Predicting the severity of a car accident

1.**Introduction**:

1.1.**Background**:

Car accidents are quite possible on the road.The consequences of road accidents might be severe like injuries for involved persons. It also causes terrible traffic jams which delays others' routines.The roads are shut down in case of severe crashes. It is important to predict the road accidents to avoid such situations.

1.2.**Problem**:

Given the road and weather conditions, we intend to warn the chances of car accidents and how severe it could be, so that the driver can drive carefully. It will also help the driver to change the travel plans if able to. Based on the car accident severity data provided, I intend to identify the chances of getting into a road accident.

1.3.**Interest**:

Obviously a car driver wants to avoid the road accidents, which has severe impacts like injuries or issues in his/her insurance. Insurance companies might also be interested in predicting the chances to work on their reserves for a specified period or for any other reasons.

2. Data acquisition and Cleaning

2.1. Data Sources:

The capstone assignment has a sample dataset provided which was used for further studies. This dataset has the details of car accident history of San Francisco.

2.2. Data Cleaning:

The dataset almost had 0.2M records. I decided to consider only the records that have reported at least 5 accidents at the same location, which reduced the size of my data to 0.15M.

The dataset had some redundant columns which I decided to remove.

Kept columns	Dropped columns	Reason for dropping
INCDTTM	INCDATE	Incident date is available as part of Incident time column
EXCEPTRSNCODE	EXCEPTRSNDESC	This feature is intended to show whether the record has enough information to explain the accident, the value -NEI says it does not have enough information. Though these NEI records

		state that these do not have enough information in one or the other fields, some of the records still have the information like how the road condition/weather was, at the time of accident. So I decided to keep these records too, planning to replace the insufficient information with mean or frequency.
SEVERITYCODE	SEVERITYCODE.1	These fields show whether the accident is injury collision (value:2) or property damage only collision (value: 1) and have the same values in both columns. So the redundant data can be removed.

There were 19 features which had missing values. I decided to replace these missing values by the mean or frequency.

2.3 Feature Selection:

I would also need to check whether the features like Weather, Location, AddressType, Junction Type, Road condition,Light Condition and Speeding have any correlation with the target severity level to identify the severity of the car accident.
----

3. Methodology:
3.1 Exploratory Data Analysis
3.1.1 Calculation of target variable
3.1.2 Relationship between variable and its target
3.2.Machine Learning
    3.2.1 Regression models
    3.2.1 Applying standard algorithms and their problems
    3.2.2 Solution to the problems
    3.2.3 Performances of different models
3.2.2 Classification models

4.Results
5.Discussion
6. Conclusions
