Learning from Route Plan Deviations from Last-Mile Delivery

Author: William Phillips & Yiyao Li Advisor: Dr. Matthias Winkenbach Sponsor: CTL

MIT Center for Transportation & Logistics

Massachusetts Institute of Technology

AGENDA

INTRODUCTION	What route plan deviation is and why it matters
DATA	Ask the driver through data
METHODOLOGY	Getting the tools ready
RESULTS	Here are the model
CONCLUSIONS	So what…
Q&A	Love to hear questions

INTRODUCTION

- 1 mile of reduction in average route distance results in **$50M** of annual cost savings for UPS in US only
- Urbanization and new costumers demands are making last-mile delivery optimization **increasingly complex and relevant** to retail companies
- Lacking tools and/or capabilities to include costumer specific or environmental constrains such us:
  - **Time windows (implicit or explicit)**
  - **Congestion patterns**
- Even for companies wiling to make **capital investment**, if the driver failed to follow the plan it forfeits the investment
- Drivers stated preference is studied but not **revealed preference**, so this project is actually asking the driver through data

Figure 1: Planned Route

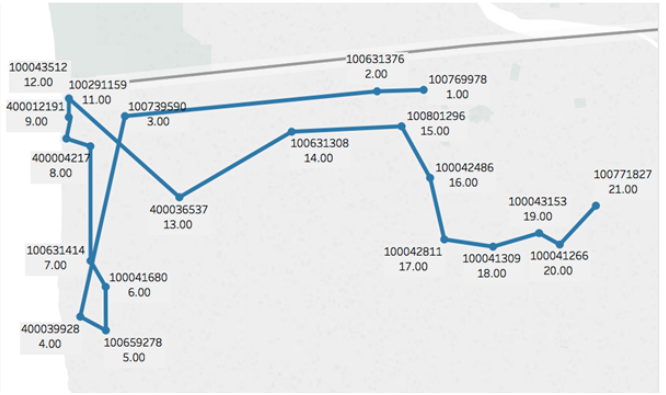
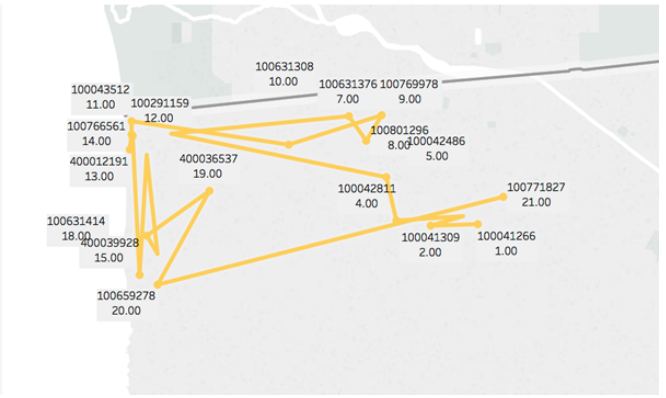


Figure 2: Actual Route



DATA

- Data Description:
  - Information about the **route instances**
  - Information about the **stops**

- Measuring deviation:
  - **Deviation**
  - **Sequence Deviation = Arcs not followed by driver / Total Arcs**
  - **Distance Deviation = Actual Distance / Planned Distance – 1**
  - **SLD Deviation = Actual Sequence SLD* / Planned Sequence SLD – 1**

