RTS preventable accidents identification

Final presentation
Team 7

Team Members

Xiaoran Li

Project Manager

Second-year MS Data Science Student

Weiran Lin

Algorithm Engineer

Second-year MS Data Science student

Melissa Chen

Algorithm Engineer

Second-year MS Data Science student

Weinan Hu

Algorithm Engineer

Second-year MS Data Science student

Agenda

- Introduction
- Data Overview
- Predictive Analysis
- Causal Analysis
- Future Work

1

Introduction

Business Understanding

Take actions

- Improve the quality of service
- decrease cost
- avoid unnecessary injuries and damages

Vision & Goals

Vision

- Reduce future preventable accidents and improve safety.

Goals

- Identify potential causes of preventable accidents.
- Deliver reusable models to predict future preventable accidents.

2

Data Overview

Data Integration

Absence
Operator ID
Absence Date
(08/24/2015 – 08/31/2019)
Absence Type
...

Scheduled Days Off
Operator ID
Effective Date (04/04/2016)
End Date (09/02/2019)
Weekday
…

Operator Activities
Statistics vs Date

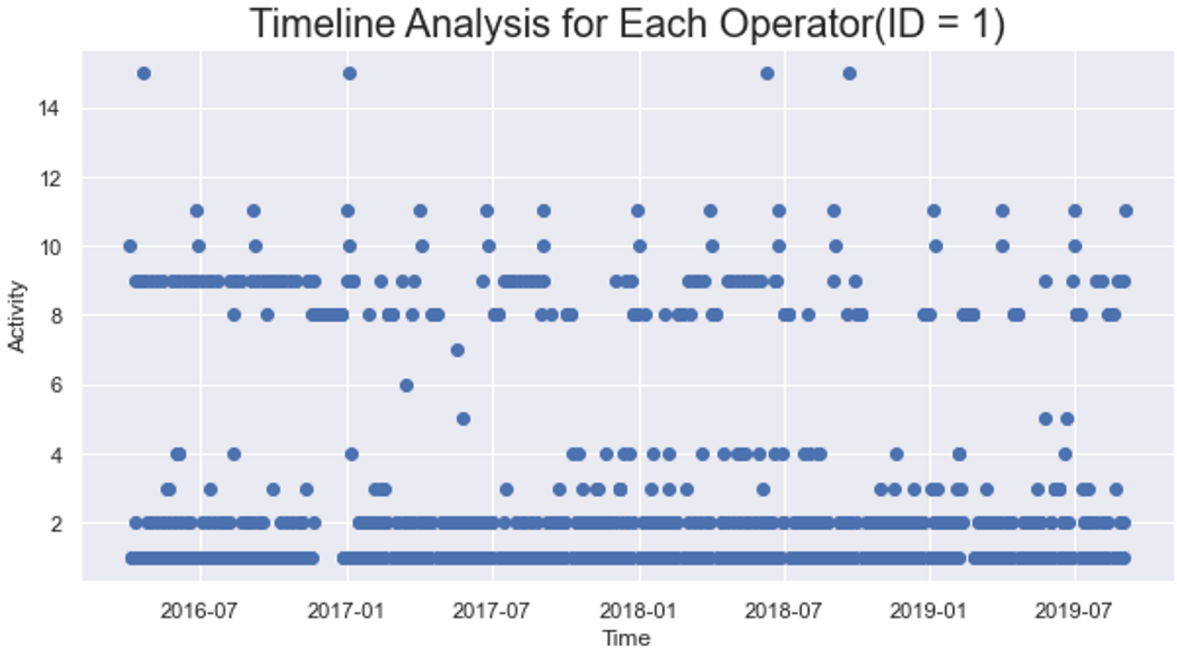
Timeline
Date (04/04/2016 – 09/02/2019)
Operator ID (1 - 430)
Activity 0
Activity 1
…
Activity 15

Timeline Analysis

Timeline Data

	1	2	3	……	430	C0	C1	C2	……	C15
2016/4/4	10	10	10		0	106	0	0	3	1
2016/4/5	1	1	1		0	107	227	28	12	0
2016/4/6	1	1	4		0	107	226	28	16	0
2016/4/7	1	1	1		0	107	228	33	12	0
2016/4/8	1	1	4		0	106	223	27	17	0
2016/4/9	9	2	2		0	107	87	176	7	0
2016/4/10	2	9	2		0	107	66	199	6	1
……										

Timeline Analysis



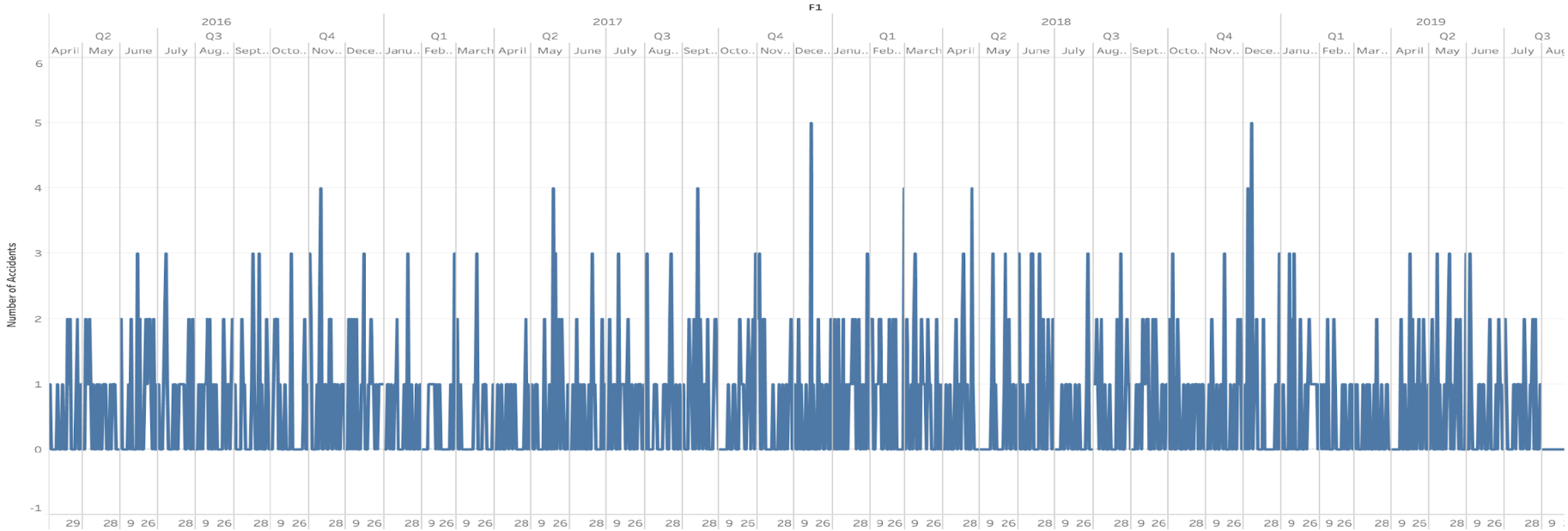
15: accident !

