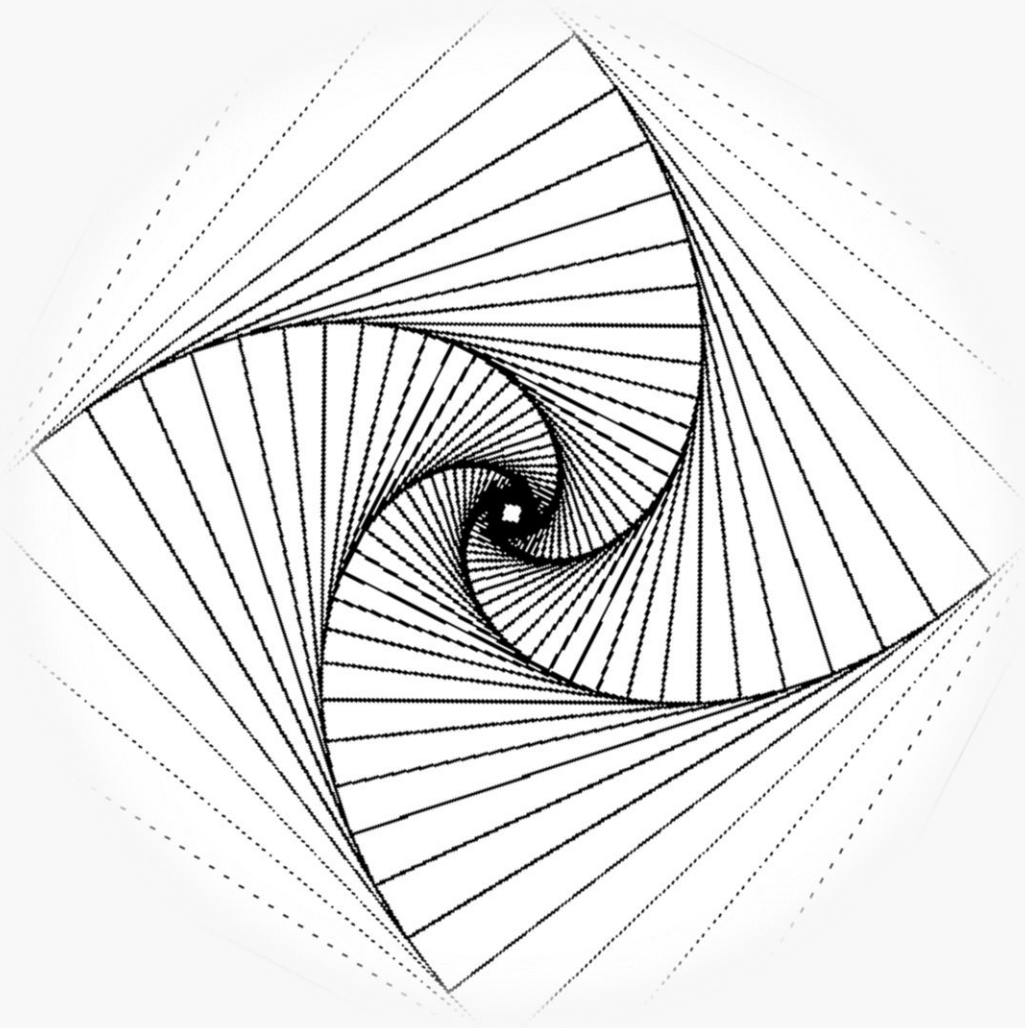


Predicting Car Accident Severity

INTEREST

Predicting car accident severity is useful to two groups:

- Emergency Response Services (ERS) – police, ambulance, firefighters, etc.
- Urban Planners

The goal is to create a generalised model which can predict the severity of a car accident before it has happened, using just a few environmental factors and locational information.