    ** SLD: Straight Line Distance*

Figure 1: Planned Route

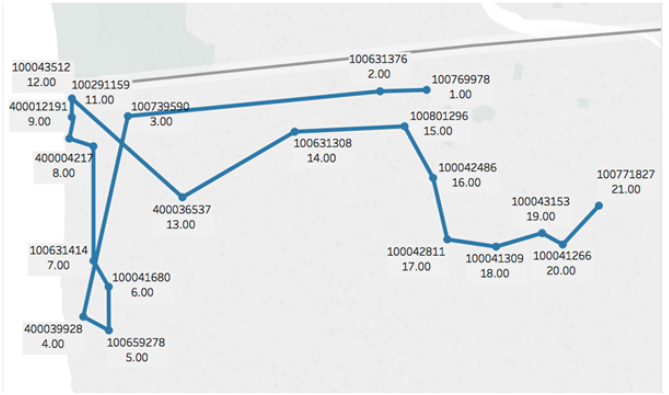
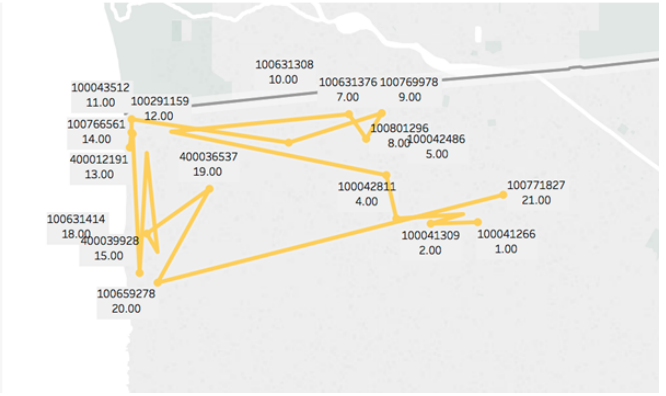
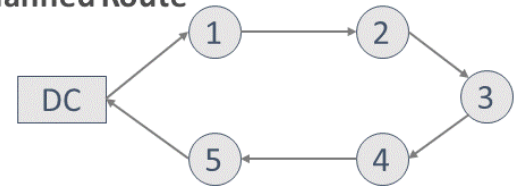


