Capstone Project

Car Accident Severity Prediction

**Sections**

* Introduction / Business Problem
* Data Overview & Usage
* Methodology
  + Data Cleansing & Structuring
  + Machine Learning Models
* Results & Recommendations
  + Accuracy Scores Across Dataset Types
  + Features Importance: DT & RF
* Conclusion

# Introduction / Business Problem

The City Municipality of Seattle (target audience) is facing a deficiency in the budget that is dedicated to addressing damage caused by car accidents, budget is mainly used for covering the cost of both human (health) recovery and property damage. One of the areas the city can improve within is the deployment and allocation of resources, it has been identified that resources have been deployed in situations where they couldn't add value in remediating or controlling the damage caused by car accidents. For example, medical staff such as EMT's cannot add value if there are no human injuries caused by an accident, therefore deploying them to the accident location is a waste of resource usage.

The Municipality of Seattle’s main goal is to allocate medical and natural damage control resources more efficiently and effectively based on the predicted severity type of an accident. If an accident severity is level 2 (human injuries outcome of an accident), then more medical resources such as ambulances & EMT's can be deployed. If an accident severity is level 1 (property damage outcome of an accident), then more of resources that address controlling damage caused by natural factors (tornadoes, storms, etc.) to physical property such as firefighters can be deployed.

# Data Overview & Usage

The Dataset is owned by the Seattle Department of Transportation who will grant Seattle's Municipality access to it, it's made up of 38 columns, 194673 rows, and comprises of attributes that describe existing features of when car accidents occurred and outcomes caused by such accidents.

Data will serve as an input to build a machine learning mechanism that predicts the severity type of an accident based on key factors causing the accident. The key factors are either human behavior or natural events causing the accidents. Human behavior factors are Drivers' Under the Influence, Speeding, & Not Granting Right of Way to a Pedestrian; these will serve as the X (independent) variables. Natural events factor entails the Weather Condition attribute, which will also serve as an X (independent) variable.

The Y (dependent) variable will be the severity type since we are trying to predict it.

If the severity level can be predicted, then the city can allocate appropriate emergency staff for efficiently addressing the damage caused by the car accident.

# Methodology

## Data Cleansing & Structuring

As mentioned in the previous section, the X features will comprise of:

UNDERINFL: Driver was under the influence of a substance, either alcohol or drugs.

SPEEDING: Driver’s speed was a cause of the collision.

PEDROWNOTGRNT: Pedestrian was not given the right of way to walk / pass (Y means not given, N means given)

WEATHER: Weather condition at time of accident.

Under the Influence feature comprise of values 0, 1 and some null values. The need for this feature is to identify if the driver’s abuse of any substance caused him/her to lose control of driving, therefore the type of substance doesn’t matter for this case study. “1” values were replaced with Y meaning the driver was under the influence of a substance, “0” values were replaced with N insinuating the driver wasn’t under the influence, and the null values were replaced with N assuming they weren’t under the influence since we don’t have data to support otherwise.

Speeding and the Pedestrian features are already in binary (Y & N) format. Observations with null values were replaced with N since we have nothing to support otherwise.

Weather feature comprises of multiple unique classes (Clear, Raining, etc.), however observations with null values were replaced with Unknown (which is an existing class) since we have nothing to suggest otherwise what the weather conditions were during those observations.

Regarding the Y feature, the dataset comprises of two unique values 1 & 2 and there were no null values. Finally, the data was hot-encoded and transformed to apply normalization.

For exploratory purposes, the dataset was balanced so the impact of inputting a balanced dataset in the machine learning algorithms can be compared with the unbalanced dataset.

The dataset in its original state was split ~70% for observations with Severity Code (Y feature) 1 and remaining ~30% of value 2. The majority sample severity code 1 was downsized to a sample count equal to severity code 2 observations.

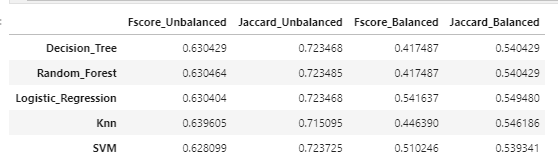
## Machine Learning Models

Decision Trees, Random Forest, Logistic Regression, Knn, & SVM were the 5 machine learning models operated. There are two main reasons for selecting these, the first is due to the dataset being of type supervised, and the second is because the datatype of the attributes inputted in the models are of class type / binary format. As mentioned earlier, all the X features are in binary format (Y vs N) except for one that comprises of multiple unique classes, while the target variable (Y) is in binary format. All of these attributes are labelled therefore utilizing these types of algorithms are more suitable since they are considered as Classification Models and specialize in supervised datasets. Each model was executed twice, first by inputting the unbalanced dataset and then by inputting the balanced dataset. Regarding evaluating model performance, F1-Score and Jaccard Index were calculated for each model operated instance (5 algorithms X 2 dataset types = 10) and X feature “importance scores” were calculated for Decision Tree and Random Forest models (2 algorithms X 2 dataset types = 4); the importance scores assess the impurity & informative levels. It seems as the importance score of a feature increases, the impurity level decreases and its informative level increases (considered more beneficial for model performance).

# Results & Recommendations

## Accuracy Scores Across Dataset Types

As mentioned previously, models were executed on two different dataset types, unbalanced & balanced. Below is a table displaying the models’ accuracy scores via inputting two different dataset types:



It’s clear each score is somewhat constant across its dataset type. To further elaborate, the F1-score is ~63% across all models for when using an unbalanced dataset while it’s ~42% when using a balanced dataset. The Jaccard Index is ~70% across all models when using the unbalanced dataset while it’s ~55% when using the balanced dataset.

