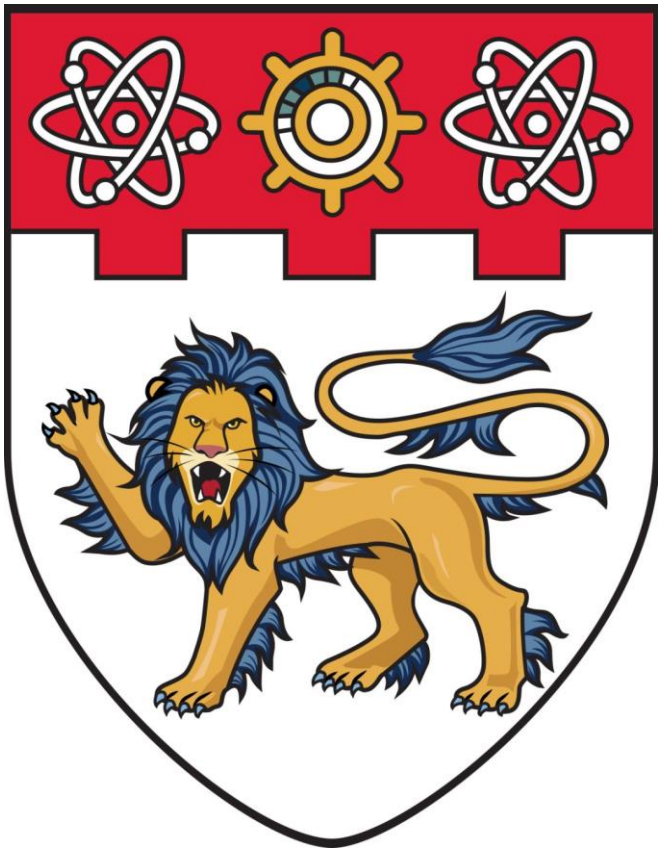

An analysis of Traffic Fatality using machine learning model

COLLEGE: Nangyang Business School

Major: Business Analytics

NAME: Wang Jingyi

MATRI NO: G2000751D

INSTRUCTOR: Teoh Teik Toe

Date: 2020.10.11

Contents

1. Motivation

1.1 Introduction

Traffic accidents are a major concern for safety in the world. According to the WHO, road accidents has caused 1.4 million fatalities annually. In Singapore, there were 7,690 road accidents involving injuries in 2018 and 120 people lost their lives. It is estimated that the road fatality rate in Singapore, 2.73 among 100,000 citizens is higher than London, Hong Kong and Tokyo. By 2030, traffic-related deaths are predicted to become the 7th leading cause of death worldwide. Therefore, it is important to raise people's attention to lower the risk of traffic accidents. There are many society groups concerned about risk of traffic accidents, like government, public health, vehicle manufacturers, insurance companies and even individuals. In this paper, I would stand on the side of government and Insurance company to figure out key factors affecting the fatality levels of traffic accidents and discuss potential directions that they can help to reduce the rate. I would use the opensource dataset, which includes details of traffic accidents from Collisions facts to individual information, based on which performs machine learning methods to help do necessary prediction and then compare the model performance.

1.2 Project Objectives

There are several key objectives this paper aiming at:

- Perform machine learning methods to build up traffic fatality model based on Government interest
- Evaluate the model and offer recommendations for government to do necessary prevention
- Perform machine learning methods to build up traffic fatality model based on Insurance Company interest
- Evaluate the model and offer recommendations for Insurance Company to work on product strategy

2. Literature Review

Many researchers have worked on traffic severity, including research on victims such as drivers, pedestrian, conduct and evaluate performance of machine learning methods, and then analyze key factors such as vehicle, geometry etc.

Joon-Ki Kima, et al (2008) explores the injury severity of pedestrians in motor-vehicle crashes. He conducted a heteroskedastic generalized extreme value model to find explanatory factors with four injury outcomes: fatal, incapacitating, non-incapacitating, and possible or no injury. The research finds out pedestrian age significantly induces probability of fatal injury, as well as male driver, traffic sign, commercial area, darkness with or without streetlights, sport-utility vehicle, truck etc. Conversely, with increasing driver age, during the PM traffic peak, with traffic signal control, the probability of a fatal injury decreased.

Considering model capability of present limited data sets, F. Rezaie Moghaddam, et al (2010) performs artificial neural network to illustrate the simultaneous influence of human

factors, road, vehicle, weather conditions and traffic features including traffic volume and flow speed on the crash severity in urban highways. Among many variables, variables such as highway width, head-on collision, type of vehicle at fault, ignoring lateral clearance, following distance, inability to control the vehicle, violating the permissible velocity and deviation to left by drivers are most significant factors that increase crash severity in urban highways.

H.M. Abdul Aziz, et al (2013) develops random parameter logit models for explaining pedestrian injury severity levels of New York City accounting for unobserved heterogeneity in the population and across the boroughs. Obtained results illustrate that, road characteristics (e.g., number of lanes, grade, light condition, road surface, etc.), traffic attributes (e.g., presence of signal control, type of vehicle, etc.), and land use (e.g., parking facilities, commercial and industrial land use, etc.) are found to be statistically significant in the estimated model.

Although there are many researches working on modeling, there are still lack of specific business analysis or solutions discussed for specific group like government, company etc. In this paper, I would offer suggestions to government and company based on the model result.

3. Problem Statement

3.2.1 To find out suitable prediction models in traffic fatality

I would select several machine learning methods to find out suitable models to describe traffic fatality, design the input and output for the models, adjust the parameters to get the best performance.

3.2.2 To find out key indicators of traffic fatality

To reduce traffic fatality rate, I would find out what kind of indicators are significantly associated with traffic fatality, leading to increase in purity between unfatal and fatal accident. In this way I can give further recommendations.