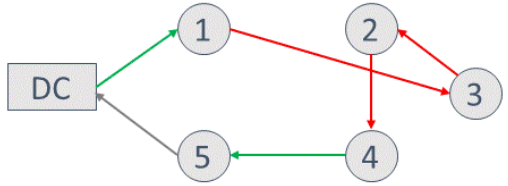
Figure 2: Actual Route



Planned Route



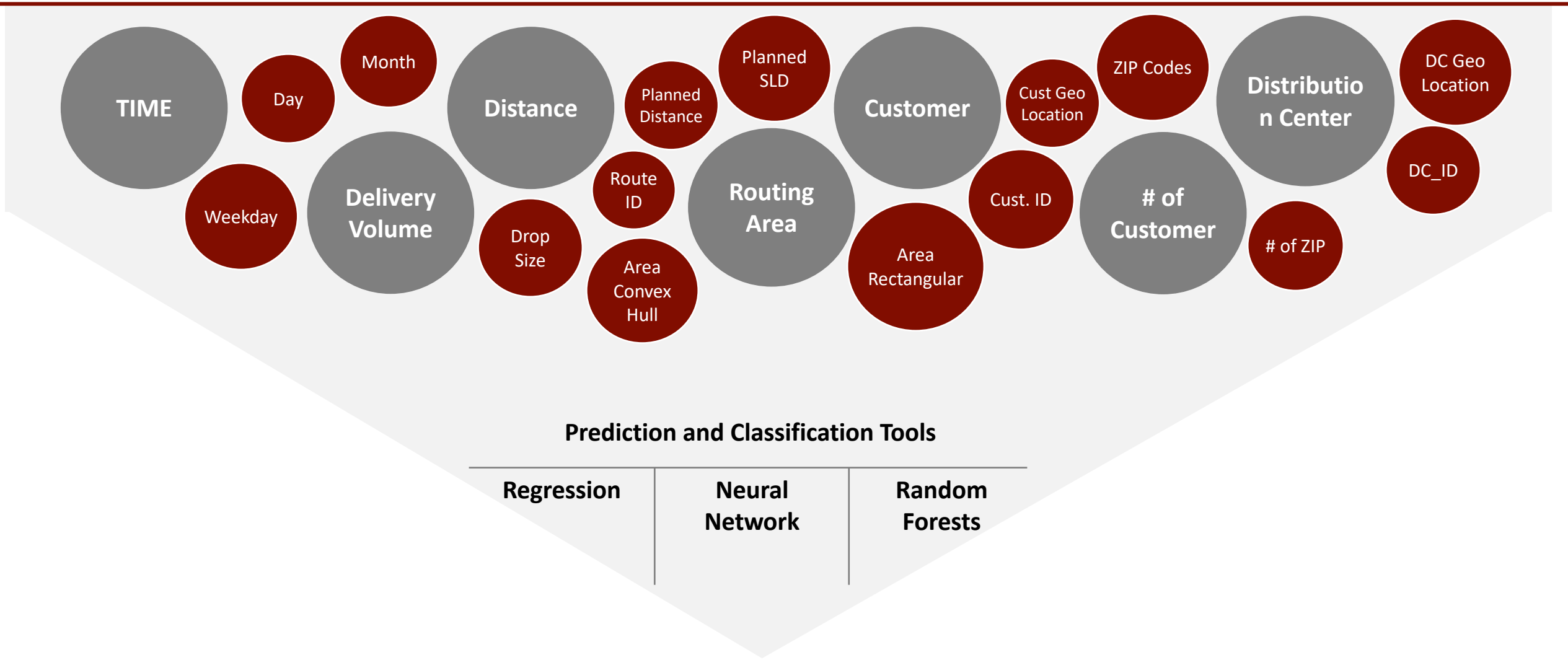
Actual Route



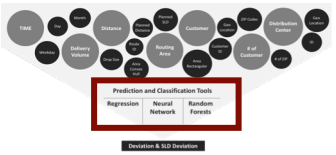
Example:

- **Deviation = 1**
- **Sequence Deviation = 3/5**
- **Distance Deviation = 15/11-1 = 36.4%**
- **SLD Deviation = 7/6 - 1 = 16.7%**

METHODOLOGY



METHODOLOGY



Prediction and Classification Tools – Performance Metrics

Regression

- Continuous Variable — **Adjusted R^2**
- Binary Variable — **Generalized R^2**

Neural Network & Random Forests

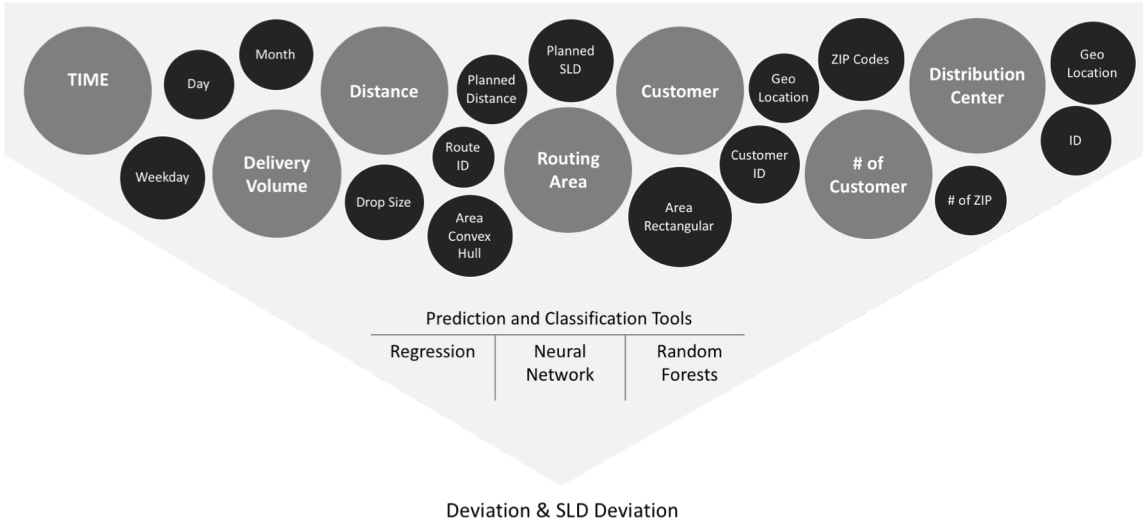
- Both Variables — **Generalized R^2**
- Binary Variable — **Confusion Matrix**
  - Specificity
  - Sensitivity
  - Accuracy
- Fixed 70% Validation Set and 30% Training Set

METHODOLOGY

	Mexico	US	All
Route Instances	7,644	47,881	55,525
Number of DCs	9	9	18
Stops per Route	17.9	12.8	13.5
Route Distance (km)	73.0	106.9	102.2
Deviation	**45.8%**	**79.8%**	**75.1%**
Sequence Deviation*	61.8%	54.7%	55.3%
SLD Deviation*	**12.1%**	**1.7%**	**2.6%**

* Only considering deviated routes

- > US valid data size is **6X** the Mexico data size
- > US deviated routes deviate more
- > US routes' deviation impact on SLD is lower
- > **Significant difference** in deviation between countries

