CAR ACCIDENT SEVERITY

1. Introduction

Traffic accidents are severe concern for most of the countries.There are many factors leading to accidents. It depends on weather, road condition, vehicle condition, driver condition ,light condition and many other factors.This project mainly aims to predict how severity of accidents can be reduced based on a some factors.
The target audience of this project is car drivers, police, government authorities.The result of this project help to advice the audience about the possibility of getting into a car accident and how severe it would be, based on the different conditions such as weather, light, road.It would help them to drive carefully.

2. Dataset

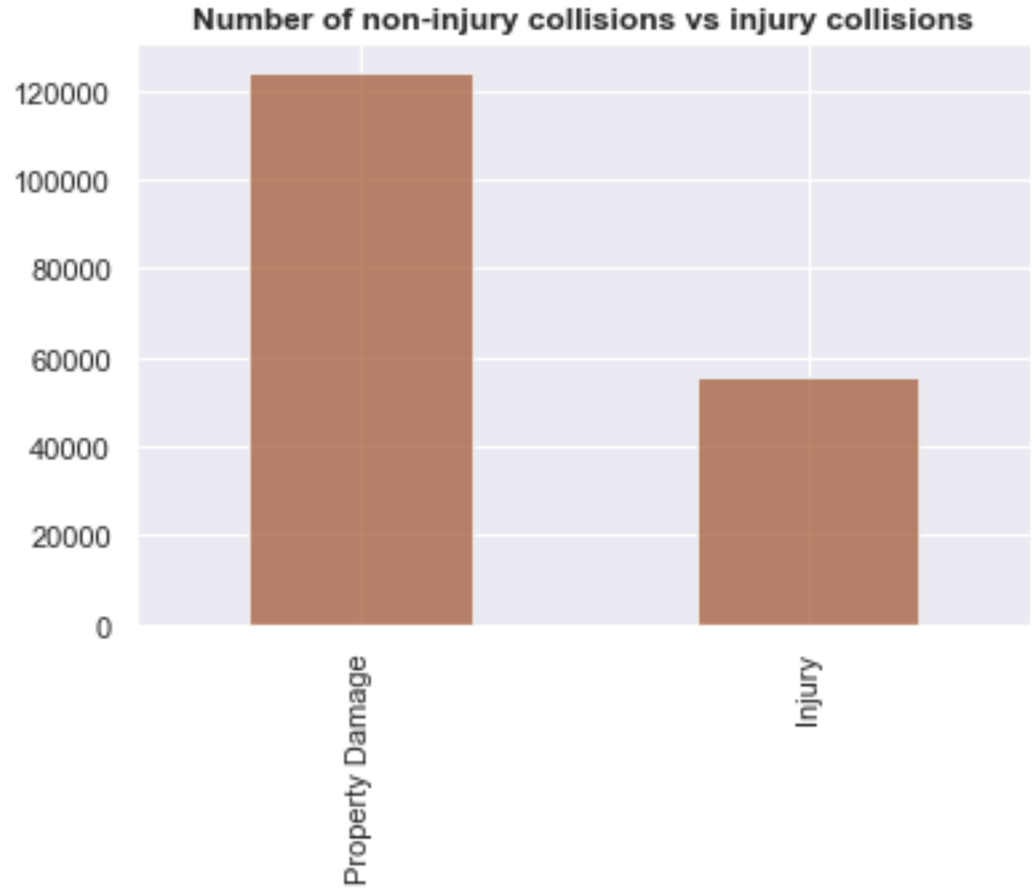
The data was collected by the Seattle police department.it contain details of car accidents which have taken place within the city of Seattle.The time period of the data is from 2004 to present.The data consists of 37 attributes or variables and it includes information such as severity, location, collision type, Weather conditions, road condition and light conditions.

