MIT Center for Transportation & Logistics

Analysis of Inefficiencies in Shipment Data Handling

Image Source: [Hamburg Port]

Table of Content

Image Source: MSC Eleni

Background and Objectives

Image Source: Safe Bee

The Sponsor

Our sponsor company is one of the world's leading providers of freight forwarding and supply chain management services.

For more than 100 years, they have been providing their customers with transportation and logistics solutions that support the way they want to do business, wherever they are in the world.

Their Global footprint and market leadership in several geographies enables them to offer their customers- new sourcing areas, customers and business opportunities with their established network

Their customer base (for this thesis scope) is split into SCM customers and Freight forwarding customers.

The Problem | Errors in Shipment Milestone Tracking

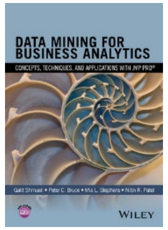
Shipment Milestones
- Key events along the journey of a shipment
- Industry standard: ~8 Milestones per shipment

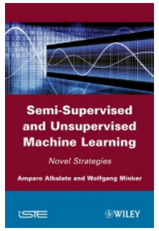
Customer Requirement
- ~18 Shipment Milestones

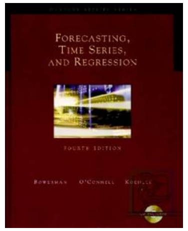
Data Errors
- System Errors
- Operational Errors

Thesis Problem Statement
- Relationship between shipment attributes and errors
- Predicting occurrences of errors in shipment data

