RTS Preventable Accidents Identification

Xiaoran Li, Weiran Lin, Meiying Chen, Weinan Hu

1. Introduction

As the field of data science evolves, not only have we found ways to utilize large datasets to explore pristine areas, but also have we possess tools and ability that can give insights into traditional industries where data had been collected and piled up.

In this capstone project generously sponsored by Regional Transit Service(RTS), we have the chance to investigate and improve public transportation in Rochester, NY. RTS is regional transportation authority providing public bus services for the Rochester city as well as nearby counties.

The goal of our project is to identify causality related to preventable accidents and to make predictions for future ones. This information will be used to help the service improvement team to take actions. And in the end, help improving the quality of service, decrease cost, and avoid unnecessary injuries and damages for RTS.

There are four members in our project team, Xiaoran Li (Project Manager), Weiran Lin (Algorithm Engineer), Melissa Chen (Algorithm Engineer) and Weinan Hu (Algorithm Engineer). We are all second-year MS Data science students at University of Rochester.

2. Dataset

2.1. Dataset Description

We got the dataset from our sponsor. There are 3 different tables: absence data, accidents data and operators scheduled days off.

Figure 1 shows the first 5 rows of the accidents data. The accidents data records each different accidents from 2014 – 2019.

- 'accdate' is the date and time of accident.
- 'Operator_ID' is the unique integer associated with each bus operator.
- 'AccTypeCodeDescr, AccSubTypeCodeDescr, acctype, BAITFISH Code' indicate the accidents' type.
- 'FileNumber' is unique identifier for each accident.
- 'AccPreventable' is our target target. It means whether the accident was deemed preventable or not.
- 'Retraining, RetrainingDate' indicate whether the operator of this accident is retrained.

- ‘vehbusno, vehmake, vehyear, VehNoOccupants’ show the situation the vehicle when the accident.
- ‘Drv_Age, CoDrvSex’ show the age and sex of the operator of the accident.
- ‘Roadway, RoadwayLightCondistions, RoadwayWeather, RoadwaySurface’ indicate the situation of the road when the accident.
- ‘LocAddr1, Addr_Type, Neighborhood’ indicate the location information.

accdate	Operator_ID	AccTypeCodeDescr	AccSubTypeCodeDescr	acctype	BAITFISH Code
8/4/14 11:57 AM	22	Vehicle Approaching From Angle	Vehicle approaches from right-vehicle straight-RGRTA turns	03 - MVA with Vehicle	General
8/5/14 11:33 AM	209	Vehicle Operating Ahead	Vehicle slows or stops after overtaking and passing for traffic or other reas	02 - MVA with Injuries	0
8/5/14 2:17 PM	187	Vehicle Passing	Vehicle cuts in and scrapes RGRTA on left side	03 - MVA with Vehicle	0
8/5/14 3:00 PM	285	Thrown by movement of RGRTA stopping	Other part of equipment-passenger standing or walking	06 - On Board Injury/Incident	0
8/6/14 11:15 AM	211	Collisions w/Fixed Objects	Miscellaneous	04 - MVA with Fixed Object	0

FileNumber	AccPreventable	Retraining	RetrainingDate	vehbusno	vehmake	vehyear	VehNoOccupar	Drv_Age	CoDrvSe	Roadway	RoadwayLightCondit
14-839	Preventable	Y	2014/10/1	523	Gillig	2009		73	M	Straight and Grade	Daylight
14-838	Non-Preventable	N		527	Gillig	2009	20	38	M	Straight and Level	Daylight
14-840	Non-Preventable	N		1295	New Flyer	2011	20	54	F	Straight and Level	Daylight
14-836	Non-Preventable	N		768	Gillig	2007		38	M		Daylight
14-842	Non-Preventable	N		910	Gillig	2011	1	37	M	Curve with Hillcrest	Daylight

RoadwayWeath	RoadwaySurfa	LocAddr1	Addr_Typ	Neighborhood
Clear	Dry	Dewey and Ridgeway	Address	West Irondequoit
Cloudy	Dry	East Main St @ Chestnut St	Intersection	Charlotte
Clear	Dry	358 East Main Street	Address	0
Cloudy	Dry	271 Greece Ridge Center Drive (Greece Ridge Ma	Address	0
Clear	Dry	950 Norton Street (Franklin High School- Hudsc	Address	Strong

Fig. 1 Accidents Data

Figure 2 shows the first 5 rows of the absence data. The absence data records all absence of each operator from 2015 – 2019.

- ‘Absence_Type’ is the general category of why absent. Exception is "Worked Day Off" - this means the person didn't take their scheduled day off.
- ‘Operator_ID’ is the unique integer associated with each bus operator.
- ‘From_Date, From_Time, To_Date, To_Time’ indicate the absence time.

Absence_Type	Operator_ID	From_Date	From_Time	To_Date	To_Time
Absent	1	2019/3/13	0:00	2019/3/13	25:00
Absent	1	2019/2/8	0:00	2019/2/8	25:00
Absent	1	2019/6/10	0:00	2019/6/12	25:00
Absent	1	2019/6/6	0:00	2019/6/7	25:00
Absent	1	2019/6/5	0:00	2019/6/5	25:00

Fig. 2 Absence Data

Figure 3 shows the first 5 rows of the scheduled days off data. The scheduled days off data shows the work shifting of each operator during 2016 – 2019.

- ‘Effective_On, No_Longer_Effective_After, Pick_Name’ indicate each pick. There are 4 picks of each year for operators to choose from.
- ‘Seniority_Date’ is the time when an operator began working at RTS.
- ‘Operator_ID’ is the unique integer associated with each bus operator.
- ‘Sun, Mon, Tue, Wed, Thu, Fri, Sat’ indicate whether this day of the week is a regularly scheduled day off during this pick.
- ‘Type_Work’ Indicates whether the operator picked regular work or is on the extra board for this pick.

Effective_On	No_Longer_Effective_After	Pick_Name	Seniority_Date	Operator_ID	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Type_Work
2016/4/4	2016/6/26	APR2016	06-20-1974	1	OFF						OFF	REGULAR
2016/4/4	2016/6/26	APR2016	08-22-1974	2	OFF						OFF	XB
2016/4/4	2016/6/26	APR2016	06-22-1977	3	OFF						OFF	REGULAR
2016/4/4	2016/6/26	APR2016	08-03-1977	5	OFF						OFF	REGULAR
2016/4/4	2016/6/26	APR2016	02-04-1980	6	OFF						OFF	XB

Fig. 3 Scheduled Days Off Data

2.2. Dataset Integration

2.2.1. Timeline Analysis

We have working and absence information available. We decide to draw a timeline of employees to dictate their everyday work activities during 3 tables common time which is from 2016-04-02 to 2019-09-02. And we link the accidents time on it so that we could draw more information from our timeline table.

Firstly, we joined the scheduled days off table and the absence table on date and generated the timeline table to dictate each employee pick time, regular work time, regular off time, different types of absence time and extra work time. Then we merge timeline table with accident table to link on each accident.

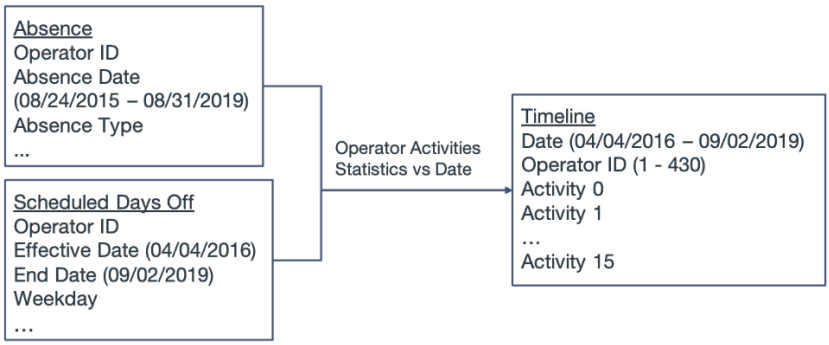

Fig. 4 Timeline Data Generation Process

Figure 5 shows the first 5 rows of the timeline data. The timeline data shows the activity of each operator and the number of times each activity happened on each day from 2016 – 2019. The corresponding activity of the activity code in the timeline data can be found in the activity code data in Figure 6.

