Predicting Severity in US Traffic Accidents

—

Jingxuan Li * Jialiang Wei †

December 13, 2020

Abstract

This research employed three different machine learning algorithms to U.S. national traffic accident dataset, which covers 49 states of the United States and is continually collected from February 2016 till now, to study the difference in different models, impacts of different factors, geographical locations, environmental stimuli and other relevant factors on the severity of accidents. The main goal of this research is to examine the significant factors that impact the severity of accidents, and to discuss how federal and local governments, as well as drivers can make changes to prevent or reduce the impact of accidents in different environments. This report also selects some states to conduct model researches, and discuss the similarities and differences among states and the entire nation.

*Cornell University, *jl4267@cornell.edu*
†Cornell University, *jw2684@cornell.edu*

1 Explanatory Data Analysis

1.1 Data Description

Our project employs the *US Accidents* [1] [2] from Kaggle.com. This is a countrywide traffic accident dataset, which covers 49 states of the United States. The data is continuously being collected from February 2016, currently there are about 3.5 million accident records. For every accident record we have data on the source of the accident report, description of the event, severity of the accident, start and end time, latitude and longitude information, zip code, time-stamp of weather observation record, temperature, wind chill, humidity, air pressure, visibility, wind direction and speed, presence of crossing, railway, traffic signal, junction, etc., and the period of the day. We will build learning models to study how different factors impact the severity of a car accident.

1.2 Data Visualization

1.2.1 Accidents Summary

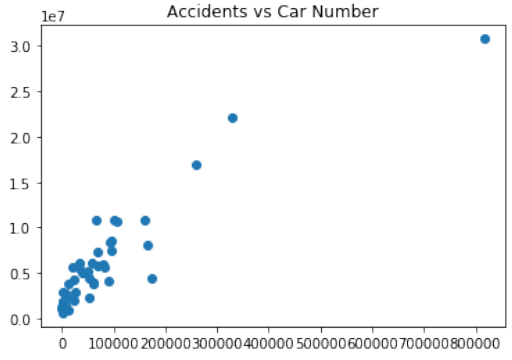

Figure 1

Figure 2 shows the number of accidents in different states recorded in the dataset. Based on the bar chart, California has much more accidents than the other states. We are interested whether the number accidents is correlated with other factors.

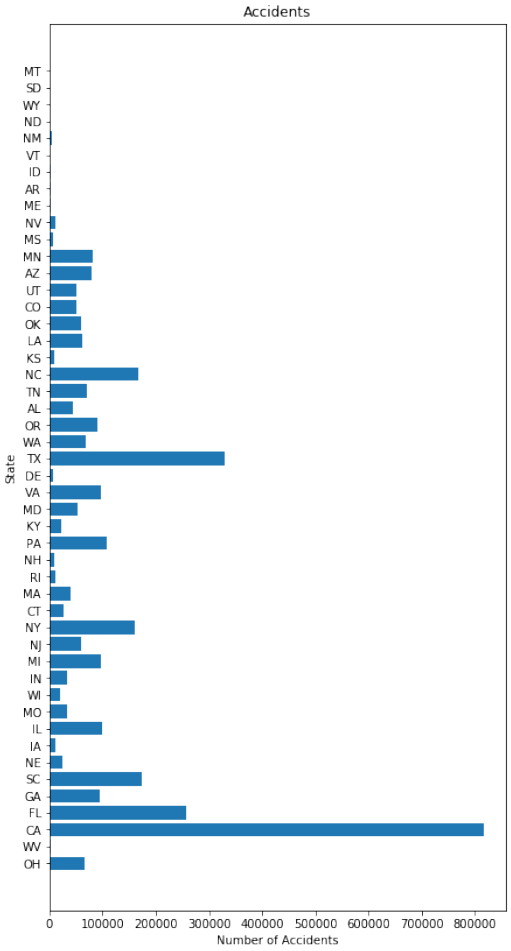

Figure 2: Number of Accidents

Figure 1 shows the scatter plot of the number of accidents versus the total number of automobiles in each state. We can see there is a strong positive relationship between these two variables and the correlation coefficient is 0.9134, which is consistent with what we expect that the more automobiles in a state, the higher possibility of traffic accidents.

