PROJECT TITLE: DETERMINANTS OF THE SEVERITY OF CAR ACCIDENTS

DESCRIPTION OF THE PROBLEM

Car accidents are a huge concern in most of societies. In the US, more than 3 million o f people are injured every year in a car accident, and more than 30,000 people die. For this reason, it is important to understand the drivers behind car accidents, so that bette r policies can be implemented to reduce the aforementioned numbers.

Yet, many are the factors that may be behind car accidents and their severity. In this p roject, I explore some of these factors and contribute towards the understanding of ho w accidents can be avoided. I develop a model to predict the severity of car accidents based on weather, road and light conditions. There two main parties that may be intere sted in a decrease in the number of car accidents. First, the government, which looks a fter its citizens, is interested in decreasing as much as possible car accidents and fatal ities. Second, the citizens, as they would like to drive as safe as possible.

DATA

To analyze the aforementioned problem and predict the severity of car accidents base d on weather, road and light conditions, the model will use data on Seattle city car acci dents.

The source of the data is the SDOT Traffic Management Division of the Traffic Record s Group. This dataset provides labelled data, contains 194,673 observations and 37 at tributes. The dataset provides rich information on characteristics regarding car acciden ts, such as severity of the accidents, type of vehicle or the weather and road conditions when the accidents occur.

The variable I am interested in predicting is 'SEVERITY', and it describes the fatality of an accident. The three attributes I will use to predict severity of car accidents are 'WEA THER', 'ROADCOND' and 'LIGHTCOND'.

After importing the dataset, I observe that the variables that I will use in the analysis ta ke the following values:

