MIT SCM RESEARCH FEST

May 25, 2017

Predicting On-time Delivery in the Trucking Industry

Authors: Rafael Duarte Alcoba, Kenneth W. Ohlund
Advisor: Matthias Winkenbach

Massachusetts Institute of Technology

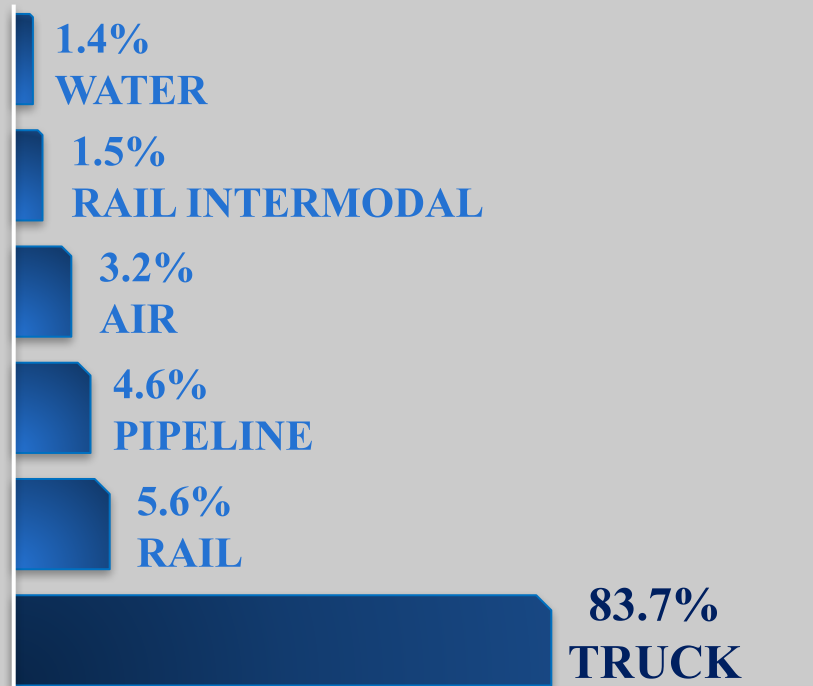
MIT Supply Chain MANAGEMENT

Agenda

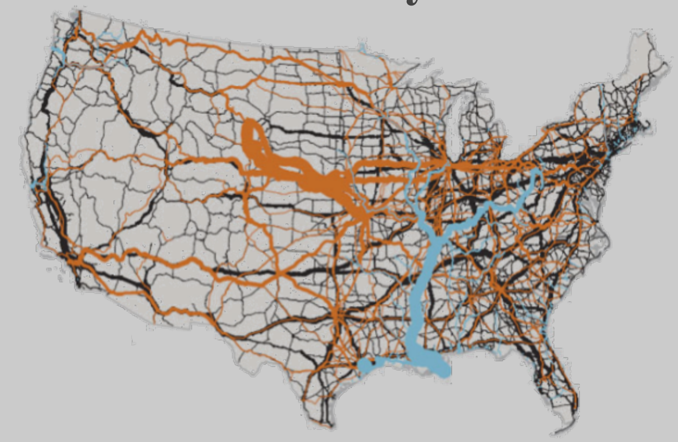
- Motivation
- Methodology
- Results
- Conclusion

