Capstone Project Report – Car accident severity

**Business Problem**

According to the wandering RV website, car accidents cause an average of 3,287 deaths per day worldwide. It also states that automotive crashes rank as the 9th leading cause of death and account for 2.2% of all deaths globally. It also states that road crashes and road traffic deaths cost USD $518 billion globally, costing individual countries 1-2% of their annual GDP.

If people could be warned about the possibility of getting into a car accident and how severe it would be, this will definitely help decrease the number of car accidents.

This project aims to help the car owners know the situations and variables that might cause them into getting into car accidents. It’s considered a classification problem, where we classify the severity of car accidents for the recorded accidents. It will also help in determining the most important features that contribute in getting into a car accident.

**Data Description**

The data set provided includes all types of collisions. Collisions will display at the intersection or mid-block of a segment. Timeframe: 2004 to Present. The Data set contained 38 features, only 15 of them will be used in our model.

The features that will be used in our classification model are:

1. ADDRTYPE: Collision address type: Alley, Block or Intersection.
2. COLLISIONTYPE: Collision type.
3. PERSONCOUNT: The total number of people involved in the collision
4. PEDCOUNT: The number of pedestrians involved in the collision.
5. PEDCYLCOUNT: The number of bicycles involved in the collision.
6. VEHCOUNT: The number of vehicles involved in the collision.
7. INATTENTIONIND: Whether or not collision was due to inattention. (Y/N)
8. UNDERINFL: Whether or not a driver involved was under the influence of drugs or alcohol.
9. WEATHER: A description of the weather conditions during the time of the collision.
10. ROADCOND: The condition of the road during the collision.
11. LIGHTCOND: The light conditions during the collision.
12. PEDROWNOTGRNT: Whether or not the pedestrian right of way was not granted. (Y/N)
13. SPEEDING: Whether or not speeding was a factor in the collision. (Y/N)
14. HITPARKEDCAR: Whether or not the collision involved hitting a parked car. (Y/N)
15. SEVERITYCODE: A code that corresponds to the severity of the collision: (3—fatality, 2b—serious injury, 2—injury, 1—prop damage, 0—unknown)

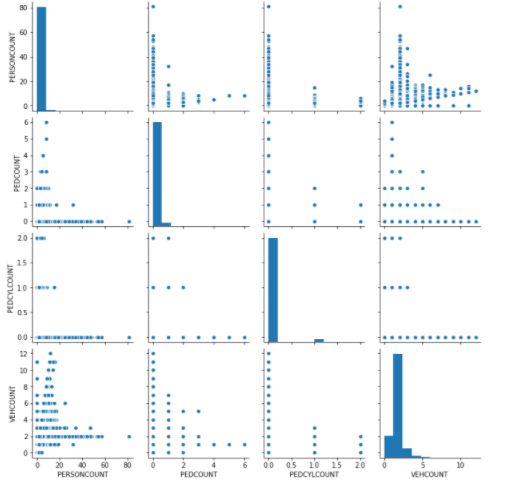
Some columns will need some pre-processing like replacing missing values and grouping relevant values together in one group, which will be discussed in the next section.

**Methodology**

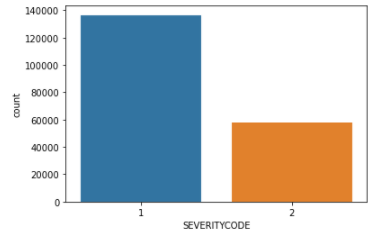
**Exploratory Data Analysis:**

When exploring the data set, distribution and count plots have been graphed for the numerical and categorical variables.

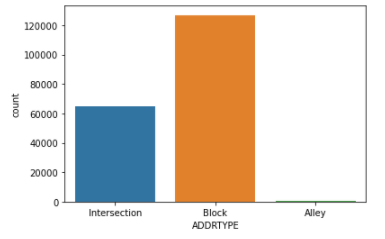
For the numerical variables, distribution plots were examined using the pair plot function in sea born library that examined the distribution of each feature and the correlation between numerical features, seen in the below figure:



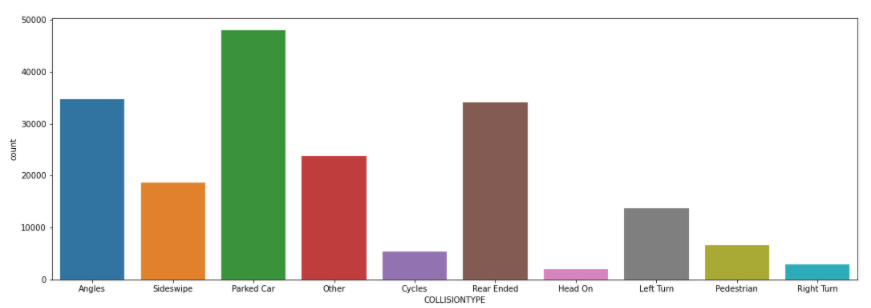
Regarding the categorical features, count plot is examined for each feature to check the different labels in each feature.