Balancing a dataset is usually applied when a dataset has a very low number of observations for a particular class because model performance may be highly biased to predicting the common class (the one with significantly higher observations), therefore adjusting the dataset to increase observations of the rare class or by reducing the number of observations of the majority class leads to even distribution of observations of unique classes. As mentioned previously, the downsizing practice was applied by reducing the majority class (severity code of 1) in order to structure a dataset comprising of equal observations between severity code 1 and severity code 2 (i.e. 50% of observations of severity code 1 and 50% of observations of severity code 2).

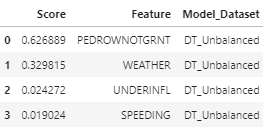
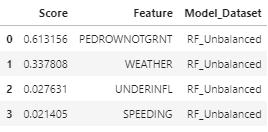
Since the accuracy scores across the unbalanced dataset are higher than the balanced dataset, it’s recommended to use the unbalanced dataset only. The problem of bias performance explained earlier doesn’t apply here due to two main reasons. Firstly, bias performance is usually identified when the accuracy scores are significantly high (greater than ~95%). Secondly, bias performance is usually tied to datasets that have a significantly higher number of observations of a particular class. For example, if the distribution of observations of the dataset were 97% of severity code 1 and 3% of severity code 2, then this could be problematic as the models don’t have enough samples of severity code 2 to structure an effective algorithm for predicting it (as mentioned earlier, dataset distribution is ~70% vs ~30% between the two different severity codes).

## Features Importance: DT & RF

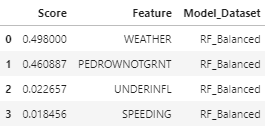
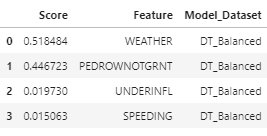
As a recap, the features importance scores assess the impurity and informative levels of input features to the Decision Tree and Random Forests models. The higher the score of a particular feature is, the higher the level of usefulness it is in predicting the Y variable.

The importance scores were computed on both models when inputting the balanced vs unbalanced datasets, below are the results.

Unbalanced Dataset (left table is Decision Tree while the right table is Random Forest):



Balanced Dataset (left table is Decision Tree while the right table is Random Forest):



For the unbalanced dataset, the Pedestrian feature (whether or not the pedestrian right of way was

not granted) is the top scorer and is significantly higher than the other features. For the balanced dataset, Weather is the top scorer however it’s not significantly greater than the Pedestrian feature which ranks 2nd in highest score. Both of these features are significantly greater than the remaining features Under the Influence and Speeding. For each dataset type, each feature’s score is consistent across model types. For example, Under the Influence feature’s score difference between the two models when using the balanced dataset is immaterial since the score under the Decision Tree model is 0.01973 and the score under the Random Forest is 0.022657. It’s interesting how the Weather feature ranks highest scorer in the balanced dataset, the sample downsizing practice may have led to balancing the number of unique Weather classes observations thus easing the algorithms’ processing.

Since it was previously recommended to utilize the unbalanced dataset, there is no need to provide recommendations regarding the balanced dataset. There is no need to review the Pedestrian feature since its importance score is relatively high. The Weather feature should be revised potentially by reviewing the distribution of samples among unique classes; this feature comprises of multiple unique classes therefore equal sample distribution among each class may improve its score. For the lowest two features Under the Influence and Speeding, their scores are concerning and its recommended to gather further samples and execute separate models with them as the only X inputs to identify if segregating them from other features has any impact on model performance.

Assuming there will be no change in the X and Y variables, Decision Tree model is the best option to adopt. The accuracy scores are constant among all models therefore running all models is extra noise (and potentially costly) to output the same score is not beneficial. The Decision Tree model significantly took less time to run in comparison to SVM & Knn, and it comes with the importance feature evaluation tool which can be useful at pinpointing which features require revision to improve model performance.

As a short recap of recommendations and reasoning, its best to apply the Decision Tree model with the unbalanced dataset due to three main reasons. Decision Tree takes less time to execute in comparison to all other models (especially SVM and Knn), it comes with the importance feature scoring tool that assesses each feature’s contribution in predicting the target, and the unbalanced dataset led to higher model scores (F1 & Jaccard Index) in comparison to a balanced dataset.

# Conclusion

In summary, the Municipality of Seattle would like to predict the severity type of a car accident so it can make more sound decisions on deploying damage control resources to the scene. Deployment resources mainly cover human and property damage caused by accidents; such damage types are the prediction target variable in the machine learning models that were operated. The X features used to predict the severity code can be classified as either human behavior or natural events contributing to the damage resulting from such car accidents. The decision has been made to adopt the Decision Tree model with an unbalanced dataset due its efficient processing time, model scores, and ability to evaluate features individually w/o adjusting the input to the model (no need to rerun the model by excluding other features to identify the impact of a particular feature). The Municipality will need to determine the resources deployment distribution based on the predicted severity code. To further elaborate, they should come up with a set percentage of resources deployment when the predicted severity type is property damage vs when its human injury. Hypothetically speaking to clarify the picture, set 70% of deployment resources to entail medical staff such as EMT’s and 30% of property damage control staff such as firefighters when the predicted severity is human injury.