- 1 – 430: The activity happened on each date for each operator (ID 1 – ID 430)
- C0 – C15: the number of times each activity(activity 0 – activity 15) happened on each date.

	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

Fig. 5 Timeline Data

Below is the activity code, which shows the activity corresponding to its code.

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

Fig. 6 Activity Code

We plotted the activities against timeline for each operator, below we took the operator whose ID = 1 as an example. In this plot, the below 2 lines show regular work information; the middle dots show different types of absence; the line equals to 9 shows extra work information; the dots equals to 10 and 11 dictate pick period; and highest points mean accident time. We can imagine each employee's absence and extra work activities have some relevant to do with accidents. This leads to our next time lag features calculation.

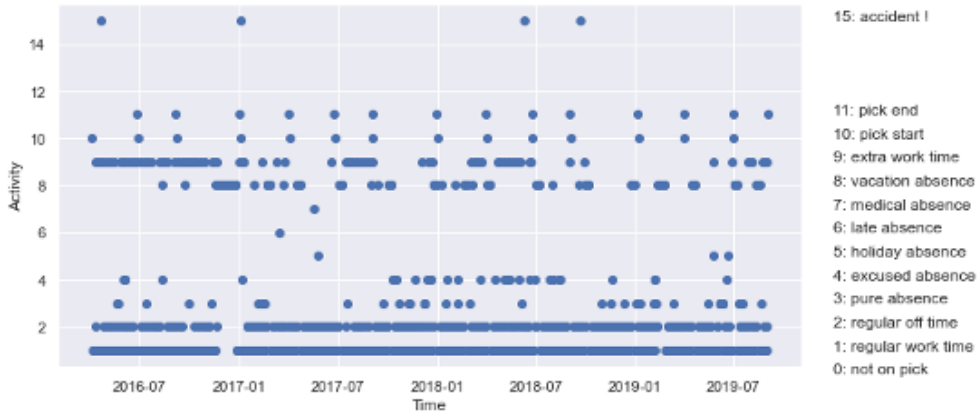

Fig. 7 Timeline For The Operator ID=1

We also plotted the aggregate activities of all operators from 2016-04-04 to 2019-09-02. Below we took the accident activity for an example. By doing this, we can find some patterns of how the seasonality impact the activities. For example, the accidents are more likely to happen in winter.

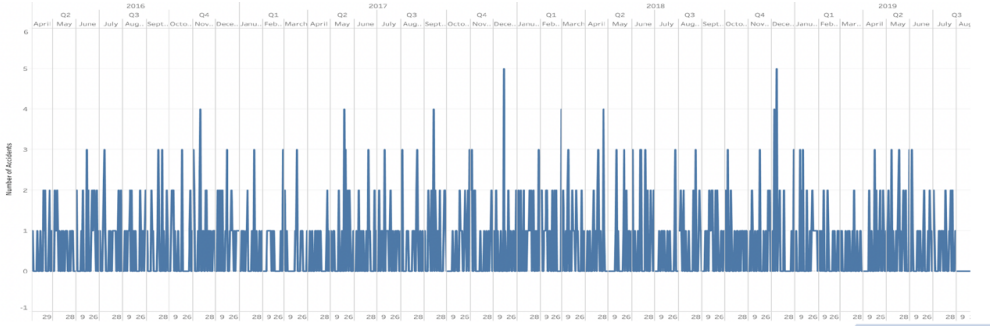

Fig. 8 Timeline For The Aggregated Accidents Of All Operators

2.2.2. Lag Variables Adding

After generating the timeline data, we can add lag variables to the accident table based on it. The intuition here is calculating the days between the activity happened date and the accident data. Since the lag variables are very important in our analysis, we will discuss how to add them in the following part. There are 2 ways to calculate lag features of each accident utilizing different information. The process is shown in figure 9.

First, from the timeline table, based on each accident date, we calculate the time lag between each activity and the accidents. Therefore we end with 7 lags which are pick start lag, absence lag, excused lag, holiday lag, late lag, medical lag, vacation lag. For example, pick start lag means how many days have passed since the latest pick starts until the accident happened. After corresponds with our sponsor, we also corporate some domain knowledge where we should try calculate last accident lag and employment lag as well. Therefore, we have 9 lags.

Another way is to calculated frequency of employee absence and frequency of employee extra work prior fixed time length of each accident. We calculate absence frequency and extra work frequency within one week, one month and two month. We also include them as our lag features, total in 15 features.

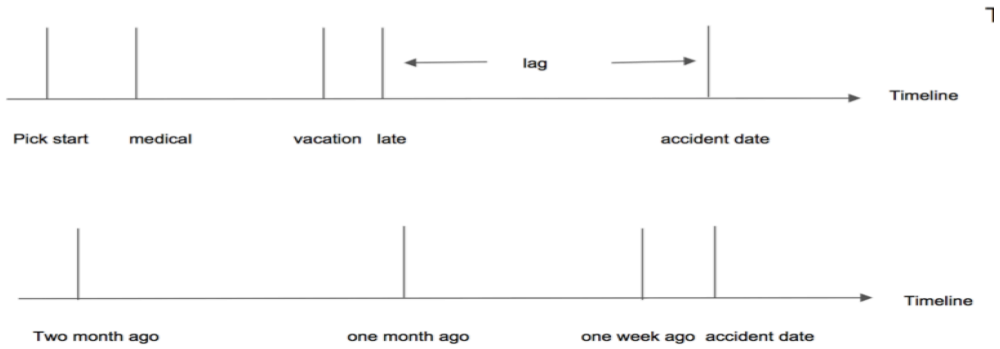

Fig. 9 Lag Features Definition Process

2.2.3. Geolocation Analysis

Geolocation analysis is targeted to extract information in the format of geolocation. The resources in our internal datasets include location on the map (latitude + longitude), the road condition, the traffic condition on map etc.

In addition to the dataset give, we have incorporated external dataset of all US accidents data[1]. The dataset recorded all traffic accidents in United States, mostly involving private vehicle. We selected entries about Rochester city and focused our study.

By researching and discussing, we recognize that traffic accidents geolocation distribution have patterns different from other types of data:

- Accidents tend to cluster around hotspots. Examples include crossings and corners of the road.
- Follow the first one, although locational data (latitude and longitude) are continuous, they are very poorly linked to accidents directly and numerical values usually have no real meaning.
- Most accidents happen on the street. Thus, the Euclidean distance metric is not so appropriate. Manhattan distance, evaluated as the simple sum of distance in two directions could be a better metric for distance on the street.
- Many other variables could be related to the section of the road and road design quality. Many of these can be reflected in the "hotspots".

With these patterns in mind, we conducted a comparative analysis between our two datasets: RTS (bus accidents) and Rochester Accidents (all vehicle, primarily private ones), in the aim to find the similarity or differences. The intuition is that, traffic hotspots (identified with all accidents) can reflect variables related to the geolocation.

In the process of analysis, we created multiple maps to explore the connection. Since we are focusing on our RTS accidents, we believe that a way to "import" information from Rochester accidents can help with our prediction as well as causal analysis. To explore these methods, we created this map to represent the distribution of RTS accidents and whether their locations are hotspots:

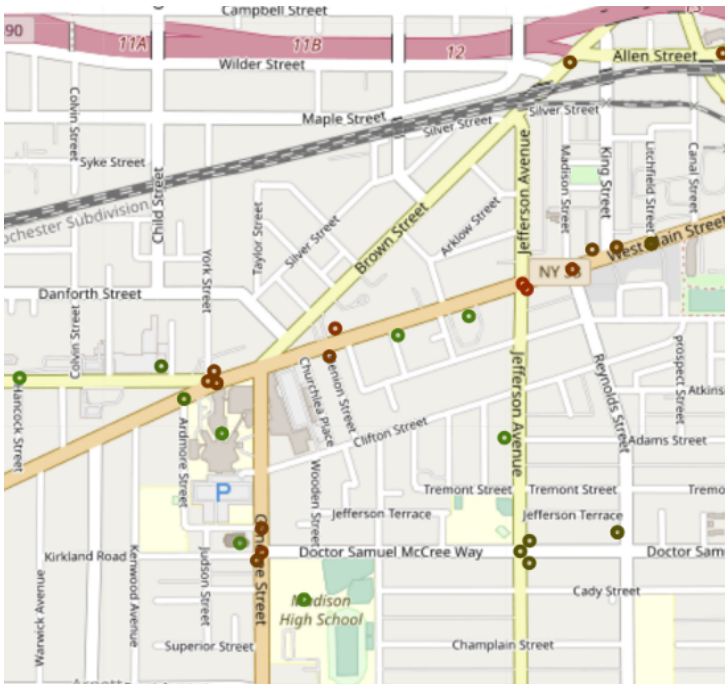

Fig. 10 Hotspot Analysis, section of map

In figure 10, each dot represents an RTS accident, the color represents how "hot" that location is, in other word, number of Rochester accidents in the place. This section demonstrates a curious fact: many accidents happen in a non-hotspot. In other word, RTS buses might be the "only" ones having an accident there!

