**Capstone Project**

**Introduction**

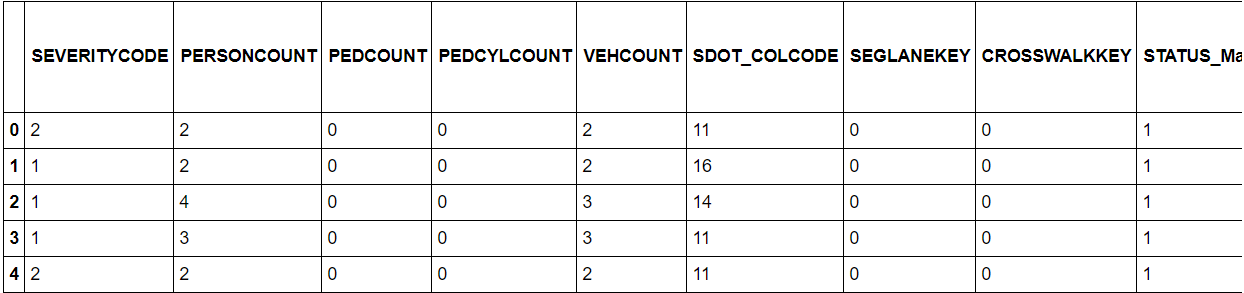
Say you are driving to another city for work or to visit some friends. It is rainy and windy, and on the way, you come across a terrible traffic jam on the other side of the highway. Long lines of cars barely moving. As you keep driving, police car start appearing from afar shutting down the highway. Oh, it is an accident and there's a helicopter transporting the ones involved in the crash to the nearest hospital. They must be in critical condition for all of this to be happening. Now, wouldn't it be great if there is something in place that could warn you, given the weather and the road conditions about the possibility of you getting into a car accident and how severe it would be, so that you would drive more carefully or even change your travel if you are able to.

People who like travelling by car, bus drivers, truckers and even ordinary persons who hate being stuck in traffic jams will be happy to know about possibility of getting into a car accident. It can save a lot of time and you will always be happy about your journey.

**Data**

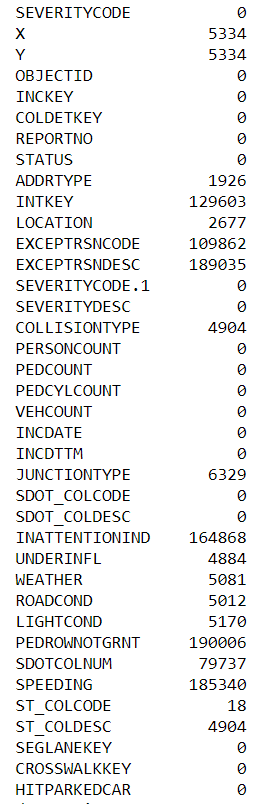
I will use data about collisions. All collisions provided by SPD and recorded by Traffic Records. This includes all types of collisions. Collisions will display at the intersection or mid-block of a segment. Firstly, I am going to preprocess the data, dealing with all the missing data. Secondly, I will change type of columns to numeric. After that I will decide which features to take by correlation. Then I am going to scale the data with StandardScale and use SVM as machine learning algorithm to predict severity of collision.

Example of final dataset:



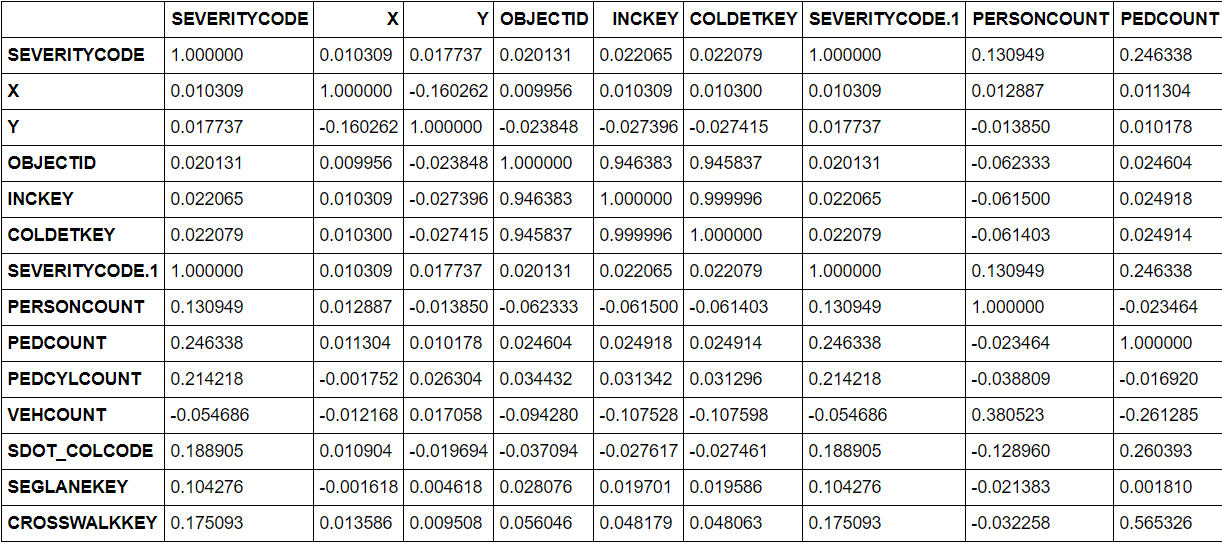
**Methodology**

To begin with, I decided to look on data types of columns. After that I checked how many null values does each column have. That’s what I get:



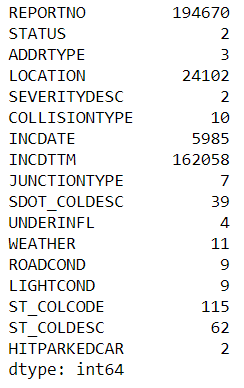
We can see that there are some columns which have really a lot of missing values. So I decided to drop such columns which have more than 70000 missing values.

After that I checked correlation between columns:



As we can see columns X, Y, OBJECTID, INCKEY, COLDETKEY has a values close to zero. It means that they are not so much correlated with column SEVERITYCODE. So I decided to drop it.

But what about columns which type is object? Let’s deal with them. I decided to look on number of unique values in each column to understand can we get dummies from it or no.



Then I exchange columns that has less or equal to 11 unique values (based on weather value) with it’s dummies. What about others? I dropped REPORTNO column because it is simply id of each row and it is useless in our research. Other columns with a large number of unique values I decided to drop because they are useless for many reasons. After that I scaled my dataset and split it into train and test sets.

**Results**

It is time to use machine learning algorithms. I tried to use 3 different algorithms: SVM, Decision trees, Logistic Regression. All of them give amazing results. To evaluate them I used F1 score and Jaccard similarity score. SVM give result of 0.9995 accuracy. But it is not computationally efficient of big datasets, so it took long time to run. But Decision trees and Logistic Regression gave results of 1.0 accuracy. It means that we can purely predict severity of collisions.

**Discussion**

During my research I dropped some columns that have nearly 5000 missing values. You can exchange missing values with mean value for numeric type or for value which has the most frequency. After this research I understand influence of data science in real life. This results can help a lot of people to save their time in journeys. Moreover, you will never occur in situation which spoil the impression of your journey because of traffic jams for example. Recently I have to drive 500 km to my siblings and during that I lost 2 hours in traffic jams. It is really awful to be in such situation. With such results you can avoid it.

**Conclusion**

In this study, I analyzed the relationship between collision’s severity and it’s features. I identified number of pedestrians involved in collision, type, weather, road and light condition among the most important features that affect a severity of collision. I built classification models to predict category of severity. These models can be very useful for people who like journeys and don’t want to lose their time in traffic jams.