Literature Review







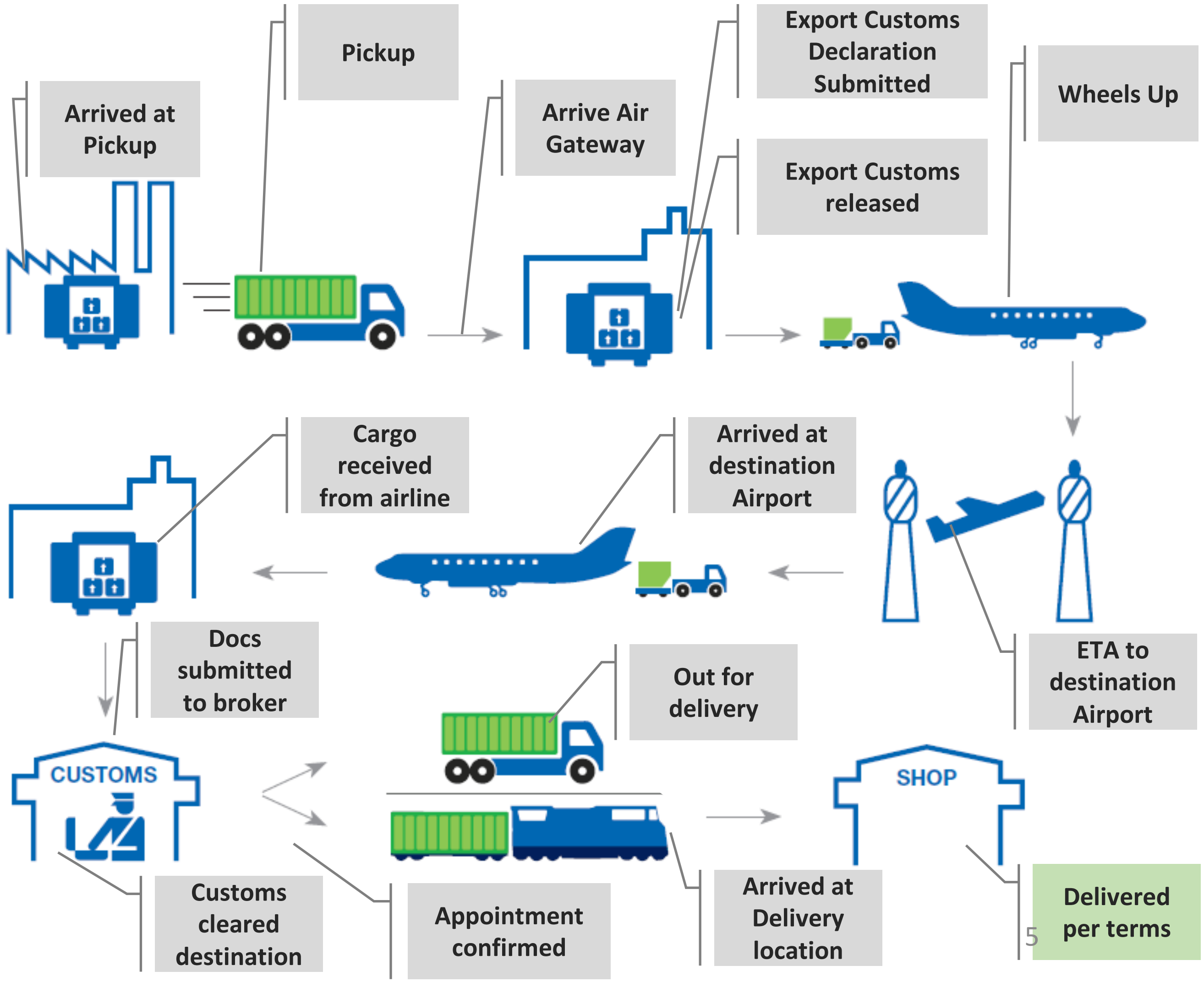


Image Source: Sponsor Company Webcontent

MIT Center for Transportation & Logistics

Thesis Methodology

Data Collection → Data Exploration → Design and Build → Model Validation

Image Source: Express World

Methodology

Data Collection → Data Exploration → Design and Build → Model Validation

Data Sources	
• Transactional Data from Legacy system • System Logs	• De-normalized Data structure • Data types

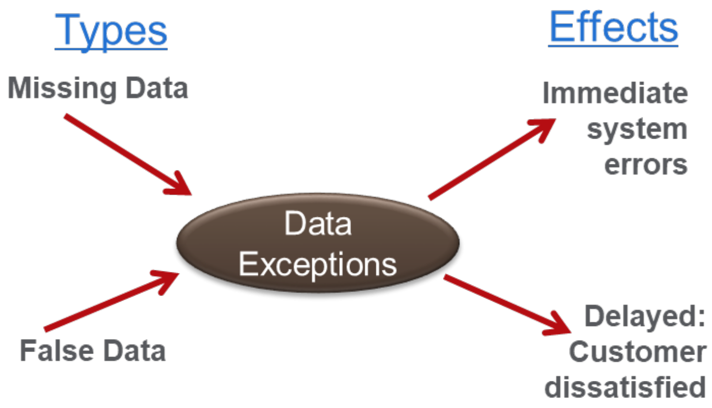
Data Preparation	
• Data Cleaning • Current exceptions identification	• Data error-correction mapping • Dimensionality reduction