11: pick end
10: pick start
9: extra work time
8: vacation absence
7: medical absence
6: late absence
5: holiday absence
4: excused absence
3: pure absence
2: regular off time
1: regular work time
0: not on pick

- Timeline range from 2016-04-04 to 2019-09-02
- Operators have different timelines
- Timelines as features

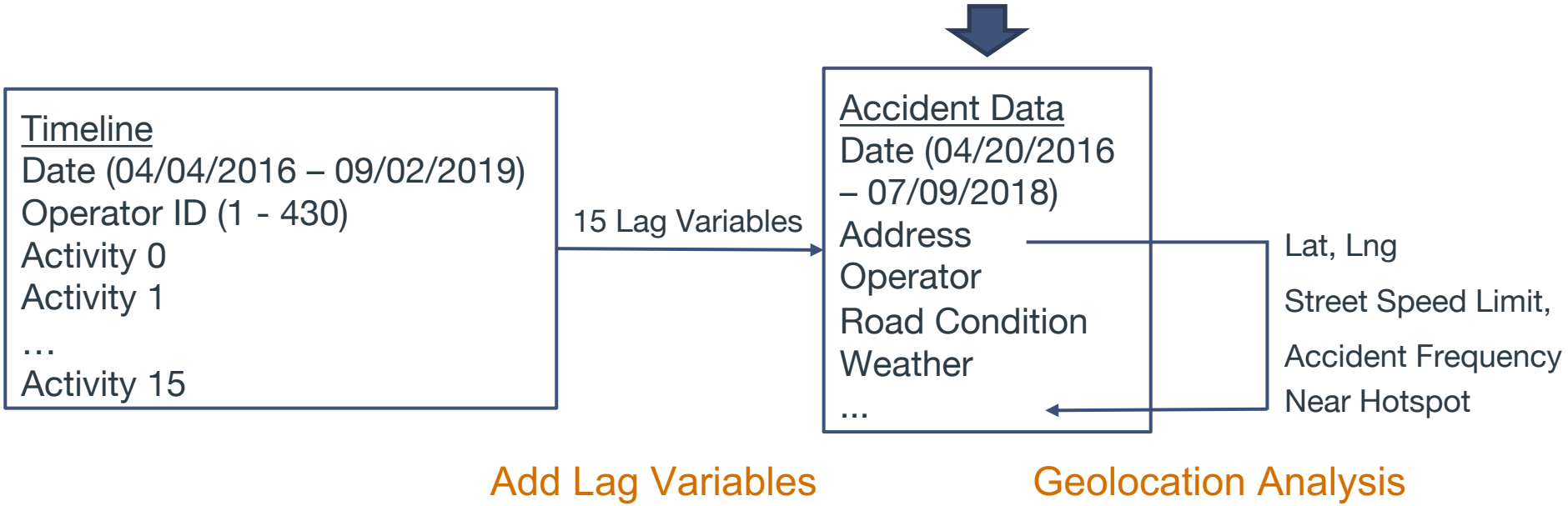
Timeline Analysis

<timeline statistics for each Accident>

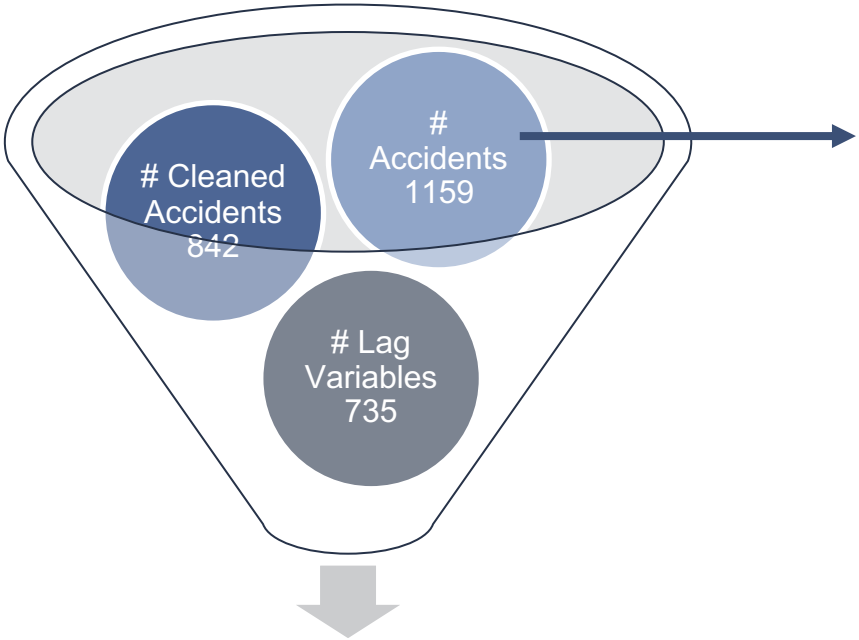


- Activity(accident) statistics on all operator

Data Integration



Data Cleaning



Merged Accidents data 397

Add geoinformation

Missing Value: 309 in Location, 617 in Vehicle Occupancy

Duplicate Value: 0

Wrong Value: 11 in Accident Year

Data Description

- 497 rows and 42 columns

Column Type	Column Name
Operator ID	
Vehicle	vehicle maker, year
Road	Rode type, light, weather, light, surface
Address	Address neighborhood, type, long, lat, speed limit, frequency near hotspot
Time	Year, Month, Date, Time(Morning, Afternoon, Night, Before Morning)
Lags	12 Activities' Lags, Frequency of absence/extra work
Target	Preventability

3

Predictive Modeling

Machine Learning Models Development Procedure

1. **Further Data Preparation**

Data preparation ⇄ Modelling
Increase cohesion of data types/scales

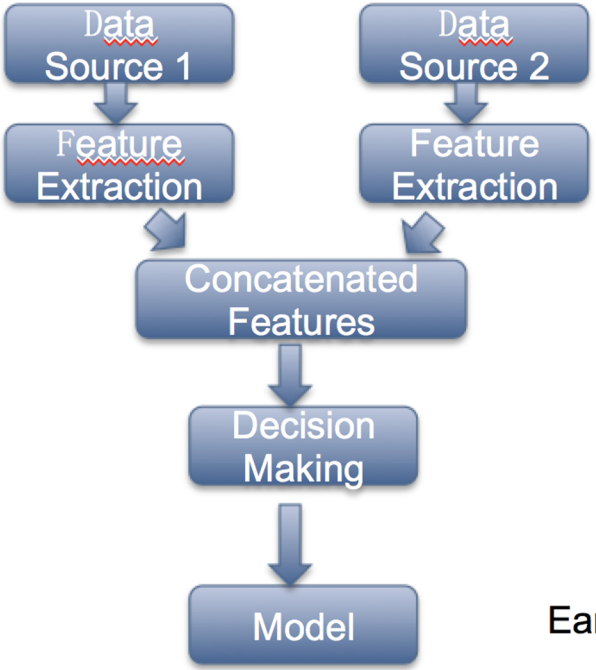
2. **Feature Engineering and Model Fitting**

