CAR ACCIDENT SEVERITY IN SEATTLE

Joseph Zahar

IBM – Applied Data Science

Table of Contents

1 Introduction

All around the world, roads are shared by many motorized vehicles that have made transportation faster and more comfortable while supporting many countries' economic and social development. However, these vehicles cause a global problem. Car accidents are responsible for 1.35 million deaths on roadways every year. Almost 3,700 people are killed globally in road traffic crashes, where more than half of those killed are pedestrians, motorcyclists, and cyclists. Road traffic injuries are estimated to be the eighth leading cause of death globally for all age groups and the leading cause of death for children and young people 5–29 years of age [1]. Car accident severity depends on many factors and usually causes a high volume of traffic.

The main business problem is the severity of these accidents, which can sometimes be fatal and critical for pedestrians, bicycles, or vehicles. The project's objective here is to get people to safety, and the best way to solve this problem is to prevent those accidents as accurately as possible. Imagine being able to predict in advance the probability of an accident happening depending on the weather, the type of road, and many other features! This would help save many lives and help drivers get to their destination in the safest and fastest route possible.

The audience involved for this kind of project would be any Weather program, News channel, Radio station, government, or mobile application developer that wants to inform drivers of the possible areas of road crashes in a community and propose alternative solutions to get to their destination.

2 Data

For this project, the Dataset was shared on Coursera as a csv file available by Clicking here.
The data is composed of 38 features that accurately describe each car accident that happened in Seattle from 2014 to 2020. They are classified in terms of severity, type of weather and road condition, location, address type, and many more. However, for this case study, not all the attributes are useful as the main objective is to predict an accident's probability and severity. Therefore, the Dataset needs deep understanding and analysis before choosing the right attributes to reach our goal. For example, *SDOTCOLNUM, X, Y, LOCATION, INCDTTM, INCDATE, REPORTNO, COLDETKEY, INCKEY* and *OBJECTID* are features that give descriptive and detailed information about an accident, and are then not relevant to predict the severity of an accident in general. Moreover, *EXCEPTRSNCODE, EXCEPTRSNDESC, PEDROWNOTGRNT, SPEEDING, INATTENTIONIND* and INTKEY have a high number of missing data that would skew and bias our predictive model. After selecting the appropriate features, the new Dataset is balanced and preprocessed before feeding it to a supervised machine learning model that will learn to predict in the future the probability of a car accident based on the chosen attributes, in the most accurate way.

To consider the problem we can list the data as below:

- After visualizing and comparing the effect and weight of many attributes on the severity of an accident, the main attributes that will be used are the following:
  - ‘WEATHER’: A description of the weather conditions during the time of the collision.
  - ‘ROADCOND’: The condition of the road during the collision.
  - ‘LIGHTCOND’: The light conditions during the collision.
  - ‘JUNCTIONTYPE: The type of junction where the accident happened.
  - ‘UNDERINFL’: If the driver was under influence of drugs or alcohol.
  - ‘DAYOFWEEK’: Representing the day when the accident happened

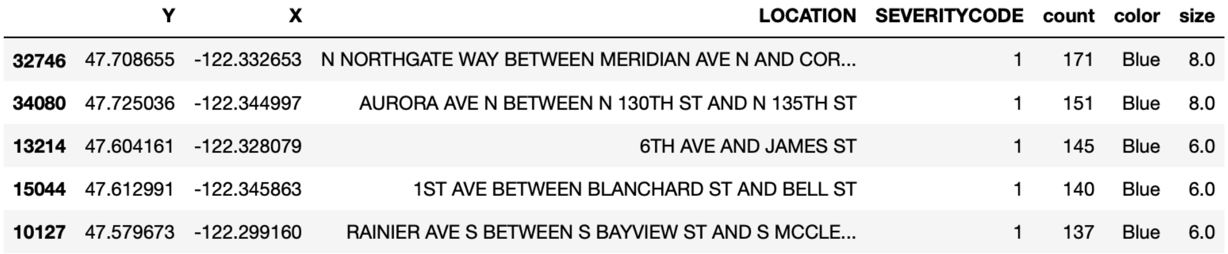
- After getting the number of labels for each type, we can clearly see that the data is unbalanced as there is a lot more of accidents of type 1 than 2. To balance the data we may either down sample the number of type 1 or duplicate cases of type 2 to get equal values of each label. Balancing the labels prevent to create a bias machine learning model.
- We also have to check if there is any missing data in any of our features column and replace them or either delete the corresponding row (in case of low number of Nan)
- Moreover, most of the features have categorical values that need to be transformed to numerical values in order to improve the predictability of our model.
- Finally, we will be using several types of supervised learning models and compare their performance to choose the best classifier for this project.