Data Cleaning

- Null Checks and blank fields
- Type casting
- Calculated fields

Current exceptions identification

- Error classification- system and operational



Data error-correction mapping

- Foreign Key relationships
- Status
- Edited Fields

Dimensionality reduction

- Drop fields
- Assess applicability of Principal component analysis

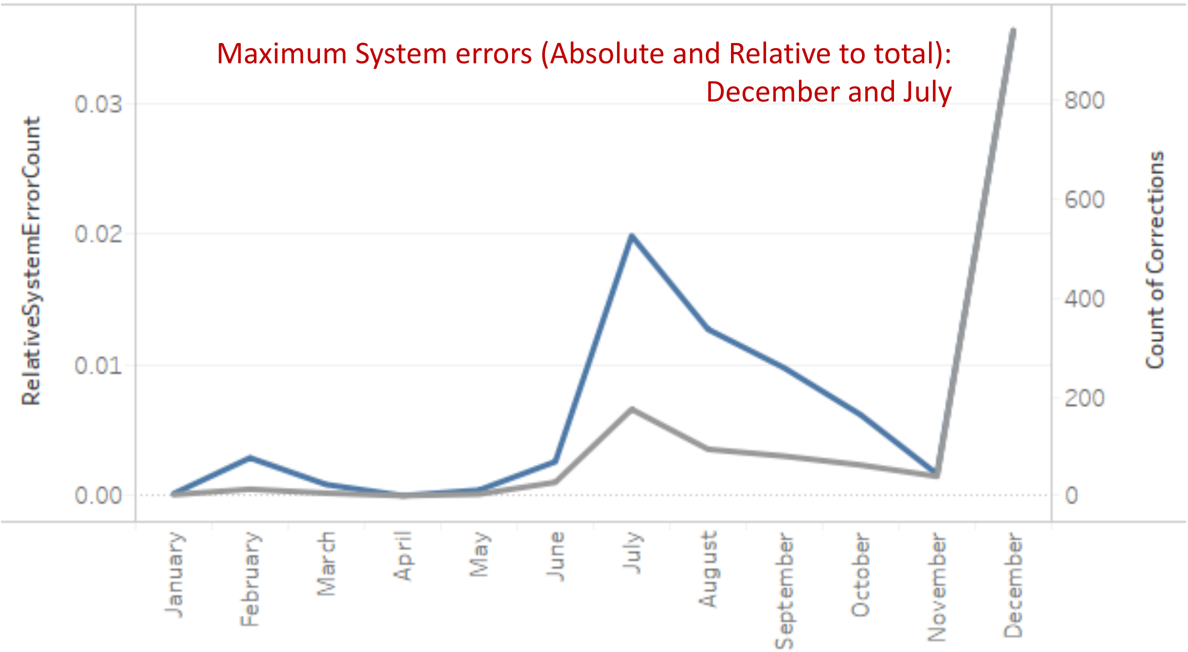
Methodology

Data Collection → **Data Exploration** → Design and Build → Model Validation

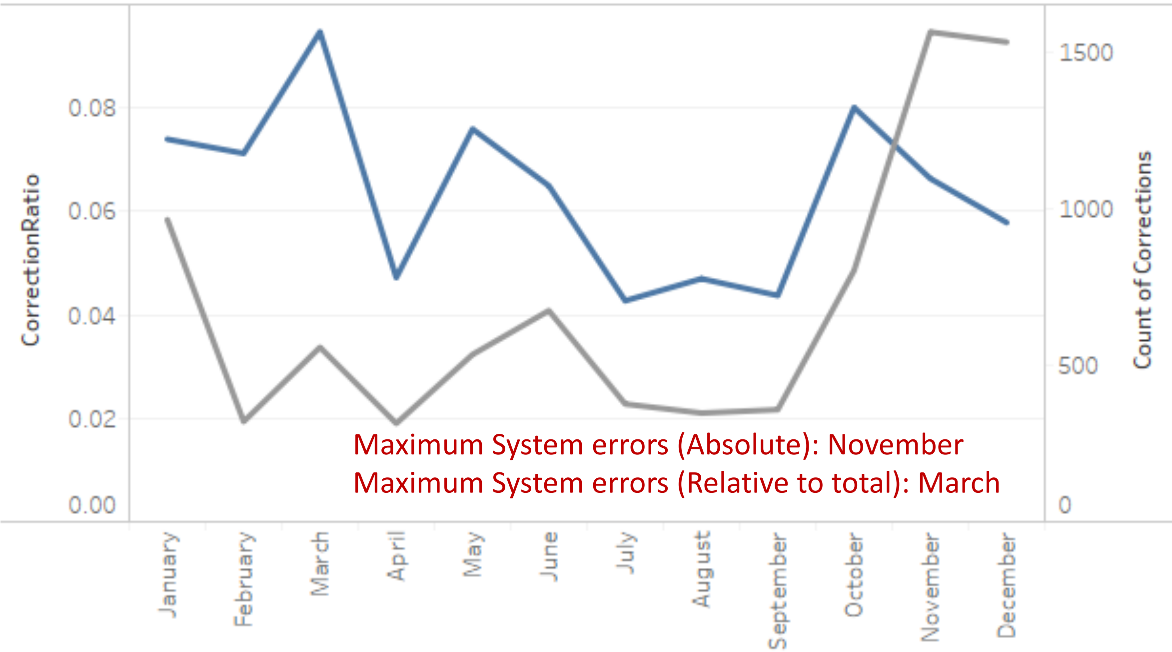
A. Descriptive evidence of hypothesis

- Temporal hypotheses
- User- and consignee-driven hypotheses
- Geo-spatial hypotheses
- Others

Absolute Distribution of System Errors (Grey) and Relative Distribution of System Errors (Blue)



Absolute Distribution of corrections (Grey) and Relative Distribution of corrections (Blue)