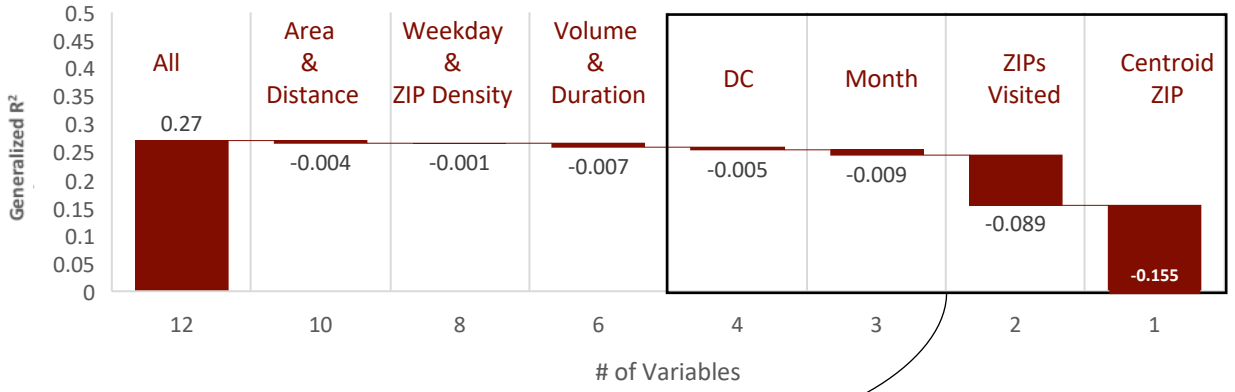


x 2

Results – *Deviation* by Regression Analysis

> Iterative Process of selecting significant variables

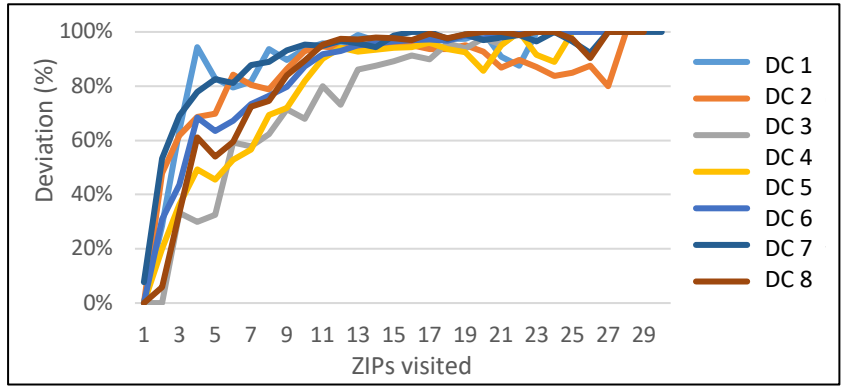
> Performance measured by Generalized R^2



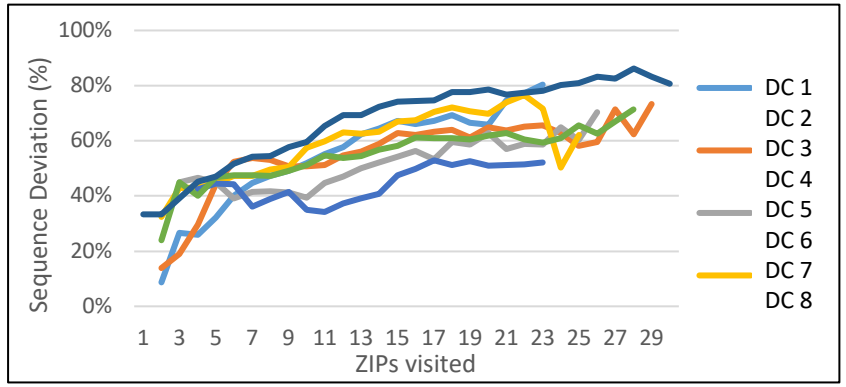
> Sensitivity Analysis

	Gen. R^2	Difference
ZIPs visited	0.167	0.091
Centroid ZIP	0.188	0.070
Month	0.244	0.014
DC_ID	0.253	0.005
All included	0.258	0.000