Figure 1 – Severity of car accidents

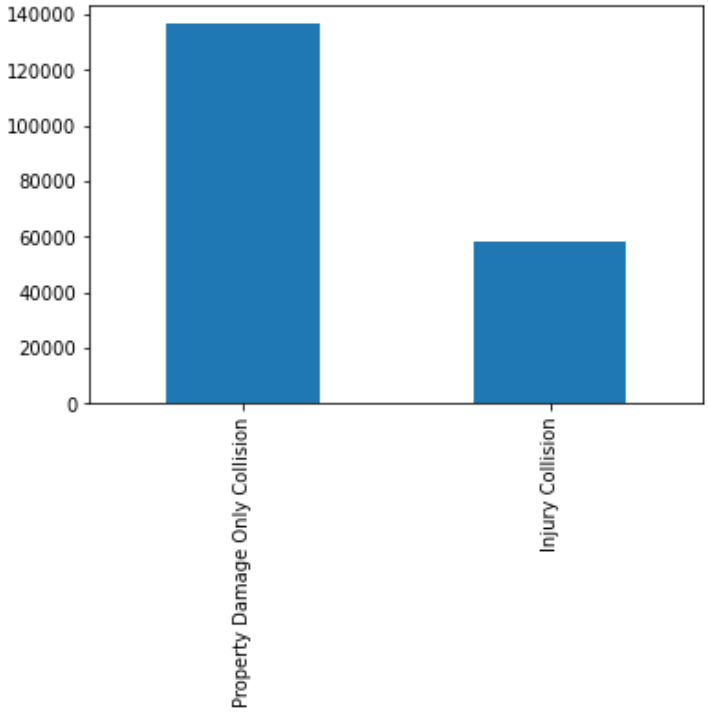


Figure 2 – Weather conditions

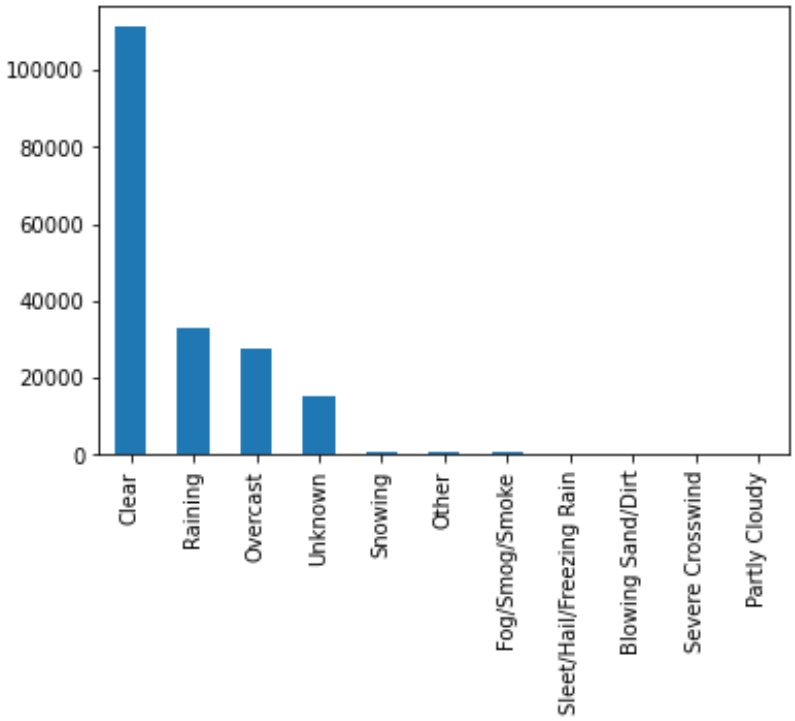


Figure 3 – Road conditions

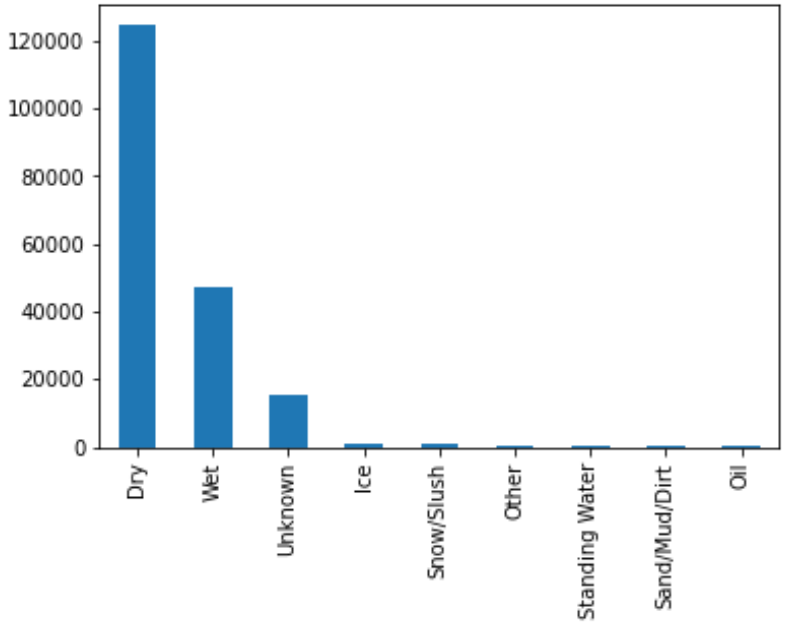
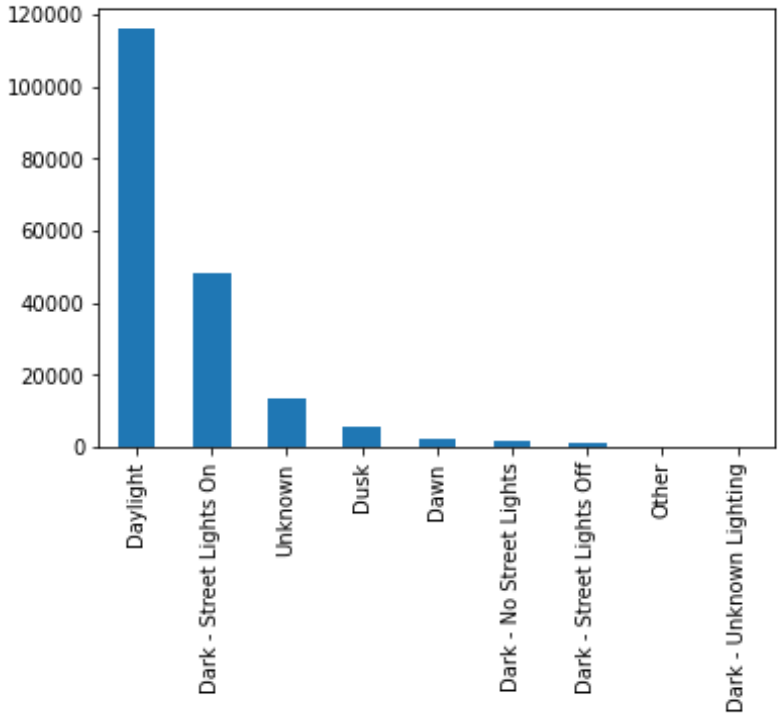