4. Data Overview

4.1 Data Description

The data table collected data from crashes for year 2014 includes 297613 records with 22 columns (including records with a number of problems like unknown, or “dummy" use) for crash cases in which 4502 cases(C_SEV) are related to fatalities and others injuries or non-injuries. Data table contains the following types of information:

1. Collision level data elements: Year, Month, Day of week, Collision hour, Collision severity, Number of vehicles involved in collision, Collision configuration, Roadway configuration, Weather condition, Road surface, Road alignment, traffic controls.

2. Vehicle level data elements: Vehicle sequence number, Vehicle type, vehicle model year.

3. Person level data elements: Person sequence number, sex, age, position, Medical treatment required, Safety device used, Road user class.

There are two variables relating to fatality in the table: the Collision severity(1-

Collision producing at least one fatality,2- Collision producing non-fatal injury) and the Medical treatment required(1-No Injury,2-Injury,3-Fatality).

One collision can generate more than one records in the table. The Vehicle sequence number is used to distinguish cars in the same collision, and Person sequence number is used to distinguish persons in the same car or in the same collision. Except for this and Year, Month, Day, Hour, Age, all the variables in the table are category variables.

4.2 Statistical Analysis

Based on the statistical analysis, I can get familiar with the data features and then come to choose reasonable variables for the model. I explore the data using excel pivot table.

In order to calculate the number of non-repeating collision and numbers of persons involved in the table, I use the Collision level data elements to generate a new column called C_NUM and also Person level data elements a new column called P_NUM. After which I find out there are 1703 collisions lead to fatalities taken away 1876 lives. Among this, 1559 collisions caused one death and 6 collisions caused four deaths at most.

For the collision level, distinct count of C_NUM is used to do calculation. Hours between 22:00PM-5:00AM have at least 1% higher fatal rate compared to other time range, which suggests that driving in the dark is one of the potential indicators for fatality. The fatal rate in weekday and month dimension have similar distribution. For Collision configuration, rear-end collision covers 22.11% of the total cases and Head-on collision causes the highest fatal rate at 10.72%, which is significantly higher than fatal rate 2.26%. For Roadway configuration, Non-intersection and at an intersection of at least two public roadways covers up to 82.01% of the total cases and tunnel or underpass causes the highest fatal rate at 6.67% and then Passing or climbing lane at 5.84%. which is significantly higher than fatal rate 2.26%. The Weather condition of Strong wind causes highest fatal rate at 3.25% although sunny and cloudy covers 77.64% of total cases. For the Road surface, Flood has highest fatal rate at 12.50%, which is significantly higher than average rate 3.37%.

For the personal level, count of P_NUM is used to do calculation. I explored the fatal features for drivers and passengers separately and find it does not have much difference. Calculating on the passengers, Purpose-built motorhome and Snowmobile as well as Off road vehicles are the three types have the highest fatal rate: 6.67%, 5% and 3.25%. When the age of passengers go higher, the fatal rate increase obviously and the age group between 90 and 99 achieves 48.12%. For safety device used, No safety device used or No child restraint used has highest fatal at 3.89% and conversely Safety device used or child restraint used at 0.32%. .

4.3 Data Visualization based on factors

With data visualization, I can observe the features of different factors better. Based on the fatal categorical variable C_SEV, I perform distribution visualization among the group of fatal and unfatal. The variables I compared here are HOUR, MONH, VEHS, CONF, RCFG, WTHR, RSUR. The result shows that mostly the distribution does not appear much difference, but I can still detect some facts that are in consistent with the statistical analysis

above. The below shows the distribution for HOUR, MONH, VEHS, CONF, RCFG, WTHR, RSUR, RALN, TRAF separately considering CSEV=1 and CSEV=2, the result shows that the statistical facts are mostly similar to the analysis above. The visualization of PISEV is also performed and the conclusion is the same.