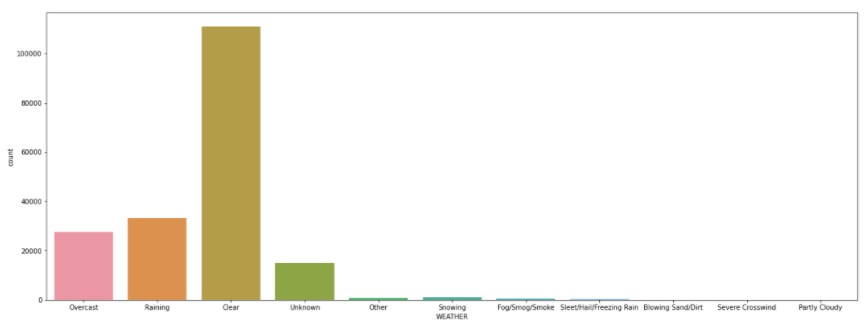
The majority of the data had accident severity with proper damage (1), while the minority led to injury (2).



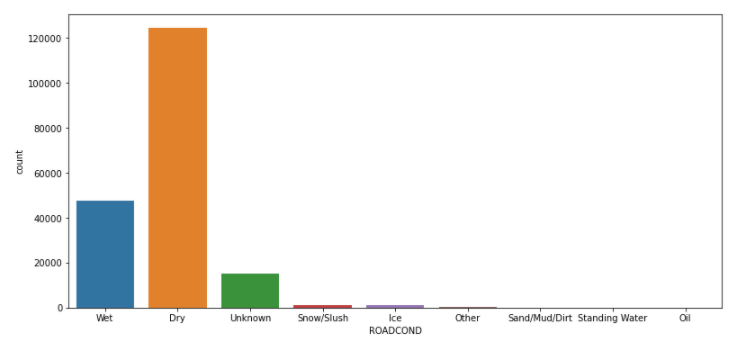
The majority collision address type is "Block" then "Intersection" while minimum in "Alley".



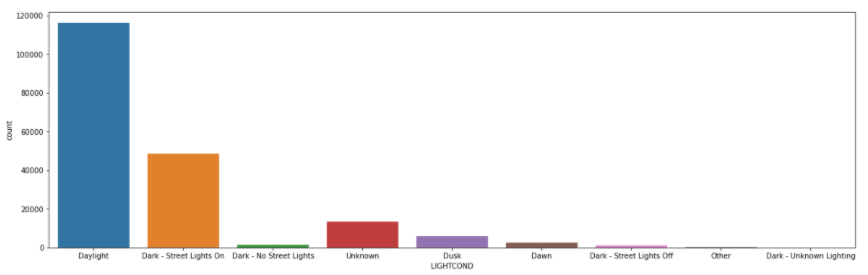
There are nine collision types in the dataset. The three majority collision types are: "Parked Car", "Rear Ended" and "Angles".



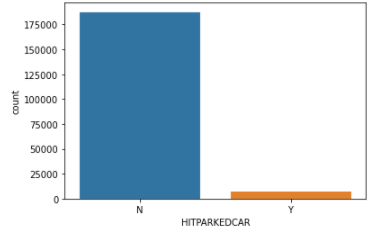
There are nine different weather conditions in the dataset. The majority three weather conditions are: "Clear", "Raining" and "Overcast".



There are seven road conditions in the dataset. The majority conditions are the "Dry" and "Wet" road conditions.



There are seven light conditions in the dataset. The majority conditions are the "Daylight" and "Dark - Street Lights on" light conditions.

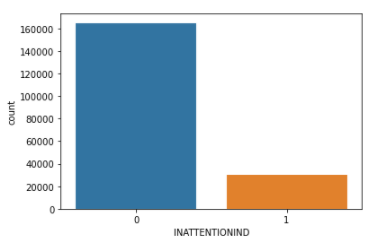


The majority of the collisions didn't involve hitting a parked car.