Deviation vs ZIPs visited, by DC

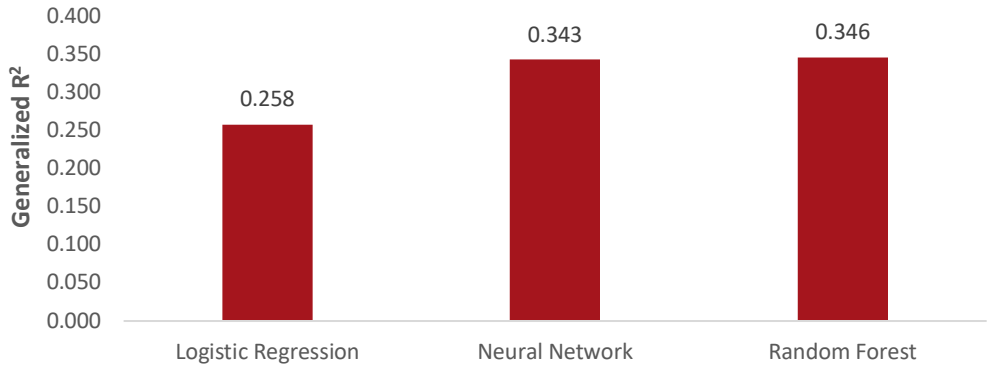


Sequence Deviation vs ZIPs visited, by DC

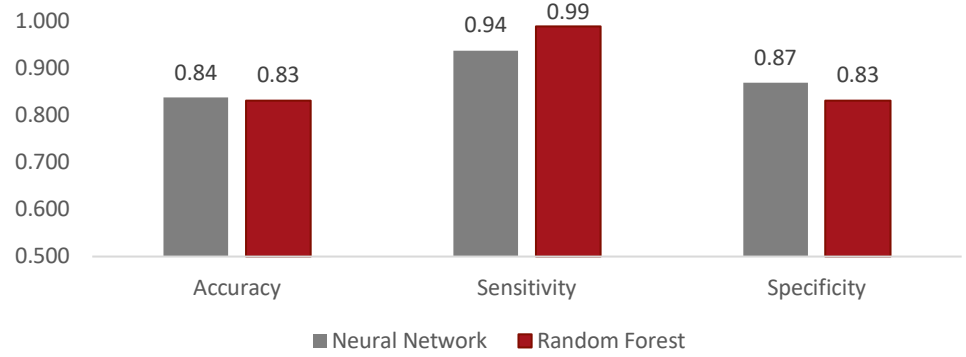


Results – *Deviation* by Classification Methods

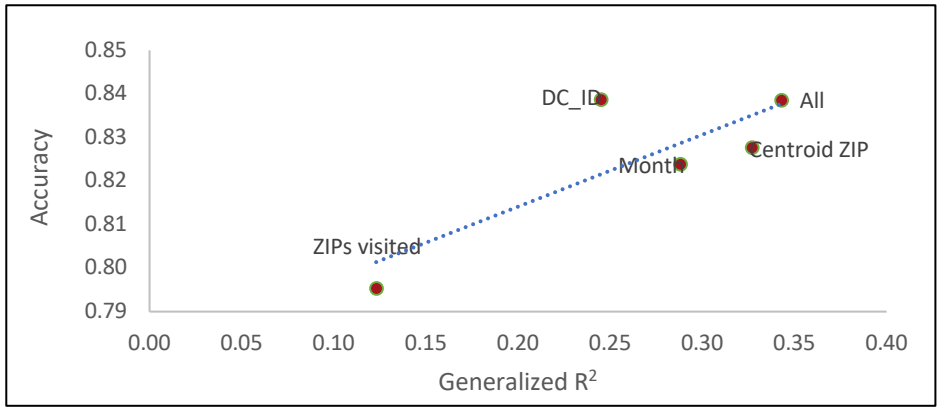
> **Significantly Higher Generalized R^2 than logistic Regression**



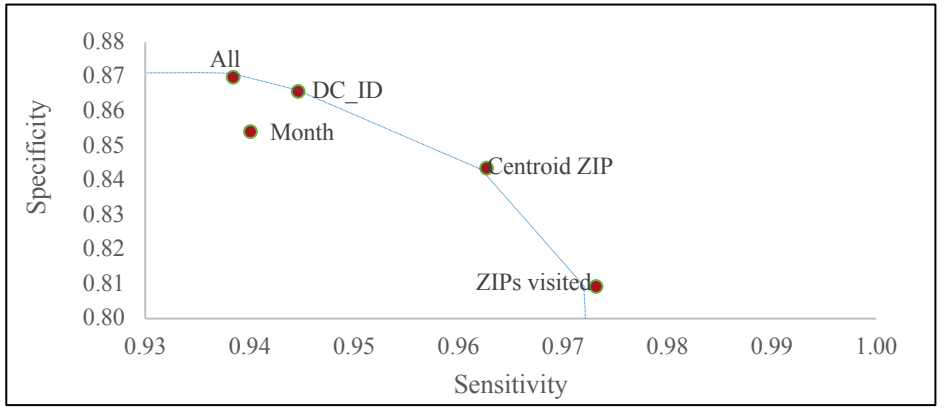
> **Random Forest has higher Sensitivity but lower Specificity**