Viewing this, we decided to incorporate these values to our models in the following process.

In the end, we created this interactive map, with all information in one page. When you hover mouse on a spot, it will show information about that accident.

Fig.11 Interactive Map

2.2.4. Geolocation Features Adding

Based on the location information in the accident table and the external data, we add geolocation information - latitude, longitude, street speed limit and accident frequency near hotspot to the accident data.

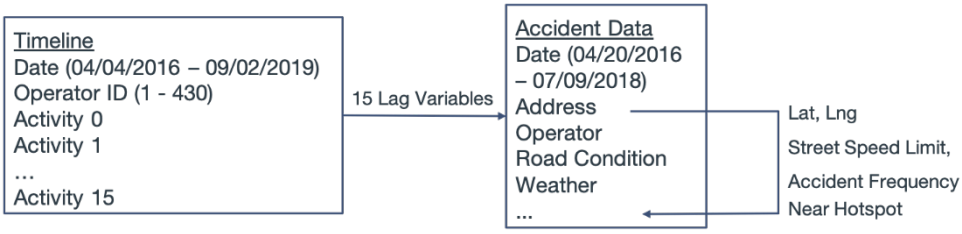

Fig. 12 Lag Variables and Geoinformation Variables Adding Process

2.3. Dataset Cleaning

The accident table is the primary one among all tables because finally we would build our models using it. So, in this step, we cleaned the accident table after adding new variables in it.

We originally had 1159 rows in the accident table. Firstly, we dealt with the missing value. There are 309 missing values in the location variables(LocAddr1, Addr_Type, Neighborhood, latitude, longitude, street speed limit and accident frequency near hotspot to the accident data) which is not a small number, so we decide to delete these rows. There are 617 missing values in the VehNoOccupants and RetrainingDate, so we decided to delete these columns. There are no duplicated values, so we do not deal with it. There are 11 wrong values in the vheyear, so we decided to impute them with the average value. After dropping some rows, we have 842 rows of cleaned accidents data. After that we added the lag features and geolocation features into the cleaned accidents data. Since the lag variables can only be calculated after 2016-4-4, we have to drop accidents during 2014 – 2016. After that, we have 497 rows of the merged accident data and finished the data cleaning.

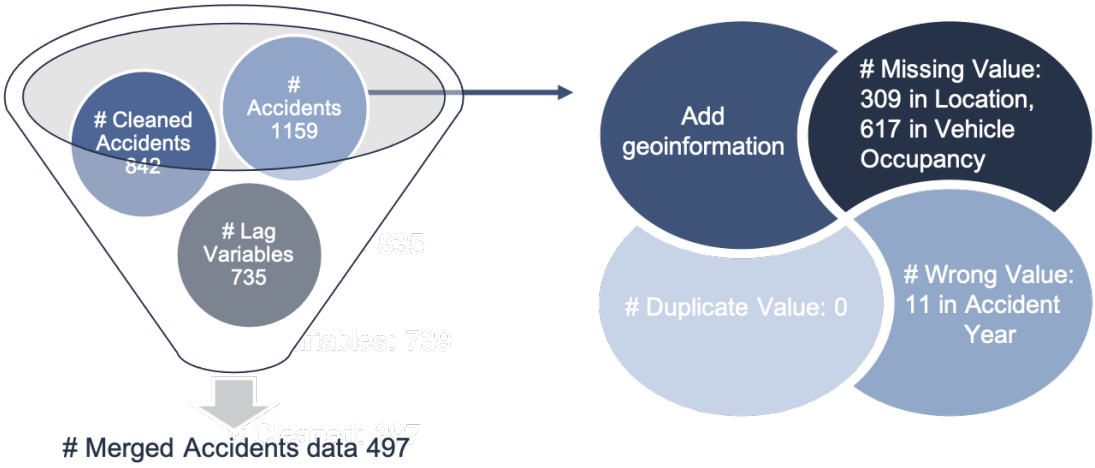

Fig. 13 Data Cleaning Process

In addition, we added the time variables – Year, Month, Date, Time(Morning, Afternoon, Night, Before Morning) because we are expecting the preventability may depends on some season, different date in a month and the time in a day in some degree. In the final accident table, we have 497 rows and 42 columns. Lastly, we encoded all the categorial features to make them available for the model.

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

Fig. 14 Columns In The Final Accident Table

3. Predictive Modeling

3.1. Machine learning models development procedure

- Data Preprocess

Before machine learning models are developed, further data preparation is applied to different models according their distinct needs and characteristics. Data preprocess is not a thing that do once and for all, choices and changes need to be made for specific model. For example, for linear models, we applied categorical data encoding, normalize numerical. But for tree-base models, as the categorical encoding is very likely to cause the explosion of model size, we leave the categorical columns as they are.

And if the data comes from multiple sources, we need a process called multi-model fusion[2]. The process is shown in figure 15. For example, if our data comes from 2 sources, we need to extract features from each data source like we did in our geo-information analysis and timeline analysis. Then if we concatenated all the features at first to let the model make the decision all at once, it is an early fusion; if we make decisions ahead to select features from each data source and combine them at last, it is a late fusion. For our data comes from multiple sources and the feature size is very large comparative to our sample size, thus using late fusion to select features from each source is optimal before fitting into our machine learning models. After the preprocess, the cohesion of both data types as data scales are increased.

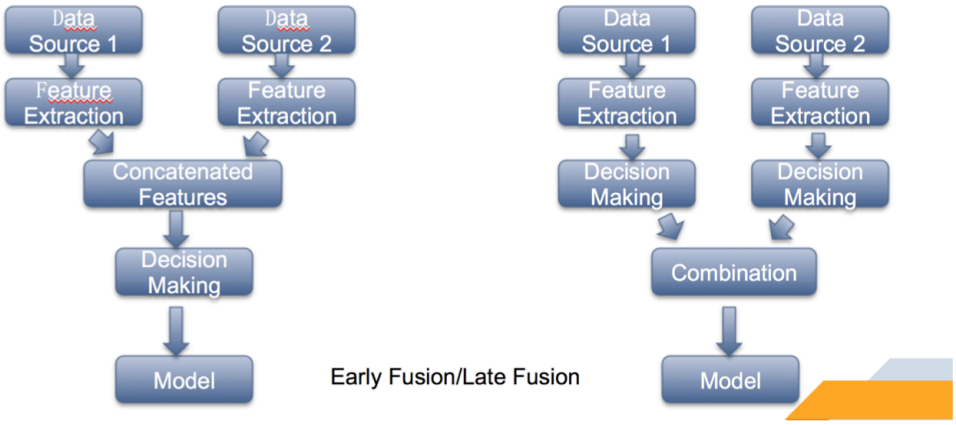

Fig. 15 Multimodal Fusion Process

- Feature Engineering

Then features with high potential contributions are selected using both statistical tests and models. This part is described in detail in the following section.

- Model Fitting

The machine learning models we use includes learning regression, decision trees, random forest, XGBoost and multi layer perceptron(MLP).

- Model Evaluation

When data are put and fitted into machine learning models, we applied matrices to compare their results and performance. Statistical indices like accuracy, ROC value Cohen's kappa are used in the model comparison part. We also calculated the time consumption of different models, as well as plotted the feature contribution and decision boundaries.

3.2. Feature Engineering

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

Fig. 16 Preventable/Nonpreventable Accident Data Features

After data cleaning, we have a table containing all accident along with their features. The 2 keys of accident date and operator ID co-decide the accident event. We have features about the accident such as accident type description,

vehicle, driver, roadway, address and our geo features and time lag features. The features size is large comparable to sample size and not all are useful to deciding our target variable which preventable and non-preventable accidents. Those features needed further selection are marked as red star in figure 16.