The US trucking industry…

Dominates the commercial transportation industry with 83.7% of the revenue



Connects the entire US territory



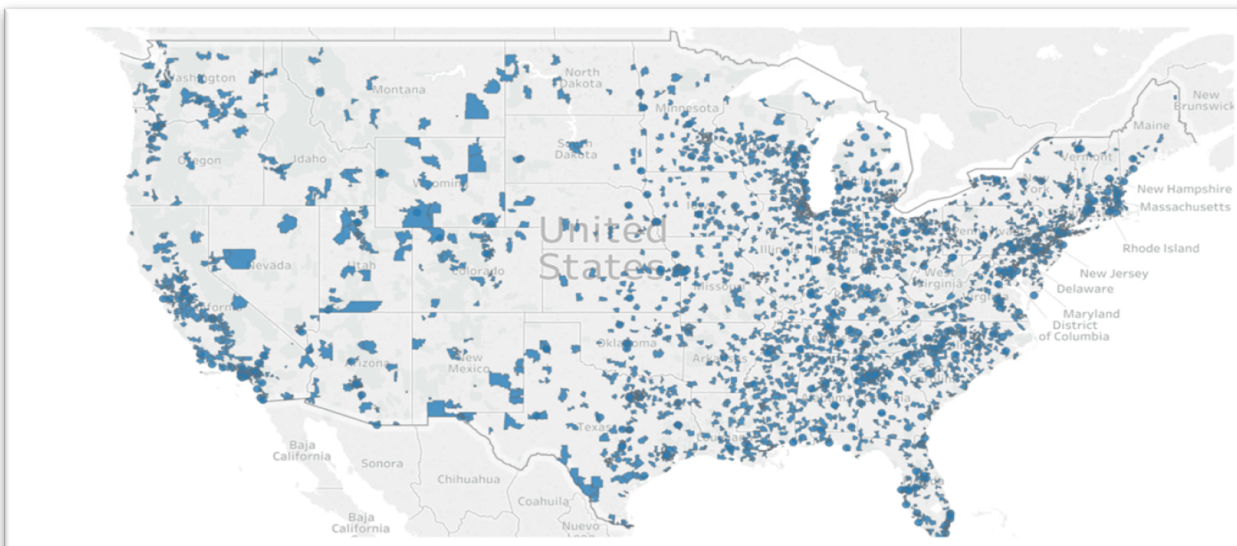
Is expected to grow **21%** over the next 10 years

Research Questions

- How can companies engaged in logistics optimize resources while improving customer service levels?
- Can on-time delivery in trucking be predicted?
- Can a predictive analytics model indicate which combinations of variables lead to delays?