Figure 3 compares the number of accidents (left) and the numbers of automobiles weighted accidents in each state (right). It shows that the except for the state of SC (South Carolina) has a higher value than the other states, other states have the similar level as shown in

[1] Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. "A Countrywide Traffic Accident Dataset.", 2019.

[2] Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights." In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.

figure 1.

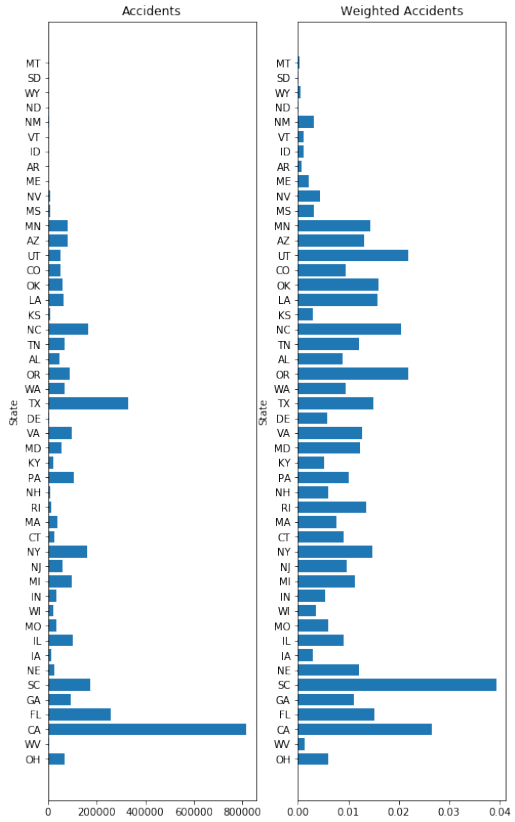

Figure 3

1.3 Data Preprocessing

1.3.1 Missing Values

In the dataset, the missing values appears in Temperature (F), Humidity (%), Visibility (miles), Weather_Condition (rain, snow, thunderstorm, fog, etc.), and Sunrise_Sunset (day or night). Since the dataset is larger enough with 3 513 617 records in total, the records with missing values in any category are dropped. We still have 3 414 253 records with complete information which is large enough for us to do analyses.

1.3.2 Ordinal Values

The response variable in the model is the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic and 4 indicates a significant impact on traffic. As severity is a ordinal value, we will choose a model that can be adapted to fit a multi-classification response variable.

1.3.3 Nominal Values

The column "Start_Time" shows start time of the accident in local time zone. We classify them into four time windows (0 AM - 6 AM, 6 AM - 12 PM, 12 PM - 6 PM, and 6 PM - 12 AM) and use one-hot encoding. The column "Street" shows the street name in address field. We believe that the accidents happened on interstate highways may be more severe than others, so we use one-hot encoding to classify the street with a name starting with "I-" which represents the interstate street. The columns "side" which contains nominal data and shows the relative side (right or left) in the accident on a street, "Weather_Condition", and "Sunrise_Sunset" contains nominal data and shows the period of day (day or night). We use one-hot encoding for each of them to make their values more expressive.

1.3.4 Additional Note

Since the data is recorded chronologically, we shuffle the data and split into training data and test data to avoid autocorrelation and overfitting. We tried with ten-fold cross-validation, however, the dataset is so large that our computational power is not capable.