First, we consider data leakage. RTS accident data contains some features that will leak future information to machine models, like accident type categories and file numbers. The accident type description is some class we assign after the accident, although they have high correlation with the target, but our objective is to find the cause, those data with possible leakage was removed from our model inputs.

Second, half of accident addresses along with longitude and latitude are missing, if we use these features, our data samples would be reduced by half which is not ideal. Since the variance of addresses is large, thus we discard the address feature and use address type and neighborhood as complement.

Third, for geo features, we used API to acquire speed limit of each accident site and found the nearby number of accidents whole Rochester within different radius. The radius we choose to use is 0.00075(20-40m), 0.0015(40-80m), 0.003(80-160m).

Fourthly, for lag features, we selected out useful features identifying all possible accidents using expletory data analysis and further selected out useful feature identifying preventable accidents using statistical tests.

- Select lag variables that are good indicator for all accidents

For all accidents, we plot lags of each activity to see if there is any pattern appearing. Below is an example of pick start lag plot. From the plot, we can see that after pick start 22 days, there is most likely to happen accident.

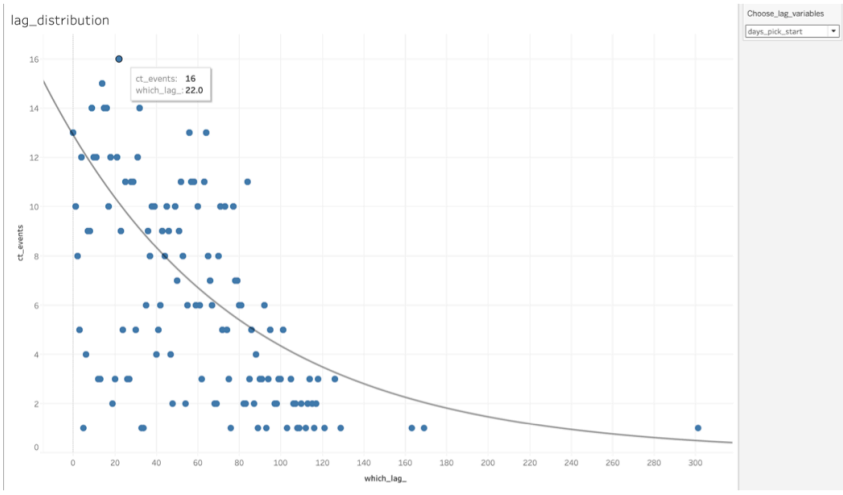

Fig. 17 Pick Start Lag Count For All Accidents

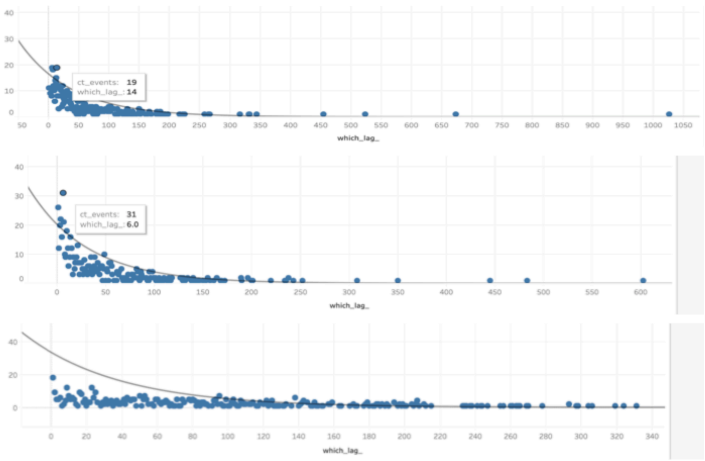

Fig. 18 Significant Lag Variables Predicting All Accidents

Figure 18 shows significant lags have exponential decreasing patterns, meaning that the likelihood of accident decrease as lag increases. Significant lags are pick start lag, absence lag, excused lag, vacation lag and all absence frequency and extra work frequency. They are good indicators for predicting accidents.

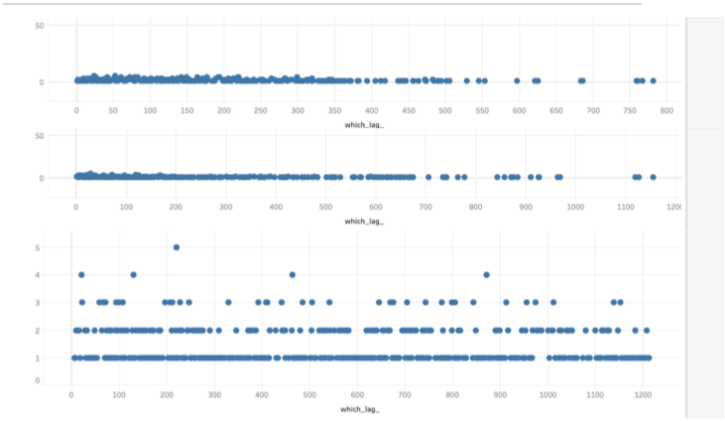

Fig. 19 Insignificant Lag Variables Predicting All Accidents

Figure 19 shows insignificant lags distributions are rather fanned out, meaning that as lag increases, there is no difference in the number of accidents. Insignificant lags are holiday lag, medical lag, late lag, last accident lag and employment lag. Those lags have nothing to do predicting accidents, thus we discard them from our lag features.

- Select lag variables that are good indicator for preventable accidents

If we want to know whether a lag variable is a good indicator for preventable and non-preventable accidents, we need to know the lag distribution between 2 groups are the same or not. We could plot selected lag variables between two groups to have an intuition first. Then use hypothesis tests as rigid procedure.

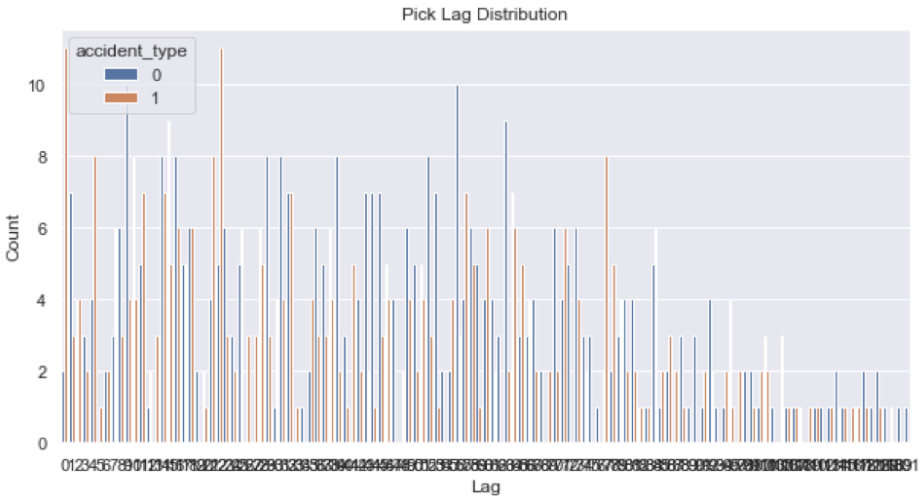

Fig. 20 Pick Start Lag Distribution Between Preventable And Non-preventable Accidents

Figure 20 shows pick start lag distribution between 2 groups where blue line means non-preventable where orange line means preventable. From the plot, the distributions are not Gaussian and the variance is large, thus we need to use non parametric tests which are Wilcoxon rank-sum test and Kolmogorov-Smirnov test(KS test).

	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

Fig. 21 Statistical Test Results of Lag Variables Predicting Preventable Accidents

Figure 21 shows the statistical results and the significant p values are marked as red under our hypothesis that lag variables distributions of 2 groups are not the same. It is good to see how two tests are aligned with each other. We find some lag variables p values are small and the distributions of 2 groups are significantly different, thus they could be used in our machine learning models predicting preventable accidents. Those lag features are pick start lag, vacation lag, frequency of absence within one week, one month and two month.

By far we complete feature engineering. Figure 22 shows the late fusion data after we made decisions of each source and before feeding into machine learning models.