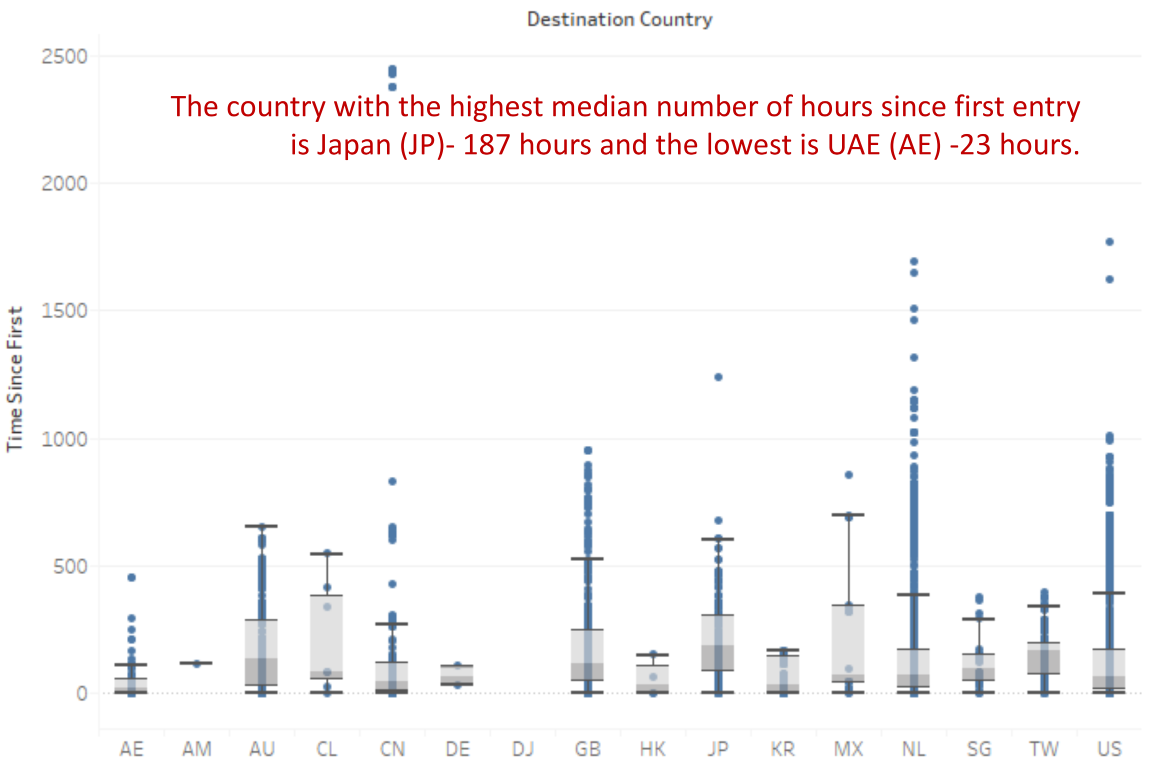
System error volumes follow the transactional volume pattern. Operational Errors (and hence inefficiencies) don't.

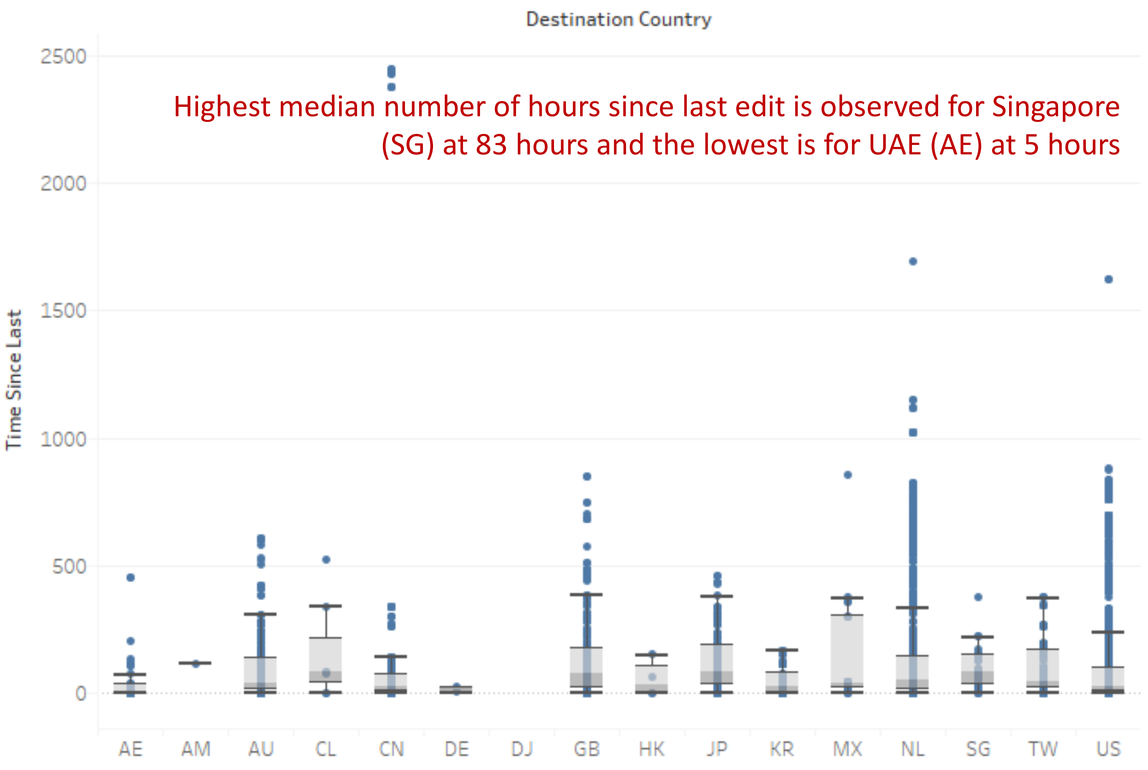
MIT Center for Transportation & Logistics

