APPLIED DATA SCIENCE - CAPSTONE PROJECT

Accident Severity Prediction

Jayasree Ramakrishnan

October 2020

Contents

1. Introduction

Road accidents are extremely common, and they often lead to loss of property and even life. Hence it is good to have a tool that can alert the drivers to be more careful depending on the weather and road conditions. If the severity is high the driver can decide whether to be extra cautious or delay the trip if possible. This tool can also help the police to enforce more safety protocols.

The goal of this project is to predict road accident severity depending on certain weather and road conditions and time of the day. The data set used for training the model is the one recorded by the Seattle Department of Transportation (SDOT) which includes all types of collisions from 2004 to present. It has around 194673 records with 38 attributes.

2. Data

2.1 Data Description

We will be using the shared data, ie. the collision data recorded by the Seattle Department of Transportation(SDOT) which is avialable at
- https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

Inorder to develop a Accident Severity Predicting Model, we will be considering the following Attributes.

- WEATHER - A description of the weather conditions at the time of the collision.
- ROADCOND - The condition of the road during the collision.
- LIGHTCOND - The light conditions during the collision.

The target is the Severity of collision which is represented by column :

- SEVERITYCODE - A code that corresponds to the severity of the collision

We have two possible outcomes for this in our data set : 1 - Property Damage Only Collision 2 - Injury Collision

2.2 Data Cleaning

Data Cleaning is one of the important steps in pre-processing stage. Eliminating unwanted, duplicate and irrelevant data helps us to have better insight of data and get better results.

Redundant and Useless Data : There are some columns which has data irrelevant to our solution development. We can remove those columns from data frame. Eg. X,Y,OBJECTID,INCKEY,COLDETKEY,REPORTNO etc.

Missing Values : We remove the record which has missing values for our Target variable or our Features. (NAN values, Unknown and Others)

2.3 Data Exploration

In order to have a better understanding of the dataset, before building the model, we first explore some of the main features of the data set.

By plotting a bar chart for the features WEATHER, ROADCOND and LIGHTCOND we understood that they had values which makes hardly any impact during the model development. Those values were deleted from the data set.

1. Partly Cloudy, Severe Crosswind, Blowing Sand/Dirt etc. from Weather Conditions (Figure 1)
2. Standing Water, Oil etc. from Road Conditions (Figure 2)
3. Dark-No Street Lights, Dark-Street Lights Off from Light Conditions (Figure 3)