2.1 Data Cleaning

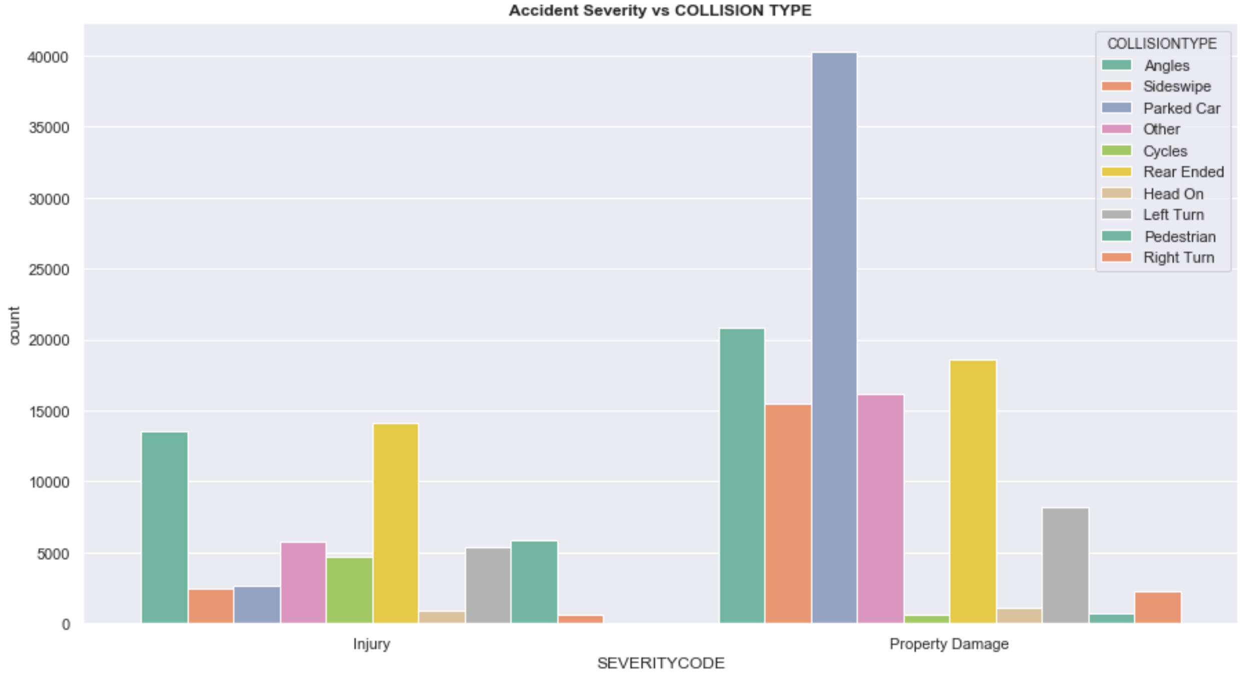
Firstly, I have checked the data types of all columns.The dataset had a lot of empty columns. After reviewing all 37 attribute columns in the datasets, I have to drop the non-relevant columns. Y was given value of 1 whereas N and nan value was given 0 for the attributes like speeding ,under the influence. Also, most of the features are of type object, when they should be numerical type. All the categorical variables are converted to numerical data using one-hot encoding. The time format is changed to default standard.

3. Methodology
3.1 Exploratory Analysis

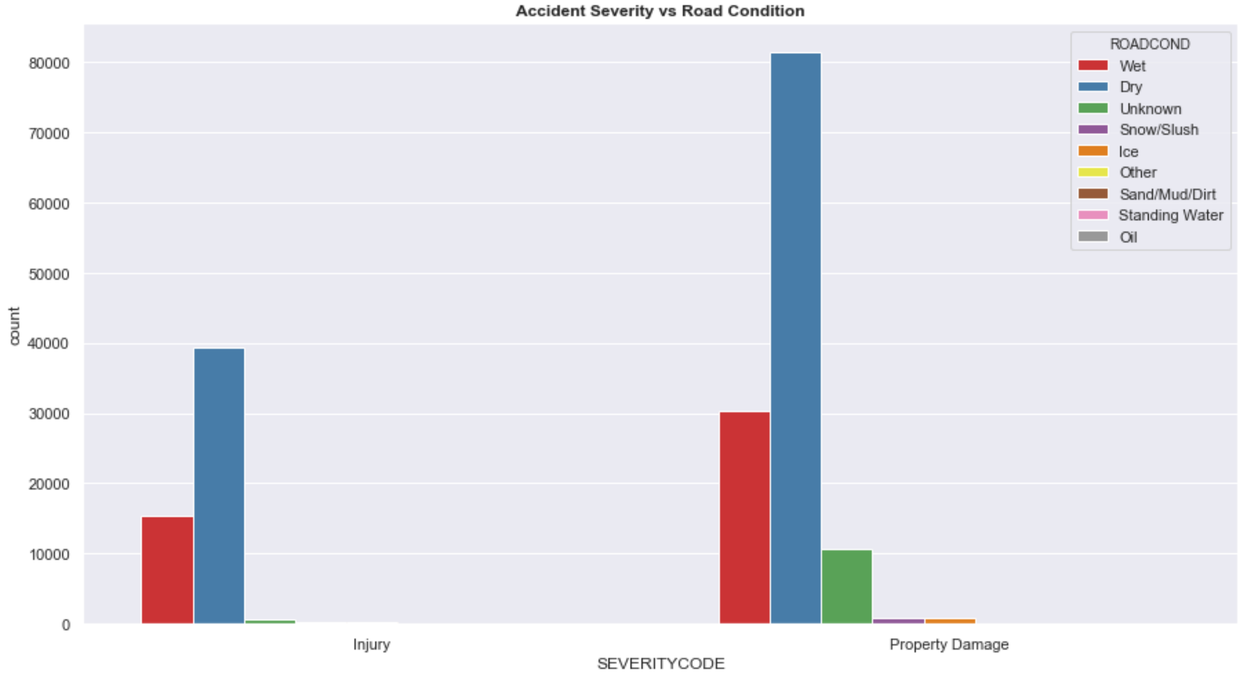
In total number of accidents in Seattle most of them are property damaged only collision are more than injured.