Methodology

A. Descriptive evidence of hypothesis

- Temporal hypotheses
- User- and consignee-driven hypotheses
- Geo-spatial hypotheses
- Others





Errors are addressed and corrected at different rates for shipments destined for different countries.

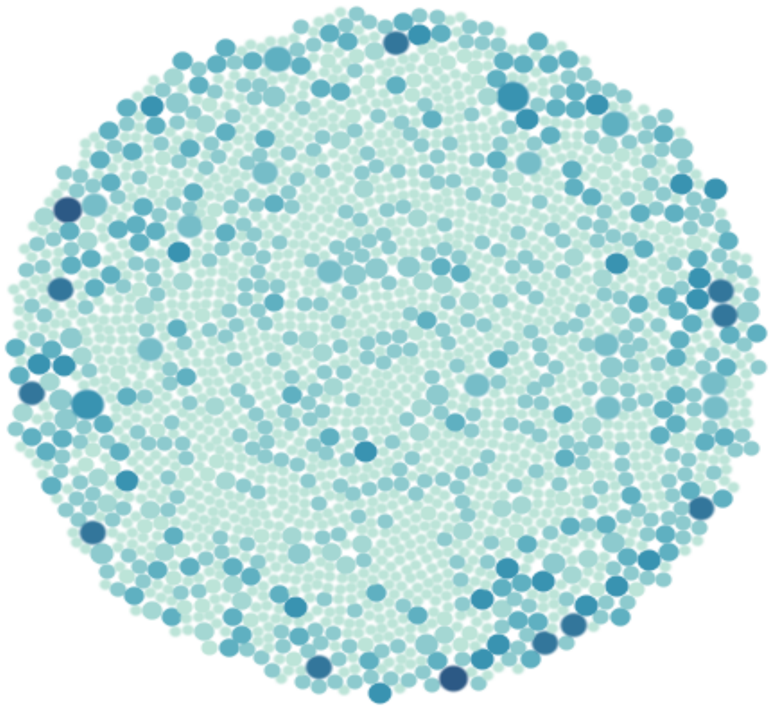
Methodology

Data Collection | Data Exploration | Design and Build | Model Validation

A. Descriptive evidence of hypothesis

- Temporal hypotheses
- User- and consignee-driven hypotheses
- Geo-spatial hypotheses
- **Others**

Update Entries per Shipment- color is Relative #Edits, Size Total no of updates



- Darker Color: Higher ratio of Transactions/Unique Shipment Milestone
- Larger radius: Greater number of transactions
- All waybills in the entire dataset.
- Problem: Small dark blue dots

"**Initial Entry**": Exactly one entry corresponding to the status "Initial entry" for each unique waybill-milestone pair.

"**Correction**": Several dark spots capturing shipments where the same event is corrected five to six times

"**Redundant**": Few dark spots capturing shipments where the same event is corrected five to six times

"**Update**": Several dark blue points with the same event updated for the same shipment up to 5 or 6 times. Not a problem.