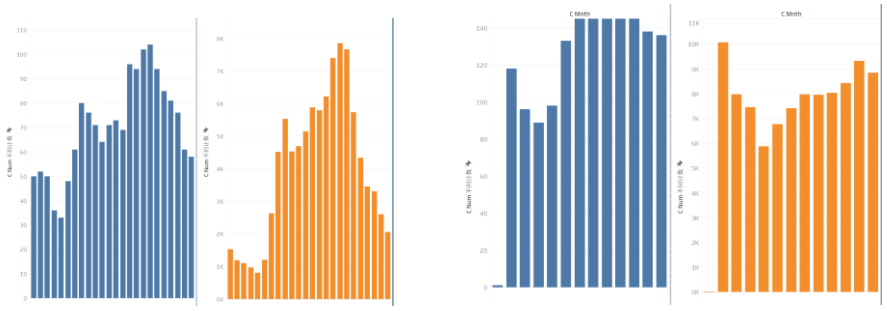

Figure 1 C_HOUR &C_MONH

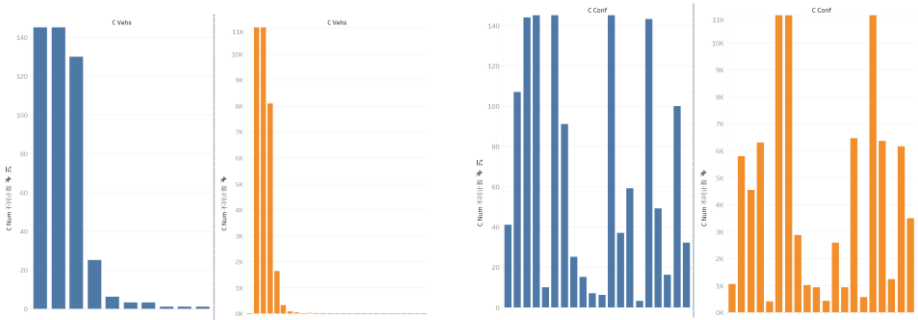

Figure 2 C_VEHS&C_CONF

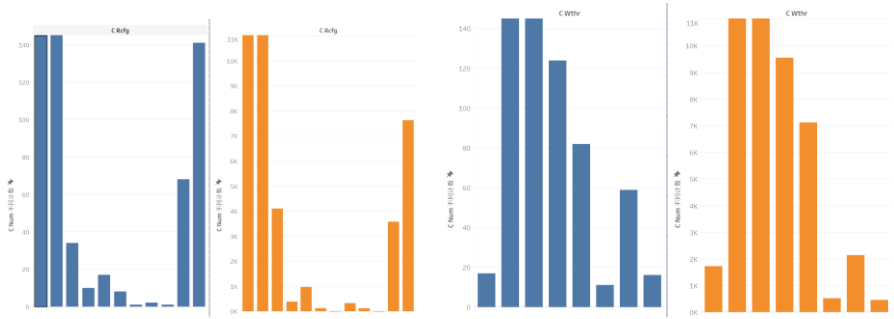

Figure 3 C_RCFG&C_WTHR

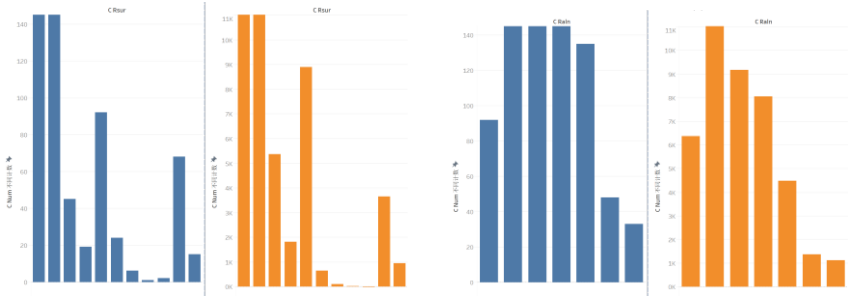

Figure 4 C_RSUR&C_RALN

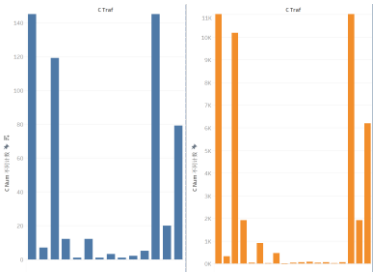

Figure 5 C_TRAF

4.4 Data Cleaning & Data Wrangling

I am going to work on prediction models for two different social groups: Government and Insurance Company. For government, what it may concern should be the peak period, roadway conditions, road surface, weather the traffic condition that government can predict or take action ahead of accident, for the benefit of citizens,thus its Y should be C_SEV. For Insurance Company, the difference is it cares more about how personal variables like sex, age, driver or not, would affect the result because the main purpose of insurance company is to promote insurance products to customers, thus its Y should be P_ISEV. Hence, I should generate different variables for these two models. I would generate two separate datasets based on the original dataset and perform data cleaning separately to remain as much information as possible.

4.4.1 Government Prediction Model

The dataset includes indicators that I think government would care about and significant through data analysis. There are 10 columns: C_MNTH, C_WDAY, C_HOUR, C_CONF, C_RCFG, C_WTHR, C_RSUR, C_RALN, C_TRAF, C_SEV.

Firstly, I checked the data format content, remove spaces, special symbols.

Secondly, I check the data logic. For example, to make sure data is consist, I check whether C_SEV is fatal when PISEV is fatal. For categorical variables, I check whether the value is out of categories, and for continuous variables, I check whether have outliers. Then, drop duplicate columns.

Third things to do is to deal with missing values valued ‘Q’,’U’, ‘X’, ‘QQ’,’UU’ ‘XX’,etc. I try to drop all the missing data records but find it fairly large amount, which would cause a information loss. So I try to fill up the missing records. Adjacent value method is to use the data in the adjacent positions to fill in, which is random for this dataset. Thus I use adjacent value method in python.

After removing all the string value, I change the datatype into numeric type. Then I cut MNTH, HOUR, C_CONF into groups and generate new range variables.

4.4.2 Insurance Prediction Model

The dataset includes indicators that I think insurance company would care about and significant through data analysis. There are 12 columns: 'C_RCFG', 'C_WTHR', 'C_RSUR', 'C_RALN','C_TRAF','V_TYPE','P_SEX','P_AGE','P_PSN','P_SAFE','P_USER','P_ISEV'.

Similar data processing method is performed for insurance model, except the original dataset contains records specially for pedestrians, which are dummy variables in vehicle level data but have significant meanings. Thus, Missing Values of these records are carefully filled up. On the other hand, dataset contains records specially for parked cars,

which are dummy variables in person level. These records have no contribution to the model and thus removed.

5. Insurance Company Prediction Model

5.1 Data preparation

5.1.1 Data Sampling

The Y variable I used for this model is P_ISEV. The original value of it is 1,2 and 3. After data cleaning and data Wrangling, I have obtained an imbalanced dataset. The amount sample of P_ISEV equal to 1,2,3 is 83456 (41.7%) and 115601(57.7%) and 1232 (0.6%) respectively. Imbalanced dataset would easily cause the machine learning algorithm to be more biased to the majority class during the training process, which consequently leads to a low sensitivity of the minority class. Data under-sampling and over-sampling are therefore designed to mitigate the influence of imbalanced dataset. Thus, I rebalanced the dataset using under-sampling methods. Under-sampling techniques remove examples from the training dataset that belong to the majority class in order to better balance the class distribution. This is different from oversampling that involves adding examples to the minority class in an effort to reduce the skew in the class distribution. Typically, under-sampling methods are used in conjunction with an oversampling technique for the minority class, and this combination often results in better performance than using oversampling or under-sampling alone on the training dataset. Since our target is predict fatal and unfatal, I would refer to the P_ISEV=3, which has 1232 records, then I select 1232 evenly from subset of P_ISEV=1 and P_ISEV=2 to get equal information . After calculation, I obtained a training set with the same set of positive and negative sample(2464 in total). The new training set was then further used for the next step.

