BUSINESS UNDERSTANDING

CAR ACCIDENT SEVERITY :

Why we need to pay attention to this topic :

Every year approximately 5 million people lose their lives in accidents, out of which over 3 million + are car accident.

Hundred of thousands people gets disabled . This issue is one of those that can be over viewed by any means because healthy life is the most important thing in this world.

This issue can be solved by taking some safety measures as well as some precautions. These precautions and meausres depends upon the factors affecting those accidents.

DATA UNDERSTANDING:

DATASET: I have selected Data- Collision Dataset from kaggle.com , It has all the past few years accident details including date and time. This dataset has many columns but the most important columns that will help in creating a machine learning model are :

Severity Code:

It has two values 1 / 2 in which 1 represents that damage is done only to vehicle and 2 represents that damage is done to the person as well. This column will be predicted by our model.

Weather:

It represents the weather condition at the time of the incident. This column play a vital role in predicting the Severity Code.

Road Condition:

It also play an important role in predicting the severity code of the incident because road condition is an important aspect when it comes to car accidents.

Lights Condition:

It represents that when was person travelling was it a sunny day with loads of sunlight or it was a dark night and he faced difficulty in driving.

So , these 3 factors are most important columns and reasons that can lead to severity.

Data Cleaning:

The data cleaning is the process of giving a proper format to the data for its further analysis. The rst step was to deal with missing values and outliers.

Initially the latitude, longitude and road number were dropped form the data frame as more than a 50% of its values where NaN or 0 which is an outlier in this case. Then keeping with replacing the missing values, the analysis was divided in two groups of features. The rst group had in all features a label which described other cases, for instance the feature describing the atmospheric conditions had a value of 9 for any other atmospheric condition not labeled with the other 8 values. Therefore, the missing values and outliers were replaced with the other cases label for the features of atmospheric conditions, type of collision, road category and the surface conditions. For the second group of features instead, the distribution of their values was analyzed. Then two features were dropped, the infrastructures and reserved lanes, as the outliers represented more than 75% of its data. Finally with the rest of the features with missing values, the tra c regime, the number of lanes, the road pro le and shape and the situation at the time of the accident, the NaN and outliers were replaced with the feature's most popular value.

Modeling:

Di erent classi cation algorithms have been tuned and built for the prediction of the level of accident severity. These algorithms provided a supervised learning approach predicting with certain accuracy and computational time. These two properties have been compared in order to determine the best suited algorithm for his speci c problem. Firstly, the 839.985 rows where split 80/20 between the training and test sets, afterwards an additional 80/20 split was performed among the training samples creating the validation set for the development of the models. Then the data was standardized giving zero mean and unit variance to all features. Four di erent approaches were used:

1- Decision Tree,

2- Random Forest,

3- Logistic Regression

4- K-Nearest Neighbour

5- Supervised Vector Machine

The same modus operandi was performed with each algorithm. With the train and validation sets the best hyperparameters were selected and using the test set the accuracy and computational time for the development of the models were calculated.

The decision tree model was upgraded to the random forest. With the default random forest the features were sorted by impurity based importance in the prediction of the severity. Thus, the 10 least important features were dropped to decrease the computation complexity for the KNN and SVM models.

Keeping with 13 features the accuracy stayed the same and the computational time decreased signi - cantly. After evaluating the parameters for each algorithm these were the models.

Random Forest: 10 decision trees, maximum depth of 12 features and maximum of 8 features compared for the split.

Logistic Regression: c=0.001.

KNN: k=16

SVM: size of the training set= 75,000 samples.

The following visualizations show how the parameters for KNN and SVM models were selected. The SVM model is computationally ine cient with huge sample sets. Therefore, an equilibrium between accuracy and computational time was a found evaluating di erent training sizes.

Results:

The metrics used to compare the accuracy of the models are the Jaccard Score, f1-score, Precision and Recall.

In this case, the recall is more important than the precision as a high recall will favor that all required resources will be equipped up to the severity of the accident. The logistic regression, KNN, and SVM models have similar accuracy, however the computational time from the regression is far better than the other two models. With no doubt the Random Forest is the best model, in the same time as the log. res. it improves the accuracy from 0.66 to 0.72 and the recall from 0.45 to 0.59.

1 Proportion of predicted severe accidents that were truly severe.

2 Proportion of truly severe accidents that were properly predicted.

Conclusion:

In this study, I analyzed the relationship between severity of an accident and some characteristics which describe the situation that involved the accident.

Initially I thought that features such as atmospheric conditions, the lighting or being a holiday would be the most relevant ones, yet I identi ed the department, the day and time of the accident, the road category and type of collision among the most important features that a ect to the gravity of the accident. I built and compared 4 different class cation models to predict whether an accident would have a high or low severity. These models can have multiple applications in real life.

For instance, imagine that emergency services have a application with some default features such as date, time and department/municipality and then with the information given by the witness calling to inform on the accident they could predict the severity of the accident before getting there and so alert nearby hospitals and prepare with the necessary equipment . Also by identifying the features that favor the most the gravity of an accident, these could be tackled by improving road conditions or increasing the awareness of the population.