For the 'Correction' and 'Redundant', number of corrections concentrated around a few shipments

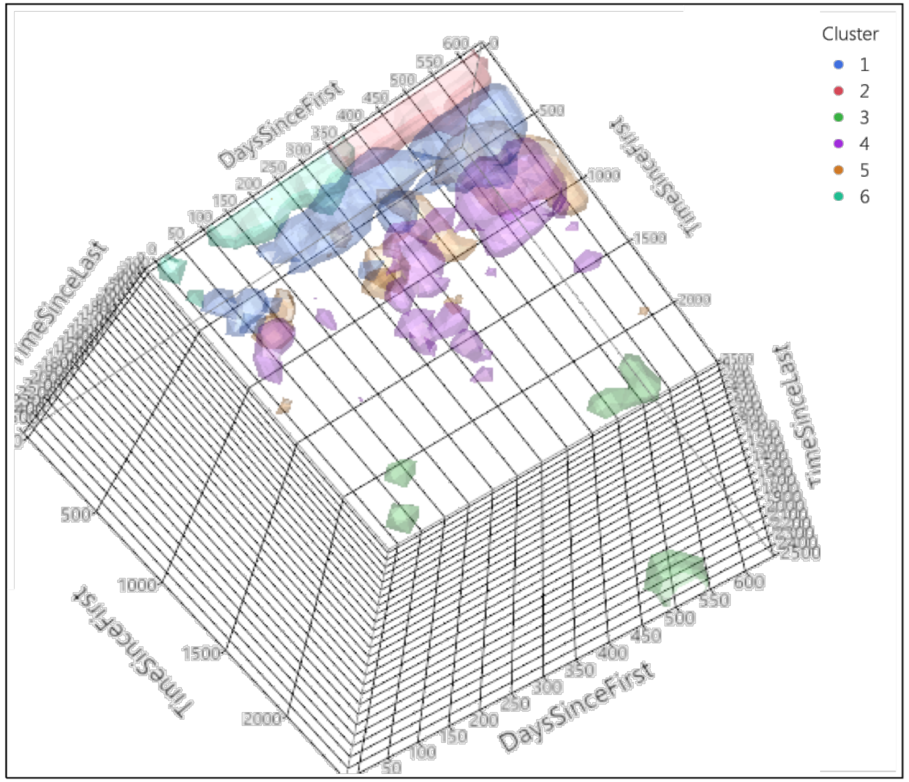
Methodology

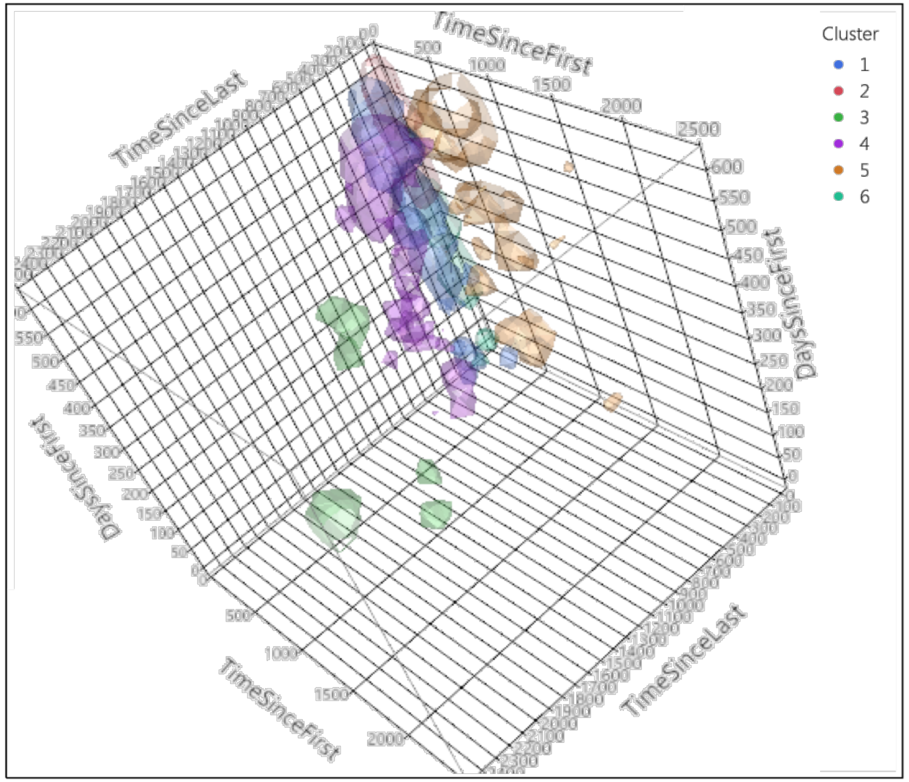
Data Collection | **Data Exploration** | Design and Build | Model Validation

B. Classification using K- Means

- K: 6 Clusters
- Y: TimeSinceFirst, TimeSinceLast, DaysSinceFirst
- Similarity : Distance between points.
- Better suited for use with larger data tables
- Limitation: Only supports numeric columns

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





2 clusters (green, brown) are distinct from the other clusters with little overlap. The green cluster corresponds to records with high value of 'TimeSinceLast' and 'TimeSinceFirst' and brown for low values of the same.