Since our target is predict fatal and unfatal, thus I change P_ISEV equal to 0 if the original value is I or 2, which makes non-fatal a baseline in the data.

5.1.2 Group Variable and Dummy Variable

The categorical x variables in the dataset have many categories that it might be difficult for programs to identify the features of variables because the matrix is too sparse. According to the data dictionary I regroup the variables by cutting them into ranges to reduce the number of categories. The new generated variables are 'P_AGEGR', 'P_PSNGR' to replace 'P_AGE', 'P_PSN' in the original dataset. After this, all the variables can be considered categorical variables and dummy method is used.

I perform one type of dummy encoding. I use “drop first” one-hot encoding to create dummy variables for the categorical variables. The base level of every categorical variable is absorbed into the intercept term for logistic regression and bias term of the input layer for neural network.

The final dataset for logistic regression and neural network model construction contains 73 attributes. Values having zero value in the original categories being the base level.

5.1.3 Train and Test Set

After processing the data, the dataset is spilt into train and test set at a ratio of 8:2, trainset is used to build the model and test set would help validate the model afterwards.

5.2 Model and Configuration

5.2.1 Logistic Regression

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

The accuracy against the testing set of the logistic regression is 76.46%, which is satisfactory. The recall is 74% and AUC 76.5%, which suggests that my model can still perform well considering the positive and negative samples.

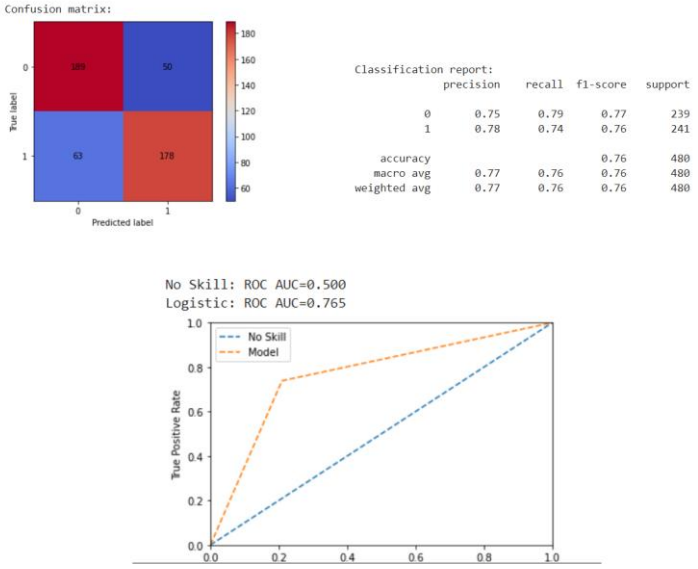

Figure 5 Result of Logistic Regression

Logistic regression model can also be used as means to indicate the statistical significance of the association between a particular feature and the target variable. The variable importance can be interpreted by coef function and I plot the result below. The result shows that are the most significant variables.

We observe that AGE_(85,99),TRAF_18,RALN_5 are highly positive significant features, which suggests that higher age group, and Traffic control of no control present, Road alignment of top of hill or gradient would increase the fatal rate when happen accidents. SAFE_2, TYPE_9, SAFE_12 are highly negative significant features, which suggests that safety device used or child restraint or other devices used, Vehicle type of school bus are indicators that would lower the risk of fatal.

All other features are statistically significant based on the p-values of the logistic regression model.

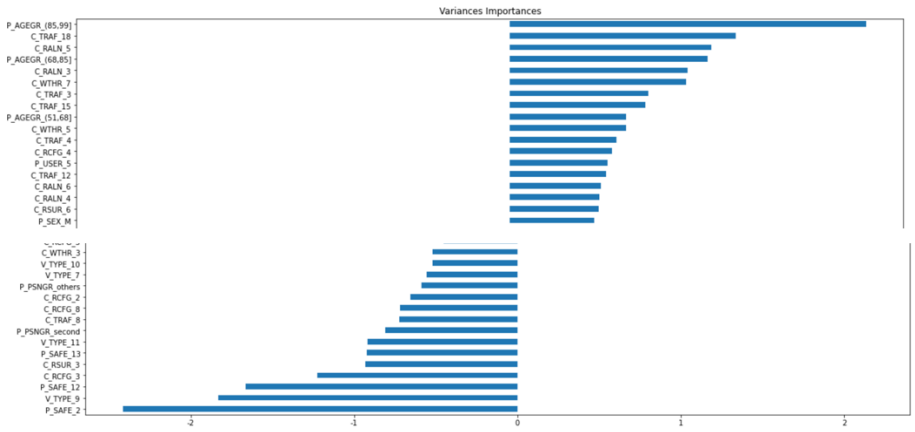

Figure 6 Result of Variable importance

5.2.2 Decision Tree

Decision Trees are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. Here Decision Tree model is used for classification and the full tree is grown and then pruned by finding the optimal cost complexity parameter. The optimal cost complexity parameter is where the average cross-validation error using 10-fold cross-validation on the training set is minimum. The results are shown below.