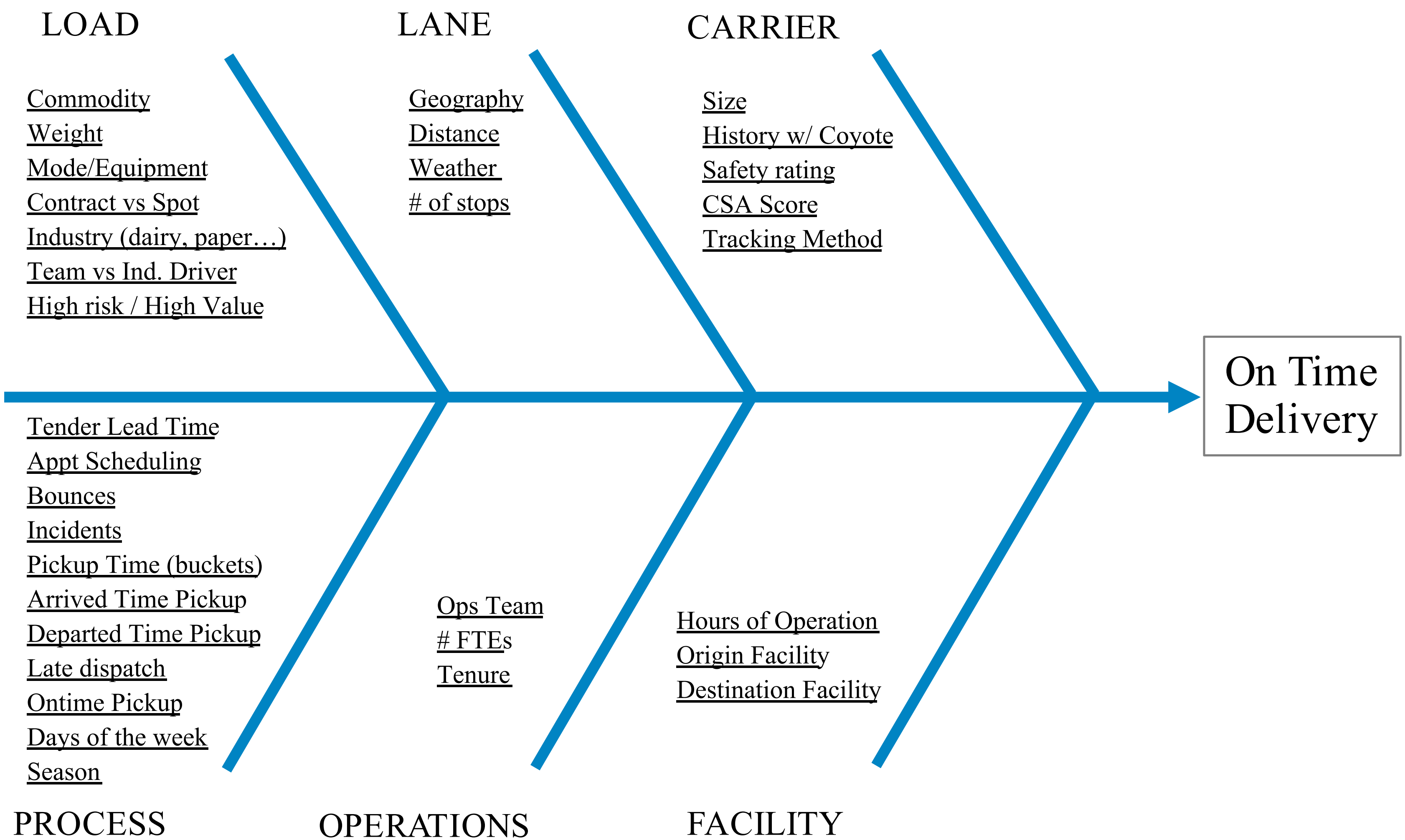
Gathering Data

- Loads within the United States (more than 6,000 locations)
- Restricted to FTL (full truckload)
- Data from October 1, 2014 to September 30, 2016
- Binary decision variable for on-time delivery (0 = delayed; 1= on-time)