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

Fig. 21 Feature Engineering Results

3.3. Model Fitting

The machine learning models we use includes learning regression, decision trees, random forest, XGBoost and multi layer perceptron(MLP). For concrete execution, please refer to Readme file and relevant codes.

3.4. Model Evaluation

The models performances are shown below.

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%

Fig. 22 Models with 10 Fold Cross Validation Results

Now we can decide which model has performed the best based on accuracy, f1 score, Cohen's Kappa and total time taken for execution. We considered F1-Score as a better metric to judge model performance instead of accuracy. Let's check which model has performed best. If we look into the table, we can see the linear regression has the best Accuracy and f1 result. It is also worth to mention that all models perform better on class 0, which is the non-preventable accidents, that class 1, which is the preventable accidents.

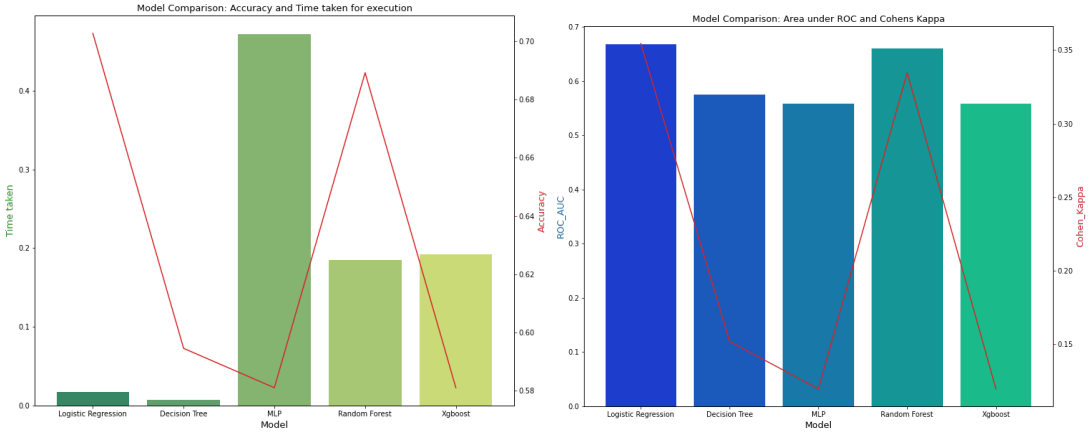

Fig. 23 Performance V.S. Time Consumption

We also compare models on time consumption vs. accuracy. The left figure shows that logistic regression and random forests have high accuracy with less time consumption. For a better decision, we have chosen other metrics like Cohen's Kappa" which is actually an ideal choice as a metric to decide the best model in case of imbalanced datasets. Form the right figure we can see the ROC and Cohen's Kappa metrics agree with each other on choosing LR and RF.

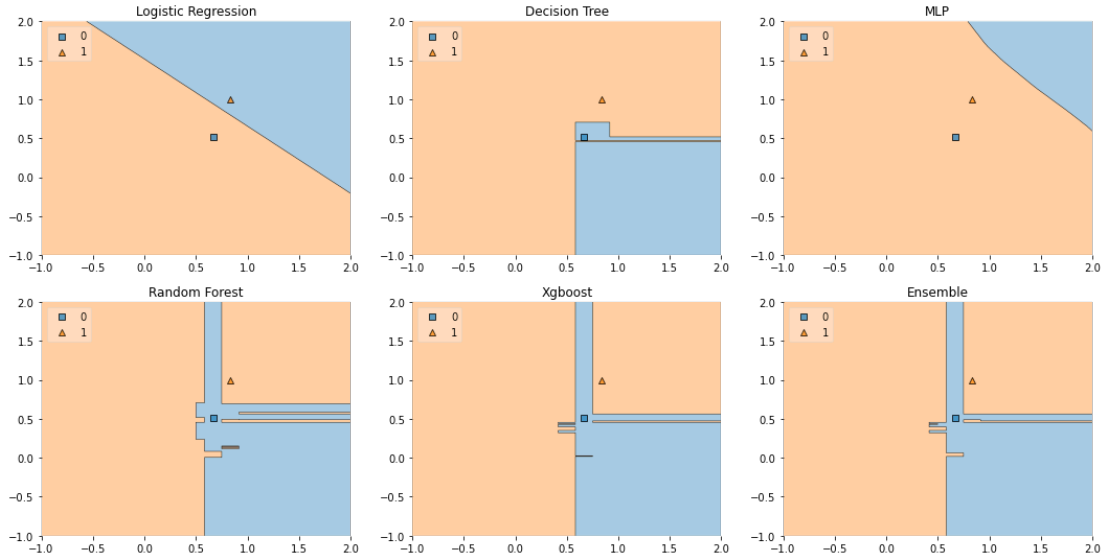

Fig. 24 Decision Boundary On Three Most Important Features

We can observe the difference in class boundaries for different models including the ensemble one. LR has the neatest regional boundary compared to all other models.

We also made some efforts on timeline models. In modeling process, we aimed to predict future accidents using only the operator timeline data. Deep learning models, a CNN and a RNN, as well as autoregression models are developed to fit the timeline data. However, we found the predictive accuracy is as low as flip a coin, and it is hard to find pattern in the operators' timeline behavior. This result shows operator timeline data only is not enough for

predicting accidents from normal days.

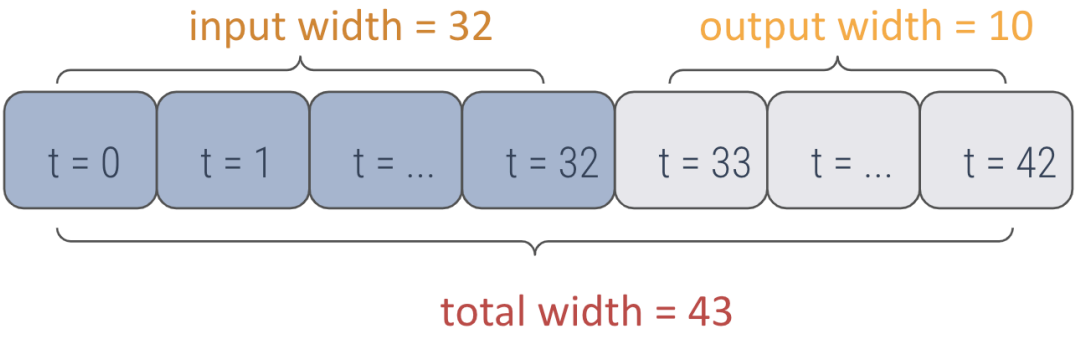

Fig. 25 An Illustration Of Deep Learning Predictive Model

4. Causal Analysis

Following our threads, first we introduce late fusion causal results using tree based models, then we dive into more details using our selected best model logistic regression.

4.1. Tree-Based Models

Since our sample is comparatively small and our objective is to find the potential cause for preventable and non-preventable accidents, thus decision tree is an optimal model to apply. 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. And the most important features can be represented as potential cause if there are no descriptive features.

We use bagging strategies to overcome the drawback of decision tree and over fitting problems. 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.

We also use two boosting strategies. One is Adaboost. AdaBoost is an 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. Another one is GradientBoost. Gradient boosting is an ensemble boosting machine learning algorithm that generalizes decision trees by allowing optimization of an arbitrary differentiable loss function.