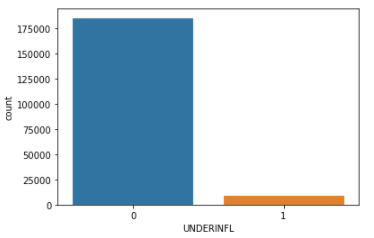
**Data Pre-processing**

**Dealing with missing values and mapping column values:**

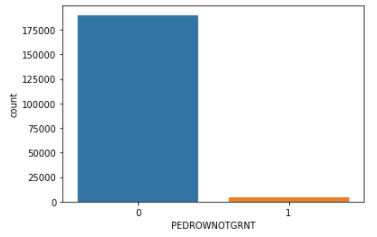
Most columns missing values were filled with the mode of each column. Other columns missing values were filled with the minority class. After filling the missing values, values for Y are mapped to 1 and N to 0.



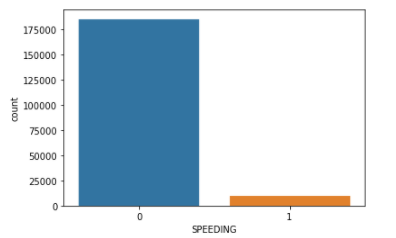
Replacing missing values in "INATTENTIONIND" column with "N" and then mapping the Y to 1 and N to 0. The majority of the collisions aren't due to inattention.



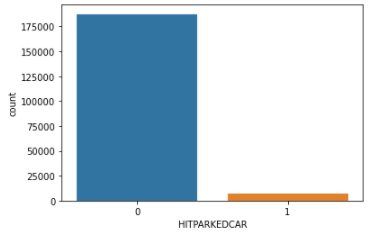
Mapping the Y to 1 and N to 0. Replacing missing values in "UNDERINFL" column with the mode of the column: N (0). The majority of collisions weren't under the influence of drugs or alcohol.



Replacing the missing values in the "PEDROWNOTGRNT" column with "N", then mapping the Y to 1 and N to 0. In the majority of the collisions the pedestrian right of way was not granted.



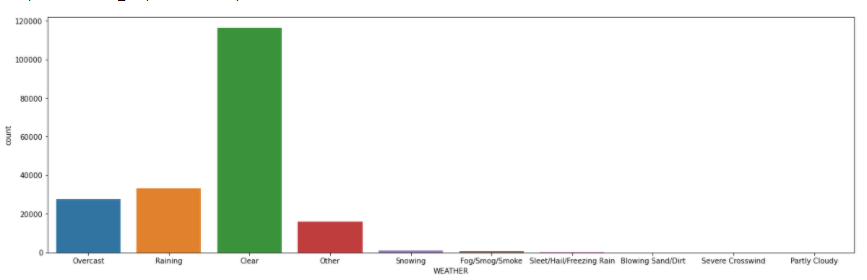
Replacing the missing values in the "SPEEDING" column with "N", then mapping the Y to 1 and N to 0. In the majority of the collisions speeding was not a factor in the collision.



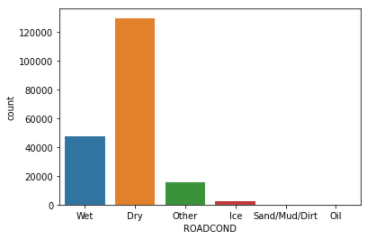
Mapping the values in the "HITPARKEDCAR" column the Y to 1 and N to 0.

**Merging column values:**

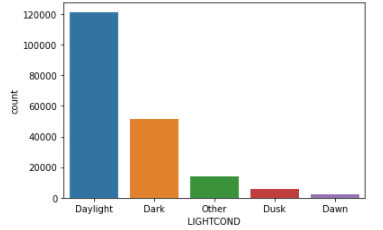
Some columns have relevant labels that can be merged together which therefore decreases the labels per column, like the “ROADCOND”, “LIGHTCOND” and “WEATHER” columns.



Merging the "Unknown" and "Other" classes together in the "Other" class in "WEATHER" column. Then replacing the missing values with the mode of the column: "Clear".



Merging the "Unknown" with the "Other" class, the "Standing Water" with the "Wet" class and the "Snow/Slush" with the "Ice" class in "ROADCOND" column. Then replacing the missing values with the mode of the column: "Dry".

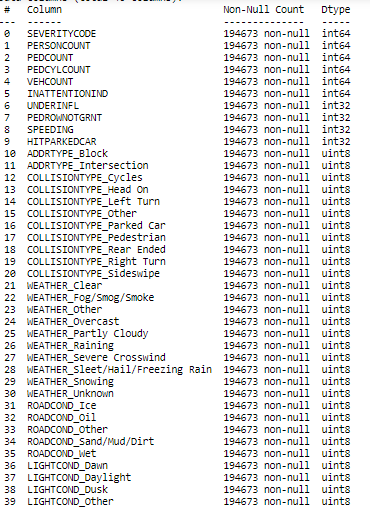


Merging the "Dark - Street Lights On", "Dark - No Street Lights", "Dark - Street Lights Off", "Dark - Unknown Lighting" classes with the "Dark" class and the "Unknown" class with the "Other" class in "LIGHTCOND" column. Then replacing the missing values with the mode of the column: "Daylight".