Figure 4 – Light conditions



As shown in the previous graphs, most of accidents are property damages without injuries. Moreover, most of the days are clear albeit rain and overcast also account for a high proportion of observations. This probably explains why most of the days road conditions are dry, albeit again it is not rare to find wet roads. Finally, the data provides information on drivers usually driving when there is daylight, albeit an important part of drivers also drive during night.

METHODOLOGY

I will build three models that predict the severity of car accidents based on weather, road and light conditions.

The first model will consist of a K-Nearest Neighbors algorithm. The K-Nearest Neighbors algorithm is a classification algorithm. Using labelled data, the algorithm learns how to label which used to be unlabelled data points. The K-Nearest Neighbors algorithm does the previous process based on the similarity of the unlabelled data point to other labelled data points.

The second model will consist of a Decision Tree. Decision Tree is a type of model that recursively splits the data so that it classifies it into more homogeneous groups that the groups before the split. Decision Trees do so by splitting the data attending to several attributes and calculating the entropy of each group after the split. Entropy is a measure of how similar is a group of data points regarding the outcome variable. A higher entropy indicates a very dissimilar group. After calculating the entropy for each group, decision trees choose the split of the attribute that provides a higher information gain, which is, that reduces the entropy more relative to the entropy of the data before the split.

The third model will consist of a logistic regression. Logistic regression is a type of statistical and machine learning model that predicts a binary/categorical outcome based on other attributes of the individual. Using a sample of data for which we have information on a number of attributes and the outcome of interest, we estimate a set of parameters that indicate how important are the different attributes to predict the outcome of interest. We then use the parameters estimated together with the attributes information of other individuals for whom we do not know the outcome of interest to predict the outcome of interest.

Given the setting of the data, and given that my outcome variable is a binary variable, the three previous models are ideal settings to give an answer to my problem. Comparing the results of the three previous models, I will provide robust evidence on whether weather, road and light conditions are important predictors of severity of car accidents.
I next prepare the data to estimate the previous three models and run the analysis.

DATA CLEANING

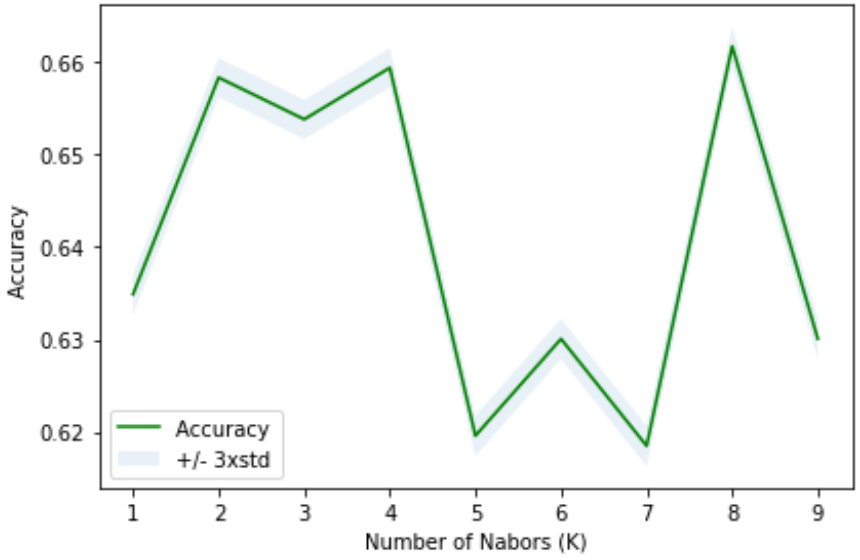
My explanatory variables of interest are string variables and I am interested in assignin g a code number to each string value. I therefore start the analysis by factorizing my w eather variable and assigning it missing values when the category of weather is unkno wn. I then drop rows where there is a missing value in any of my explanatory variables of interest. My dependent variable is already an integer variable so I do not need to do anything.