3 Methodology

3.1 Exploratory Data Analysis

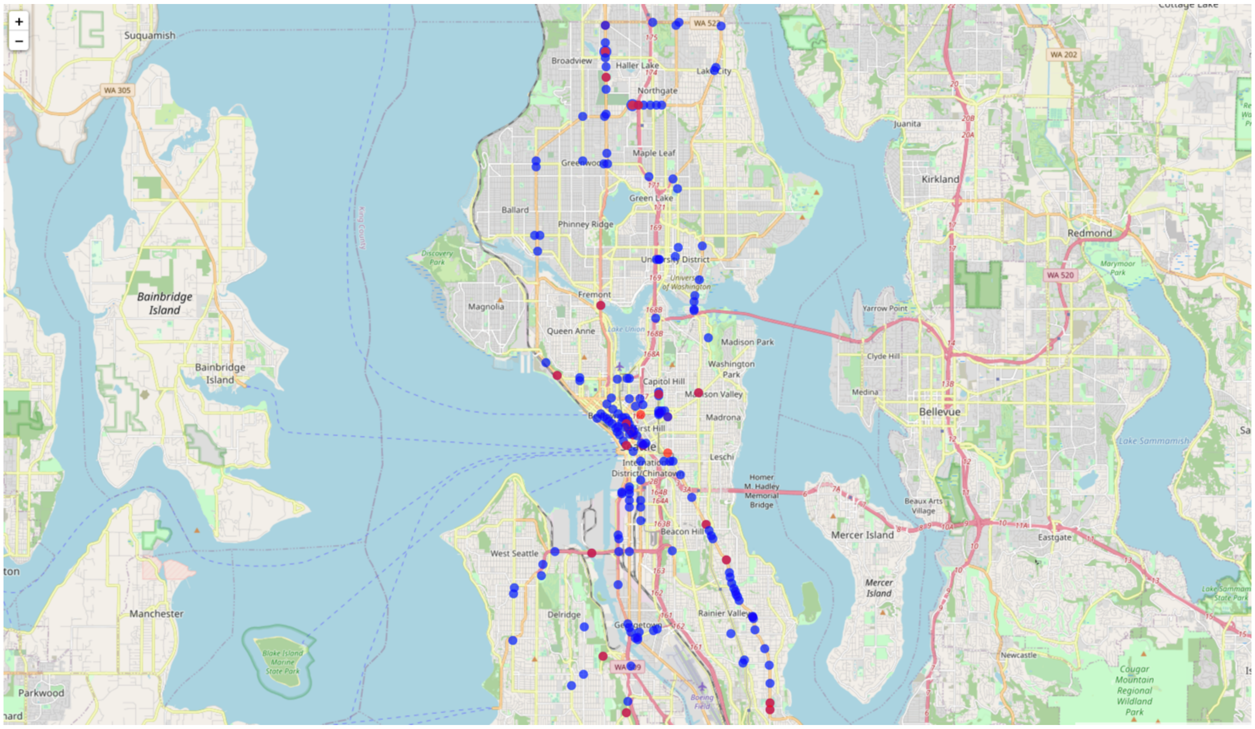
3.1.1 Map Visualization

In order to clearly visualize the dataset with its labels proportion, I used **folium** library to obtain geographic details of Seattle and its accidents. So I created a map of Seattle with markers indicating the location of the accident and its severity superimposed on top. The markers in Blue represent the accidents of type 1 severity, where property damage occurred, and the markers in Red represent accidents of type 2 severity, in place of injury. However, to more clearly visualize the map, I decided to only plot the locations where the number of accidents were higher than 50 at the same place. My data for this type of visualization was composed of SEVERITYCODE, LOCATION, X, Y and the total count of accidents at each location calculated using the count() function in python. Two more features were added to this data in order to get colored markers for each type of severity and their size dependently to the number of the number of accident at each place:



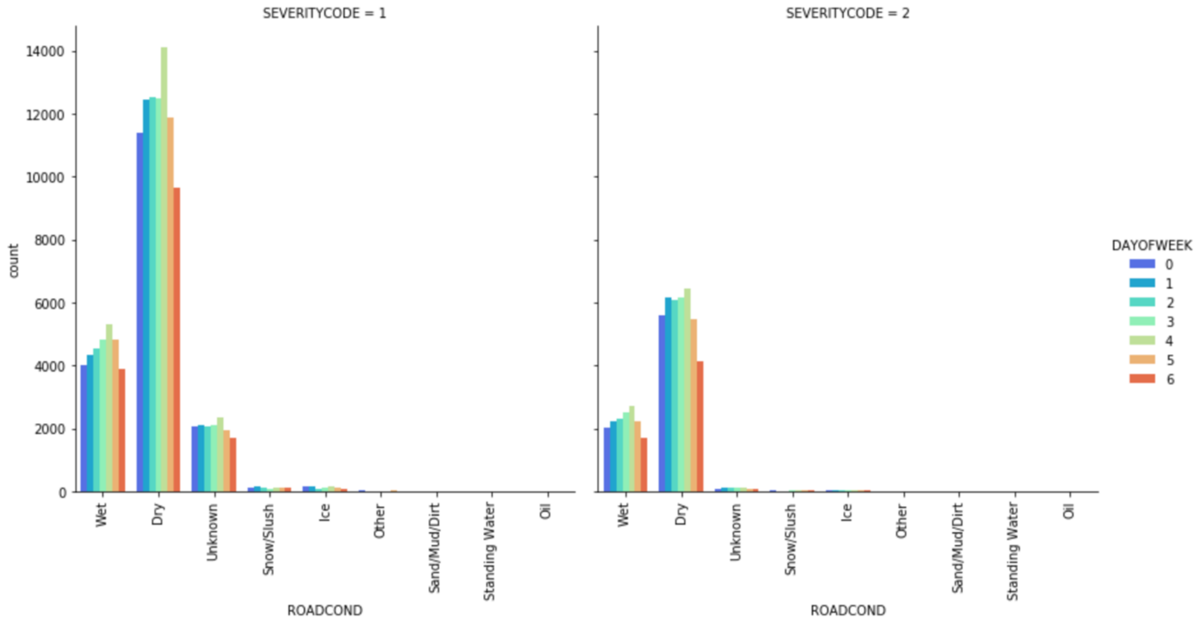
	Y	X	LOCATION	SEVERITYCODE	count	color	size
32746	47.708655	-122.332653	N NORTHGATE WAY BETWEEN MERIDIAN AVE N AND COR...	1	171	Blue	8.0
34080	47.725036	-122.344997	AURORA AVE N BETWEEN N 130TH ST AND N 135TH ST	1	151	Blue	8.0
13214	47.604161	-122.328079	6TH AVE AND JAMES ST	1	145	Blue	6.0
15044	47.612991	-122.345863	1ST AVE BETWEEN BLANCHARD ST AND BELL ST	1	140	Blue	6.0
10127	47.579673	-122.299160	RAINIER AVE S BETWEEN S BAYVIEW ST AND S MCCLE...	1	137	Blue	6.0

We can clearly see from the map below that the labels are unbalanced, there is a dominance in type 1 of accident and this need to be fixed before feeding the data to the machine learning models or we will obtain bias results. Moreover, most of these accidents seem to be located at the center of Seattle for both types.



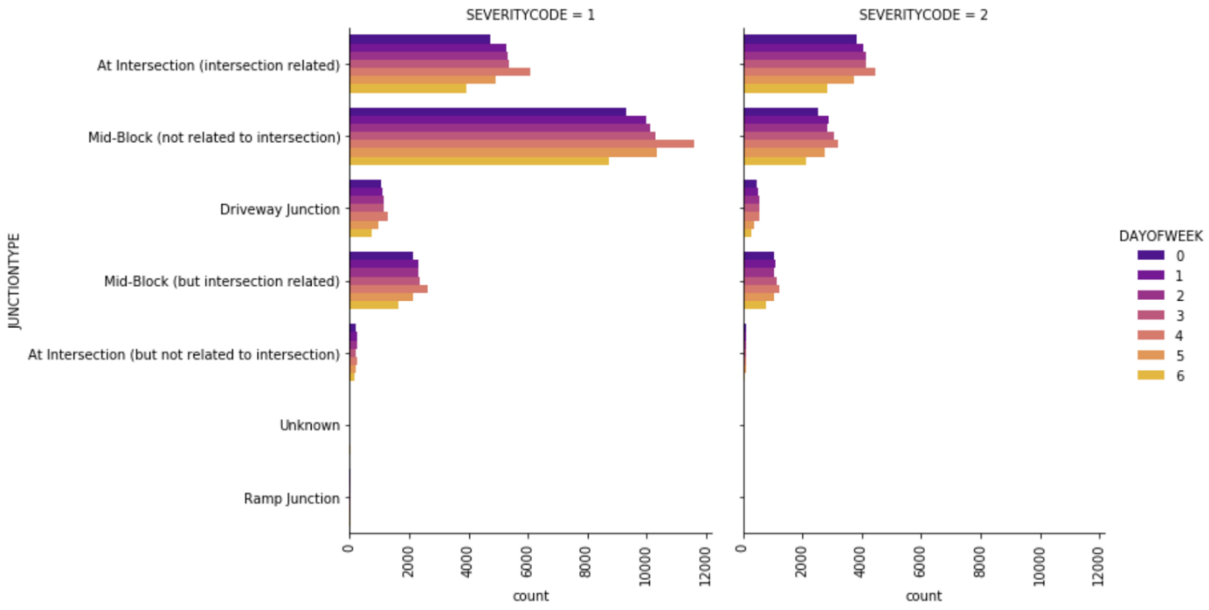
3.1.2 Relationship between Severity of an accident and the Road condition in term of day of the week