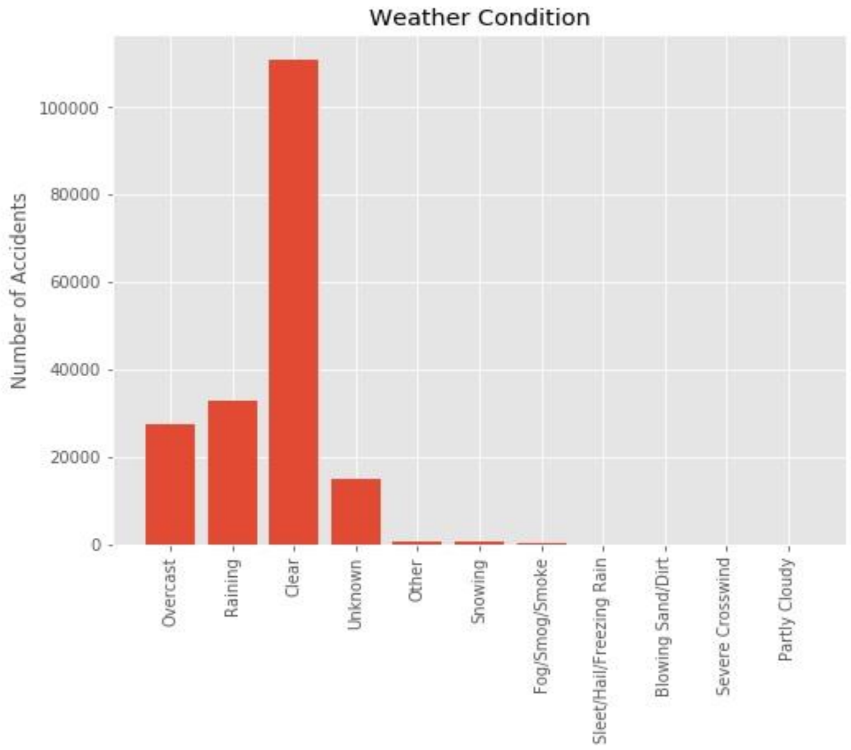

Figure 1

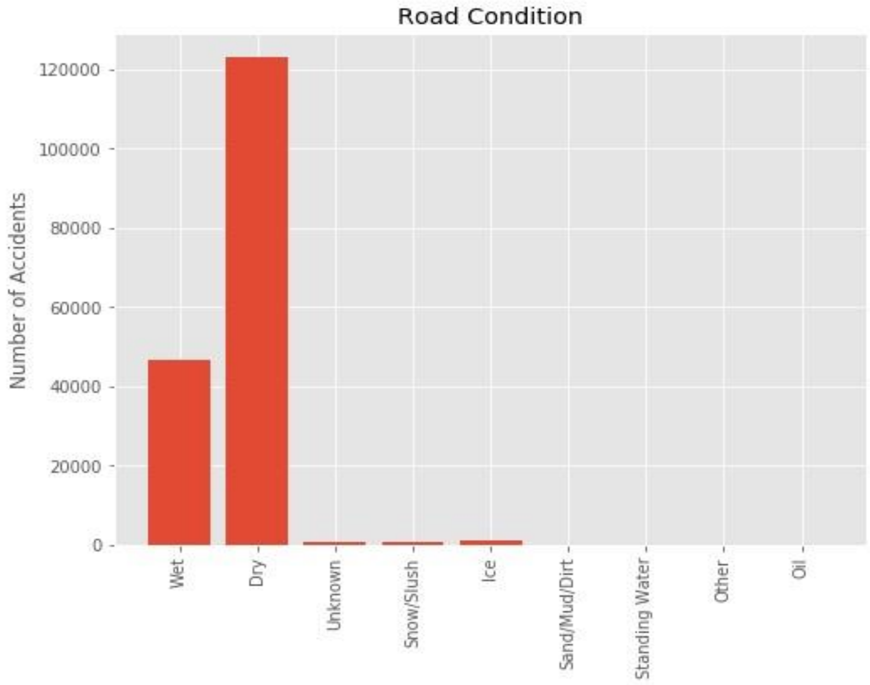

Figure 2

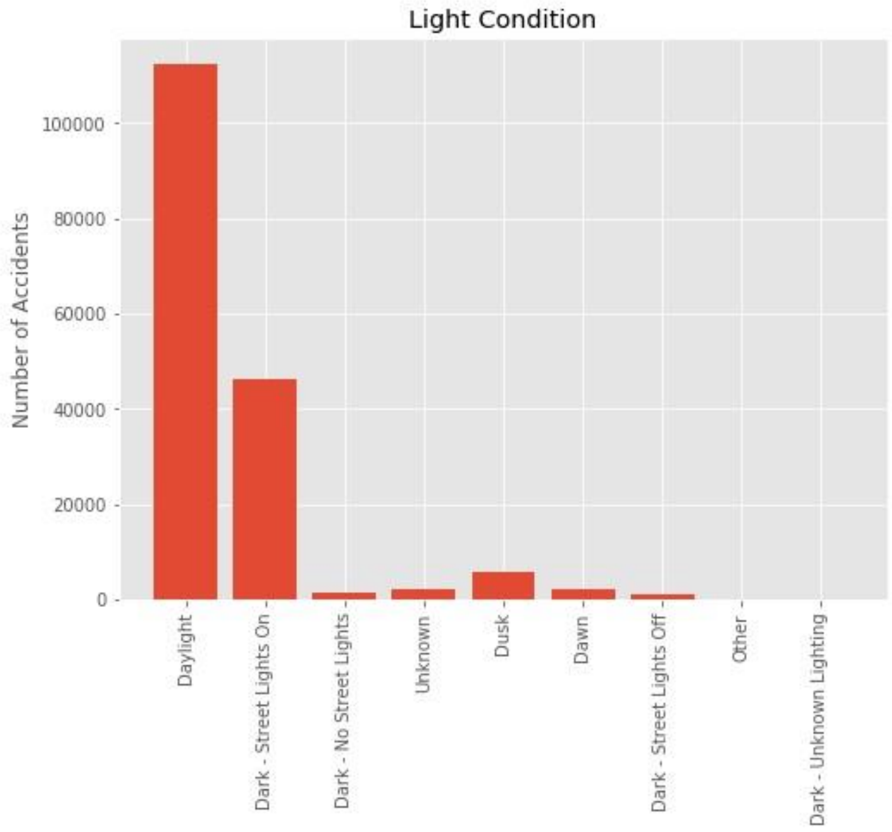

Figure 3

3. Methodology

The aim of this project is to effectively classify the severity of collision at a given time using the weather condition, road and light conditions.

As part of pre-processing of data, the categorical variables(WEATHER,ROADCOND and LIGHTCOND) are converted to binary variables.

The data is also split into training and testing subsets where 80% is used for training and 20% for testing (test_size=0.2).

Since this is a predictive modelling, we have used different Supervised Machine Learning Techniques – both regression and classification models.

1. K-Nearest Neighbour
2. Decision Tree
3. Logistic Regression

4. Evaluation and Results

The models were tested using the test data and were evaluated by calculating their Accuracy, Jaccard Similarity Score and F1 Score. We also found the best K that can give us the best K-Nearest Neighbour Model, which is 8.

The following table gives the evaluation results.

Model	Accuracy	Jaccard Score	F1 Score	Log Loss
KNN	0.67	0.67	0.54	
Decision Tree	0.67	0.67	0.54	
Logistic Regression	0.67	0.67	0.54	0.65

5. Discussion

The above results show that almost all the three models perform the same way. But while execution K Nearest Neighbour Model took a lot of time. So the recommendation would be to implement either Decision Tree or Logistic Regression.

6. Conclusion

The purpose of this study is to develop a model which can predict the severity of road accident which can then alert the drivers so that they can be cautious while driving during certain conditions. The findings can also help departments which are involved in road accidents such as Emergency Services and Police Department who can then plan ahead and implement extra safety protocols to prevent future accidents.