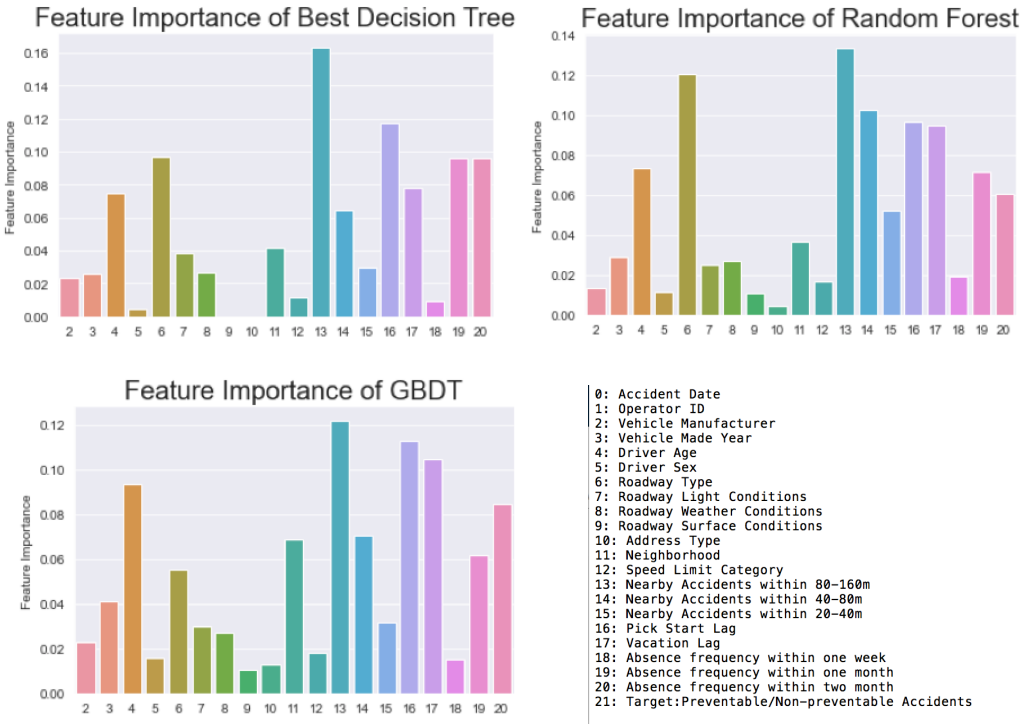

Fig. 26 Tree Based Models Causal Analysis

Figure 26 shows feature importance of three tree based models. It's exciting to see how the three models yields the similar importance rank of all features. The features 4, 6, 13, 14, 16, 17, 19, 20 are noticeably high and they are the potential cause that we are looking for.

Feature 4 is driver sex, which influence the accident type. Feature 6 is roadway type(straight/curve and level/grade/hillcrest) which influence the accident type. Feature 13 & 14 features are the nearby number of accidents within 20-40m and 40-80m. We see the Feature 15 when the radius is 80-160m is less important, meaning that as radius gets larger, the number of accidents becomes less relevant to our accident type. Feature 16 is pick start lag and Feature 17 is vacation lag. After the pick starts or employee completes their vacation, we need to pay more attention to possible accidents. It is interesting to note that the holiday lag does not influence much. Feature 19 & 20 is frequency of employee absence within one and two month. We see the Feature 18 that is frequency of employee absence within one week is not significant. If the employee starts to be absent within one week, it is okay; but as the absence accumulates in one month, which is something we need to pay attention to.

Those are potential cause for preventable accidents that we suggest our sponsor to take consideration in the future based on our tree-based models. For logistic regression models and more detailed coefficients of each cause, we would illustrate in the following section.

4.2. Logistic Regression

4.2.1. Further Data Preprocessing

We conducted one-hot encoding on our data. One-hot encoding is a technique for treating categorical variables. It simply creates additional features based on the number of unique values in the categorical feature. Every unique value in the category will be added as a feature. We can conduct the one-hot encoding here firstly because we want to use logistic regression to build the causal model later, which don't want to treat the numeric values in the categorical features as numerical values but categories. Secondly, the categorical feature is not ordinal, which would not cause multi-collinearity problem.

Below we used the Roadway Weather column as an example, in this case, the RoadwayWeather can be transferred to 4 columns, each one represents a unique value in the RoadwayWeather column.

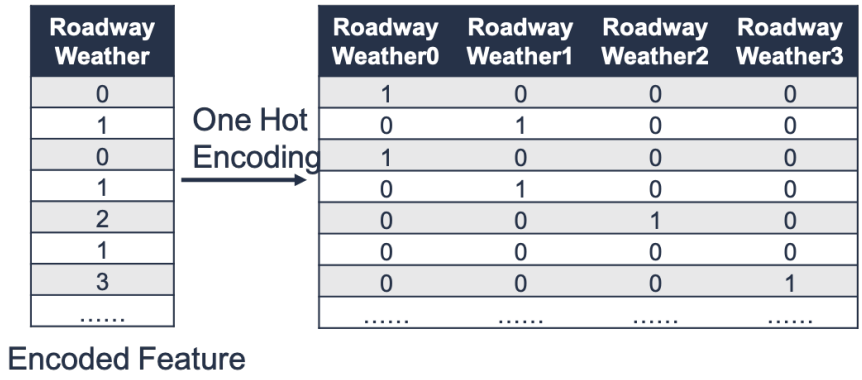

Fig. 27 One-hot encoding example

After the one-hot encoding, we have 1099 columns, which lead to the feature explosion. We will discuss how to deal this problem later.

The data we used is split into training data and test data. The training set contains a known preventability, and the model learns on this data in order to be generalized to other data. We have the test dataset in order to test our model's performance on this subset.

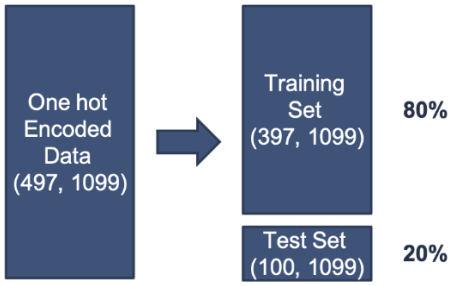

Fig. 28 Train/Test Split

4.2.2. Logistic Regression Model Fitting

Firstly, we would like to briefly introduce the logistic regression model. Logistic Regression is used when the dependent variable(target) is categorical. Logistic regression is used to describe data and to explain the relationship between one dependent categorical variable and one or more nominal, ordinal, interval or ratio-level independent variables. The idea behind the process of training logistic regression is to maximize the likelihood of the hypothesis that the data are split by sigmoid.

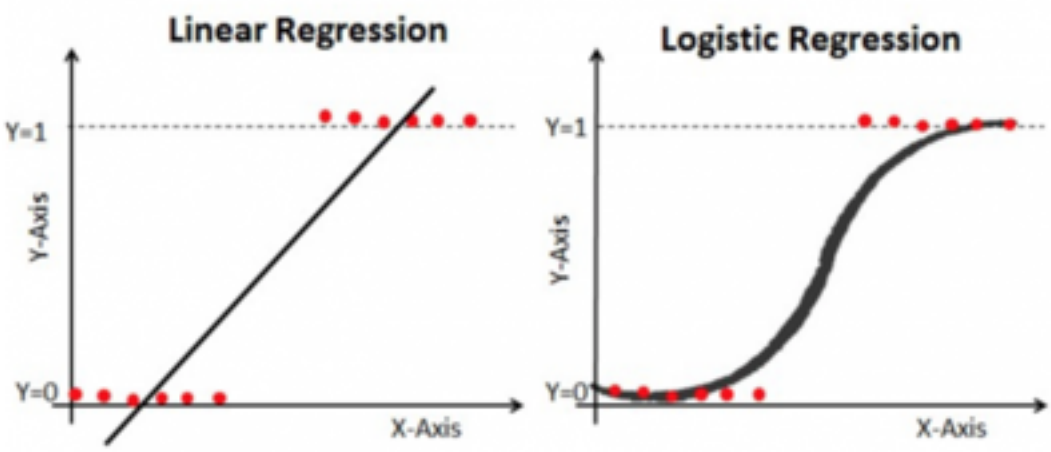

Fig. 29 Logistic regression introduction

To deal with the problem of feature explosion, we performed RFE in a cross-validation loop to find the optimal number of features. The x-axis is the number of features selected, and the y-axis is the "accuracy" score that is proportional to the number of correct classifications. Here after the recursive feature elimination applied on logistic regression with automatic tuning of the number of features selected with cross-validation, we can see the optimal number of features is 183. And we put these selected features in our logistic regression model.