Fishbone Diagram

Variables Potentially Affecting On-time Delivery



Sampling & Partitioning

On-Time Delivery

Data | Model

Imbalanced

95% on-time
5% delayed

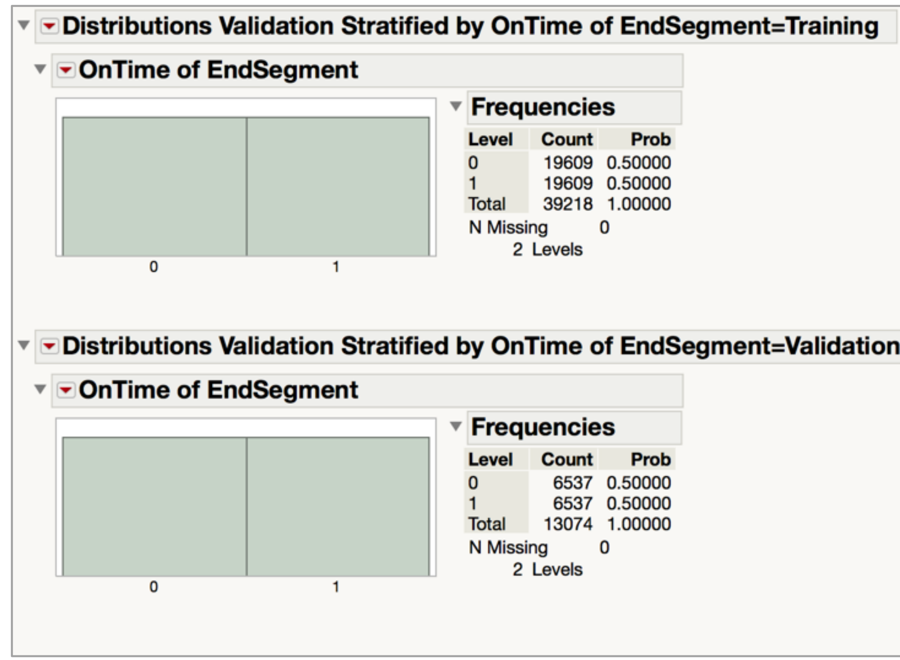
Undersampling

50% on-time
50% delayed

Overfitting

Rule of Thumb

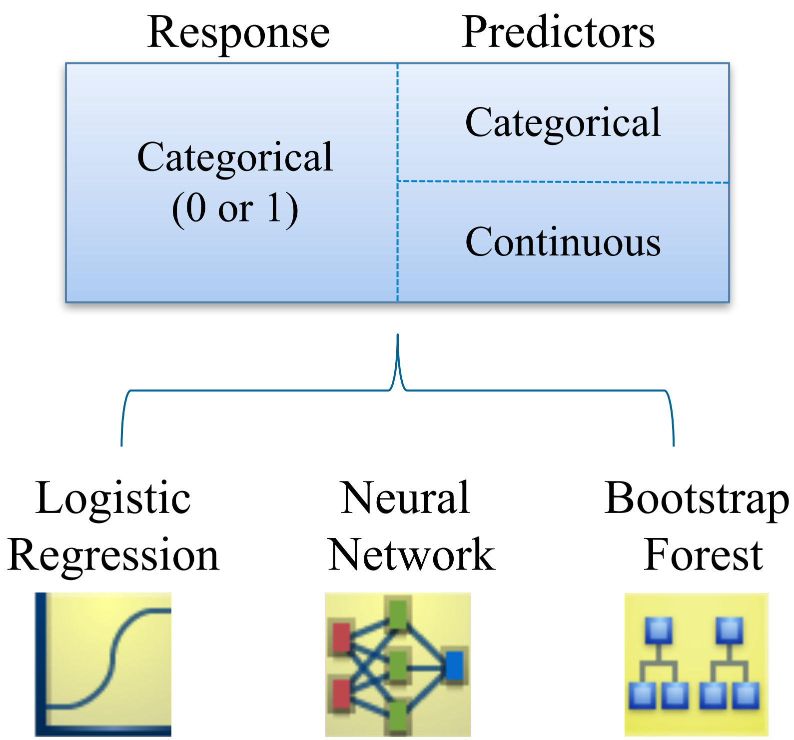
75% Training
25% Validation



Level	Count	Prob
0	19609	0.50000
1	19609	0.50000
Total	39218	1.00000

Level	Count	Prob
0	6537	0.50000
1	6537	0.50000
Total	13074	1.00000

Model Selection



- **Goal: find an explanatory model with high interpretability**
- **Main model: LR**
- **Assess Performance: NN and BF**

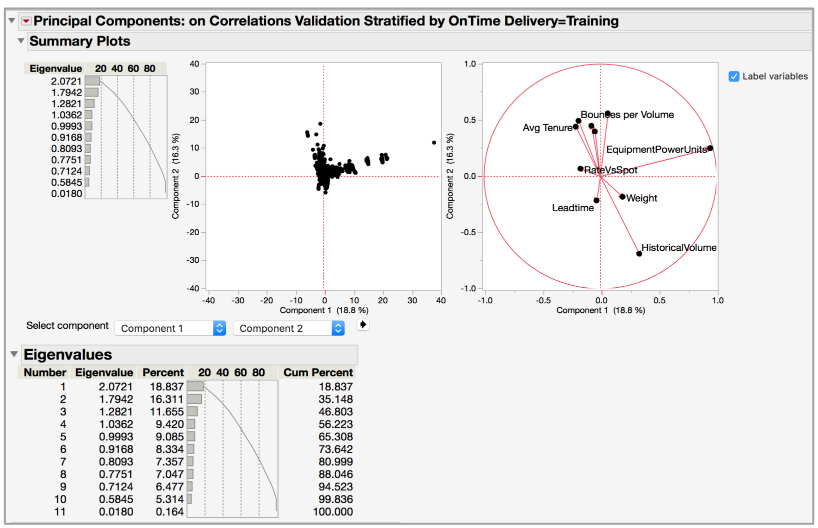
Variable Selection

Multi-Collinearity

- Correlation Matrix

Correlations	Contract-Spot	Duration at StartSegment	Historical Volume	Incidents per Volume	OnTime of StartSegment	FacilityType Appt of EndSegment
Contract-Spot	1.00	-0.07	0.13	0.26	0.05	0.07
Duration at StartSegment	-0.07	1.00	-0.06	-0.01	0.02	0.03
HistoricalVolume	0.13	-0.06	1.00	-0.23	-0.34	-0.09
Incidents per Volume	0.26	-0.01	-0.23	1.00	0.09	0.02
OnTime of StartSegment	0.05	0.02	-0.34	0.09	1.00	0.03
FacilityType Appt of EndSegment	0.07	0.03	-0.09	0.02	0.03	1.00

- PCA / MCA



Stepwise Regression Output

- Standard forward search
- Starts from an empty model
- At each step the model selects a variable that increases maximum likelihood fit.