The hypothesis is that the more damaged the road, the higher the probability of an accident and the higher its severity. Although we would think that the more dangerous and sliding the road is, the higher the number of accidents, the truth is that dry roads are more dangerous than wet ones. What's even more strange is that car crashes are 80% more likely to happen on dry than on snowy and icy roads. We can see from both histograms below that for both type of severity, the count of accident follow the same trend and are dominated by three road conditions: First for dry conditions, where the count is the most elevated, followed by wet condition and unknown conditions. The rest of the accidents are mostly in icy and snowy conditions. Furthermore, for each conditions the number of crashes is higher on day 4 of the week → Friday, especially for type 1 severity on dry conditions, where the count is in average 2,000 more than for the rest of the days. The lowest number of car crashes occur on day 6 of the week → Sunday, with 4,000 less accidents than on Friday on a dry road for type 1, and 2,000 less for type 2 almost half less! For the rest of the days we observe an arrow shape on top of each categories, meaning that the count per week increased form Monday to Friday before decreasing again form Friday to Sunday. This strange results may be due to the increased care that people apply when driving in poor conditions, while they tend to be more negligent and reckless when the conditions are safe. Another reason is that Friday is the end of the week and most of the people celebrate and go out on Friday night rather than on other days. This may also include crashes due to people under influence of drug and alcohol especially on Friday nights, but this graph cannot prove this hypothesis.



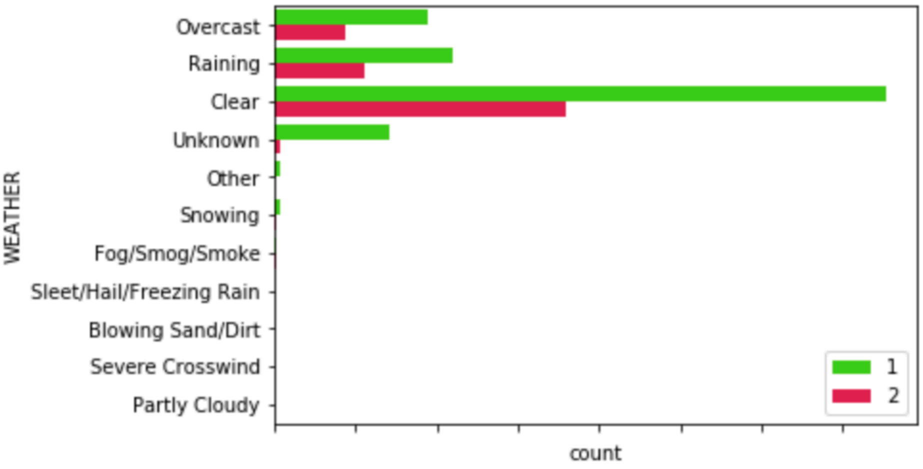
3.1.3 Relationship between Severity of an accident and the Type of junction in term of day of the week

For this case we seem to have more variety, the data looks more dispersed around the categories. For the type of severity 1, most of the accident happened at mid-block, not related to intersection) while for type 2 the ascendant category is at an intersection (intersection related). The gravity of car-crashes is then related to the type of junction as it varies depending on the junction. For the rest of the categories, the tendency look the same for both type of severity but still with a higher count for type 1 being the most dominant label. Mid-block, intersection related is the third highest count followed by a driveway junction and at an intersection (not intersection related). By comparing the count with the day of week we can see a similar tendency with the histogram analyzed beforehand on road conditions, with the highest count on Friday and the lowest count on Sunday.



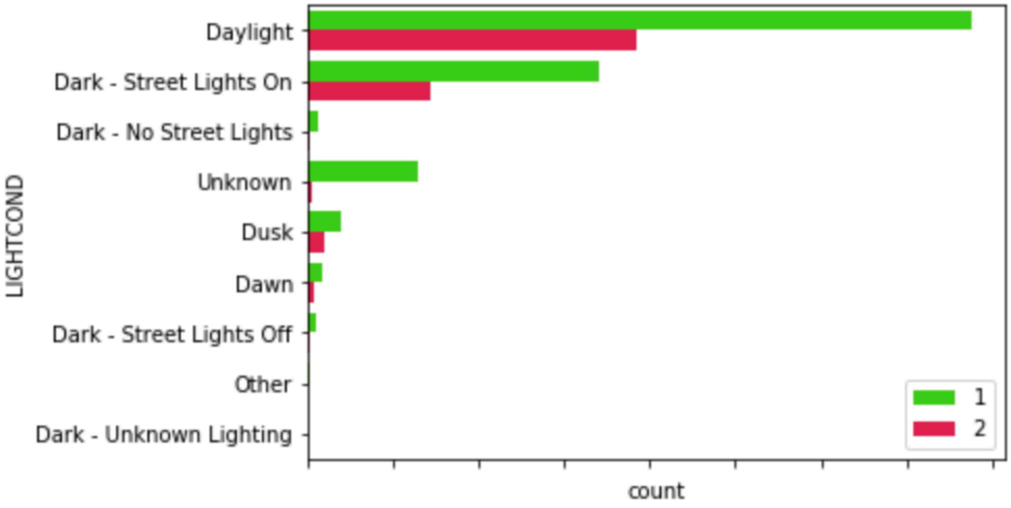
3.1.4 Relationship between Severity of an accident and the Type of Weather