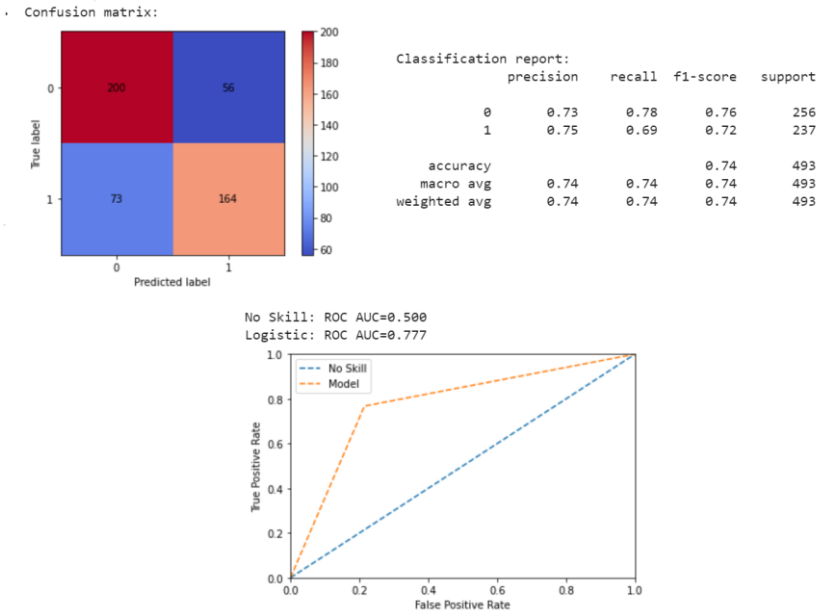

Figure 6 Result of Decision Tree

The accuracy against the testing set of the decision tree is lower than logistic regression at 75.42%. However, the recall is lower than logistics model for positive samples at 72%. The AUC is similar to logistics model at 76.5%. Thus in total the performance is not as good as logistic regression model.

5.2.3 Artificial Neural Network (ANN)

Artificial neural networks (ANN) a massively parallel distributed information processing system, based on nature of human brains, are capable of approximating any finite nonlinear models to determine the relation between dependent and independent variables.

In fact the function is based on errorback propagation so that the error between out put of the network and desired output (target) is required low. Neural networks are parts of Artificial Intelligence which have been applied in different areas successfully. Considering this high capability researchers are researching on new generations of ANN with more power and precision.

The architecture of ANN is a three-layer ANN with one hidden layer. The first and second layer have dimensionalities of the output space of 12 and 6 respectively, with ReLU activation. A sigmoid layer is used in the final layer. The optimizer of the network is Nadam with binary cross-entropy as the loss function. Nadam is also called Nesterov-accelerated Adaptive Moment Estimation which combines Adam with Nesterov momentum.

The ANN model can include all possible permutations of variables, discrete or continuous. Here can be used in my model to handle categorical variables.

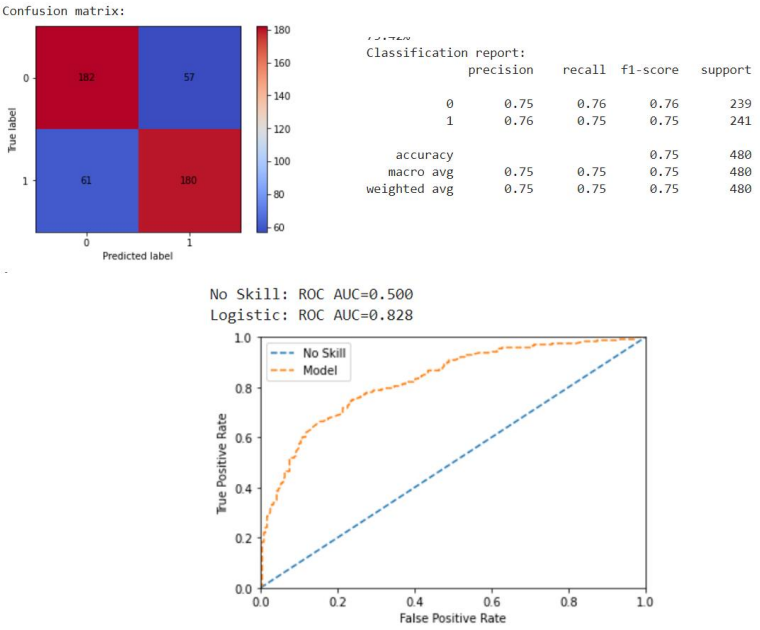

Figure 8 Result of ANN

The accuracy against the testing set of ANN is lower than logistic regression at 75.42%. The recall is 75%. The AUC is better than logistics model at 82.8%.

5.3 Model Comparison

Since this is a classification problem, here I use two common methods for model evaluation indicators of the sub-type model:confusion matrix (also called error matrix, Confusion Matrix) ,recall and AUC area.

The confusion matrix is to separately calculate the classification models for the correct and wrong category, and then display the results in a table:

	Actual Positive	Actual Negative
Predict Positive	TP	FP
Predict Negative	FN	TN

Based on the confusion matrix, Recall is calculated as: TP/(TP+FN), which measures of all the positive examples (P), how many are correctly predicted. Thus in this case, it is

suitable to use recall to compare the model because what I care about is how many fatality can I correctly predict.

The AUC is calculated based on the confusion matrix.AUC (Area Under Curve) is defined as the area under the ROC curve and the coordinate axis. The closer the AUC is to 1.0, the higher the authenticity of the detection method; when it is equal to 0.5, the authenticity is the lowest and it has no application value.