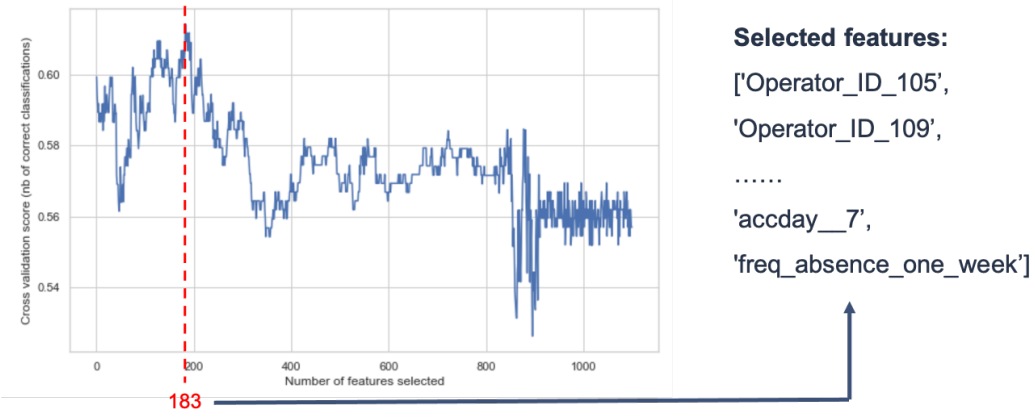

Fig. 30 Cross-validation scores VS number of features using RFE

AUC - ROC curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC is a measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at classification.

Using a threshold of 0.165 guarantees a sensitivity of 0.958 and a specificity of 0.192, i.e. a false positive rate of 0.81. The test accuracy is 0.711 and the AUC is 0.719 that indicates the model is pretty good.

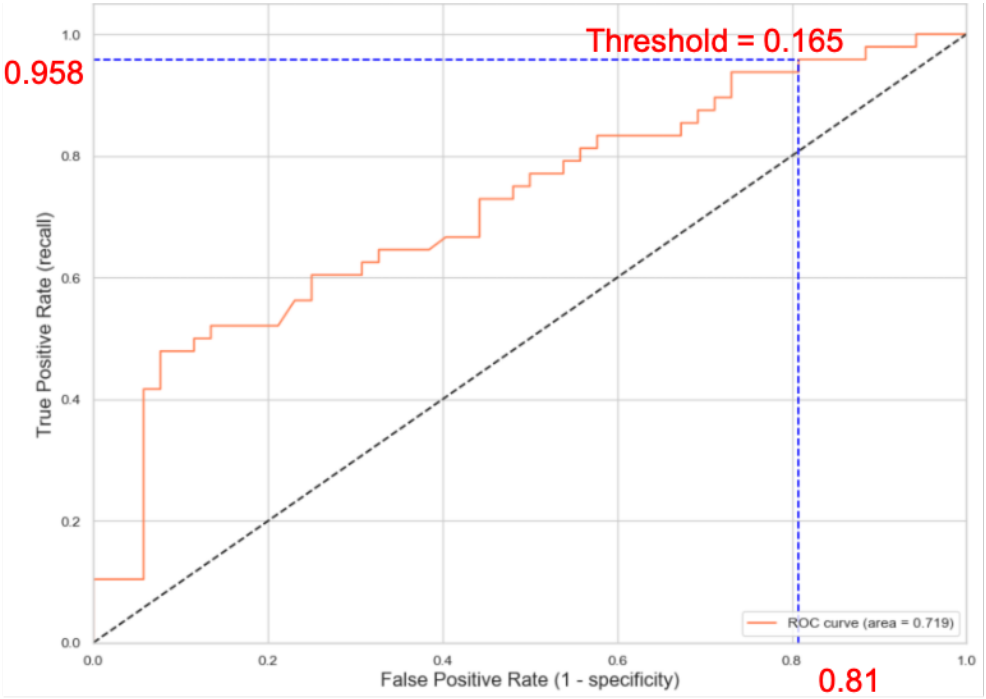

Fig. 32 ROC curve of the logistic regression

4.2.3. Logistic Regression Model Causal Analysis

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

Fig. 33 Logistic Regression Causal Analysis Result

We use the coefficients returned by the logistic regression model as weight and find out causes of preventable accidents. In figure 33, feature column is the features after the one hot encoding and note column indicates the corresponding value of the feature in the original data, which indicates its meaning in the real life.

We can see from the result that the top feature is Neighbourhood_25, unfortunately it is a missing value because it is 0 in the original data, so not a lot meaning of that. But when we look at the followings, we can find some useful insights. The low street speed limit (< 11km/h) can cause the accident preventable, so RTS should tell the operators to care more when they are driving on the street that the speed limit is low. The grade and curve, curve and level roadway can also cause the accident preventable, RTS should tell the operators to care more when they drive on these type of roads. Furthermore, the operators with ID = 249 and 182 can cause the accident preventable, so RTS can investigate on them and maybe train them more on how to avoid the preventable accidents. In addition, RTS can tell their operators to watch out when they drive on the 1372 E. Main Street.

Although the result tells us some insights, there is still a limitation. The weight (the coefficients returned by the logistic regression) is low overall, which means the causal relationship between our variables and the preventability is very low. We cannot find a strong causality that causes the result. Since that, we tried the logistic regression after covariance term in the following, hoping to get a more robust result.

4.3. Logistic Regression After Covariance Term

There are usually two perspectives to address causality. One is used above, regarding it as a more insightful analysis based on statistical and machine learning models and used primarily in natural science. The other one, which is used more prevalent in social science, regards causality as a more delicate relationship that cannot be evaluated without pre-assumptions. This section of the work looks at the second approach primarily.

4.3.1. Causality Definition and Potential Problems

Under the social science context, causality is usually defined as the treatment effect of a certain incident (estimated as an “independent variable X”) on another dependent result (estimated using “dependent variable Y”). Hold all else equal or irrelevant.

$$Treatment\ effect\ of\ X\ on\ Y\ =\ [Y\ when\ X\ =1]\ -\ [Y\ when\ X\ =\ 0]$$

Thus ideally, in the RTS scenario, we care about causal effect of any variable Vi in terms of:

$$\text{causal effect of Vi on P} = [Probability\ of\ \ P\ if\ Vi\ = 1]\ -\ [Probability\ of\ \ P\ if\ Vi\ = 0]$$

where *P =*P*reventable accident

That leads to the first problem that we are not provided with non-accidents data. Therefore, we would define causality on preventable accidents in term of comparison to non-preventable accidents.

$$\text{causal effect of Vi on P} = \frac{\text{Probability of P} \mid \text{Vi} = 1}{\text{Probability of non} - \text{P} \mid \text{Vi} = 1} \Big/ \frac{\text{Probability of P} \mid \text{Vi} = 0}{\text{Probability of non} - \text{P} \mid \text{Vi} = 0} - 1$$

To illustrate our definition, for example, we might think that "road condition is slippery" can cause accidents (both preventable and non-preventable). We can simply assume these numbers:

Case	Probability of Accidents	Do we have the data?
Preventable accidents on a slippery road [P | Vi=1]	1%	No
Non-Preventable accidents on a slippery road [not-P | Vi=1]	2%	No
[P | Vi=1]/ [not-P | Vi=1]	Ratio 1= 0.5	Yes
Preventable accidents on a not slippery road [P | Vi=0]	0.5%	No
Non-Preventable accidents on a not slippery road [not P | Vi=0]	0.5%	No
[P | Vi=0]/ [not-P | Vi=0]	Ratio 2 = 1	Yes

Fig. 34 Causality Definition Example

Then we can estimate using the definition above. The causal effect of slippery road is -50%. The intuition is that, "slippery" resulted in much more non-preventable accidents than preventable ones. Thus, there is very few things we can do about. We should instead focus on scenarios with more preventable accidents.

4.3.2. Independent Variables Selection

A huge problem in our causal analysis is the covariance terms. A typical example is the "road lighting condition", "road weather", and "road condition". We can

easily foresee that the road gets darker and more slippery on a snowy day. This can lead to a totally biased estimation of the effect of any one of them (correlated variables).

To address this problem, a typical measure and the practical one we have here, is to reduce number of input variables. We would remove variables correlated with other variables. In the end, we decided to remove variables with variance inflation factors greater than 5. This leaves us with independent variables and an overview of features is shown in figure 35.