Deal with Data leakage problem
Fit into popular and suitable machine learning models

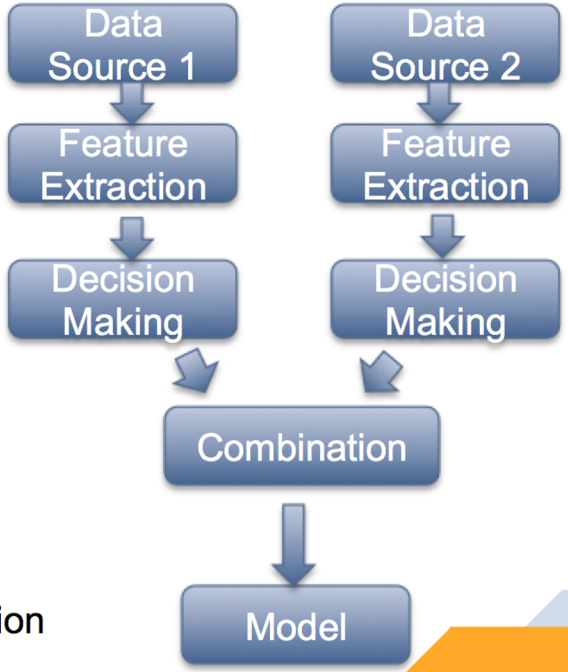
3. **Model Evaluation**

Accuracy, ROC, Cohen's kappa, Time consumption
Feature contribution and Decision boundary

Feature Engineer



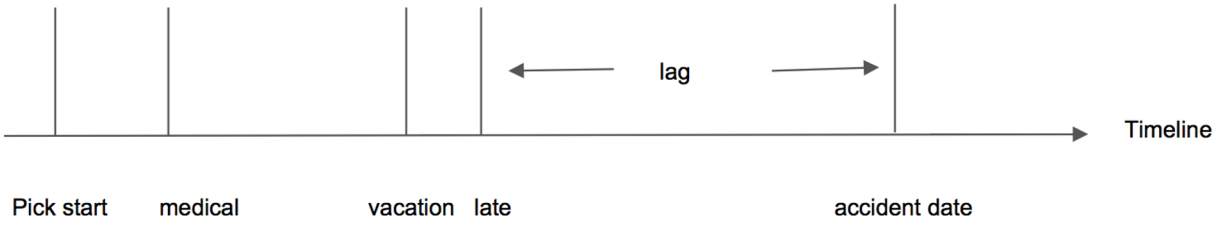
Early Fusion/Late Fusion

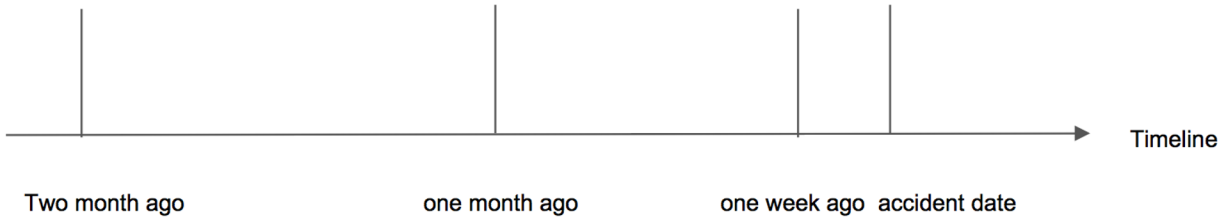


Feature Engineer

Column Type	Column Name
Keys	Accident Date, Operator ID
Accident Description*	Accident Type Code, Accident Subtype Code
Vehicle	vehicle maker, vehicle year
Driver	Age, sex
Roadway	Road type, Light Conditions, Road Weather, Road Surface
Address*	Address Type, Neighborhood, Address, Longitude, Latitude
Geo Analysis*	Speed limit, Hot spot accidents within 0.00075(20-40m), 0.0014(40-80m), 0.003(80-160m)
Timeline Analysis*	12 Activities Lag, Frequency of absence/extra work within one week, one and two month
Target	Preventable/Non-preventable

Lag Variables Definition

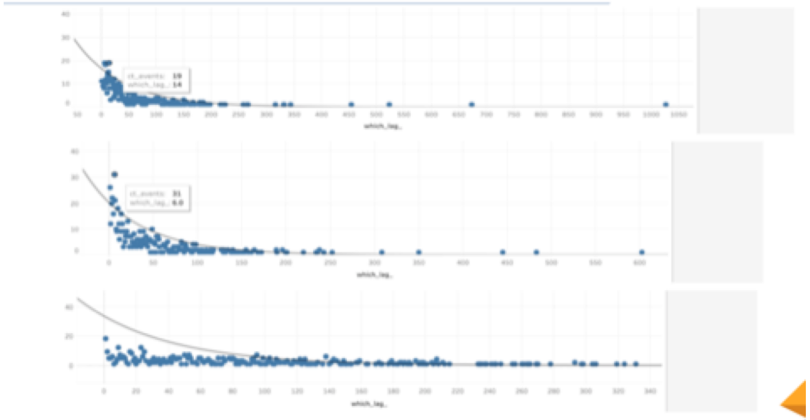


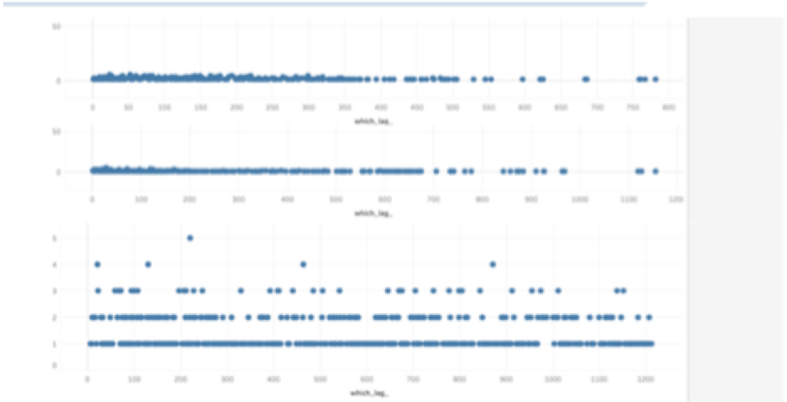


Two ways of definition

- Time lag of each activity(activity fixed)
- Domain knowledge
- Frequency of absence/ extra work of certain time period
(time lag fixed)