- ERS benefit have limited resources and will therefore benefit by having more efficient resource allocation strategies, in turn saving them money.
- Urban planners will be able to use locational data to improve the infrastructure around the most optimal locations

DATA SOURCE & CLEANING

- The data used in this project is the Seattle City Collisions with records from 2004 to present date. (Dataset can be downloaded here and the metadata can be downloaded here)
- It is provided by SPD and recorded by Traffic Records.
- In total, there are 194673 rows with 38 columns
- IDs, Codes, Location coordinates and other features with unnecessary information were dropped. The only features left were the one containing information about the external conditions as well as the features which provide general locational information.
- The final cleaned data contains 6 features, whose missing value rows were removed.
  - Features : weather, lighting condition, road condition, time and date, junction type and address type.

THE TARGET VARIABLE

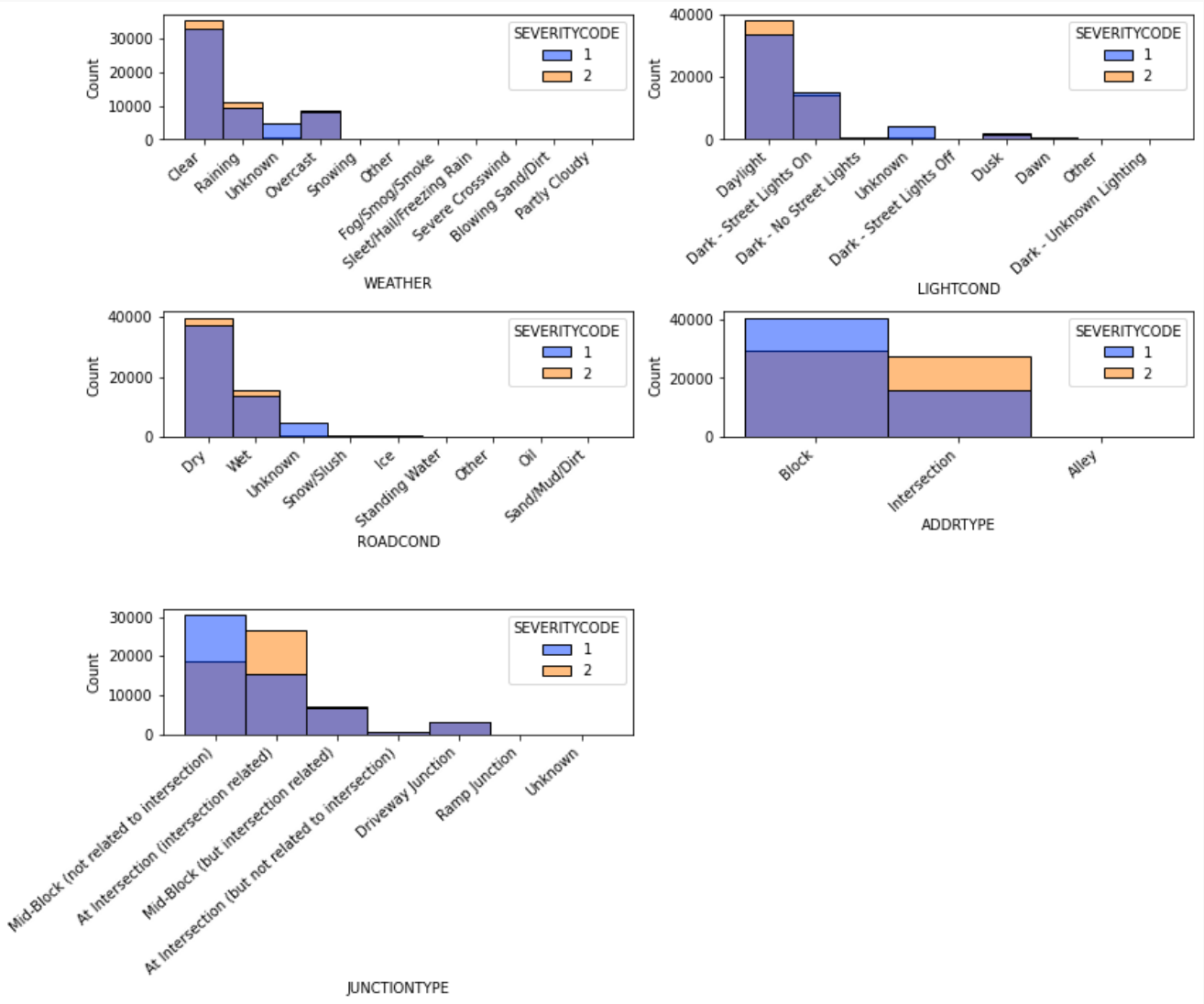
- The target variable is the severity of the accident
- There are two classes of severity present in the dataset:
  - Class 1 : Property Damage
  - Class 2 : Injury