Allegedly, the type of weather is related to the condition of the roads, in a rainy day the road conditions will be wet and in a snowing day the road conditions will be icy and snowy. Consequently the count and type of accidents will be similar to the relationship between severity and road conditions. In a matter of fact, the histogram below indicate a strong correlation and similarity with the past graph. The number of collisions of type 1 maintain a same count for Raining and Overcast weather next to a clear weather where the count is the highest. A similar inclination is perceived for type 2 collisions. The overcast weather here is a cloudy weather not related with the road condition since it does not affect it, but with the decrease in light and visibility affect appear to affect the count of accident. Also, we notice that most of the crashes in unknow weather is mostly of type 1 identically to unknow road condition with a majority of type 1 severity. Finally, the rest of the categories have a really low count throughout the years theoretically caused by the increase in care and attention of drivers in this weather.



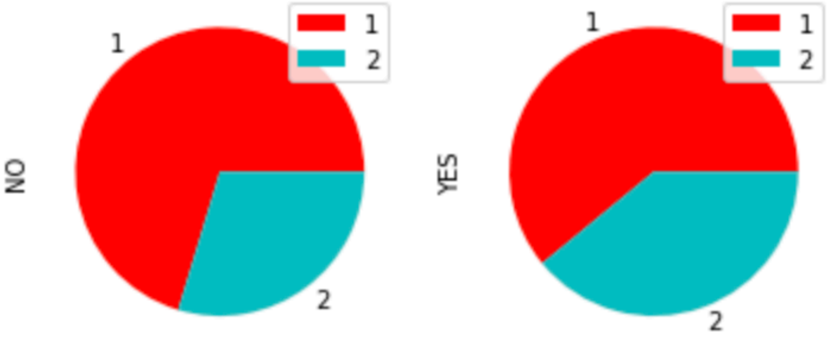
3.1.5 Relationship between Severity of an accident and the Light condition

I believe that the severity of an accident depend on the lighting and visibility on the road and that with a decreased visibility the more accidents are mean to occur. Following the histogram below, it turned out that the majority of accidents happened at daylight, in the clearest condition possible. This result is correlated with the same results and interpretation of the previous graphs were most of the accidents in both types of severity took place in the safest conditions possible whether it’s in road, light or weather condition. Moreover, Dark with street lights on condition is the second highest count of accident for both type succeeded by unknown conditions, Dusk, dawn and dark with street lights off.



3.1.6 Relationship between Severity of an accident and the driver’s condition

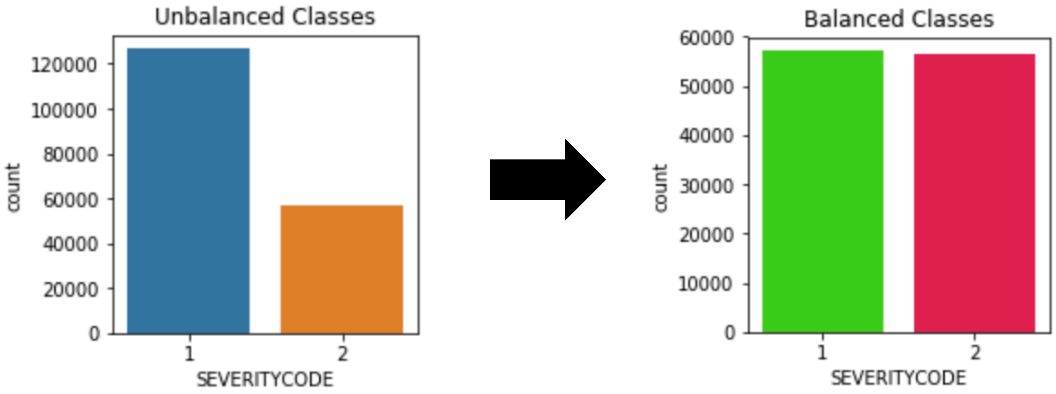
Before being able to visualize the relationship between the *UNDERINFL* attribute and the severity of the accident the data needed some processing. In fact, the feature was cut into four categories: 0, 1, Yes, No. I decided to rearrange these categories and assemble them into two main categories: YES and NO, where YES represent if the driver was under the influence of either drug or alcohol and NO represented the opposite. To do that, I grouped the categories 1 and Yes into 1 category and the categories 0 and No into another. The Data for this feature is now cleaner and ready for analysis. It appears that the severity of type 2 accidents have a higher proportion when the driver is under the influence of any drug or alcohol causing more probability of injury than when the driver is sober. Driving under the influence is the crime or offense of driving, operating, or being in control of a motor vehicle while impaired by alcohol or other drugs, to a level that renders the driver incapable of operating a motor vehicle safely. The pie chart below support this fact as more injuries happens.