Methodology

Data Collection | Data Exploration | Design and Build | Model Validation

1

Naïve Bayes Classifier

Model Selection
Predictors and Response
Predictive Performance
Conclusion

Model Selection:

- Categorical predictors
- Answers the question - 'What is the propensity of belonging to a particular class?'
- For a new predictor value, it will classify the probability of the predictor occurring as zero

Predictors and Response:

- Model to Predict 'Status' using Event code, Mode and Destination city

Predictive Performance:

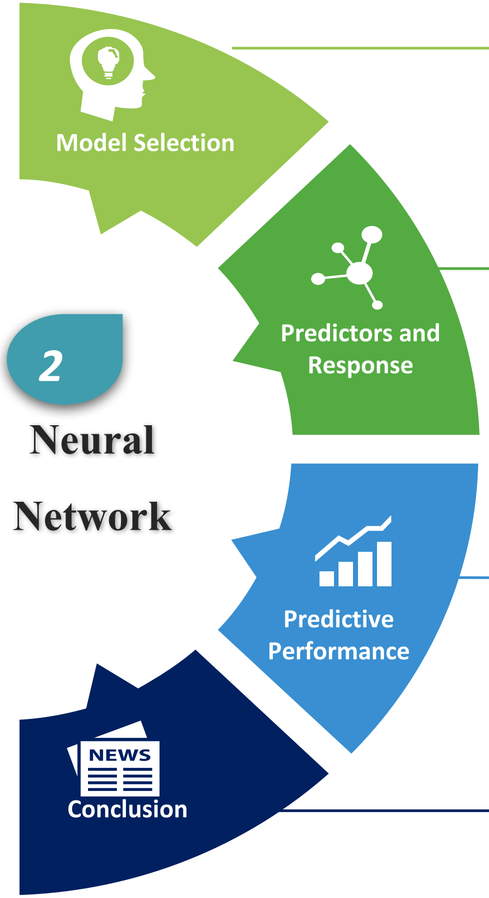
- Model has a marginally better misclassification rate (6.505%) than the naïve approach (6.506%).

Conclusion:

- Performance unsatisfactory for the constructed models.
- But the approach is promising – especially for categorical predictors
- Predictive performance expected to improve with the inclusion of more predictor variables

Methodology

Data Collection → Data Exploration → Design and Build → Model Validation



Model Selection:

- Categorical predictors
- One-third data: validation dataset
- Tendency to over-fit the data

Predictors and Response:

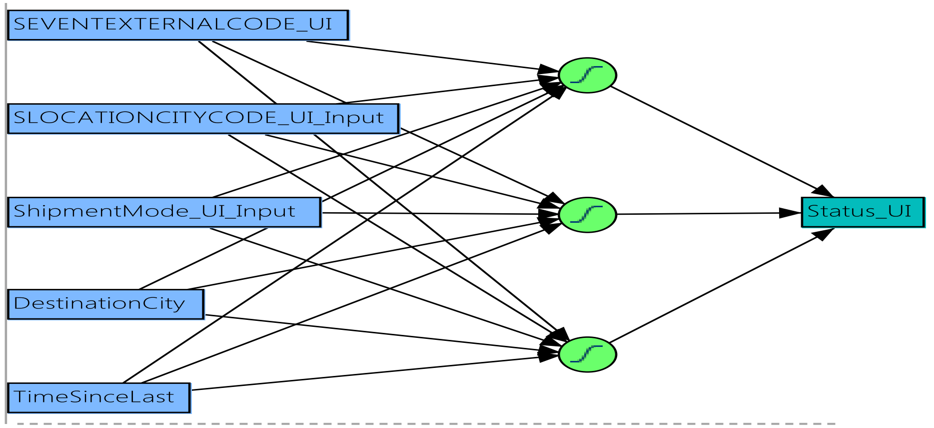
- Predict 'Status' using Event-code, Event-city, 'TimeSinceLast', Shipment-Mode and Destination-City