2 U.S. Car Accident Modeling

2.1 Model Selection

Based on our parameter selection from data preprocessing and data encoding, we predict that there will be significant overfitting issues with the data. To validate our thought, we implemented in-class algorithms of ordinary least square linear regression (OLS) as well as multinomial LASSO regression. Indeed, we discover there are significant overfitting issues, and LASSO is

ineffective of reducing overfit and it cannot perform parameter selection. When we run LASSO regression on our dataset, the lambda yields meaningless results (i.e. we get our optimal lambda = 0, which is the same as OLS). Due to overfitting, the OLS/LASSO only achieved an accuracy of 0.587 in our test dataset (refer to table 1). As we only have 4 different severity classes as our response parameter, using only random permutation would achieve on average an accuracy of 0.25. Thus, we can see that OLS/LASSO do not improve the prediction accuracy by much. In addition, due to the large volume of data (1.98 Gigabytes, 481,409,673 data points), we can only consider models that are extremely efficient and more robust in preventing overfitting. Thus, we decide to use the gradient boosting method (R package: xgboost) to fit our data.

2.2 Gradient Boosting Algorithm (XGB)

1: Initialize

$$\mu^{(0)}(x) = \arg\min_{\theta} \sum_{i=1}^{n} L(y_i, \gamma) \tag{1}$$

2: For m = 1 to M

a: For i = 1,2,...,n compute

$$r_i^{(m-1)} = -\frac{\partial L(y_i, \mu(x_i))}{\partial \mu(x_i)}\bigg|_{\mu=\mu^{(m-1)}} \tag{2}$$

b: Fit a regression tree to the current residuals $r_i^{(m-1)}$ giving terminal regions R_{jm},j = 1,...,J compute

c: For j = 1,...,J compute

$$\gamma_{jm} = \arg\min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, \mu^{m-1}(x_i) + \gamma) \tag{3}$$

d: Update $\mu^m(x) = \mu^{m-1}(x) + v \sum_{j=1}^{J} \gamma_{jm} I(x \in R_{jm})$

3: Output $\hat{\mu}(x) = \mu^M(x)$

Due to page limit, for more information, please refer to *The Elements of Statistical Learning pp 359-361.* [3]

[3]Hastie, T., Tibshirani, R., & Friedman, J. (2020). The

Thus, we can see that to fit a gradient boosting algorithm, we need tuning process. Given limited computational power, we tuned the below factors: max_depth (the complexity of each fitting tree, positively correlated with J in above algorithm), eta (learning rate, v in the algorithm), subsample (subsample a portion of the data in our training data in each iteration), colsample_bytree (subsample a portion of input parameters in each iteration).

2.3 XGBoost Tuning Parameter Selection

As our response variable (severity) is an ordinal value, we defined two error functions for multi-classification models that calculate the number of differences between our model prediction output and observed data, the percentage number of errors and percentage size of errors.

We tuned a total of 9 models, with the tuning parameter as follows: (max_depth, eta, subsample, colsample_bytree) = (5, 0.3, 0.9, 1), (11, 0.3, 0.9, 1), (20, 0.3, 0.9, 1), (20, 0.3, 0.9, 0.8), (20, 0.3, 0.9, 0.5), (20, 0.3, 0.5, 0.8), (20, 0.5, 0.5, 0.8), (20, 0.5, 0.9, 0.8), (20, 0.7, 0.9, 0.8). Figure 4 shows the percentage number of errors and percentage size of errors on the test dataset. The fitted models are in the same sequence as the sequence from above.

We can see that model 8 with tuning parameters (20, 0.5, 0.9, 0.8) has the smallest in both number of errors and size of errors. Thus, given our limited computational power, we chose this model to study the dataset's relative importance factors (to tune these 9 models, we ran for 15 hours with 3.8GHz CPU and 32G memory).

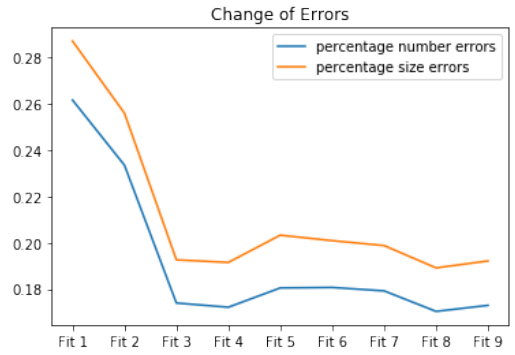

Figure 4: Errors of Different Tuning Models

US Total

Model	Level of Severity	Evaluation Metric		
		Overall Accuracy	Balanced Accuracy	F1 Score
Using OLS	1	0.587	—	—
	2		0.488	0.72414
	3		0.494	0.25091
	4		—	—
Using XGB	1	0.83	0.636	0.399
	2		0.805	0.883
	3		0.822	0.745
	4		0.616	0.355