3.2 Machine Learning Models

In predictive modelling there is two types of models, Regression and Classification, that can be used to predict the class of severity of an car accident. However for this project, we will only implement classification models that focus more on the probabilities of accidents type due to a binary data set (0,1). In effect, most of the features were object type, meaning that the values of the attributes were categorical instead of numerical, to solve this problem we need to transform these categories into binary set. I created a one hot coding function that change categories inside a feature into new features with numerical values, by insertion the categorical features into the function we will get a new data frame with new columns and only numerical values. The classification algorithms we are going to use will be fed our training set to fit the model, to later be able to classify and predict the severity of an accident as 1 or 2 using the previously fed algorithm.

After pre-processing the data, we still need to fix one major detail: balancing the labels. The majority of the data set is associated with the label of type severity 1, causing an unbalance of the classes that will create bias machine learning model. In order to prevent this of happening we have two options, we either under sample the data and randomly cut the majority class or oversample the data by adding new data labeled with the minority class. Despite that, the easiest and faster way to solve this problem is to by randomly cut out the accidents of severity type 1 to the point where both labels have the same count. The following bar charts clarify the procedure:

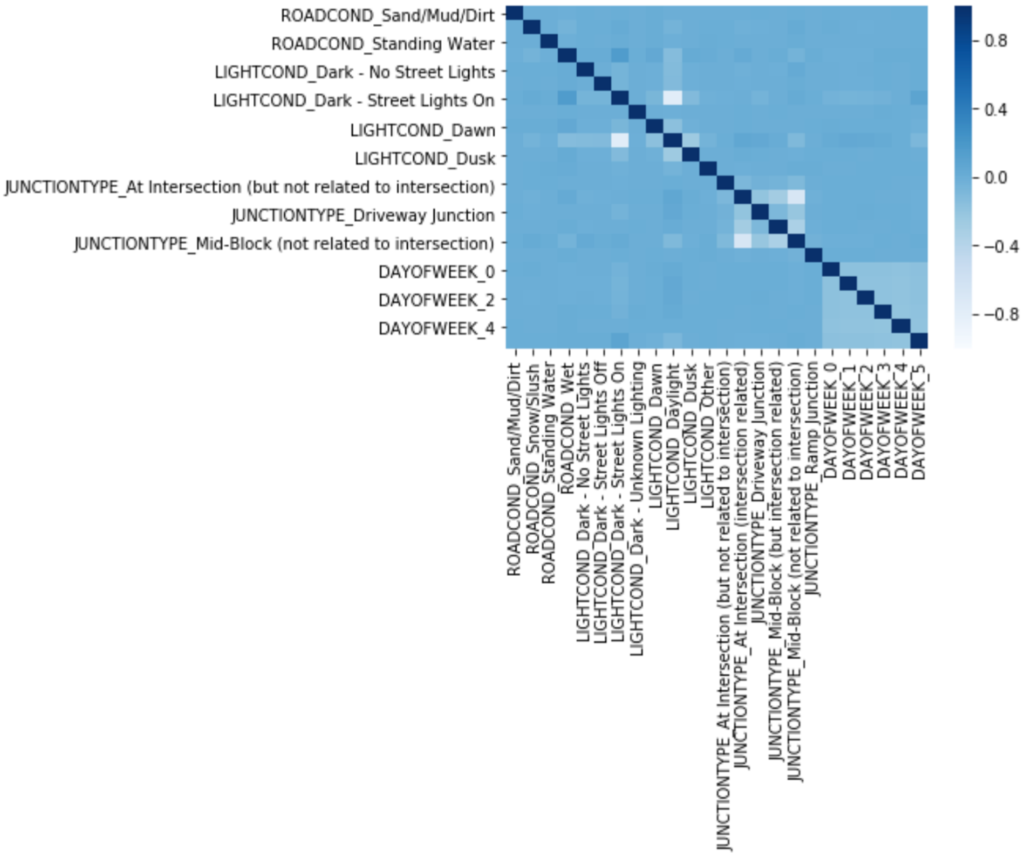


Having our data set ready we can now implement different classification problems:

- K-Nearest Neighbors
- Decision Trees
- Support Vector Machine
- Logistic Regression
- Random Forest Classifier