- Since there are only two classes, the problem becomes a binary classification one

DEALING WITH AN UNBALANCED DATASET

- The cleaned dataset consists of 126276 entries with a class 1 severity
- However, there are only 56638 entries labelled as class 2.
- To correct this, the data was downsampled so that the there are an equal number labels.
  - To ensure fairness, the data was downsampled randomly.
  - The dataset was also shuffled after concatenation to ensure fair training in the later stages.
- The final dataset contained a total of 113276 rows with equal number of labels.

FEATURE ANALYSIS



A histogram for each feature and its categories, separated by the severity class.

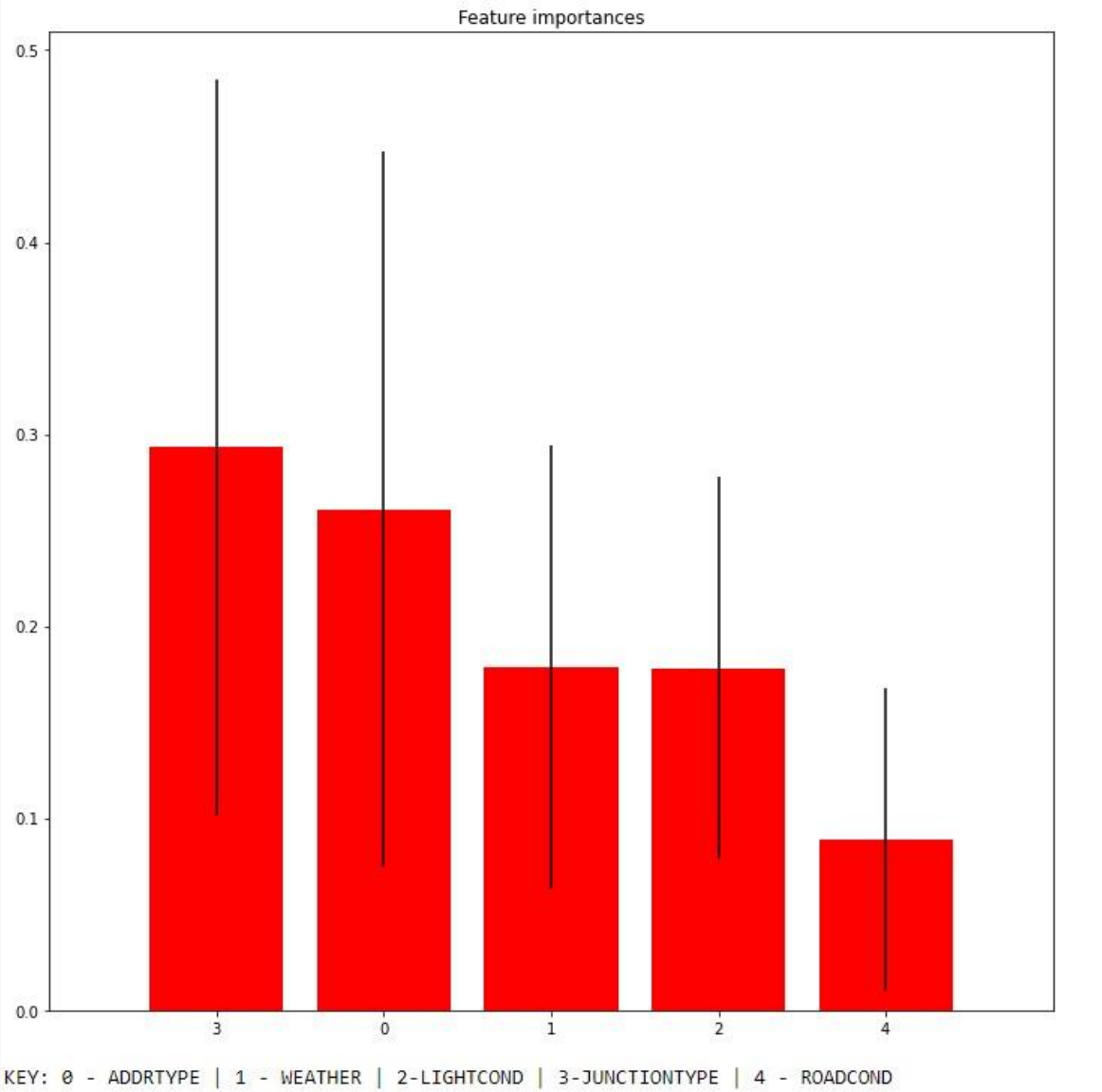
- Well separated distributions such as the JUNCTIONTYPE and ADDRTYPE suggest they are strong classifiers.
- Similarly, distributions which are almost identical, e.g. WEATHER and ROADCOND, have less classification strength.