*Note the LASSO is not included as we do not have enough computational power

Table 1

2.4 Model Comparison

From the above discussion, we can see that using XGB can achieve a test set accuracy of 0.83, compared to the 0.587 accuracy that OLS and LASSO achieves. For more comprehensive comparisons between OLS/LASSO and XGB, please refer to table 1. Note that in every evaluation criterion, XGB outperforms OLS/LASSO by a lot. In addition, due to the fact that there are only few data that have severity level 1 and 4 in the dataset, OLS/LASSO cannot make prediction in such cases, whereas XGB still can make good predictions and achieve a relatively stable results across all severity levels.

2.5 Relative Importance Factors

By fitting with our tuned gradient boosting model, we can see the relative importance of each factor. Figure 5 shows the percentage gains in loss function for each input parameters to the severity of car accidents.

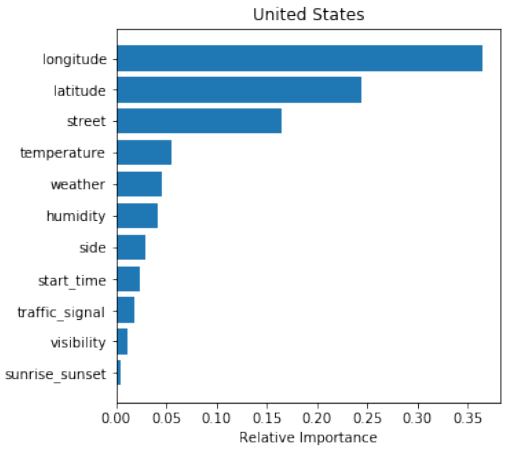

Figure 5: Relative Importance of Input Parameters

We can see that geographical location (latitude and longitude), types of streets (Interstate highway or others), temperature, weather and humidity are some of the most important factors to the severity of a car accident. However, some of other factors that we intuitively regard also important, such as sunrise_sunset (day/night), visibility and traffic signal are not as important. This shows that drivers already pays close attention when the road condition is complicated (in dark or foggy conditions), and drivers follow traffic signals closely. However, the high importance in location and weather shows authorities should invest more in certain regions for road maintenance such as snow removal. The high importance in types of streets also shows that drivers should pay close attention with high speed limit roads.

3 Car Accident Modeling in Five States

3.1 State Selection

From the previous discussion in explanatory data analysis, we discover that there are significant disparities between states. Thus, we examine five different

Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics) (2nd ed.). Springer.

states and try to see if there is any new findings. We select South Carolina, New York, California, Texas and Minnesota to investigate. We select South Carolina as it has the highest weighted number of car accidents. For the other four states, our selection is based on the geographical locations they are in the United States, as we have seen from the relative importance factors analysis of the U.S. modeling that location plays an important role in determining the severity. We select New York as the representative state of the northeast region, California as the representative of the west region, Texas as the representative of the south region and Minnesota as the representative of the middle region.

3.2 South Carolina

3.2.1 Model Comparison

South Carolina

Model	Level of Severity	Evaluation Metric		
		Overall Accuracy	Balanced Accuracy	F1 Score
Using OLS	1	0.689	—	—
	2		0.497	0.809
	3		0.499	0.185
	4		—	—
Using LASSO	1	0.865	—	—
	2		0.769	0.92
	3		0.77	0.65
	4		—	—
Using XGB	1	0.918	0.575	0.25
	2		0.89	0.95
	3		0.893	0.812
	4		0.56	0.207

Table 2

We run OLS, LASSO and XGB models to the car accident data in South Carolina, and LASSO here outputs different result compared to the OLS, as indicated in table 2. We can see that for all three models, they all achieve better results compared to the results of the entire country. However, both OLS and LASSO still cannot make predictions on data that have severity level 1 and 4. For XGB, we can see that it achieves better results in balanced accuracy and F1 score in severity level 2 and 3, but less robust results in severity level 1 and 4. This might due to the less data available in severity level 1 and 4.