Feature Selection(0)

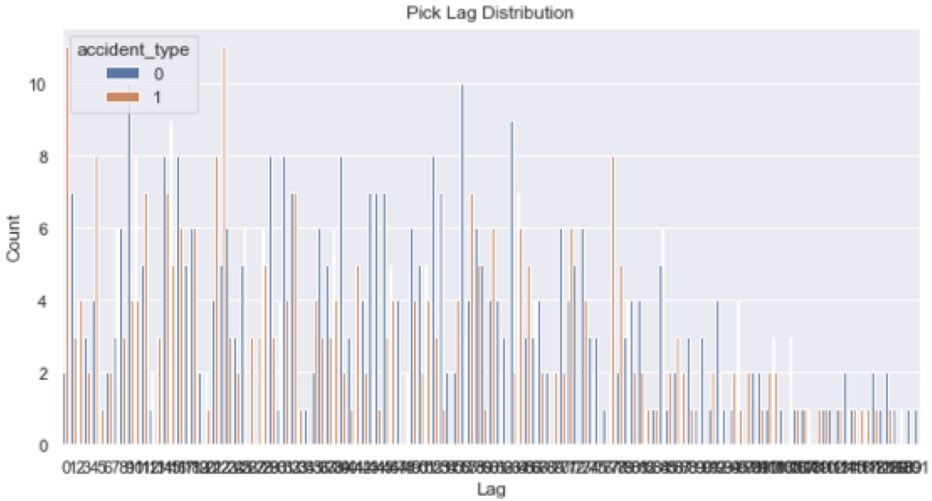


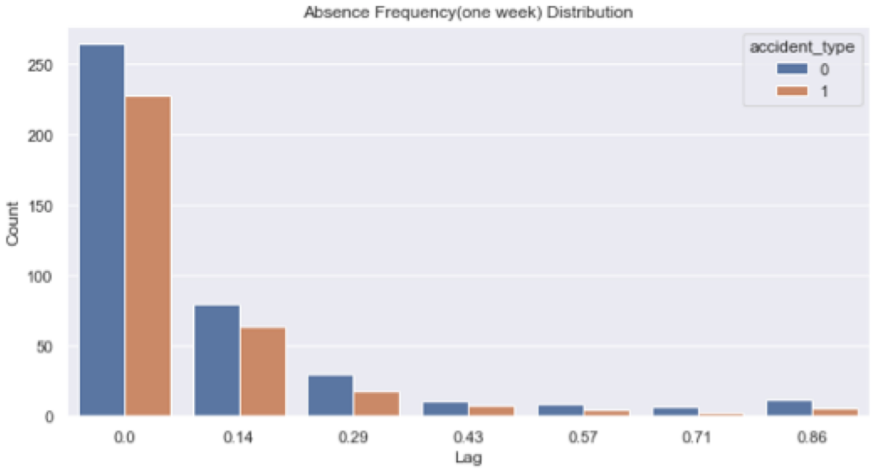


Significant variables:
Pick lag;
Excused lag;
Absence frequency;
Extra Work frequency;
Absence lag;
Vacation lag;

Insignificant variables:
Holiday lag;
Late lag;
Medical lag;
Last accident lag;
Employment lag;

Feature Selection(1)





- Distribution of two groups eyeballs the same
- Distribution is not normal
- Statistical Tests

Feature Selection(1)

- Wilcoxon rank-sum test:

A nonparametric test that allows two groups or conditions or treatments to be compared without making the assumption that values are normally distributed.

- Kolmogorov–Smirnov test:

A nonparametric test of the equality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability, or to compare two samples

Feature Selection(1)

	Wilcoxon Test		Kolmogorov Test	
	statistics	P values	statistics	P values
Pick lag	62489.0	0.044*	0.1021	0.041*
Absent lag	65126.0	0.214	0.0699	0.315
Excused lag	63582.0	0.092	0.0606	0.488
Vacation lag	60249.5	0.00605**	0.1048	0.033*
freq_absence_one_week	63043.5	0.035*	0.0533	0.652
freq_extra_one_week	66877.5	0.284	0.0080	1.000
freq_absence_one_month	58437.0	0.00082***	0.1299	0.0037**
freq_extra_one_month	66193.0	0.224	0.0309	0.991
freq_absence_two_month	59989.5	0.00497**	0.0913	0.088
freq_extra_two_month	66035.5	0.241	0.0418	0.888

Feature Engineering

25 columns: 2 keys + 22 features + 1 target
735 rows

Column Type	Column Name
Keys	Accident Date, Operator ID
Vehicle	vehicle maker, vehicle year
Driver	Age, sex
Roadway	Road type, Light Conditions, Road Weather, Road Surface
Address	Address Type, Neighborhood
Geo Analysis	Speed limit, Hot spot accidents within 0.00075(20-40m), 0.0014(40-80m), 0.003(80-160m)
Timeline Analysis	Pick start lag, Vacation lag, absence work within one week, one month, two month
Target	Preventable/Non-preventable

Models Fitting

- Decision Tree

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label. The paths from root to leaf represent classification rules.

- Adaboost

AdaBoost is a ensemble boosting machine learning meta-algorithm. The output of the learning algorithms ('weak learners') is combined into a weighted sum that represents the final output of the boosted classifier.

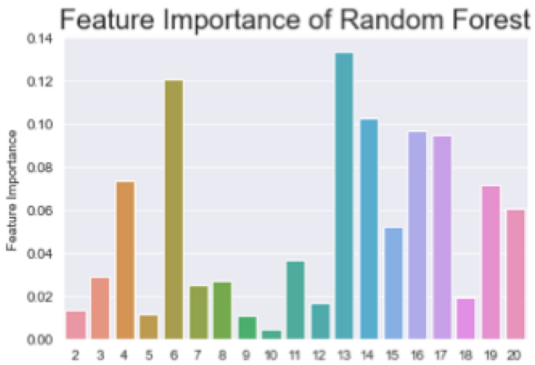
- Gradientboost

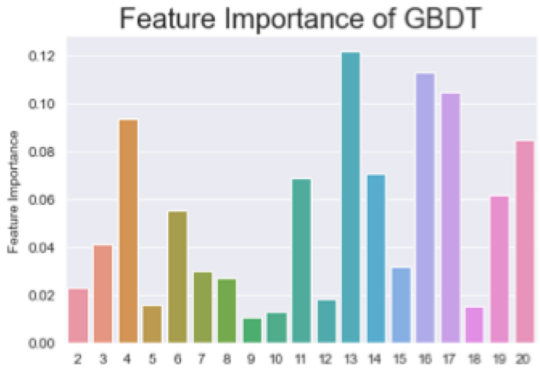
Gradient boosting is a a ensemble boosting machine learning algorithm that generalizes decision trees by allowing optimization of an arbitrary differentiable loss function.