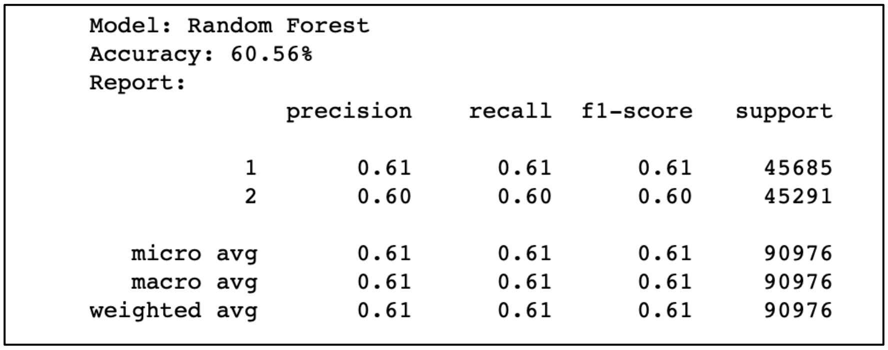
3.2.1 Classification Algorithms

I started by using the Random Forest Classifier model to get the correlation between each feature to find some kind of relationship between them. After splitting our balanced data set into test and training sets, the training set was fitted into the model. A random forest is an estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting. This model also happen to have the 'feature_importances_' function that will classify the importance of each feature depending on the weight it brings to the prediction. Furthermore, we can use 'X_train. corr()' and plot the result in a heat map to obtain the correlation of each feature with the other. The following map is the result we obtain. We can see that there is not a lot of high correlation between features and then low linear relation between some of them.



3.2.2 Evaluation Metrics

I applied classification models support vector machines (SVM), random forest, logistic regression, decision trees and K-Nearest Neighbors models to the dataset, using precision, recall, f1-score and mean accuracy as evaluation metric. However, The results all had the same problems. The predicted values had much narrow range than the actual values, and as a result, the prediction errors were larger as the actual values deviated further from zero. These results were not acceptable, Having larger errors on those predictions was obviously not desirable. For example this is what we obtained as evaluation matrix for the Random Forest:

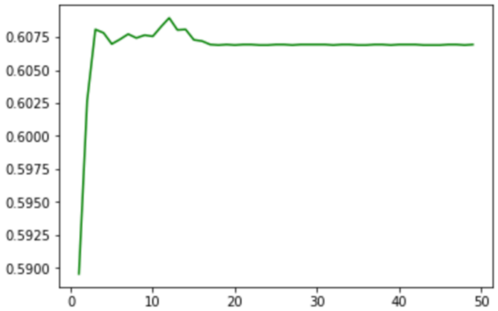


```
Model: Random Forest
Accuracy: 60.56%
Report:
              precision    recall  f1-score   support

           1       0.61      0.61      0.61     45685
           2       0.60      0.60      0.60     45291

   micro avg       0.61      0.61      0.61     90976
   macro avg       0.61      0.61      0.61     90976
weighted avg       0.61      0.61      0.61     90976
```

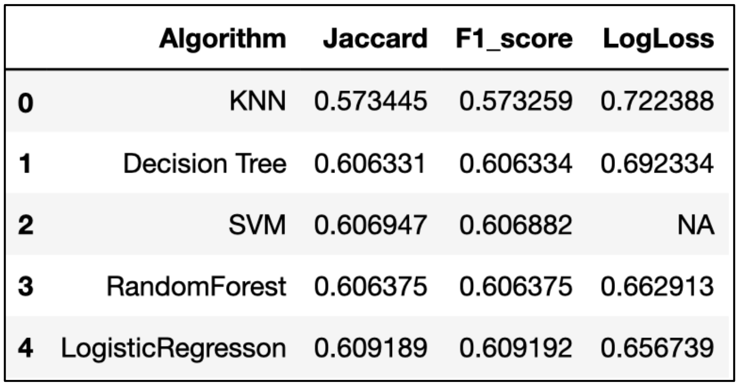
Let's keep in mind that the other models obtain approximately the same result of evaluation metric. A low F1 score is an indication of both poor precision and poor recall. The machine-learning annotator generates erroneous annotations and fails to find annotations that it should have found.

For the K-Nearest neighbors algorithm, one of the step to achieve an accurate model is to find the best 'k' or number of neighbors to the algorithm. In order to run the algorithm a lot of times with different value of k to find the one that gives the highest mean accuracy. In this project the best value of k is 20, that means that it's going to take the 20 nearest neighbors of the point to form a class.

