# SC1015 Mini Project

Video Transcript

Hi everyone. This is Kevin, Liv and Yong Jian, from ECAD 1 group 8. The title for our SC1015 Mini project is the Prediction on Severity of Road Accidents.

## 1.0 Sample Collection & Practical Motivation

According to the Association for Safe International Road Travel, every year in the US, car accidents have resulted in more than 38,000 people dying and 4.4m people seriously injured. In fact, road crashes are one of the top 5 leading causes of death in the US.

In order to reduce occurence of car accidents, we must first study the factors affecting it. This leads to our problem: How do different factors affect severity of road accidents in the US?

In this project, we explore the US Accidents (2016-2021) dataset by Sobhan Moosavi found on Kaggle.

## 2.0 Data Preparation

After having a quick glance of the dataset, we notice that there are 47 columns and around 2.8m data. Our response variable will be distance and severity.

We first extract 20 relevant columns and further remove 3 more after finding out most data in those columns converges towards a value.

We continue data cleaning by removing or replacing NULL inputs. For example, we replace numerical variables with the mean or median depending on its skewness to maintain the distribution.

After realizing that the categorical variable “Weather conditions” has 128 unique categories, we group similar ones together until we are left with 15 of them.

## 3.0 Exploratory Analysis & Analytic Visualization

Moving on to EDA, we decide to focus on the data in 2016 due to the large dataset size.

For univariate statistics, we plot the box plot, histogram and violin plot for distance. As can be seen, the outliers are too much so we decide to remove // data with distance>5000. The data looks much better now.

We also plot a cat plot for severity using the original data after removing outliers for distance and find out that most accidents happen with severity of level 2 to 3 to 4 and no accident has severity of level 1.

For all numerical predictors, we plot the graphs, removed outliers and plotted the new graphs.

We also plotted cat plot for each categorical predictors.

To investigate the relationship between both response variables, severity and distance, we plot a box plot and observe a positive relationship.

By importing folium library, we also did geospatial visualization to observe where car accidents happen more frequently.

After that, we plot a heatmap and find out that correlation between each numerical variables and distance is extremely low. Perhaps this can be explained from the pairplot, where we can deduce that there is no linear relationship between the numerical variables and distance.

Considering time as a factor, we plot histograms and find out accidents happen more frequently at peak hours during weekdays, when people commute to and back from work.

Most accidents also happen when the weather is clear or cloudy, which make sense since these are the weather during most of the time.

In tableau, we made a summary for the whole EDA part. Do check out the tableau file for a more interactive and insightful EDA experience.

## 4.0 Machine Learning & Statistical Inference

Moving on to machine learning, we will be using a few ML models from scikit learn and imblearn.

First, we perform a multivariate linear regression, which is the Ordinary least squares LinearRegression(OLS).

From the graph plotted, we can see that this model is unsuitable. Even after using Standard Scalar // to rescale the data, the problem remains. This aligns with the EDA part where we found out the low correlation between the numerical variables and distance.We decide to shift our focus to the categorical response variable, severity.

Using OneHotEncoder, we first transform variables from categorical to numerical and obtain a new encoded dataset.

Then, we move on to decision tree classifier with max depth of 4.

After applying the data to the trained model, we notice a high true rate for class 2, but high false rate for 3 and 4, probably due to highly imbalanced data.

We try to improve the prediction by upsampling our sizes of severity level 3 and 4 to the size of level 2 and apply it to the model again.

This is the new model. Although it yields lower accuracy, we now have a less biased result. We will now deal with the issue on accuracy with cross validation, a statistical method to estimate accuracy of ML models.

Using grid search CV, we calculate the accuracy of the Decision tree classifier with different max depths.

We perform 10-fold cross validation to get the best model after considering the tradeoff between time consumption and accuracy of data. This is the model we get.

To obtain better results, we try the random forest classifier. You can imagine that we previously have a tree, but now we have a forest, improving accuracy.

We first build a model using n\_estimators of 100 and max depth of 4.

Similarly, we did 5-fold cross-validation and perform random forest classifier again using the best parameters we get. This is our final result, which yields a higher accuracy than what we obtain in the previous model.

Now, we try another machine learning model using our encoded and before-upsampled data just now. To deal with imbalanced classes, we are going to use The Balanced Bagging Classifier and the Balanced Random Forest Classifier from imblearn ensemble library.

Bagging is an ensemble learning method which selects random samples of data in a training set with replacement and train weak models independently. Ensemble learning aggregates weak learners to form a strong learner with less bias and variance.

Although the test data did not do as well as the training data probably due to the smaller test data size making the bagging process more inefficient, we can observe a huge improvement in the accuracy of this Balanced Bagging Classifier model compared to scikitlearn DecisionTreeClassifier models.

A balanced random forest randomly under-samples each bootstrap sample to balance it.

Using the same method, we build the model and the results show higher accuracy compared to the previous scikitlearn RandomForestClassifier models used.

All in all, despite the fact that we have upsample data, and cross validate for the previous scikitlearn models, the two models from imblearn ensemble library perform better.

Our last ML model is the logistic regression model which predicts the probability of an outcome.

After applying our data to the trained model,we have a classification report which examines the precision, recall, f1-score and support, which are very useful indicators to evaluate the usefulness of our data. Similar to the previous classification models, the accuracy is higher for level 2, lower for levels 3 and 4. The balanced accuracy of the train model is 0.40 and 0.41 for the test model, though it is less ideal.

We also used eli5, stands for Explain Like I'm 5 to show the feature of relative importance of each factor in this table here.

## 5.0 Intelligent Decision & Outcome

To conclude, we have built several models to predict severity of car accidents, and the best model was built using imblearn.

Our model takes inputs like temperature, weather condition and time and will predict the severity if an accident happens. Furthermore, most accidents happen during peak hours on weekdays, especially in California, when the car is on the right side of the street, and when there is no crossing, no junction, and no traffic signals. Perhaps this can be improved by reducing the number of cars allowed on the street and to have more traffic signals.

So what have we learnt from this mini project. Scikit-learn has taught us linear regression, decision tree classifier and random forest classifier. As for imblearn, we have gained better insights through balanced bagging classifier and balance random forest classifier. Some utility models which were very useful of course simple imputer, standard scaler, one hot encoder, resample, grid search cross validation, and eli5 particularly in allotting weightages on factors affecting car accidents. Tableau is undoubtedly a very useful visualization software which aids in communicating our data with our audience. As we know, road accident is one of the most prominent human killers. In fact, severity of road accident can be lowered or they can even be prevented when we are clear about the factors that caused the accident. To end with our presentation, we hope that we have a safer road for all road users in future.

Thank you.