**One hot encoding:**

Categorical features are then one hot encoded in order to fit them in the classification models.

**Changing data types to integers:**



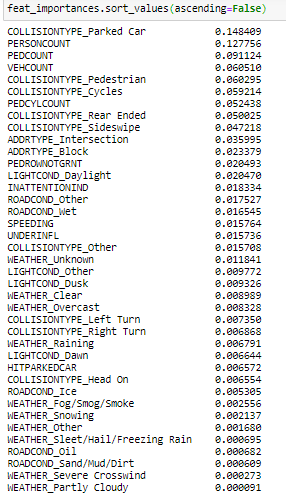
**Imbalanced data set**

The presented data set is an unbalanced data set were the label class is unbalanced towards the “1” class which represents severity: prop damage. To handle this unbalanced dataset, we did two approaches: oversampling the minority class and under sampling the majority class.

Oversampling the minority class produces a dataset of 272,970 rows, while Undersampling the majority class produces a dataset of 116,376 rows.

**Feature Selection**

Feature selection was done using fitting random forest classifier and choosing features that have importance greater than “0.01”.



**Models fitting**

After choosing the features set, we fitted multiple models: random forest, decision tree, logistic regression and K-nearest neighbors. We fitted the models on three stages, the first with the unbalanced data, the second using the oversampled dataset and the third using the under sampled dataset.

**Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Unbalanced data** | Random Forest | Logistic Regression | Decision Tree | KNN Classifier |
| Class\_1 F-Score | 84% | 85% | 84% | 82% |
| Class\_2 F-Score | 44% | 40% | 43% | 47% |
| Accuracy Score | 76% | 76% | 76% | 73% |
| AUC score | 0.79 | 0.79 | 0.78 | 0.72 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Under sampled** | Random Forest | Logistic Regression | Decision Tree | KNN Classifier |
| Class\_1 F-Score | 68% | 68% | 68% | 65% |
| Class\_2 F-Score | 72% | 72% | 72% | 69% |
| Accuracy Score | 70% | 70% | 70% | 67% |
| AUC score | 0.79 | 0.79 | 0.78 | 0.73 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Over sampled** | Random Forest | Logistic Regression | Decision Tree | KNN Classifier |
| Class\_1 F-Score | 68% | 67% | 68% | 65% |
| Class\_2 F-Score | 74% | 73% | 74% | 70% |
| Accuracy Score | 71% | 70% | 71% | 68% |
| AUC score | 0.8 | 0.79 | 0.8 | 0.73 |

**Discussion**

Accuracy scores dropped after balancing our data set from 76% to 70% and 71%. However, when looking at the F-Scores for each class, we find out that in the unbalanced case although the accuracy score is high, the F-Score for the minority class is very low around 40%, while that of the majority class is around 80%. Therefore, these models won’t perform well when predicting the minority class.

On the other hand, when the data set is balanced either by under sampling or oversampling, the overall accuracy scores decreased. However, the individual F-Scores of both classes increased to be around 70%.

Therefore, we can’t rely only on the accuracy score when measuring the performance of the classification models, especially when the data is unbalanced.

Another useful metric that evaluates the classification problems, is the AUC score, the closer the AUC score to 1 the better our model is. Our AUC score in most of the models if not all is around 0.8 which proves that they are performing well.

In the under sampling case, the three models: random forest, logistic regression and decision trees, have the same performance while that of the KNN classifier has lower performance.

In the oversampling case both random forest and decision tree models have the same and highest performance followed by logistic regression and then the lowest performing model, KNN classifier.

In general, the data set had some quality issues that had to be fixed before applying the models. If we changed the methods done to fix these issues, results may differ. Some columns that had values: Y (yes) and N (no), had only values of Y and the rest were Nans. We assumed that the Nan values are the N. This assumption led to the fact that these features had their majority values as N while the minority were Y. Changing this assumption will definitely change the distribution of these features.

Another point that if handled differently might have produced different results is merging the columns labels to have fewer labels per column for example only three or four values per column instead of nine and ten labels.

Finally, the feature selection method is only on of multiple feature selection methods and criterion which might yield different results than the ones reached if used.

**Conclusion**

Predicting the severity of car accidents is a crucial problem that is of great benefit to not only individuals but also to governments that lose huge amounts of money due to them. This problem was approached in this paper where we classified the severity of collisions after processing the data, selecting the most important features and best performing model to yield accuracy rate of around 70%.