- Ramdom Forest

The random forest is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

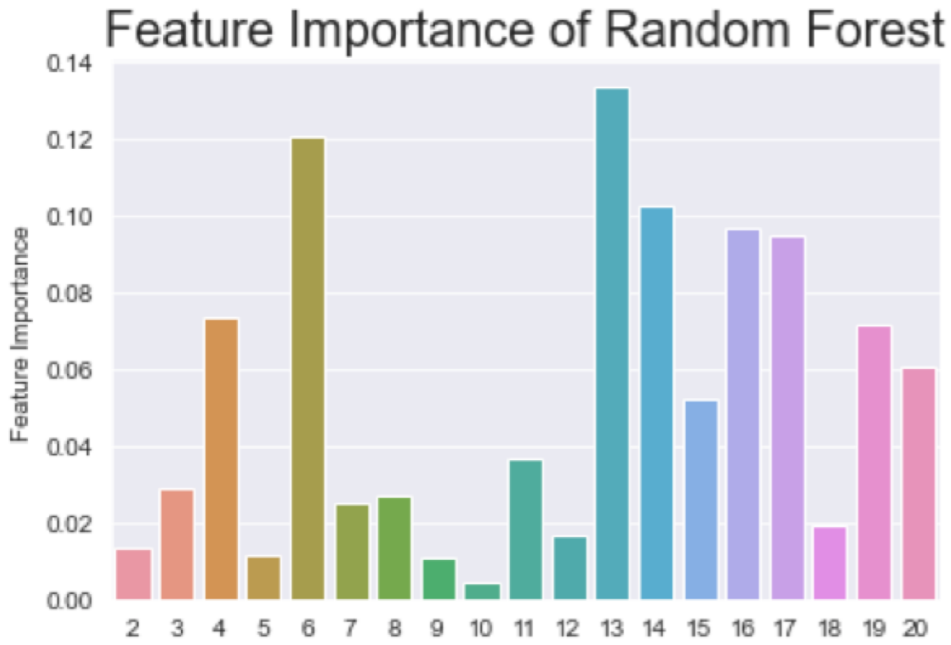
Results & Causal Analysis





- 2-3 vehicle; 4-5 driver; 5-9 roadway; 10-11 address;
- 12-15 geo features; 16-20 lag features
- 4, 6, 13, 14, 16 , 17, 19, 20 can be potential causes

Results & Causal Analysis



- 4 driver sex
- 6 Roadway Type

straight/curve and level/grade/hillcrest

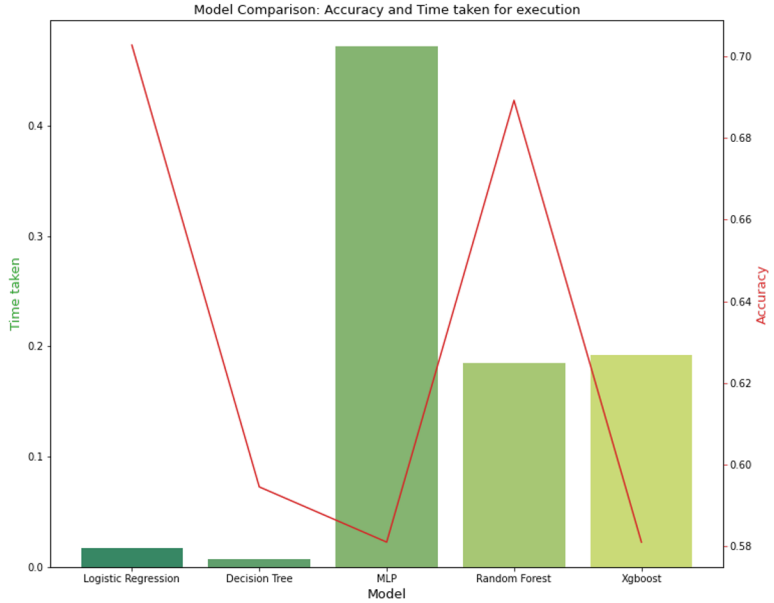
- 13 Hot spot accidents 20-40m radius
- 12 hot spot accidents 40-80m radius
- 16 pick start lag
- 17 vacation lag
- 19 absence frequency one month
- 20 absence frequency two month

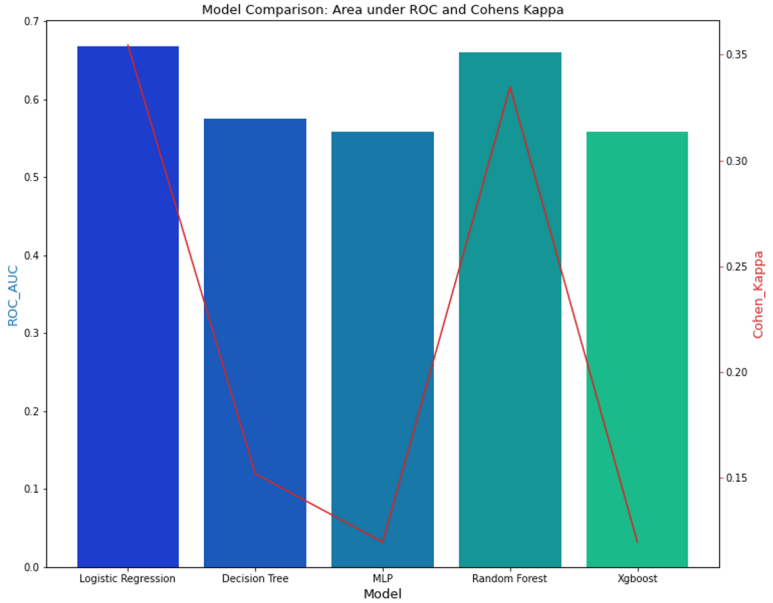
Model Performance

Models with 10-fold cross-validation results

Models	Accuracy	f1 @0 on Test	f1@1 on Test
M1: LR+l1	70.27%	77.55%	56.00%
M2: RF	68.92%	75.59%	56.60%
M3: DT	59.46%	66.67%	48.28%
M4: MLP	59.45%	67.39%	46.42%
M5:XGB	58.10%	65.93%	45.61%

