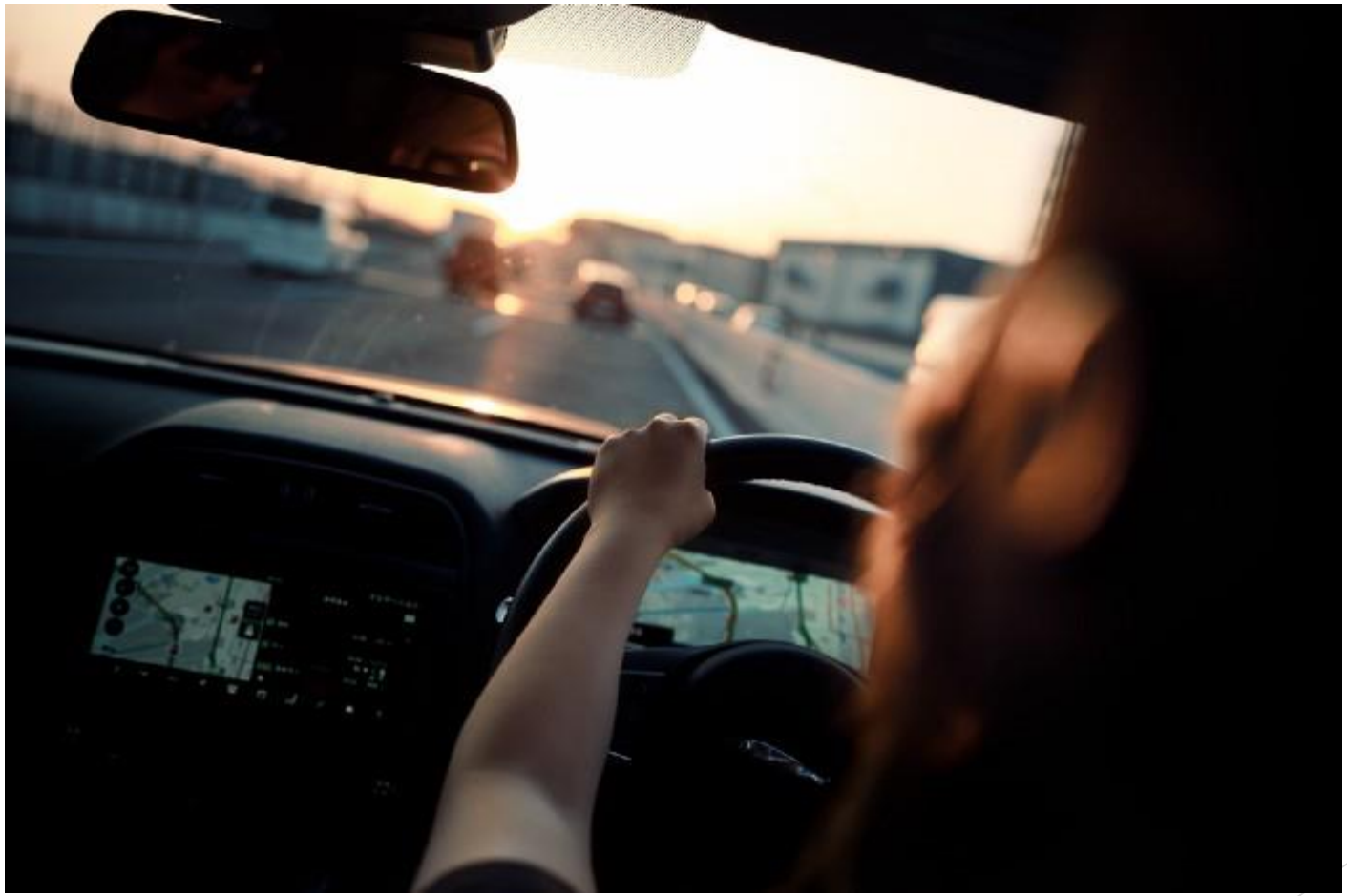
Predictive Model to estimate the influence of various factors on accident severity



Introduction:

- Accident severity is of special concern in traffic safety sector since this model is aimed not only at prevention of accidents but also at reduction of their severity.
- One way to accomplish the latter is to identify the most probable factors that affect accident severity. This model aims at examining not all factors, but some believed to have a higher potential for serious injury or death, such as accident location, type, and time; collision type; and age and nationality of the driver at fault, his license status, and vehicle type.
- Other factors were also examined such as drug or alcohol influence, driver inattention.

Business Problem:

- The goal of this model is the analysis of the collision dataset of the city of Seattle, to find patterns, relationships and determine features such as weather conditions, road and traffic conditions, light and alcohol influence on driver, driver inattention to provide the best road traffic accident severity prediction.
- It will use some analytical techniques and machine learning classification algorithms such as logistic regression, decision tree analysis, k-nearest-neighbors, support vector machine, etc. This model can mainly help transportation sector.
- Car rental or insurance companies are also among the target groups of this analysis because they can classify potential customers and design different services deliverable based on customers driving habits and other conditions.

Data Section:

- The dataset is the Car Collision dataset which contains collision data, address, and other different conditions. The dataset includes 38 columns and 194673 observations.
- **Data cleaning:** There are some problems with this dataset. First, for missing values, there are 10527 records and because most of the features are categorical data so I think I could build different models with/without the missing values and compare the accuracy later to decide whether I will exclude missing value. Secondly, the major task is to predict the severity of the car accident and the target variable should be SEVERITYCODE which was 1 (damage only) and 2 (injury). And the number of records 2 is twice as many as record 1. So I used resampling to down-sample the majority 2 to the same amount as 1 to eliminate the unbalance. Thirdly, to build the classification model, I will need to convert categorical values to numerical. I encode ADDRTYPE, COLLISIONTYPE, WEATHER, ROADCOND, LIGHTCOND to numeric values. Among WEATHER, ROADCOND, LIGHTCOND which has both 'Other' and 'unknown' values, I group these two into one value.

- **Feature selection:** I selected ADDRTYPE, COLLISIONTYPE, WEATHER, ROADCOND, LIGHTCOND as the features.
- I intended to use SPEEDING, UNDERINFL (under the influence of drugs or alcohol when driving), and INATTENTIONAND (not paying attention when driving) because there is a high correlation between these and car accidents, but the data is not representative enough.
- They only contain the value ‘Y’ and ‘Unknown’. So, I did not select these as my features.

Methodology:

- a. Exploratory data analysis: After understanding what each feature represents, I think exploratory can help me explore the relative value to understand their distribution.
- The following two visualizations represent the dataset with/without missing values. Based on the different distribution, we can see that the unbalanced data of target variable ‘severity’ affects the number of collisions among different address types.
- So, I decided to drop all the missing values and carry out my remaining modeling.

Result:

- The result of the logistic regression that we performed on the data collision dataset can be seen from the picture below:

	precision	recall	f1-score
1	0.63	0.65	0.64
2	0.64	0.62	0.63
Accuracy			0.63
Macro avg	0.63	0.63	0.63
Weighted avg	0.63	0.63	0.63

Conclusion:

- From the picture we can conclude that our model performed well. Based on the model and visualization, the light condition and road condition should be improved especially near First Hill.
- And also drivers can pay more attention to those factors when driving in Seattle to stay focused and safe.

Future direction:

The prediction of car accident severity is not completely finished. Based on the result, the dataset is under fitted, which means I will need to collect more data and features such as speeding and alcohol using to better train the model. In addition, the dataset only contains binary data for severity, however, there could be more scenarios for a car accident and also the people involved. I can stay tuned with the data and keep improving my models.