My feature variables are weather conditions, light conditions and road conditions. I sta ndardize my feature variables so that they have a mean of 0 and standard deviation of 1. This makes more sense as some of the models during the analysis will predict value s in the outcome of interest through calculating distances between the attributes of the individuals we are predicting and those who we use in the train model. Moreover, stan dardizing variables makes more comparable the estimates of all the attributes, and so i mproves the interpretation of the model.

MODELS

The first model I estimate is the KNN model, where I train the model using 70% of the s ample and then use the estimated parameters and the attributes of the sample I have l eft for testing to predict the outcome variable for the sample that has been left for testin g the model. I compare the accuracy of the model for different numbers of K-nearest n eighbours. I will select the model that provides with the highest accuracy, which in this case is K-nearest neighbours=8.

Figure 5 – Performance of KNN model



After having estimated the model using the value of K-nearest neighbour with which th e model performs the best, I calculate the Jaccard Score and Log loss Score to compa re the model the machine learning models I will estimate later, which are a decision Tr ee model and a Logistic regression model. I obtain the following scores for my KNN m odel:

Jaccard Score KNN model: 0.661720720192364
Log loss Score KNN model: 2.0428966782106426

I next estimate the decision tree model with different depths and again compare the pe rformance of each decision tree model and choose the depth that provides with the hig hest accuracy. I obtain the following accuracies for different depths:

DecisionTrees's Accuracy depth=1: 0.6732742947627706
DecisionTrees's Accuracy depth=2: 0.6732742947627706
DecisionTrees's Accuracy depth=3: 0.6732742947627706
DecisionTrees's Accuracy depth=4: 0.6732742947627706

As shown, independently of the depth of the decision tree model, I obtain a decision ac curacy of 67% using my decision tree model. I also calculate the Jaccard Score and Lo g Loss Score of my Tree decision model, in order to compare the performance of this model with the performance of my KNN model and Logistic regression model. These a re the scores I obtain for my decision tree model:

Tree decision andJaccard Score Tree Decision: 0.6732742947627706
Log loss Score Tree Decision: 0.6313378435663851

Finally, I estimate a Logistic Regression, which in our setting is ideal since my depend ent variable of interest is a dummy variable providing information on the severity of the car accident. I estimate the logistic regression using the train data, and use the estimat ed parameters together with the data that I have left for testing in order to predict outco mes regarding the severity of accidents for the test data.

I also calculate the Jaccard Score and Log Loss Score of my logistic regression in ord er to compare this model with the previous to machine learning models. These are the scores I obtain for the logistic regression model:

Jaccard Score LR: 0.6732742947627706
Log loss Score LR: 0.6315159471562749

RESULTS

The analysis has consisted of a KNN model, a tree decision model and a logistic regression that have been used to predict the severity of car accidents based on weather, road and light conditions. All models have predicted that weather, road and light conditions are important predictors of the severity of car accidents.

Regarding the performance of each model, the Jaccard Score and Log loss score of all models were the following:

Jaccard Score KNN model: 0.661720720192364
Log loss Score KNN model: 2.0428966782106426

Jaccard Score Tree Decision: 0.6732742947627706
Log loss Score Tree Decision: 0.6313378435663851

Jaccard Score LR: 0.6732742947627706
Log loss Score LR: 0.6315159471562749

Therefore, not only the different models predict that weather, road and light conditions are important predictors of the severity of car accidents, but also all models predict so in a very similar way, which provides robustness the results obtained.

DISCUSSION

The results show that weather, road and light conditions are important drivers of the severity of accidents. This is important for both policy-makers and drivers.

For policy-makers, the results lead to the conclusion that more money should be invested to improve road and light conditions as well as to improve weather forecasts so that the severity of accidents is reduced.

For drivers, the results lead to the conclusion that they should take into account weather, road and light conditions when driving in order to reduce the severity of car accidents.

CONCLUSION

The analysis, based on three machine learning models, has concluded that weather, road and light conditions are important drivers of the severity of accidents.