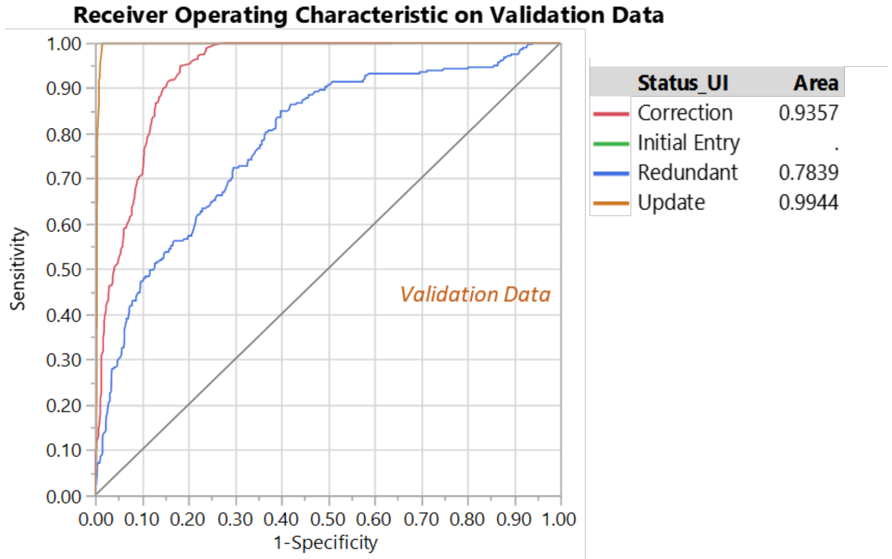
Predictive Performance:

- Satisfactory goodness-of-fit: Generalized RSquare 88%
- Good prediction accuracy: 8.7% Misclassification rate

Conclusion:

- ROC curve is close to the top-left with high Area under the curve: 96% training data AUC, 93% validation data AUC





Results and Discussion

Image Source: [Smart-Cities]

Results and Discussion

System errors

- Frequency of System errors by **Month**
    - Absolute maximum: December
    - Relative maximum: July
- Frequency of System errors by **Day of Week**:
    - Maximum: Monday. Followed by Wednesday and Friday
- Frequency of System errors by **Shipment Milestones**:
    - Maximum- 'Arrived at destination airport'

Operational errors

- Most operational errors on Mondays
- Most frequent events with Errors:

Sundays	Mondays	Rest
Pick-up	Container on Board	Delivery Appt. or Appt. Confirmed

- Month with Absolute Max: November
- Month with Relative Max: March
- 74 hours (median) to correct the operational errors
- Time is maximum for 'Arrived at Destination Hub' – 448 hours and minimum for 'Cargo received from airline' – 15 hours
- Delays driven by 'Late delivery due to Customer request'

Models

- **Naïve Bayes**: Feasible approach for categorical predictors
    - Performed no better than a Naïve approach
    - Performance expected to improve with addition of predictors
- **Neural Network**: supports categorical predictors
    - ✓ Goodness of fit
    - ✓ Predictive performance
    - Risk of overfitting
- **Conclusion**: Neural net model with predictors- (Event-code, Event-city, 'TimeSinceLast', Shipment-Mode and Destination- City) can predict Status of the record.

MIT Center for Transportation & Logistics

Limitations and Future Roadmap

Image Source: *[Travel and Leisure]*

Limitations and Future Roadmap

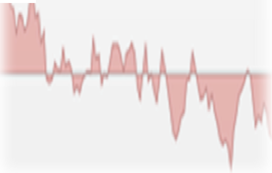
Limitations

Data

- Type
- Duration

Impacted Fields

- Time-Stamp
- City Codes



Business Rules

- Prioritization approach
- Shipment itinerary

Short term steps

System Errors

- Additional data for Root cause analysis:
  - System response rate
  - Performance
  - Geographical reasons
  - Outages

Migrate from Legacy System

- Data Bottleneck
- Process Bottleneck

Predictive Performance

- Numerical Data
- Stratified sampling approach
- Overcoming computational limitations for Naïve Bayes

Long term steps

Cloud and Big Data Enablement

- Prevention vs reaction to errors
- Data Triangulation

Reusable Methodology

- Results may be limited but the approach is extendable


Errors using inadequate data are much less than those using no data at all.
Charles Babbage

Thank You

- Rohini Prasad
- Gerta Malaj

MIT SCM 2017

Special Thanks to our
Thesis Advisor:

- Matthias Winkenbach