FEATURE ANALYSIS - CONTINGENCY TEST

	FEATURES					
Statistic	WEATHER	LIGHTCOND	ROADCOND	JUNCTIONTYPE	ADDRTYPE	STATUS(Benchmark)
Critical Value (CV)	18.307	15.507	15.507	12.592	5.991	3.841
χ^2 Statistic $(CHI2)$	3666.709	3636.771	3729.823	8056.589	6804.426	0.170
Independence criteria: $(CHI2 \leq CV)$	Dependent	Dependent	Dependent	Dependent	Dependent	Independent
P-value	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p = 0.681$
Independence criteria: $(p \geq \alpha = 0.05)$	Dependent	Dependent	Dependent	Dependent	Dependent	Independent

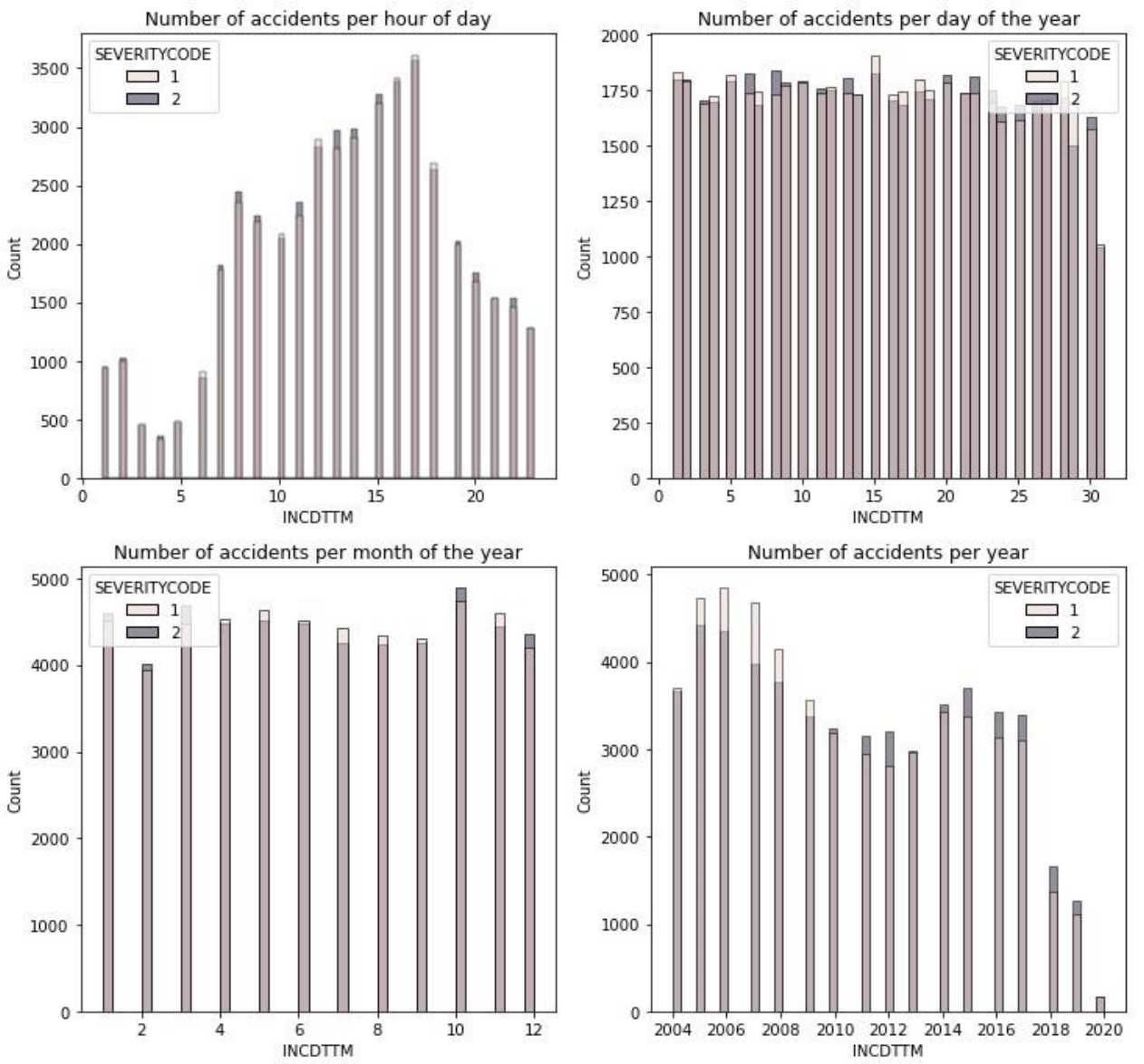
- The χ^2 contingency test used to test for correlation/dependency between categorical data
- All the features passed with high statistical significance p<0.001, meaning we are more than 99.9% sure these values are dependent on each other.