Accuracy vs Generalized R2 for Neural Network (US)



Specificity vs Sensitivity for Neural Network (US)



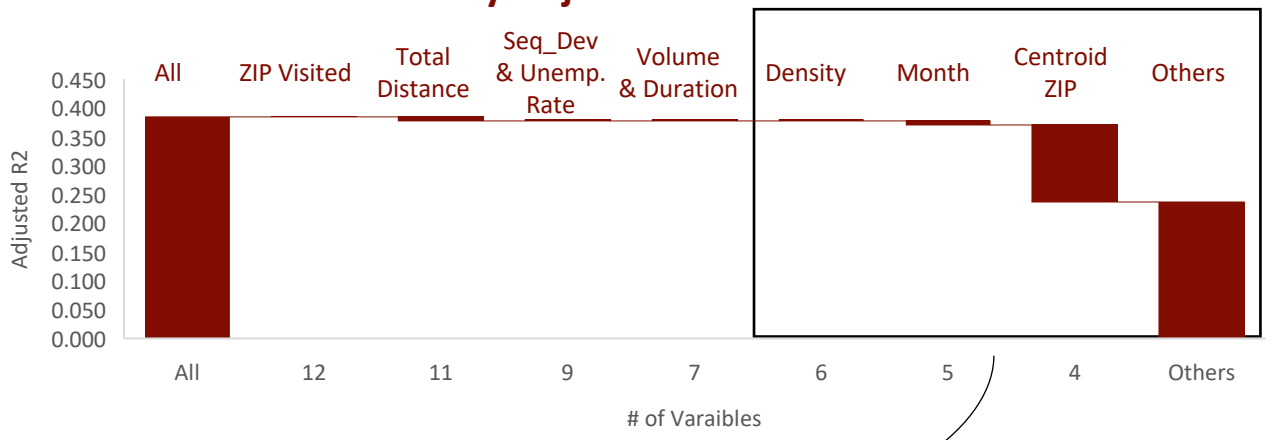
		Predicted 1	Predicted 0
Actual	1	a	b
	0	c	d

$$Accuracy = \frac{a+d}{a+b+c+d} \quad , \quad Sensitivity = \frac{a}{a+b} \quad , \quad Specificity = \frac{a}{a+c}$$

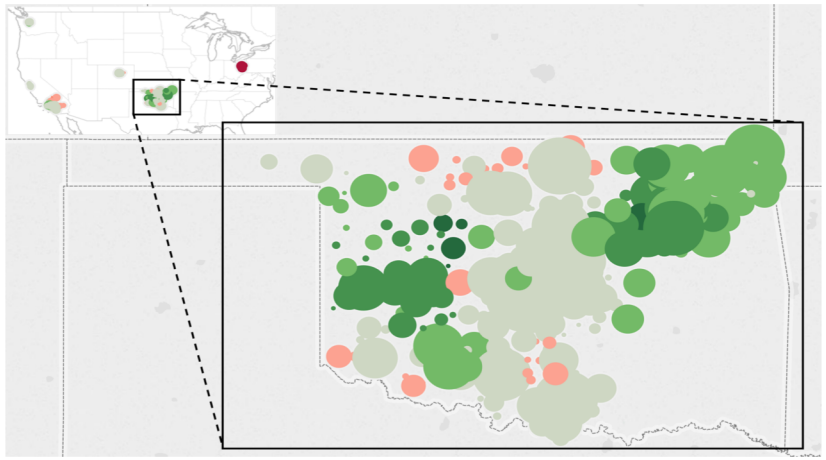
Results – *SLD_Deviation* by Regression Analysis