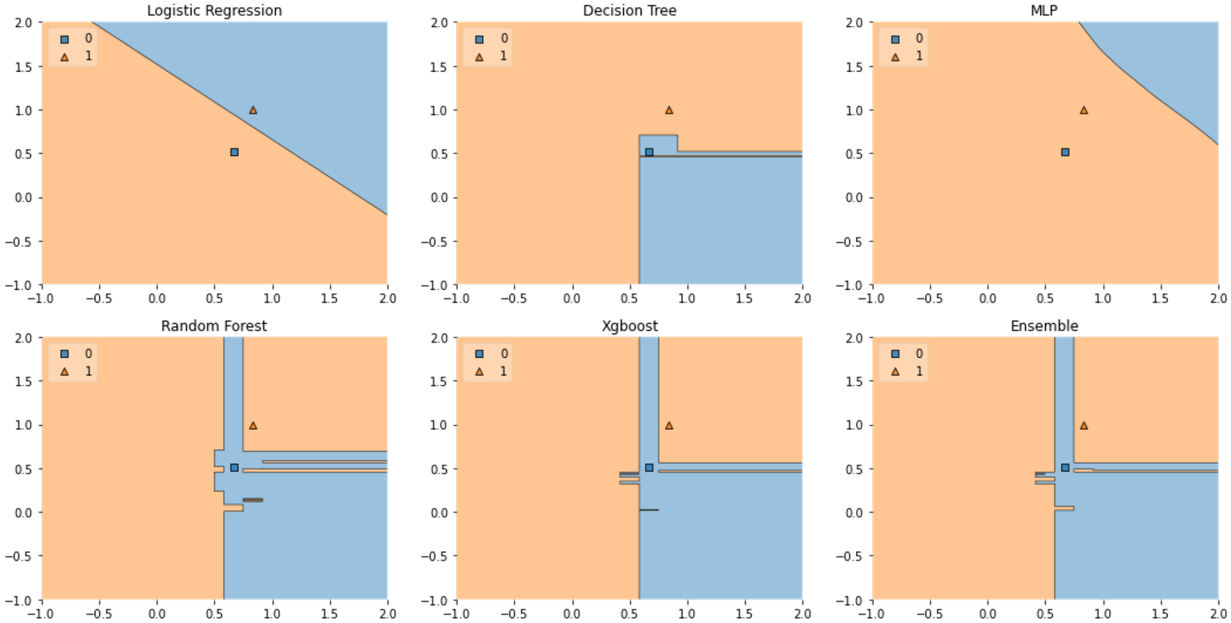
Model Comparison





Model Comparison

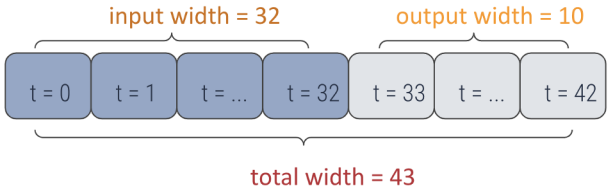
Decision boundary on three most important features



Timeline models： Failed Efforts

Goal: Find Patterns in Operator Data

- Use only operator timeline features to predict accidents
- Correspond with real-life application
- Deep Learning models(CNN/RNN), autoregression



Figure：An illustration of deep learning predictive model

Result: No Correlation Found

- Low accuracy, slightly better than flip a coin. DL agree with autoregression.
- Maybe there is **no explicit pattern** in operator timeline data.

4

Causal Analysis

4.1

Causal Analysis – Logistic Regression

Drop Useless Features

Drop 6 features:

'AccTypeCodeDescr',

'AccSubTypeCodeDes

cr', 'acctype',

'BAITFISH Code' ,

'FileNumber' ,

'Retraining'

Can't know before happen

Data leakage

One Hot Encoding

1099 Columns!

Roadway Weather
0
1
0
1
2
1
3
……

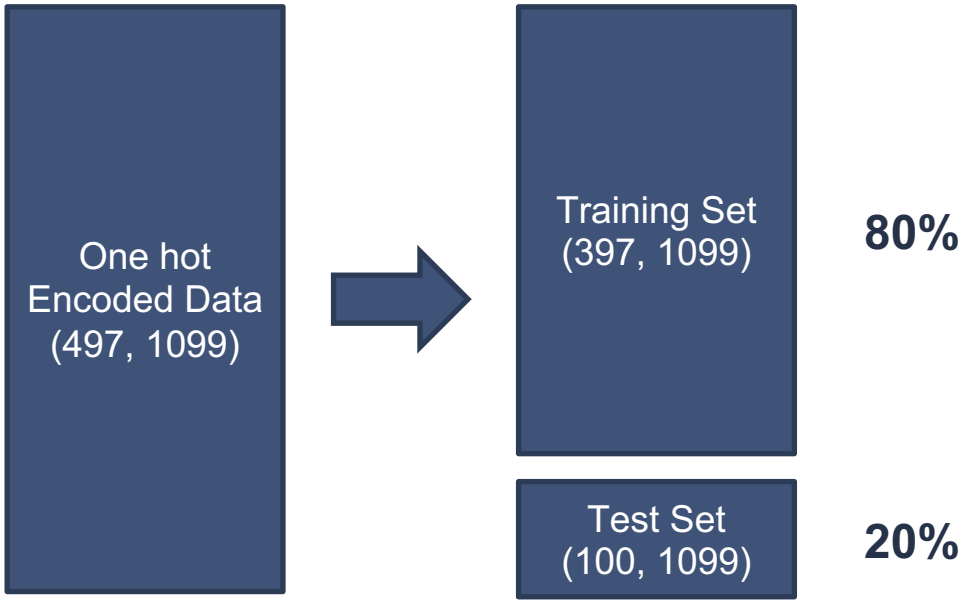
Encoded Feature

One Hot Encoding

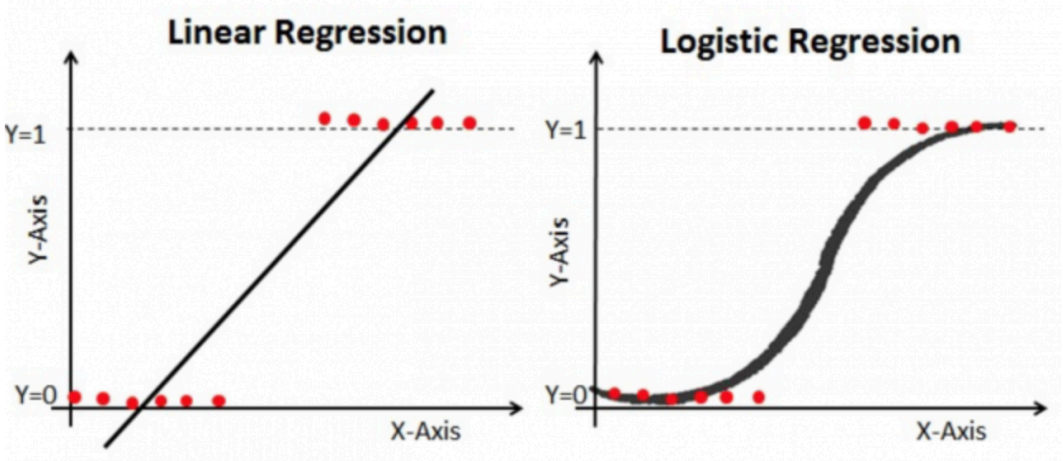
Roadway Weather0	Roadway Weather1	Roadway Weather2	Roadway Weather3
1	0	0	0
0	1	0	0
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	0
0	0	0	1
……	……	……	……

