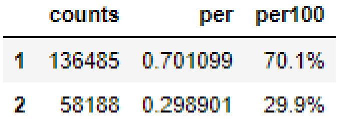
Predicting the Occurrence of High Severity Level Road Accident
Data Section

1. Data sources

Coursera shares the data which are downloaded from Seattle government's open data platform. The data includes all types of collisions from 2004 to 2020-May. There are total 195K collision records with 37 variables. The data contains many useful information including the severity level, the accident datetime, collision type, weather condition, road condition, the number of vehicles involved, the number of pedestrians involved, etc. To predict the occurrence of a high severity level accident. The variable "SEVERITYCODE" is selected as our target variable. There are two available values "1 – prop damage" and "2 – injury". In this project will define "2 – injury" as a high-level severity road accident and predict its occurrence.

1.1. Data cleaning

Firstly, the distribution of the target variable "SEVERITYCODE" is studied. The "2 – injury" only occupies 30% of the total records which shows it is an unbalanced data. Data balancing may need to be performed before building a predictive model.



	counts	per	per100
1	136485	0.701099	70.1%
2	58188	0.298901	29.9%

Secondly, the distributions of the rest variables are observed. There are some variables with >=80% records contain NaN or missing value. Some of them are indicator variables and the rest of them are not. We can identify them based on the distribution of the available values and their business meaning. For the indicator variables, we will keep the inductor with its original value "Y" and impute the NaN values to "N". For the rest of NaN variables, we will drop them from our dataset because they are not informative and should not be selected as the predictors in our model.

Finally, after dropping the columns with lots of NaN values, we check the dataset and remove the records will NaN in the remaining variables. There are 109K clean records and 35 variables are left. The distribution of the target variables "SERVERITYCODE" is checked again. The "2 – injury" still occupies 30% of the total clean records which indicates the clean dataset is still an unbalanced dataset. Therefore, we need to perform data balance before fitting a model.

1.2. Data Balancing – Oversampling

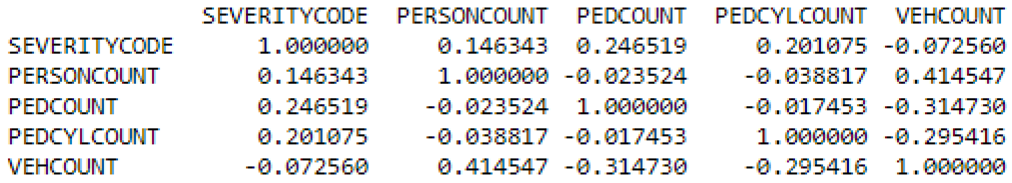
The dataset is an unbalanced dateset. The response records only occupied 30% of the total records. To balance the data for model building, the oversampling technique will be applied on the Train set. Oversampling will re-sample the minority response records with replacement to up-sample the number of response records to meet the number of majority records. This technique can emphasize the characteristics of significant predictors for building model.

2. Exploratory Data Analysis

The cleaned dataset contains 35 variables, only 4 variables are numeric and the rest of them are categorical. For the numeric variables, we will calculate the correlation matrix and select the variables which are highly correlated to the target variable and independent among each other to avoid the collinearity problem. For the categorical variables, we will perform the chi-square test between each categorical variable and the target variable and select the ones which are significant. The selected numeric and categorical variables will be the predictor for model fitting.

2.1. Numeric Variables – Correlation Matrix

As mentioned above, a correlation matrix will be created to study the relationship among each numeric variables and target variable.



```
              SEVERITYCODE  PERSONCOUNT  PEDCOUNT  PEDCYLCOUNT  VEHCOUNT
SEVERITYCODE      1.000000     0.146343  0.246519     0.201075 -0.072560
PERSONCOUNT       0.146343     1.000000 -0.023524    -0.038817  0.414547
PEDCOUNT          0.246519    -0.023524  1.000000    -0.017453 -0.314730
PEDCYLCOUNT       0.201075    -0.038817 -0.017453     1.000000 -0.295416
VEHCOUNT         -0.072560     0.414547 -0.314730    -0.295416  1.000000
```

We can find that “PERSONCOUNT”, “PEDCOUNT” and “PEDCYLCOUNT” have positive relationship with the “SEVERITYCODE”. Only “VEHCOUNT” has negative relationship.

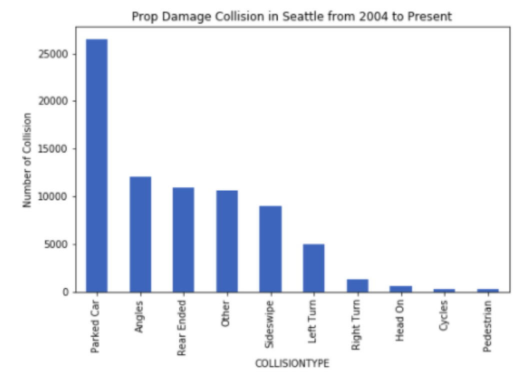
By comparing the magnitude, “PEDCOUNT” and “PEDCYLCOUNT” have a larger correlation with the “SEVERITYCODE”, but they are also correlation to each other. To avoid the collinearity problem, we should keep the predictors to be independent to each other and prevent select highly correlated variables in the same model. In this case, we would only select “PEDCOUNT” which has the highest correlation value as the predictor.