3.2.2 Relative Importance Factors

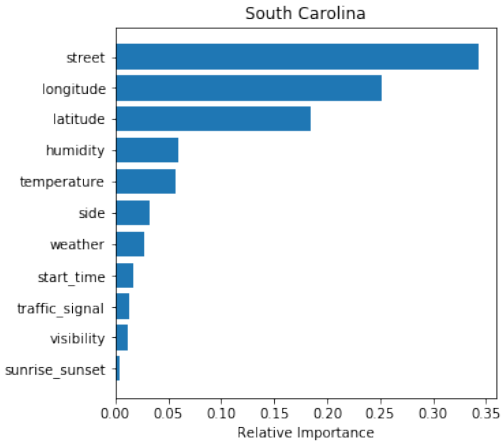

Figure 6: Relative Importance of Input Parameters

From figure 6, we can see that it deviates from the national's modeling results. The type of street is the most important factor in determining the level of severities. The longitude and latitude play the second and the third important factor in determining the level of severities. Humidity, temperature, side of the road, weather, time of the accident, traffic signals, visibility and sunrise/sunset are other factors that are less important. Thus, the South Carolina state government should fully investigate its Interstate highways, its speed limits and road quality to ensure safety. In addition, the state government should examine it fully and thoroughly of all the Interstates, to reduce the effect by latitude and longitude.

3.3 New York

3.3.1 Model Comparison

We run OLS, LASSO and XGB models to the car accident data in New York. We can see that for OLS has worse results compared to the results of the entire country, but LASSO and XGB has better results. This might indicate that the severity of car accident may be impacted by other factors more than these we investigate. However, both OLS and LASSO still cannot

New York

Model	Level of Severity	Evaluation Metric		
		Overall Accuracy	Balanced Accuracy	F1 Score
Using OLS	1	0.515	—	—
	2		0.487	0.628
	3		0.492	0.347
	4		—	—
Using LASSO	1	0.653	—	—
	2		0.6	0.757
	3		0.614	0.446
	4		—	—
Using XGB	1	0.832	0.714	0.561
	2		0.827	0.868
	3		0.845	0.803
	4		0.69	0.523

Table 3

make predictions on data that have severity level 1 and 4. For XGB, it still has the same property as shown in South Carolina's modeling, which it has more predictive power in level 2 and 3 than in level 1 and 4.

3.3.2 Relative Importance Factors

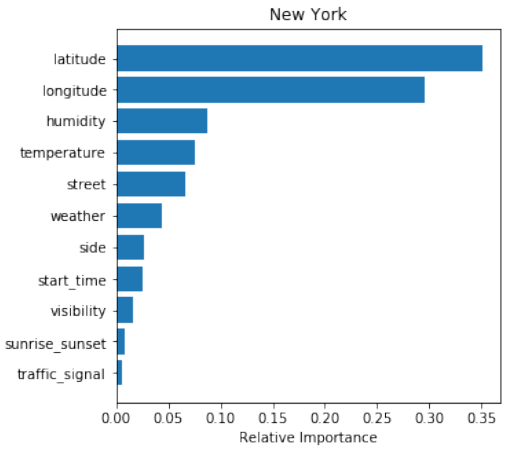

Figure 7: Relative Importance of Input Parameters

From figure 7, we can see that it is similar but not exactly the same as the national's modeling. Comparing to national's modeling and the South Carolina's modeling. For New York state, latitude has a larger relative importance impact than longitude, which is not the case for the national's and South Carolina's model results. Thus, it is susceptible that the latitude and longitude's relative importance are influenced by the geographic shape of the region. In addition, we see that the street type is not as important as in national and South Carolina's model. This shows that the local government should examine more on its local roads than Interstate roads, as with different speed limits, local roads still has a comparable level of severity when comparing Interstate and local roads accident. However, we see that humidity, temperature and weather has a high relative importance than in South Carolina's model. Thus, it is susceptible that north region's level of severity is more influenced by natural causes.