For the Decision trees classifier we also need to plug a value for the number of dept of the tree with the 'entropy' criterion. It controls the maximum depth of the tree that will be created. It can also be described as the length of the longest path from the tree root to a leaf and is primordial to obtain a model with the best accuracy possible. We can find the max depth by following the same procedure to find the 'k' value previously, and plot a graph showing the f1 score value for each max depth. We obtain the plot below that gives the best score at max depth = 12:



4 Results and Discussion



	Algorithm	Jaccard	F1_score	LogLoss
0	KNN	0.573445	0.573259	0.722388
1	Decision Tree	0.606331	0.606334	0.692334
2	SVM	0.606947	0.606882	NA
3	RandomForest	0.606375	0.606375	0.662913
4	LogisticRegresson	0.609189	0.609192	0.656739

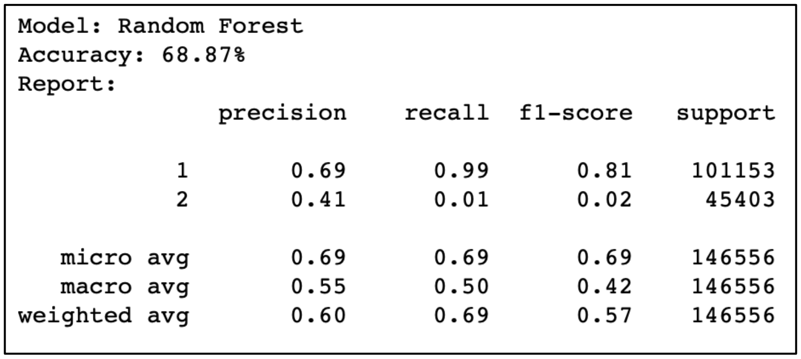
These are the results of the inferential statistical testing performed using Jaccard score, f1 score and log loss as evaluation metrics. The model with the highest performance is the Logistic regression with a log loss of 65.7% and f1 score approximately similar to its Jaccard score with a value of 60.9%. Overall, there was still significant variance that could not be predicted by the models in this study, the algorithms have low scores and need some more preprocessing. This trend in the results may be due to a low tuning of the model's hyperparameters or to the need to preprocess more the features beforehand. Another reason for this is possibly the similarity of the features for both type of accidents. In effect, the highest type 2 accident happen in the same light, road, weather condition of a type 1 severity accident. This similarity can cause the models to not be able to clearly define the strong attributes as most of them have the same tendency for both labels. This resemblance cause difficulty for the model to clearly define and classify a type of accident. Another source for this problem is the absence of other type of severity that would help the models better rank the accident, this code corresponds to the severity of the collision given by the metadata:

- 3—fatality
- 2b—serious injury
- 2—injury
- 1—prop damage
- 0—unknown

As we can see, the two types provided in the dataset (1,2) are not very different in their description and the cause of those types of accident would clearly be related, however, if the type of accidents given would be 1 and 2b or 2 and 3, it would be easier for the classifier to rank the accidents by severity. In fact, an accident of type 2b or 3 would have different conditions or unique attributes that would help more distinguish between different types. Another factor can also be the big amount of missing data present in the given data, as most of the missing data correspond to very useful attributes that would have helped defining the two types of accidents. For example, the 'SPEEDING' attribute have **185,340** missing values of the 194,673 available, meaning more than

95.2% of the accident didn’t specify if the driver was speeding or no. Theoretically, the faster the car, the biggest the damage and the higher the probability of injuries/fatalities. This argument would have a great impact on defining the type of accident if they were no missing value.

The table below represent the classification report of the Random Forest if the classes were **not balanced:**



```
Model: Random Forest
Accuracy: 68.87%
Report:
              precision    recall  f1-score   support

           1       0.69      0.99      0.81    101153
           2       0.41      0.01      0.02     45403

   micro avg       0.69      0.69      0.69    146556
   macro avg       0.55      0.50      0.42    146556
weighted avg       0.60      0.69      0.57    146556
```

We observe a higher accuracy than with balanced classes except that the score obtained are really high for type 1 accidents and very low for type 2 accidents. The recall and f1 score approach 0, a recall of 0 means that most of the cases have been predicted as type 1 → really bias model! This is due to unbalanced labels causing bias predictions that favorizes type 1 accidents for their high number in the training data set compared to type 2 accidents. In order to fix this problem, we cut out randomly half of the available rows that were labeled as 1, causing the loss of really valuable information that would have also helped improve the models accuracy.

In conclusion, I think the models could use more improvements on capturing severity type 1 and 2 individual characteristics by either improving hyperparameters or adding new data having less missing values and more different types of severity (mainly different than 1).

5 Conclusion

In this study, our goal was to predict accurately the severity type of an accident depending on the weather, road, and light conditions, depending on the day of the week, and if the driver was under the influence while driving. With the given dataset I retrieved the attributes that best qualified the problem and pre-processed them to later obtain the best predictions possible. I built five classification models and fed our data set to it to train it and measure their performance. These models can be very useful in helping weather stations or news program alert drivers of the probabilities of car crashes and its type of severity (damage, injuries, fatality,…). For example it would predict the severity of an accident when it is rainy, dark and slippery on specific areas.

6 References

1. World Health Organization (WHO). Global Status Report on Road Safety 2018. December 2018.[cited 2019 April 8]. https://www.who.int/violence_injury_prevention/road_safety_status/2018/en/external icon