Effect Summary

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

$$LogWorth = -\log_{10}(p - value)$$

Performance Evaluation

Build models using six explanatory variables with statistical significance

- Inc. per Volume
- Historical Volume
- Cont. (0) vs Spot (1)
- Facility Type Appt
- Duration Start Seg.
- On-time Start Seg.

Confusion Matrix to assess the predictive power of the models

	Predicted Class	
Actual Class	C_0	C_1
C_0	$n_{0,0}$ = number of C_0 cases classified correctly	$n_{0,1}$ = number of C_0 cases classified incorrectly as C_1
C_1	$n_{1,0}$ = number of C_1 cases classified incorrectly as C_0	$n_{1,1}$ = number of C_1 cases classified correctly

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

$$missed\ delays = \frac{n_{0,1}}{n}$$

Predictive Performance (Validation dataset)

Main model: LR

LOGISTIC REGRESSION

Actual		Predicted 0	Predicted 1	Σ
	0	219	209	**429**
	1	1,805	6,337	**8,142**
	Σ	**2024**	**6546**	**8570**
err = $(n_{0,1} + n_{1,0})/n$				23.50%
missed delays = $n_{0,1}/n$				2.44%

- Model interpretations vs "Black Box" approach
- High visibility of the predictors
- Robust results

Assess Performance: NN and BF

NEURAL NETWORK

Actual		Predicted 0	Predicted 1	Σ
	0	254	175	**429**
	1	2,075	6,067	**8,142**
	Σ	**2329**	**6241**	**8570**
err = $(n_{0,1} + n_{1,0})/n$				26.25%
missed delays = $n_{0,1}/n$				2.04%

BOOTSTRAP FOREST

Actual		Predicted 0	Predicted 1	Σ
	0	243	186	**429**
	1	2,058	6,084	**8,142**
	Σ	**2301**	**6270**	**8570**
err = $(n_{0,1} + n_{1,0})/n$				26.18%
missed delays = $n_{0,1}/n$				2.17%

Predictive Performance (Testing dataset)

New dataset to to gauge model's accuracy and robustness

- Validation

Actual \ Predicted	0	1	Σ
0	219	209	**429**
1	1,805	6,337	**8,142**
Σ	**2024**	**6546**	**8570**
err = $(n_{0,1} + n_{1,0})/n$			23.50%
missed delays = $n_{0,1}/n$			2.44%

Actual \ Predicted	0	1
0	2.6%	2.4%
1	21.1%	73.9%

- Test

Actual \ Predicted	0	1	Σ
0	23	50	**73**
1	452	1,448	**1,900**
Σ	**475**	**1498**	**1973**
err = $(n_{0,1} + n_{1,0})/n$			25.44%
missed delays = $n_{0,1}/n$			2.53%

