that can be used by most freight forwarders and supply chain organizations
3. For categorical variables, which are the dominant datatype in shipment data, neural networks provide the most robust predictive models.

Introduction

Our thesis sponsor, Damco, is a freight forwarding and supply chain management service provider. One key task that they perform is the tracking of each shipment for their customer. They classify the shipment transit lifecycle across multiple milestones, such as 'Dispatch', 'Arrival at port', and 'Awaiting customs clearance'. This thesis focuses on the application of analytics to determine the probability and likely cause of data entry errors while recording the milestones associated with a shipment. Real-time tracking is essential for supply chain visibility. Some of Damco's customers request the tracking of a number of non-standard milestones. This is a manual process, especially for air shipments. These entries have an increased probability of error due to the manual intervention involved, which can result in missing updates or data entry errors. The data analyzed is from one of Damco's non-standard customers who requires tracking of multiple shipment-events in addition to the industry standard set. The repercussions of data errors can be detrimental for Damco's clients. This thesis explores some the descriptive and predictive techniques that Damco could utilize in this process. We look at two types of errors in our analysis- system errors and operational error. System errors arise from violations of business rules that are explicitly enforced by the software system. Whereas operational errors violate business rules or requirements which are defined outside the software.

Data errors can be costly, both from a human rework perspective as well as from the perspective of increased risk due to supply chain visibility loss. The results of this thesis will enable companies to focus their efforts and resources on the most promising error avoidance initiatives for shipment data entry and tracking. By using a hybrid model that utilizes both descriptive and prescriptive analysis, our thesis develops a reusable framework for data entry error detection and correction. This framework will help improve supply chain visibility, particularly for the logistics function, where the cost of missing data and data error is high.

Methodology

We used a four-phase approach. We began with data collection and cleaning in phase one. Since the primary purpose of this thesis is – analysis, which requires reading existing data, we de-normalized the data-tables to read-optimize them. Phase two is the data exploration phase, wherein we perform descriptive analysis. Using Tableau, we visualize the data and observe any patterns that relate shipment attributes to the error occurrences. Additionally, in the data exploration phase, we use the K-means clustering algorithm to cluster the data using the time impact of errors as the clustering criterion. Phases three and four are the model building and validation phases respectively. We build predictive models using the Naïve-Bayes and neural networks. Although we identified a series of alternative predictive models (including regression models) in our literature review, they are not used in our analysis. This is because the data set has a large number of categorical variables which cannot be used as predictors for these models. However, Naïve-Bayes and neural nets are two models that can use categorical variables as predictors. In the model validation phase we compare the predictive performance of the models on the validation datasets.

Descriptive Analysis

We used a three-phase approach for the descriptive analysis starting from 'Data Exploration' to 'Descriptive evidence of hypothesis' and finally 'Classification using K- Means'.

During the 'Data Exploration' phase we first assign a status to each data point. The status is one of four categories: 'Initial Entry', 'Redundant', 'Correction' or 'Update'. We devised the business rule to classify the records into one of these four based on the inputs provided by Damco. We then plotted the data to understand the distribution of the transactions based on destination countries, shipment milestone and the status of the record. We find that the maximum number of records corresponds to the events X3 (Arrived at Pickup) and AF (Pickup). Further, we see that 45.6% of the transactions are for the destination country US.

We use a hypothesis driven approach in the 'Descriptive evidence of hypothesis' phase. We looked at four broad categories of hypothesis- 'Temporal hypotheses', 'User and consignee-driven hypotheses', 'Geo-spatial hypotheses' and 'Other hypotheses about corrections and reason codes'. For the given dataset, we plotted the variation of error occurrence over time. We find that relative to the total number of transactions made, the highest number of system errors occur in the month of July. As shown in Figure 1, for operational errors, the highest number of occurrences relative to total transactions take place in the month of March. Further, for most events, these are almost equally concentrated on Mondays, Wednesdays and Fridays. The behavior is different for the events 'Cargo received from airline', 'Estimated Time of Arrival to destination airport', 'Transport Confirmation' and 'Forecasted Estimated Time of Arrival)' for Mondays where the error occurrence is significantly lower.

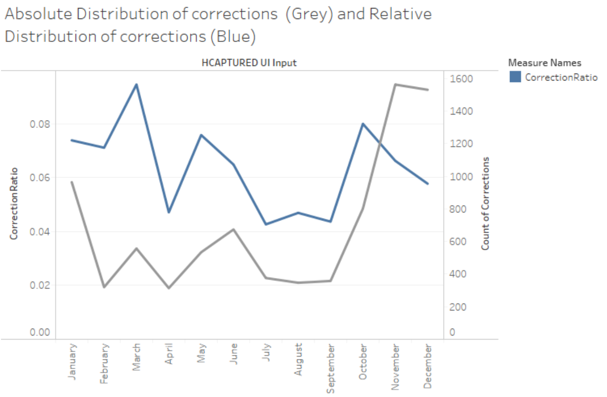
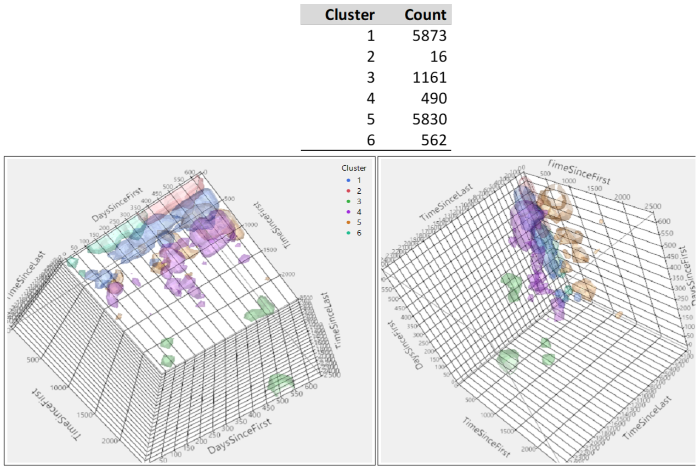