From the result we can see that Logistic Regression achieves highest accuracy at 76.46% . However, neural network achieves highest recall at 75% with accuracy 75.42%. Since our goal is to train the classification model that can predict traffic fatality that happens actually, I need to focus on recall and auc. Therefore, we can conclude that Artificial neural networks is better than decision tree and Logistic Regression.

6. Government Prediction Model

6.1 Data preparation

6.1.1 Data Sampling

The Y variable I used for this model is C SEV. The original value of it is 1 and 2. Thus I change C_SEV equal to 0 if the original value is 2, which makes non-fatal a baseline in the data. After data cleaning and data Wrangling, I have obtained an imbalanced dataset. The amount of positive and positive sample is 1702 (1.7%) and 97070 (98.3%) respectively. Imbalanced dataset would easily cause the machine learning algorithm to be more biased to the majority class during the training process, which consequently leads to a low sensitivity of the minority class. Data under-sampling and over-sampling are therefore designed to mitigate the influence of imbalanced dataset. Thus, I rebalanced the dataset using under-sampling methods. The ratio of positive and negative samples is set to be 2:8. After calculation, I obtained a training set with the same amount of positive sample(1702) and negative sample (6808 for each). The new training set was then further used for the next step.

6.1.2 Group Variable and Dummy Variable

The categorical x variables in the dataset have many categories that it might be difficult for programs to identify the features of variables because the matrix is too sparse. Similar to the previous method, according to the data dictionary I regroup the variables by cutting them into ranges to reduce the number of categories. The new generated variables are 'C_MONTH' , 'C_TIME', 'C_CONFC' to replace 'C_MNTH', 'C_HOUR', 'C_CONF' in the original dataset. After this, all the variables can be considered categorical variables and dummy method is used.

I perform one types of dummy encoding. I use “drop first” one-hot encoding to create dummy variables for the categorical variables. The base level of every categorical variable is absorbed into the intercept term for logistic regression and bias term of the input layer for neural network.

The final dataset for logistic regression and neural network model construction contains 62 attributes. Values having zero value in the original categories being the base level.

6.1.3 Train and Test Set

After processing the data, the dataset is spilt into train and test set at a ratio of 8:2, trainset is used to build the model and test set would help validate the model afterwards.

6.2 Model and Configuration

6.2.1 Logistic Regression

A logistic regression model is constructed by fitting all predictors available. The results are shown below.

The accuracy against the testing set of the logistic regression is 71.37%, which is not satisfactory. The recall is 72% and AUC 71.4%.

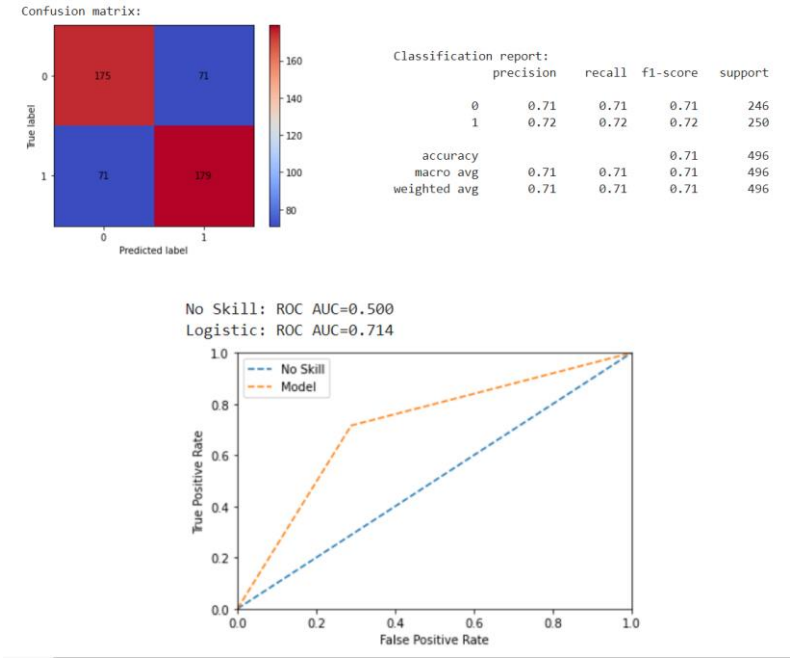

Figure 9 Logistic regression

Logistic regression model can also be used as means to indicate the statistical significance of the association between a particular feature and the target variable. The variable importance can be interpreted by coef function and I plot the result below. The result shows that are the most significant variables.

We observe that C_TRAF_2 ,C_TRAF_18, C_RCFG_4, C_RCFG_6, C_WTHR_7

C_RALN_6 are highly positive significant features, which suggests that Traffic signals in flashing mode, No control present , Railroad level crossing, unnel or underpass, strong wind, Bottom of hill or gradient increase the fatal rate when happen accidents.

C_CONFIC_TwoVehiSame,C_RCFG_3, C_RSUR_3, C_CONFIC_TwoVehiPar, C_RSUR_4, C_RSUR_5, C_RCFG_8 are highly negative significant features, which suggests that Ramp, Two Vehicles in Motion - Same Direction of Travel, Intersection with parking lot entrance/exit, private driveway or laneway, Ramp, Snow (fresh, loose snow), Two Vehicles - Hit a Parked Motor Vehicle, Slush ,wet snow or icy, are indicators that would lower the risk of fatal.

All other features are statistically significant based on the p-values of the logistic regression model.