Actual \ Predicted	0	1
0	1.2%	2.5%
1	22.9%	73.4%

Application - Results

Using model results to prioritize loads requiring attention

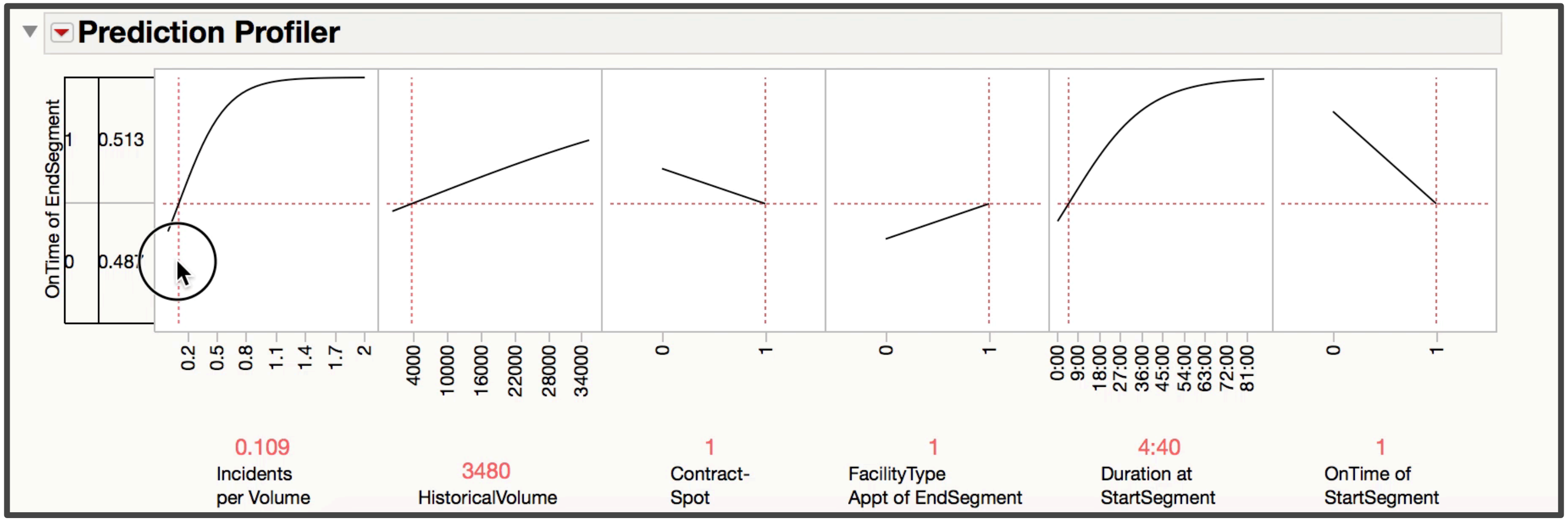
Critical Loads - Dashboard

Import Load
Sort Prob.

LoadStopID of StartSegment	LoadStopID of EndSegment	Contract-Spot	Duration at StartSegment	Historical Volume	Incidents per Volume	OnTime of StartSegment	FacilityType Appt of EndSegment	Prob [On-time]
XXX1	YYY1	1	14:39	111	0.12	0	1	10%
XXX2	YYY2	1	0:08	2011	0.07	0	1	20%
XXX3	YYY3	1	16:55	1010	0.08	1	1	42%
XXX4	YYY4	1	5:30	1349	0.07	1	1	57%
XXX5	YYY5	1	1:30	654	0.03	1	1	66%
XXX6	YYY6	1	2:30	1077	0.06	1	0	74%
XXX7	YYY7	1	1:40	6	0.00	1	0	80%
XXX8	YYY8	1	0:15	4	0.00	1	0	81%
XXX9	YYY9	1	0:01	85	0.00	1	0	82%

Application - Results

Using model results to drive actions



Inc. per Volume	Historical Volume	Cont. (0) vs Spot (1)	Facility Type Appt	Duration Start Seg.	On-time Start Seg.

known when the load is tendered | known after pick-up

Conclusion

1. **Resources can be optimized using the Logistic Regression Model**
2. **On - time delivery can be predicted**
3. **Using a combination of six variables with high statistical significance can deliver predictive power**

Prediction

Actual		
	2.6%	2.4%
	21.1%	73.9%

- **Tracking 23.7% of the loads**
- **Missing only 2.4% of loads that will be late**