- Logistic regression won't treated them as numerical values
- The categorical feature is not ordinal (No multicollinearity)

Train/Test Split

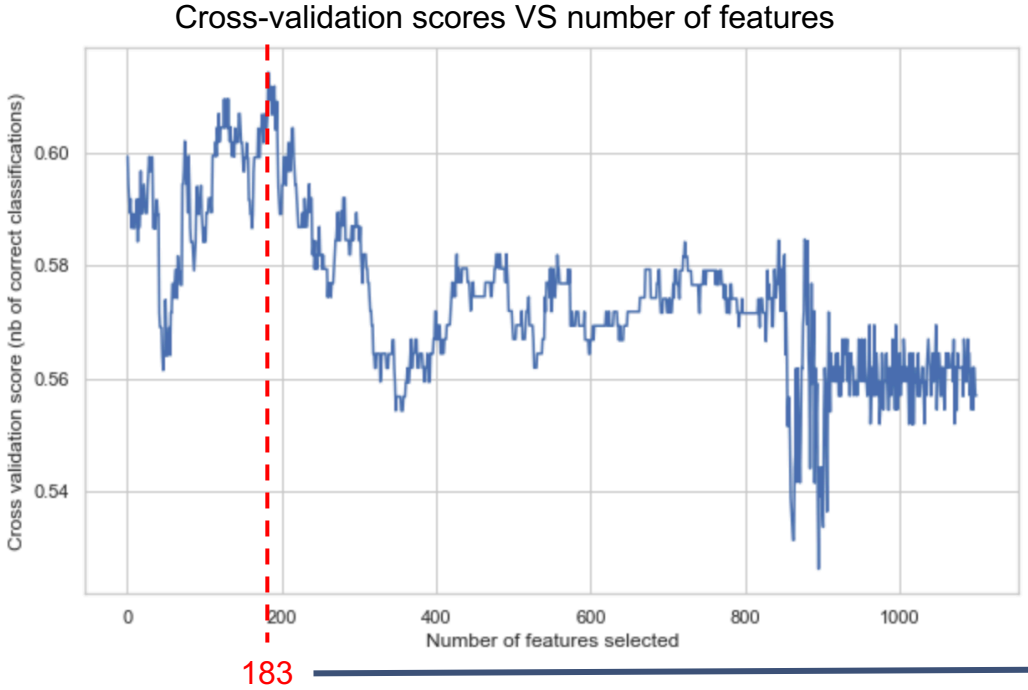


Introduction to Logistic Regression



- use when the dependent variable(target) is categorical.
- maximize the likelihood of the hypothesis that the data are split by sigmoid

Recursive Feature Elimination



Selected features:

['Operator_ID_105',

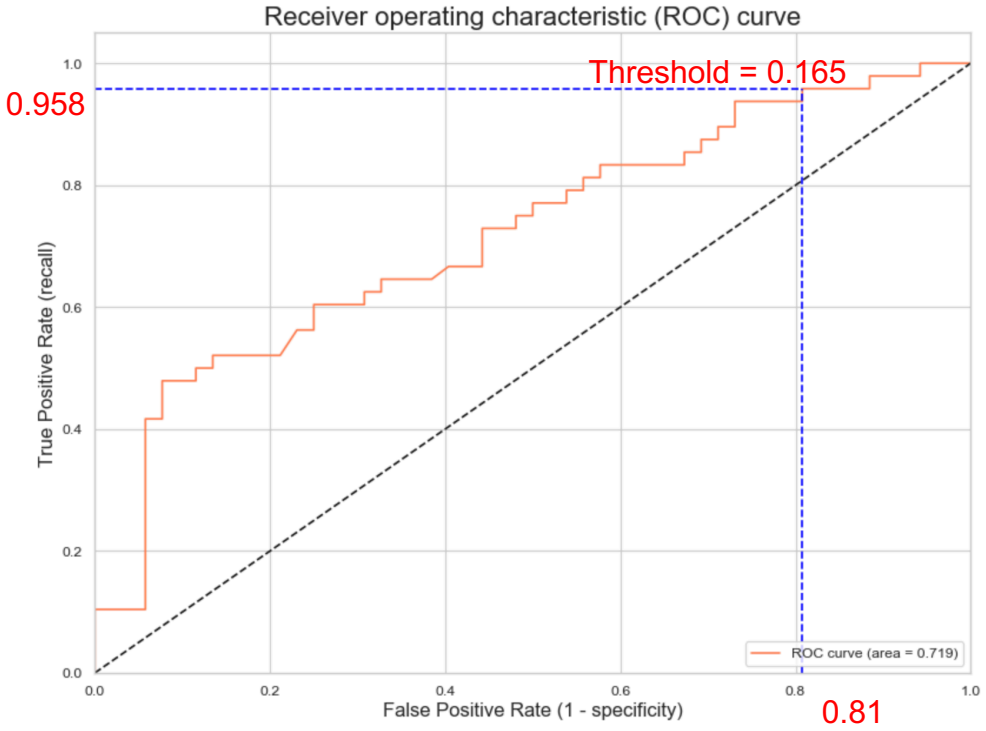
'Operator_ID_109',

……

'accday__7',

'freq_absence_one_week']

ROC – AUC Curve



$$\text{TPR /Recall / Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{FPR} = 1 - \text{Specificity} = \frac{FP}{TN + FP}$$

Test score: 0.711

AUC: 0.719

Causality Result

Weight	Feature	Note
0.0320 ± 0.0344	Neighborhood_25	0
0.0220 ± 0.0080	SpeedLimitCategory_6	SC8 (<11 km/h)
0.0180 ± 0.0294	Roadway_5	Straight and Level
0.0160 ± 0.0271	Roadway_3	Other
0.0160 ± 0.0098	SpeedLimitCategory_3	SC5 (51-70 km/h)
0.0140 ± 0.0098	LocAddr1_128	1372 E.Main St
0.0140 ± 0.0098	Roadway_0	Grade and Curve
0.0120 ± 0.0080	Operator_ID_249	Id = 249
0.0120 ± 0.0080	accday__7	7th data every month
0.0120 ± 0.0196	Operator_ID_182	Id = 182
0.0100 ± 0.0000	Roadway_1	Curve and Level

Limitation!