> Iterative Process of selecting significant variables

> Performance measured by Adjusted R^2



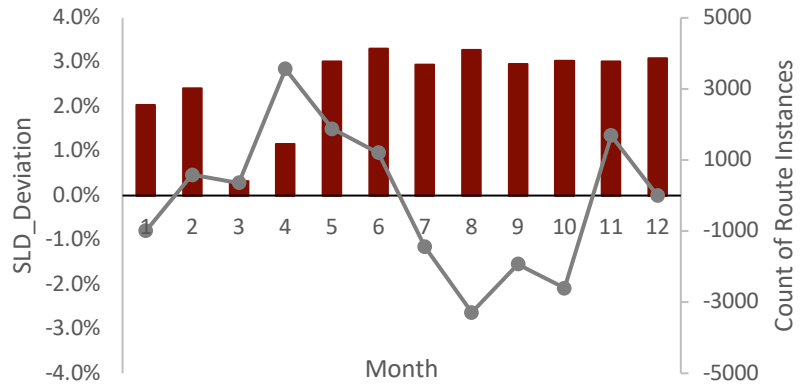
Centroid ZIP vs SLD_Deviation



> Sensitivity Analysis

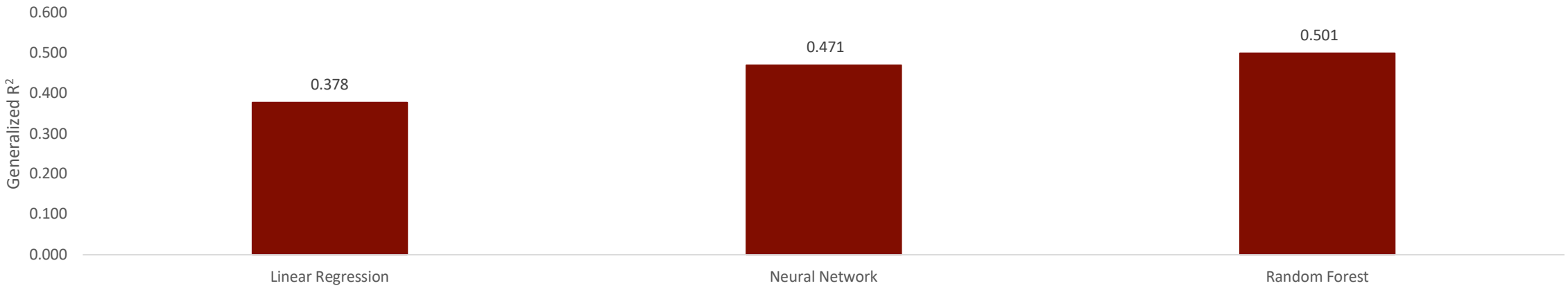
Variables	Adjusted R^2	Difference
Planned_SLD	0.220	-41.8%
Centroid ZIP	0.250	-33.9%
Route_ID	0.294	-22.2%
Number_Customers	0.367	-2.9%
Month	0.370	-2.1%
Area_Rectangular	0.370	-2.1%
All Included	0.378	0.0%

Seasonality of SLD_Deviation



Results – *SLD_Deviation* by Classification Methods

> Significantly Higher Generalized R^2 than Linear Regression



> Classification method, more variables ≠ better R^2

> Centroid ZIP and Route_ID are among the very significant variables in the linear regression but are the least significant in the neural network

Variables	Adjusted R^2	Difference
Planned_SLD	0.220	-41.8%
Centroid ZIP	0.250	-33.9%
Route_ID	0.294	-22.2%
Number_Customers	0.367	-2.9%
Month	0.370	-2.1%
Area_Rectangular	0.370	-2.1%
All Included	0.378	0.0%

Variables	Generalized R^2	Difference
Planned_SLD	0.231	-50.8%
Area_Rectangular	0.356	-24.3%
Month	0.435	-7.7%
Number_Customers	0.436	-7.3%
Route_ID	0.478	1.5%
Centroid ZIP	0.478	1.6%
All Included	0.471	0.0%

Variables	Generalized R^2	Difference
Route_ID	0.449	-10.3%
Centroid ZIP	0.453	-9.6%
Area_Rectangular	0.490	-2.2%
Planned_SLD	0.501	0.0%
Month	0.535	6.9%
Number_Customers	0.536	7.0%
All Included	0.501	0.0%

Conclusion

> Using environmental variables that describe the route, drivers' decision to deviate from the plan can be predicted with an accuracy of **84% in the US and 71% in Mexico**.

> The impact on distance of the deviations can be predicted with a coefficient of determination **R^2 of 0.54**.

> Drivers are more likely to deviate and increase the route's distance when **more customers** are visited.

> Customers' **geographical locations**, reflected in the ZIP codes and group of customers, are useful to predict deviations.