Conclusion

Trade-off: Resource Reduction vs Missing Error

CUTT OFF = 0.5

Prediction / Actual

2.6%	2.4%
21.1%	73.9%

- Tracking 23.7%
- Missing 2.4%

CUTT OFF = 0.6

Prediction / Actual

4.1%	0.9%
53.8%	41.2%

- Tracking 57.9%
- Missing 0.9%

CUTT OFF = 0.7

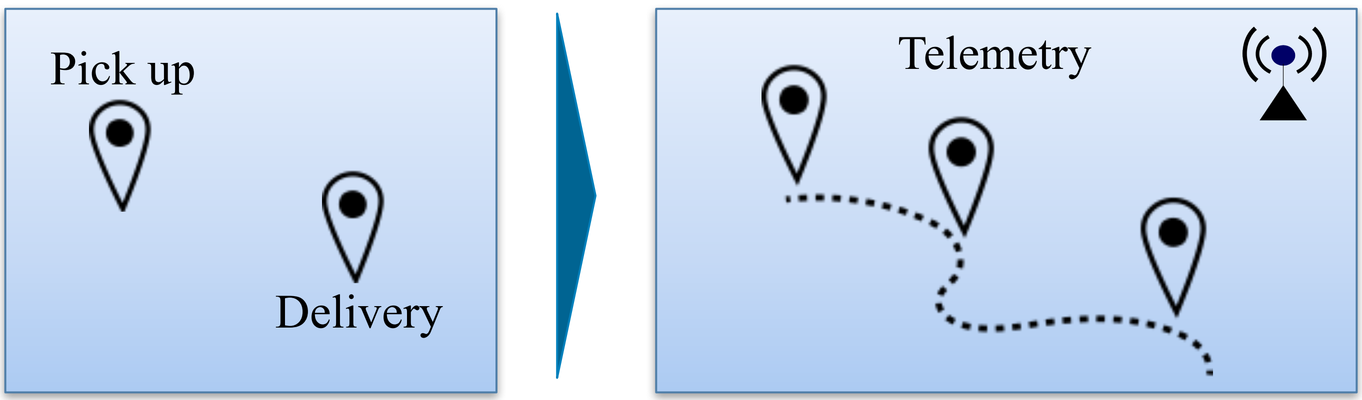
Prediction / Actual

4.9%	0.1%
86.6%	8.4%

- Tracking 91.5%
- Missing 0.1%

Suggestion for Future Research

- Increased availability of online information through new technologies
- Readiness to store records on remote servers using (cloud servers)
- Predictive model able to capture information from online records could bring new insights and complement the analysis presented in this study



Q&A

backup slides

Variables

Build models using six explanatory variables with statistical significance

	Inc. per Volume	Historical Volume	Cont. (0) vs Spot (1)	Facility Type Appt	Duration Start Seg.	On-time Start Seg.
Prob. [0]	+	+	0 +	1 +	+	0 +
Prob. [1]	−	−	1 +	0 +	−	1 +

Reweighted Confusion Matrix

	Reweighting	
	0	1
Original Data	5%	95%
Undersampling	50%	50%
	10.00	0.53

LOG REGRESSION

Original Confusion Matrix

Actual \ Predicted	0	1	Σ
0	2,192	2,093	**4,285**
1	950	3,335	**4,285**
Σ	**3142**	**5428**	**8570**
err = $(n_{0,1} + n_{1,0})/n$			35.51%
Err for predicting 1 and actual = 0			24.42%

Divided by 10

Divided by 0.53

Reweighted Confusion Matrix

Actual \ Predicted	0	1	Σ
0	219	209	**429**
1	1,805	6,337	**8,142**
Σ	**2024**	**6546**	**8570**
err = $(n_{0,1} + n_{1,0})/n$			23.50%
Err for predicting 1 and actual = 0			2.44%

of observations

Total Sample: 522,920

Excl. Outliers or missing values: 342,800

Undersampling: 34,280

Validation: 8570

LOGISTIC REGRESSION

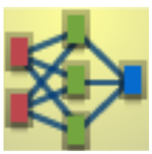
		Predicted		
		0	1	Σ
Actual	0	219	209	**429**
	1	1,805	6,337	**8,142**
	Σ	**2024**	**6546**	**8570**
err = $(n_{0,1} + n_{1,0})/n$				23.50%
missed delays = $n_{0,1}/n$				2.44%

Predictive Models

Logistic Regression

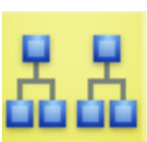
- **Simply saying, it works with the same ideas as linear regression, but for a categorical output.**
- **Relies on mathematical equation relating predictors with the outcome.**

Neural Network



- **Machine Learning technique. It mimics the activity in the brain, where neurons are interconnected and learn from experience**

Bootstrap Forest



- **Variation of Random Forests. It combines results from multiple trees to improve predictive power**