4.2

Causal Analysis – Logistic Regression After Covariance Term

Defining causal analysis

When road is slippery

Preventable Accident = 1%
Non-preventable Accident = 2%
ratio 1 = Preventable / Non-preventable = 0.5

When road is not slippery

Preventable Accident = 0.5%
Non-preventable Accident = 0.5%
ratio 2 = Preventable / Non-preventable = 1

We will focus on the **ratio** lift here: causal effect = ratio 1 - ratio 2 = 50% less

Translation: Slippery road causes 50% less likely to be preventable accidents

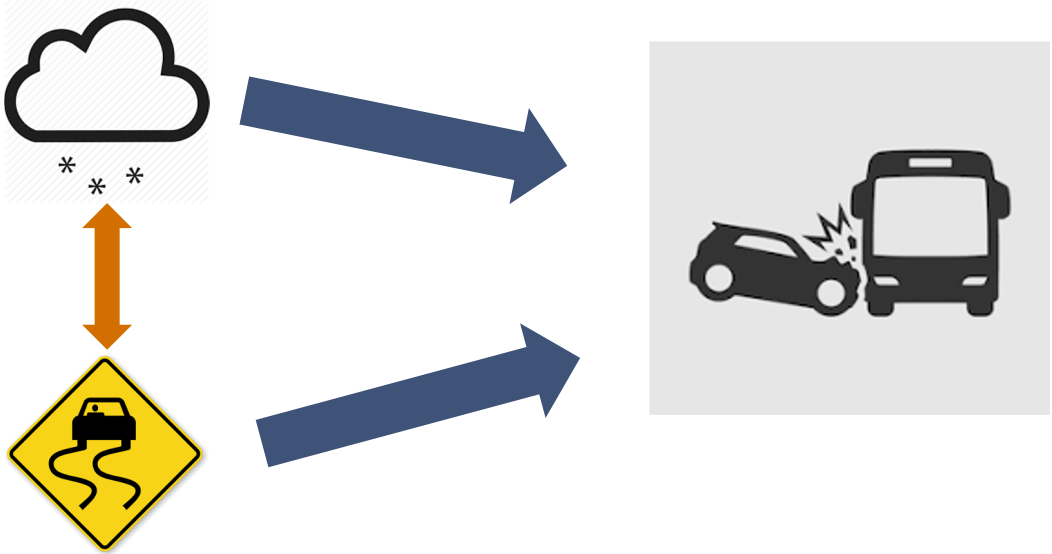
Causal vs Predictive Modelling

In causal analysis:

We hate **covariance** between variables

We hate **hidden variables** we didn't observe:

- New drivers on the road?
- Rochester becoming NY state capital?
- ...

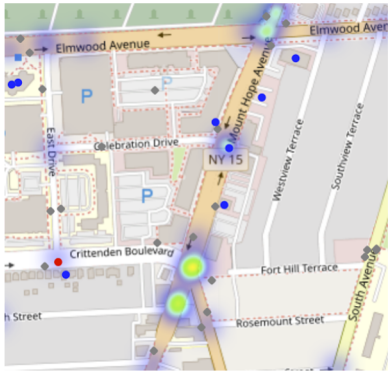


Assume hidden variables stay stable

Method: Use independent variables

Geolocation:

Hotspot
Bus stop distance
Neighborhoods



Criteria: Variance_inflation_factor << 5 (standard)

Driver age
Drive gender
Lag variables (of the driver):
- Days since absence

etc ...

Month **or** seasons
Hour of the day
isWeekend
Bus age
etc...

Roadway condition
Roadway light
Roadway Surface
~~Roadway Weather~~

Causal model: logistic regression

How big the effect & Prevalence

How big the effect is

How confident we are

Very Likely causes: Neighborhood	Correlation	**Causal effect**	P_value
Neighborhood__33_vs_25	0.049975	**Totally preventable**	0.999741
Neighborhood__12_vs_25	0.070746	**Totally preventable**	0.997555
Neighborhood__39_vs_25	0.070746	**Totally preventable**	0.996898
Neighborhood__27_vs_25	0.03109	**27.8% more preventable**	0.809414
Neighborhood__42_vs_25	0.009661	**26.5% more preventable**	0.833052
Neighborhood__25 (Baseline)	NA	**NA**	

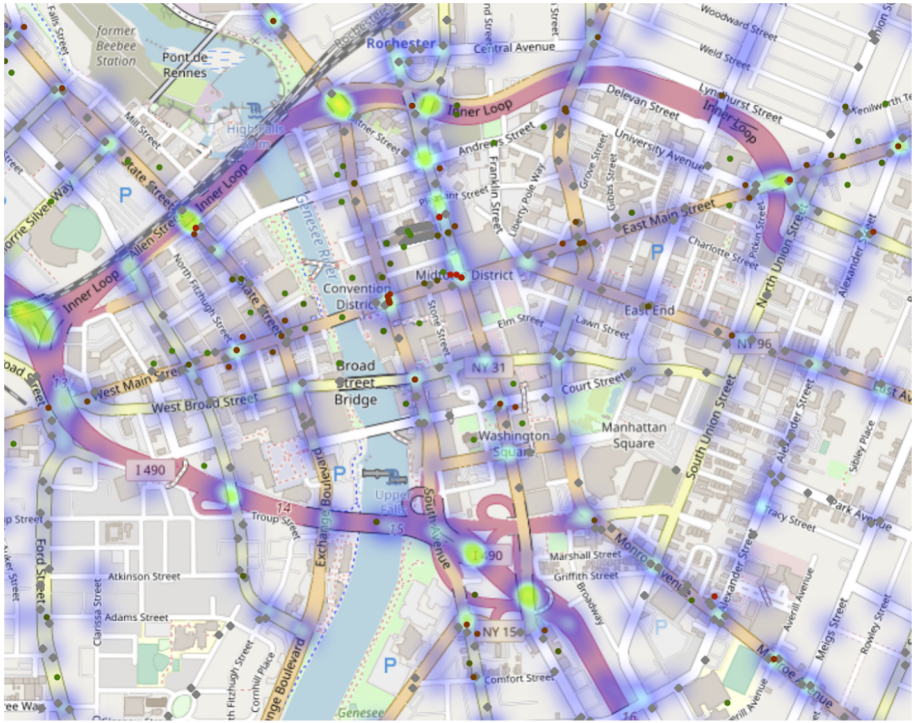
* Causal effect = exp(coefficient) -1

Causal model: logistic regression

Likely causes	Correlation	**Causal effect***	P_value
SpeedLimitCategory__0 (comparing to 4)	0.049975	**Totally preventable**	0.999721
SpeedLimitCategory__1 (comparing to 4)	-0.04034	**Not preventable**	0.999748
RoadwayLightConditions__4_vs_2	0.034483	**23.7% more preventable**	0.908607
Roadway__2_vs_5	-0.00947	**16.3% less preventable**	0.86083

* Causal effect = exp(coefficient) -1

Geoinformation Output



An interactive map to show:

- Accident hotspots
- Bus stops
- Past accidents

Hotspot is correlated with **not preventable**

5

Future Work

Future Work

- Wrap up code and dashboard
- Come up with business suggestions
- Compile report

Acknowledgement

Thanks for

RTS Quality Service Team

Professor Ajay

Professor P.J.

THANKS!

Any questions?