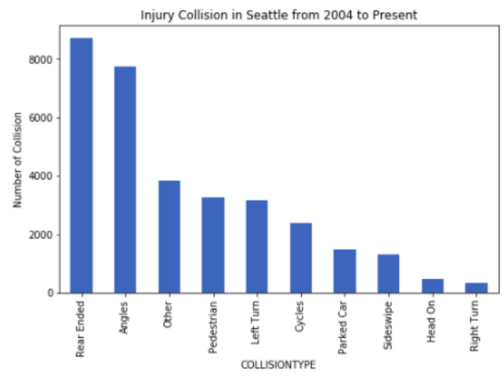
2.2. Categorical Variables – Chi-Square Test

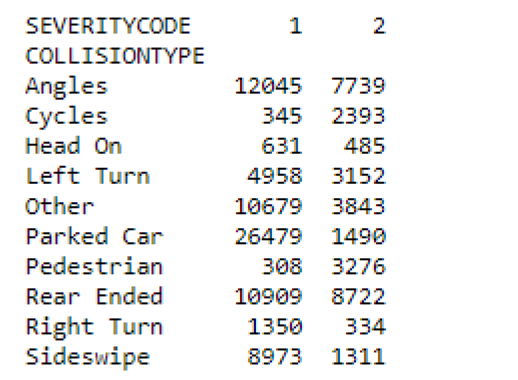
To find out significant categorical variables, we will perform chi-square test of independence between each categorical variable and the target variable.

The following is the example of Chi-Square test of independence between “COLLISIONTYPE” and “SEVERITYCODE”. The Chi-Square test of the rest variable will be covered in the report

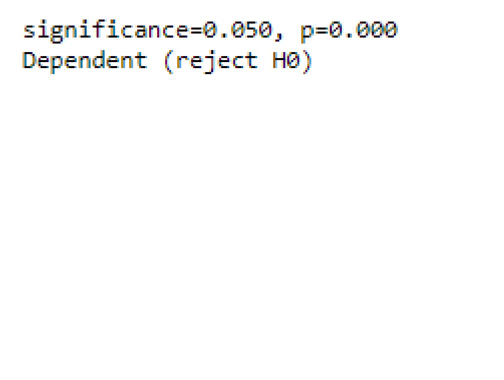
Distribution Plot and Chi-Square Test of “COLLISIONTYPE”







```
SEVERITYCODE        1     2
COLLISIONTYPE
Angles          12045  7739
Cycles            345  2393
Head On           631   485
Left Turn        4958  3152
Other           10679  3843
Parked Car      26479  1490
Pedestrian        308  3276
Rear Ended      10909  8722
Right Turn       1350   334
Sideswipe        8973  1311
```



```
significance=0.050, p=0.000
Dependent (reject H0)
```

From the distribution,

- For “1 – prop damage”, the highest collision type is “Parked Car”
- For “2 – injury”, the highest collision type is “Rear Ended”

The differences of the distribution imply that “COLLISIONTYPE” may be able to classify a high severity level accident. Then, a Chi-Square test is performed to proof our hypothesis.

Chi-Square Test:

- H0: There is no relationship between “COLLISIONTYPE” and “SEVERITYCODE”.
- H1: H0 is not true.

Chi-Square test of independence is performed. The result p-value is 0.00 which are less 0.05 significant level. We can reject the H0 and conclude that there is a relationship between “COLLISIONTYPE” and “SEVERITYCODE”. “COLLISIONTYPE” will be selected as one of our predictors for model fitting.

3. Predictive Modeling

It is a supervised learning problem to predict the occurrence of the high severity level accident. We would split the data into Train Set (80%) and Test Set (20%). Train Set would be used to build logistic regression to predict the probability of event occurrence. Test Set would be used to evaluate the model performance.

3.1. Classification Model – Logistic Regression

Logistic regression is a machine learning algorithm to model the probability of the respond records. The predicted probability is between 0 and 1. It can be used in classification problem. If it is larger than 0.5, we would classify the prediction would be “Will be happened”. If the predicted value is smaller then 0.5, the prediction would be “Will not be happened”. The threshold value 0.5 can be adjusted based on the business problem’s objectives. For some special problem, like fault detection, the business users are used to prefer a smaller threshold to reduce the false negative and capture fault case as more as they can. In this project, for simplicity, we will use 0.5 as the threshold value.

3.2. Model Performance Evaluation

Three metrics are calculated using Test set to evaluate the model performance.

- Accuracy Score
- Confusion Matrix
- Area Under ROC Curve