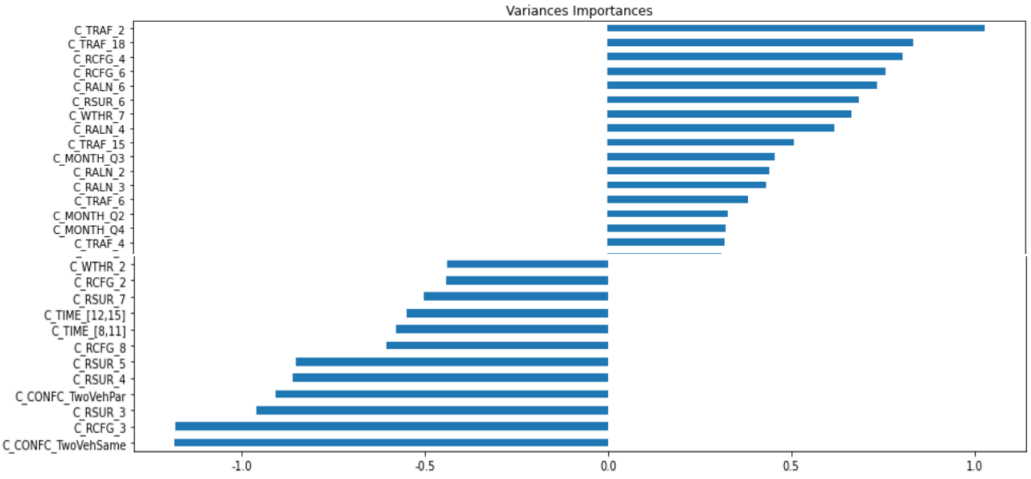

Figure 9 variable importance

6.2.2 Decision Tree

Then I use decision tree models. The results are shown below.

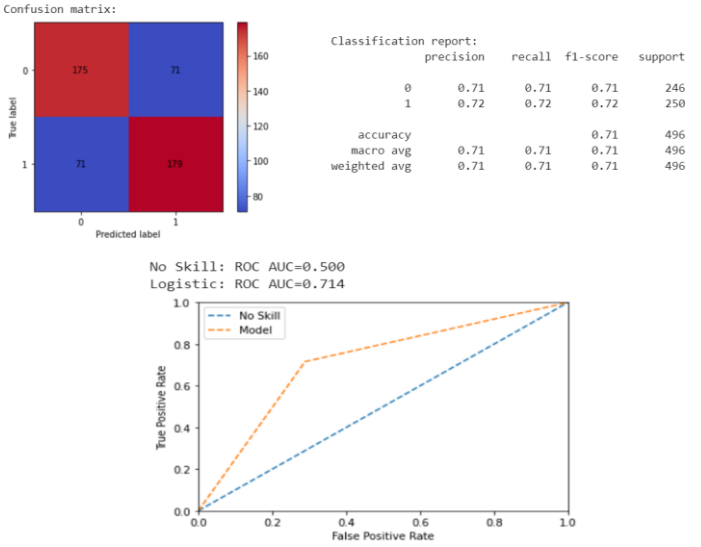

Figure 10 Result of Decision Tree

The highest validation accuracy of the decision tree is higher than logistic regression at 73.79%. The recall is 72%, and AUC is the same as logistics model Thus in total the performance better than logistic regression model.

6.2.3 Artificial Neural Network (ANN)

Artificial neural networks (ANN).

Then I use the ANN model. Below are the results.

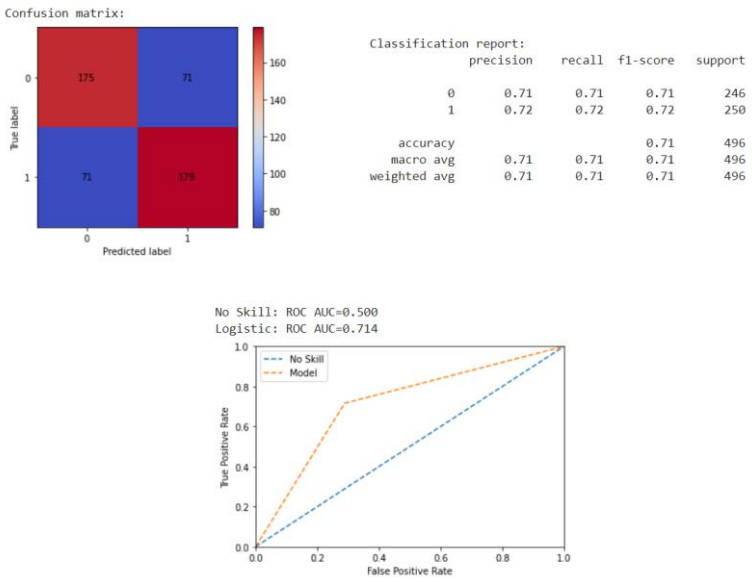

Figure 11 Result of ANN

The accuracy against the testing set of ANN is lower than logistic regression at 71.37%. The recall and AUC are also similar to logistics.

6.3 Model Comparison

From the result I can see that Decision Tree achieves highest accuracy at 73.79%. Since our goal is to train the classification model that can predict traffic fatality that happens actually, I need to focus on recall and AUC. Considering the recall and AUC are the same among these three models, I recommend decision tree as the final models used for fatal prediction.

7.Discussion

The models finally serve for government and insurance company and each has own interest. From the variable importance result from logistics regression model, we can further discuss the indicators of traffic fatalities and offer suggestions.

7.1 Government

Government would care more about road construction, signal control traffic rules which are under charge of government so that officers know what measures they can put forward to improve traffic safety. From the Government model, we can see Traffic control, road alignment, roadway configuration are influential factors.

7.1.1 Traffic Control

The traffic control situation of the collision cases offered shows it a critical variable to fatalities. There are three key findings. First, the road with no control present would cause a high rate, which suggests that set up traffic control starting from crossings with higher fatal rate. Second, collisions happen at pedestrian crosswalk caused a high fatal rate. Explanation might be over-speed due to drivers or the pedestrian failed to obey the traffic rules, thus government might need to consider how to better restrict citizens to obey the

rules. On the one hand, government may give heavier fine for crossing red lights or speeding. On the other hand, government may better design the waiting time for traffic lights to reduce impatience.