FEATURE ANALYSIS - RANKING



- The features were ranked using an extra-random tree classifier
- The feature importance shows the classification strength of each feature
- As expected, JUNCTIONTYPE and ADDRTYPE are the strongest classifiers and therefore have the highest importance
- On the other hand, as predicted, ROADCOND has the lowest classification strength and will therefore be excluded from the feature set

FEATURE ANALYSIS – TIME AND DATE



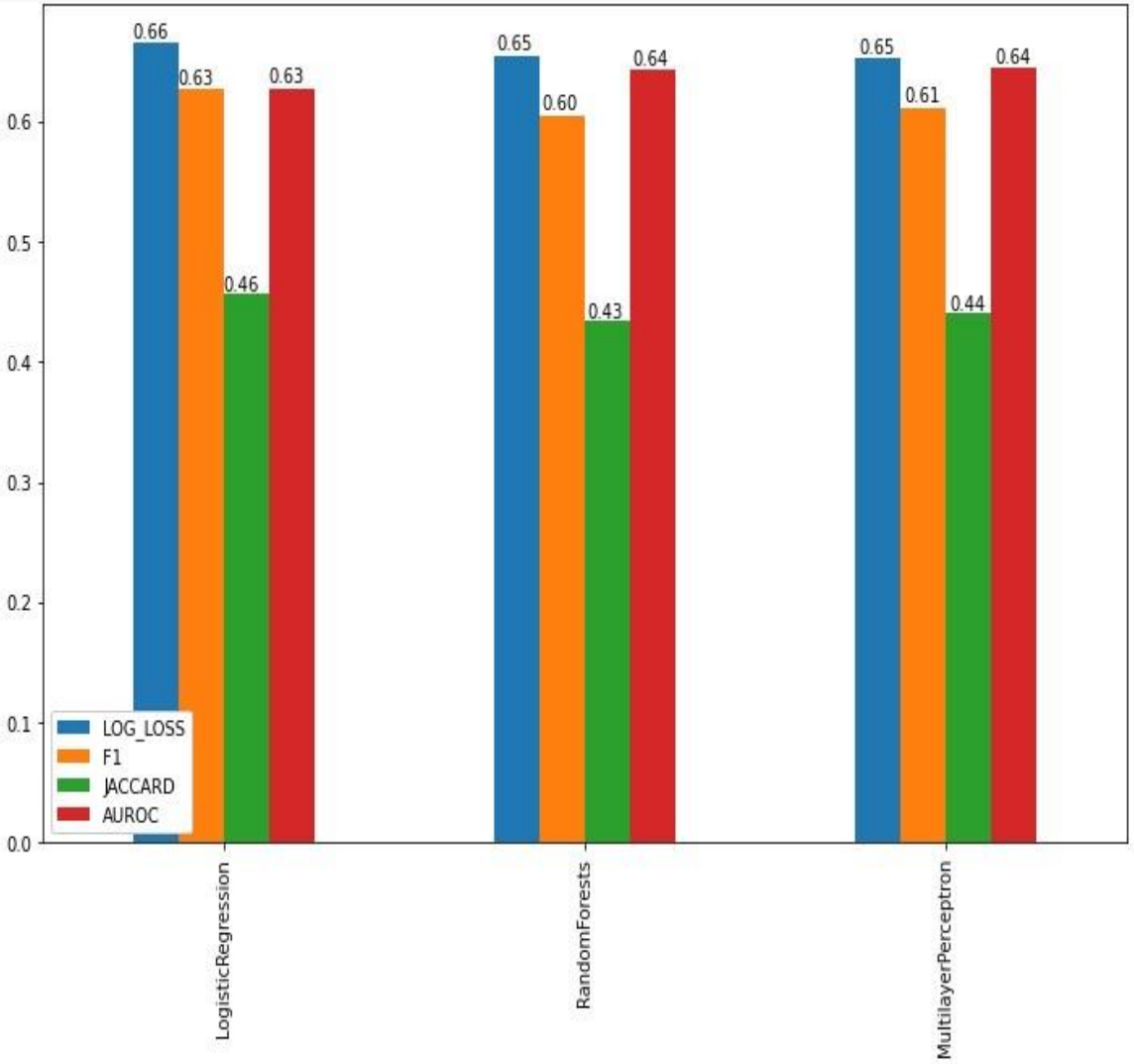
- Histograms of the components of the date and time, separated by the severity class
- The components are:
  - Hour of the day (1am to 23pm). Midnight was excluded due to an anomalous count, probably caused by negligent recording.
  - Calendar day
  - Month
  - Year
- The flat distributions, as well as the little to no difference in the shape of the separate class distributions, means that date and time is not a good feature to use in a classifier

MACHINE LEARNING - RESULTS

	LOG_LOSS	F1	JACCARD	AUROC
LogisticRegression	0.664819	0.626692	0.456349	0.627344
RandomForests	0.654687	0.604969	0.433673	0.642596
MultilayerPerceptron	0.651927	0.610904	0.439831	0.643755

- Three classifiers were trained and hyper-parameter optimised using GridSearchCV.
  - Logistic Regression
  - Random Forest Classifier
  - Multilayer Perceptron
- The best performing classifier was chosen to be one with the smallest logarithmic loss, due to the probabilistic nature of the problem.
- That is the **Multilayer Perceptron** classifier.

MACHINE LEARNING - DISCUSSION



All the models have relatively low accuracy (slightly above 50%) and a low Jaccard score.

The possible reasons for this are (not an exhaustive list):

- Only 6 features were used. We might need a more complex model with more features.
- Deeper parameter optimisation is needed. The parameter combinations in the GridSearchCV were kept low, due to insufficient memory.

However, there is one factor which is arguably the most important one. Vehicles are driven by humans and therefore the controlling factor behind a car accident is human error, which is unquantifiable.

CONCLUSIONS

- Built a model which can predict with more than 50% accuracy whether a car accident will be severe or not based on only 4 features.
- The accuracy however has a huge room for improvement
- It is owed to the fact that car accidents are random by nature and their probability and severity is controlled by human error, which is unquantifiable.
- The model succeeds in classifying the severities due to the features (external factors and locational information) affecting the probability of human error occurring.