Figure 1: Absolute and relative distribution of corrections by day of month

We also find that, on average, the operational errors are resolved within 74 hours of the erroneous entry first being made. We also notice variations in time taken to correct an error based on the destination country- The country with the highest median time to correct an error is Taiwan and the lowest is UAE. An analysis of the data based on the transaction status shows that the number of changes in the 'Correction' and 'Redundant' entries, show a strong tendency of clustering around a few shipments.

In the 'Classification using K- Means' phase, we perform the clustering of the dataset using the three variables that capture the 'Time Since First version of a transaction', 'Time Since Last version of the transaction' and a calculated field 'Days Since First entry in the dataset'. The version of transaction here refers to a shipment milestone and shipment waybill number pair. K-means is a non-hierarchical clustering method that aims to find k records in the training dataset that are similar to a new record. The similarity is computed using distance between points. Figure 2 shows the results of the clustering procedure.



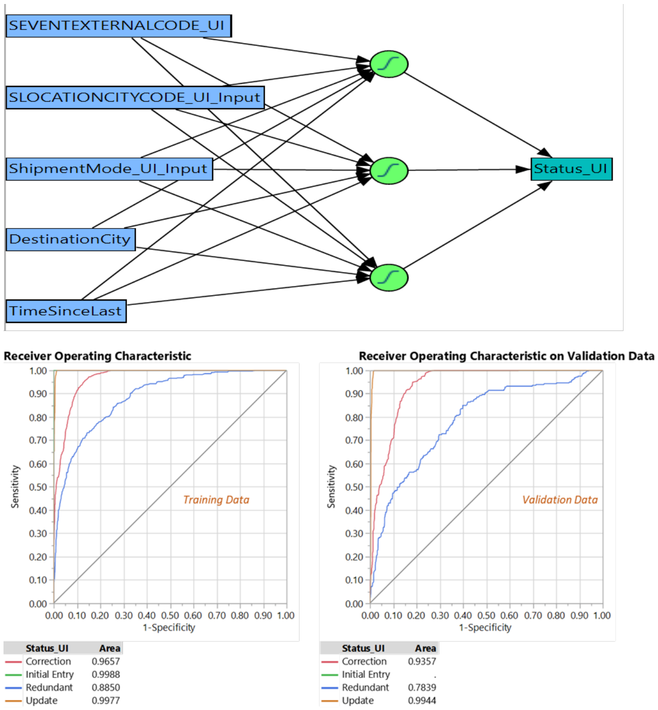
Cluster	Count
1	5873
2	16
3	1161
4	490
5	5830
6	562

Figure 2: Scatterplot of 6 (largely indistinguishable) clusters created using K-Means

Predictive Analysis

We used the Naïve- Bayes classifier and neural networks to build predictive models. These modeling techniques can take predictors that are categorical in nature. We built four Naïve-Bayes models- one using Event-code, Event-location, User, shipment mode and a flag indicating whether there is a system error to predict whether a record is the final entry. The predictive power of this model, however, is inferior to a naïve approach. Similar results were obtained for other naïve-Bayes predictive models, with the exception of the model to predict the status of a record using as Event code, Mode and Destination city predictors. This last model has a misclassification rate of 6.505% as compared to the misclassification of 6.506% for a naïve approach.

We built two sets of neural network models – one to predict whether the record has an error and the other to predict the status of the record. The best predictive performance was observed for the neural net model that predicts the status of a record using Event-code, Event-city, Time Since Last edit, Shipment-Mode and Destination- City. The neural net model and its Receiver output characteristics (ROC) curve that shows model performance are shown in Figure 3.



Status_UI	Area
Correction	0.9657
Initial Entry	0.9988
Redundant	0.8850
Update	0.9977

Status_UI	Area
Correction	0.9357
Initial Entry	.
Redundant	0.7839
Update	0.9944

Figure 3: Neural net model and ROC Curve- Correction entries have an Area under the curve of over 93% for validation dataset

Conclusions

The K-means clustering method does not reveal any meaningful insights about the attributes of the records, and the predictive performance of the Naïve-Bayes and the neural net models is largely unsatisfactory in predicting whether a record is final. However, for our specific data set, and operating under computational limitations, we did not find a Naïve-Bayes model which significantly out-performs a naïve approach to classification (to the largest class). Despite this result, the Naïve-Bayes approach is promising for our data set given the categorical nature of the variables. The predictive performance of the model is expected to improve with the inclusion of more predictor variables.

The neural net performs better than the Naïve approach. The Naïve model misclassifies 39% of the records, whereas our model misclassifies 8% of the training data and 11% of the validation data.

One of the main constraints we encountered is the data’s time span, which is less than a year and thus limits the potential inferences related to seasonality or for a high degree of confidence in the error forecasts. Likewise, the data does not identify entries as correct or incorrect, and does not indicate the applicable corrections for the incorrect entries. Further, the data integration is difficult as it is extracted from several differing legacy systems.

Despite these limitations, Damco could leverage these insights in ways ranging from redistributing its resources towards paying closer attention to certain events, to addressing the sources of errors correlated with specific users or events. In this way, Damco can improve in cost, time, and quality of service to its clients.