3.4 California

3.4.1 Model Comparison

California

Model	Level of Severity	Evaluation Metric		
		Overall Accuracy	Balanced Accuracy	F1 Score
Using OLS	1	0.63	—	—
	2		0.495	0.76
	3		0.498	0.214
	4		—	—
Using LASSO	1	0.756	—	—
	2		0.646	0.842
	3		0.656	0.493
	4		—	—
Using XGB	1	0.855	0.721	0.596
	2		0.813	0.901
	3		0.822	0.751
	4		0.573	0.246

Table 4

We run OLS, LASSO and XGB models to the car accident data in California. We can see that for all three models, they all achieve better results compared to the results of the entire country, showing the impact of local fits. However, both OLS and LASSO still cannot make predictions on data that have severity level 1 and 4. For XGB, it still has the same property as shown in other states' modeling, which it has more predictive power in level 2 and 3 than in level 1 and 4.

3.4.2 Relative Importance Factors

Similar to above findings and inference, the latitude has larger relative importance than longitude with California, as suspected due to the more north/south distance than east/west of the state. Here, types of street has larger impact than natural conditions (weather, temperature, humidity), which means that the state government should examine more regulations (speed

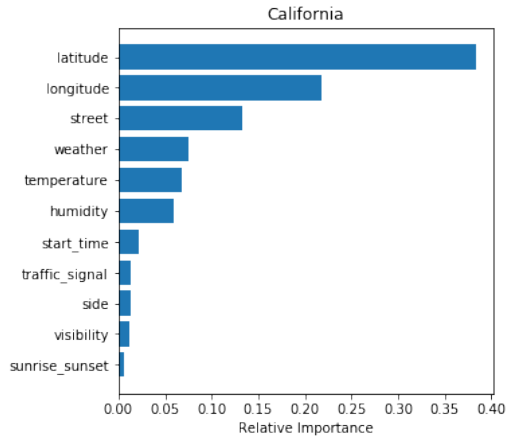

Figure 8: Relative Importance of Input Parameters

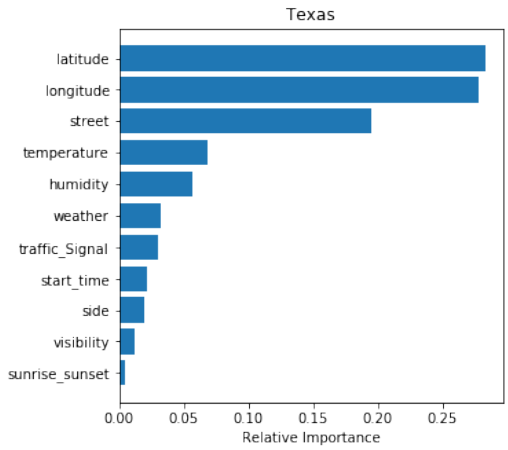

Figure 9: Relative Importance of Input Parameters

limits, road condition) on its Interstate roads. In addition, we see that among natural conditions, weather has the most impact which we have not seen in previous models, showing California has more severe-weather-related severe car accidents.

3.5 Texas

3.5.1 Model Comparison

Similar to the model fit of South Carolina, all three models achieve better results compared to the results of the entire country, but to a less extent than South Carolina's model.