Thirdly, it is surprised that the data shows where railway crossing with signals, or signals and gates, have higher fatal rate than Railway crossing with signs only. However, we should consider the real situation. Railway crossing builds up gate to make sure safety, but collisions might be caused in case of the traffic sign at one right of either pedestrian or drivers are not that obvious, thus government might consider need to rebuild obvious signals near the railway crossing in the area.

7.1.2 Road alignment

Road alignment is linked with an increase in the probability of fatal injury, especially for Bottom of hill or gradient, curved with gradient, curved and level. This result indicates that crashes occurring on a curve and downgrades were found positively correlated with fatality, generally because at these alignment drivers are not easy to slow down which leads to crashes. Government might build up deceleration zone and complete the navigation automatic reminders to suggest drivers to slow down in advance.

7.1.3 Roadway configuration

The roadway configuration is associated with fatality. Model shows that at railroad level crossing, Tunnel or underpass has higher rate of fatality and conversely at intersection with parking lot entrance/exit, private driveway or laneway, and ramp. Traffic accidents happens related to railroad are always severe and inside tunnel or underpass is always more difficult to carry out rescue. While crashes happen at parking lot or private driveway, drivers drives at a low speed. Government should make effort on emergency rescue strategy.

7.1.4 Other variables

Several other factors such as weather conditions, and traffic peak are also found to be statistically significant in the three estimated models.

7.1.5 recommendations

What governments can do to improve road safety? I think mainly three aspects. Firstly , governments should develop on Institution, making road safety a political priority and make it publicly accountable. Secondly, set appropriate road safety construction targets and establish national road safety plans to achieve them. The road and signal construction would need budgets thus should achieve step by step. Thirdly, establish data collection systems designed to collect and analyze data and use the data to improve safety. Through this can collect more related data to apply machine learning models and improve accuracy.

7.2 Insurance Company

Insurance Company would care more about personal factors on the accident, because the main objective is to promote products for specific groups. From the Insurance Company model, age, person position and vehicle type are influential indicators.

7.2.1 Age

Age is a powerful discriminator on traffic fatality. Considering the age group is older, the probability of severe fatality increases. According to statistical analysis, the fatal rate of passengers over 90 years old even reaches 48.12%, compared to the rate 5.13% of youngsters no more than 20 years old. Thus, elder people are specific group that need protection. Elders also have a higher cost for injuries. According to the result, insurance company can segment customers into target group with age ranging from (85,99], (68,85], (51,68] and offer insurance products with higher price for elder groups.

7.2.2 Person Position

The result shows that persons sitting at the third rows has higher fatal rate. One of the explanation may arise from the fact that elders are more likely to take the seat in the front row. Besides, sitting outside passenger compartment have a high fatality. Although the drivers' position does not yield much significance, however, drivers covers almost 68% of the total cases, which make it a vulnerable group needed special support. The suggestion for insurance company is to offer fair price products or discount for drivers to attract more customers.

7.2.3 Vehicle Type

Inside the model school bus and Urban and Intercity Bus are found to be safe traffic tools. The Motorcycle and moped, snowmobile, farm equipment are not obviously significant in the model but shows a great effect on fatality in statistical analysis. Thus, different kinds of insurance should cater to these target groups.

7.2.4 recommendations

The effect of traffic fatality is more related to government level control factors although features can be found on individuals. Firstly, considering government's action on traffic governance can affect the risk of deadly road accident, insurance company can work with government on accident prediction. Secondly, elder groups and drivers and motorcycle users can be target groups to promote insurance products and company can also take the responsibility to advise on the importance of insurance.

8. Conclusion

This study sets a background to analyze the factors of traffic fatality and prediction model and use opensource dataset to investigate traffic fatality. The whole process includes background understanding, data exploration and visualization, modelling, result analysis and business insight through our work.

Based on my understanding of the interest government and insurance company cares, I decide to build up two separate models. Firstly, I did data exploration and statistical analysis on tableau, and select potential factors, then I perform data processing to make sure cleaned data used in the model.

I use logistic regression, decision tree and artificial neural network methods to build up the model and evaluate the accuracy, recall and AUC. The result shows that for government use, logistic regression model achieves the highest accuracy and recall. For insurance company, decision tree model is the best.

From the Government model, we can see Traffic control, road alignment, roadway configuration are influential factors and for Insurance Company model, age, person

position and vehicle type are influential indicators. Then I give analysis on the significant factors and offer suggestions for government and insurance company.

9. Future Study

Along this study working on traffic indicators, traffic speed is significant that has affect many other indicators to cause traffic fatality. However, in this data source there are no information about traffic speed. Thus, further analysis can be performed on the effect of speed on traffic accident. Besides, there many machine learning methods like neural networks, random parameter models used to study this problem. Considering the model accuracy at most 77%, other typical models used by researchers in the area can be tried out to optimize the accuracy.

10. Reference

[1] Kim J K, Ulfarsson G F, Shankar V N, et al. Age and pedestrian injury severity in motor-vehicle crashes: A heteroskedastic logit analysis[J]. Accident Analysis & Prevention, 2008, 40(5): 1695-1702.

[2] Rezaie Moghaddam F, Afandizadeh S, Ziyadi M. Prediction of accident severity using artificial neural networks[J]. International Journal of Civil Engineering, 2011, 9(1): 41-48.

[3] Venkataraman N, Ulfarsson G F, Shankar V N. Random parameter models of interstate crash frequencies by severity, number of vehicles involved, collision and location type[J]. Accident Analysis & Prevention, 2013, 59: 309-318.