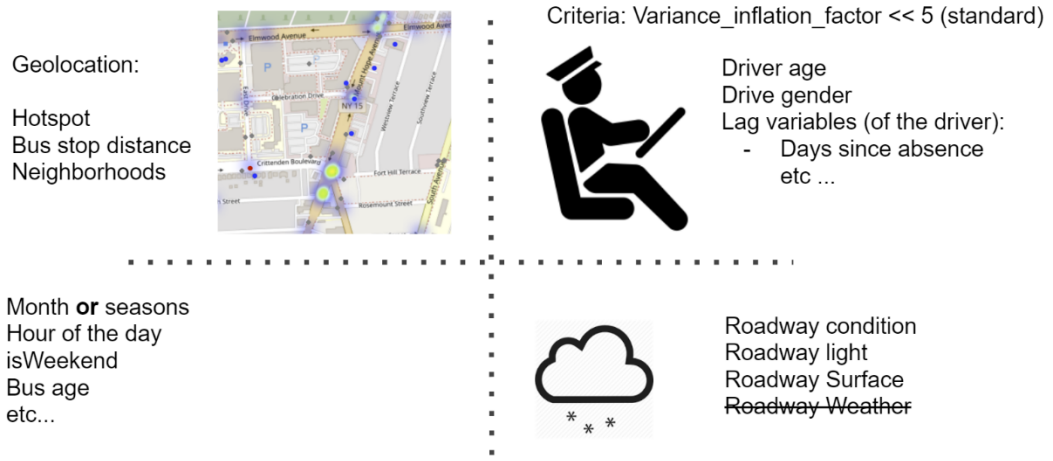

Fig. 35 Independent Variables Overview

The side effect is losing "accuracy". In other word, we are considering less variables and the model is less complex.

4.3.3. Estimation of Causal Effect

Using also a logistic regression, we would not be using any feature selection process. And the weights of the model can be interpreted almost directly as our defined causal effect.

The full variable weights and causal effect of each variable can be seen in the appendix. If certain variable is of particular interest, please refer to the appendix.

5. Conclusion and Future Work

Through this project, our team have analyzed the data provided by our sponsor and external data that we can find. We explored data in timeline analysis and geolocation analysis. And based on the analysis, we developed predictive models and causal models. We have given our best effort in help improving RTS services based on the data we can obtain.

Many parts of our results are quite restrictive, due to the limited amount of data, correlations among variables, and sparse records. Well we wouldn't want more accidents, but small dataset did limit our ability to create precise models and quantitively address the problem.

For the future works, we suggest circumventing the problem of small dataset. Approaches like transferred learning (learn patterns of bus accidents somewhere else and adopt it to our setting), few shot learning (learning with very few examples) in the data science field, as well as a deeper dive in the social science literature might help with similar studies.

6. Acknowledgements

In this year of 2020, our capstone project has faced many challenges like the access of more information (library resources), online communication as well as impossible field trips. Yet our team, mentors, sponsors all worked together to tackle all these difficulties.

First, we would like to thank our sponsor RTS to join our capstone program and offer us an opportunity to serve the company's need and improve Rochester public transportation together. It is a pleasure to work with Wray Robert, Swift Adam and Dobson Christopher D. We shared the same integrity to RTS, professional attitudes as well as the enthusiastic for the intersection area of data science and public transportation.

Mostly, we would like to thank Professor P.J. Fernandez, Professor Ajay Anand as well as our teaching assistant Wade Bennett who carried us along the whole course. Your attentive caring, professional instruction as well as thoughtful arrangement are something we would remember after graduation.

Lastly, our team member Xiaoran Li, Weiran Lin, Melissa Chen and Weinan Hu, we are glad we made it. Besides all the requirements, we managed to dictate our team own document tracking all valuable details. Hardships we went through yet success we share, on the moment of completion, Meliora is our best award.

7. Reference

[1] Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. "A Countrywide Traffic Accident Dataset", 2019.
"This is a countrywide car accident dataset, which covers 49 states of the USA. The accident data are collected from February 2016 to June 2020, using two APIs that provide streaming traffic incident (or event) data. These APIs broadcast traffic data captured by a variety of entities, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks. Currently, there are about 3.5 million accident records in this dataset. Check here to learn more about this dataset."

[2] Timeseries Analysis Guest Letcure 'Supervised and Unsupervised Temporal Multimodal Fusion '. Professor Ajay Anand & Professor Edgar.A.Bernal
Causal effect: exp(coef)-1

8. Appendix

	Coef	Causal Effect: exp(coef)-1	Pvalue
Neighborhood__11_vs_25	-38.0955	-100%	1
Neighborhood__1_vs_25	-28.8566	-100%	0.999982
Neighborhood__6_vs_25	-23.0623	-100%	0.999717
Neighborhood__9_vs_25	-22.7109	-100%	0.999721
SpeedLimitCategory__0_vs_4	22.67543	704388122894%	0.999721
Neighborhood__40_vs_25	-22.5325	-100%	0.999723
Neighborhood__30_vs_25	-21.6269	-100%	0.999734
Neighborhood__29_vs_25	-21.5774	-100%	0.99965
Neighborhood__31_vs_25	-21.2006	-100%	0.999739
Neighborhood__33_vs_25	21.0412	137428154363%	0.999741
Neighborhood__32_vs_25	-20.8897	-100%	0.999314
SpeedLimitCategory__1_vs_4	-20.4647	-100%	0.999748
Neighborhood__8_vs_25	-20.1112	-100%	0.998777
Neighborhood__26_vs_25	-18.7616	-100%	0.997792
Neighborhood__22_vs_25	-18.4747	-100%	0.997901
Neighborhood__12_vs_25	18.11598	7373404738%	0.997555
Neighborhood__39_vs_25	17.64783	4616947537%	0.996898
Neighborhood__13_vs_25	-13.2611	-100%	0.980854
freq_extra_one_week	-2.91804	-95%	0.40597
RoadwayLightConditions__3_vs_2	2.732165	1437%	0.040893
Roadway__3_vs_5	2.402995	1006%	0.000778
Roadway__0_vs_5	2.370803	971%	0.044162
RoadwaySurface__2_vs_0	-2.08105	-88%	0.0136
Neighborhood__16_vs_25	-1.61257	-80%	0.187994
Neighborhood__41_vs_25	-1.43883	-76%	0.236617

Neighborhood__7_vs_25	1.371539	294%	0.318244
Neighborhood__37_vs_25	-1.35841	-74%	0.142645
SpeedLimitCategory__5_vs_4	1.322696	275%	0.075554
Neighborhood__10_vs_25	1.232273	243%	0.198972
isWeekend__True_vs_False	1.182071	226%	0.050522
Neighborhood__34_vs_25	-1.09753	-67%	0.396622
RoadwaySurface__1_vs_0	-1.05821	-65%	0.507356
Roadway__6_vs_5	-1.01848	-64%	0.214463
SpeedLimitCategory__2_vs_4	-0.98962	-63%	0.311941
Neighborhood__24_vs_25	0.879617	141%	0.490868
SpeedLimitCategory__6_vs_4	-0.71789	-51%	0.18064
Roadway__1_vs_5	0.672749	96%	0.223978
RoadwayLightConditions__0_vs_2	0.648518	91%	0.178473
Neighborhood__2_vs_25	-0.62469	-46%	0.100388
Neighborhood__3_vs_25	0.58447	79%	0.452442
Neighborhood__14_vs_25	-0.57853	-44%	0.309277
Neighborhood__0_vs_25	0.514802	67%	0.507359
Neighborhood__36_vs_25	-0.38368	-32%	0.581761
freq_absence_one_week	-0.36172	-30%	0.556219
Neighborhood__17_vs_25	-0.35713	-30%	0.522337
SpeedLimitCategory__3_vs_4	-0.31224	-27%	0.399783
Neighborhood__27_vs_25	0.245657	28%	0.809414
Neighborhood__23_vs_25	0.24072	27%	0.669758
Neighborhood__42_vs_25	0.234813	26%	0.833052
RoadwayLightConditions__4_vs_2	0.212557	24%	0.908607
Roadway__2_vs_5	-0.17842	-16%	0.86083
RoadwayLightConditions__1_vs_2	-0.14554	-14%	0.683601
Roadway__4_vs_5	-0.0427	-4%	0.886542
n_accidents_in_0.0015	-0.03131	-3%	0.002083
Neighborhood__35_vs_25	-0.00772	-1%	0.992345
n_accidents_in_0.006	0.00605	1%	0.000947
days_pick_start	-0.00546	-1%	0.175126
days_absent	0.003137	0%	0.224491
days_vacation	-0.00269	0%	0.112299
days_excused	-0.00269	0%	0.237716
days_last_accident	-0.00179	0%	0.023322
days_late	-0.00039	0%	0.636741
meters_from_bus_stop	-0.00021	0%	0.565625