Texas

Model	Level of Severity	Evaluation Metric		
		Overall Accuracy	Balanced Accuracy	F1 Score
Using OLS	1	0.644	—	—
	2		0.496	0.773
	3		0.498	0.197
	4		—	—
Using LASSO	1	0.784	—	—
	2		0.667	0.865
	3		0.672	0.517
	4		—	—
Using XGB	1	0.866	0.654	0.419
	2		0.844	0.912
	3		0.847	0.772
	4		0.579	0.256

Table 5

3.5.2 Relative Importance Factors

Both latitude and longitude showed a large relative importance with no significant difference in between, which coincides with Texas north/south distance to east/west distance ratio. In addition, types of street has a larger relative importance than weather, temperature and humidity. Thus, it also coincides with our suspicion that north region is more influenced by natural causes than south region.

3.6 Minnesota

3.6.1 Model Comparison

We can see that for all three models, they all achieve worse results compared to the results of the entire country. Thus, it means that for local fit, there are other factors that impact the severity of car accidents in Minnesota than these we investigate.

Minnesota

Model	Level of Severity	Evaluation Metric		
		Overall Accuracy	Balanced Accuracy	F1 Score
Using OLS	1	0.573	—	—
	2		0.495	0.698
	3		0.495	0.282
	4		—	—
Using LASSO	1	0.708	—	—
	2		0.642	0.794
	3		0.644	0.506
	4		—	—
Using XGB	1	0.845	0.5	—
	2		0.821	0.885
	3		0.824	0.771
	4		0.601	0.319

Table 6

3.6.2 Relative Importance Factors

In Minnesota, latitude outweighs longitude in relative importance, which coincides with our suspicion

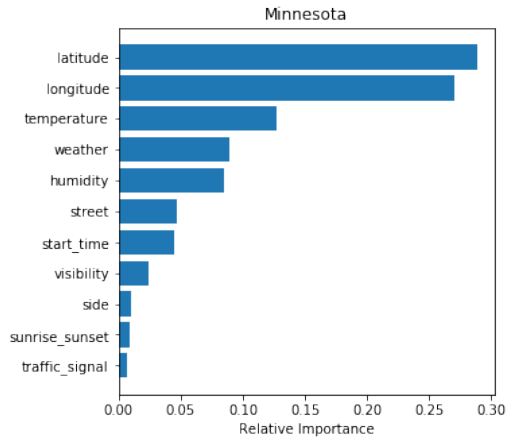

Figure 10: Relative Importance of Input Parameters

as Minnesota has a larger north/south distance than east/west. However, when comparing to other states, natural causes (temperature, weather and humidity) plays a more important role than other states, which can be inferred from its poor maintenance in the state, as well as more investment in natural cause mitigation such as snow removal.

3.7 Findings Across States

Firstly, we find out that there is strong correlation with accuracy rate of each state and their weighted car accident rate, as correlations are 0.7375, 0.7944, 0.8647 for OLS, LASSO and XGB respectively. We deem this due to the data completeness. With more complete data, the predictions are more accurate. In addition, this is another evidence that XGB fits the data better.

Secondly, in order to validate our suspicion that the north/south to east/west distance ratio has correlation with latitude and longitude relative importance factor, we calculate the correlation is 0.8381, which indeed shows strong positive correlation.

Thirdly, in order to validate our suspicion of the relationship between north/south states and its relative importance in natural causes, we calculate the correlation between each states' geographic center's latitude with natural cases relative importance, and the correlation is 0.9019, which also validates our inference.

4 Conclusion and Recommendations

1. Federal government should invest more in northern state's road maintenance and natural case mitigation (such as more investment in road quality check and severe weather prediction).

2. State government should invest more in either Interstate's or local roads regulation oversight and road maintenance, based on models from each state. The determination method is already stated in above discussion.

3. For individual drivers, northern regions' drivers should pay extra attention to extreme weather and temperature, and all drivers should pay close attention when driving on Interstate highways.

5 Weapon of Math Destruction and Fairness

The model is not a Weapon of Math Destruction. The dataset used is large enough and well recorded, so the prediction of severity in traffic accidents is measurable. Also, the prediction is independent with the severity of future accidents and will not cause more severe accidents happened, so it doesn't affect results or create negative feedback cycles.

The dataset doesn't include any information about the drivers, so there is no fairness problem in the algorithms design.