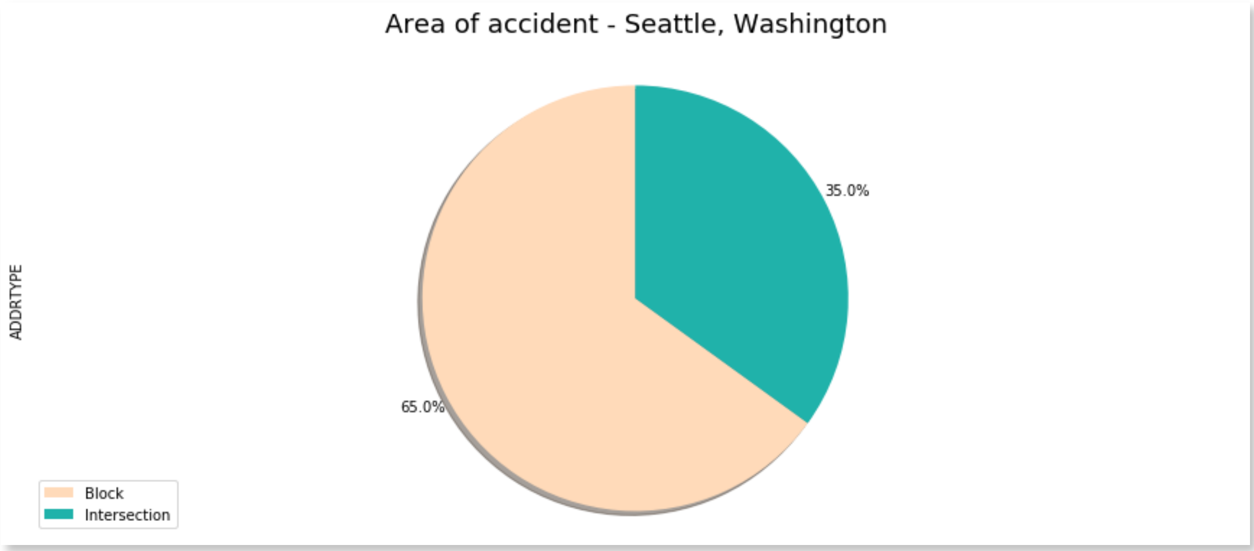
Relationship between Collision type and severity of accident:



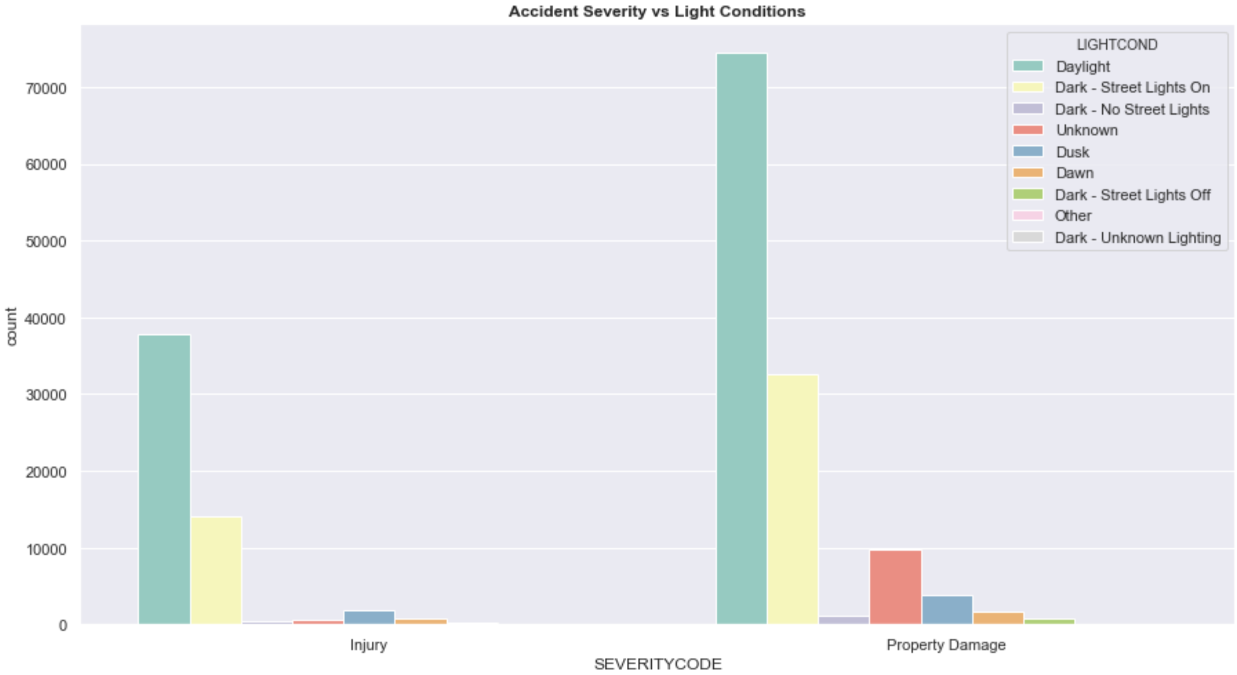
Relationship between Road conditions and severity of accident:



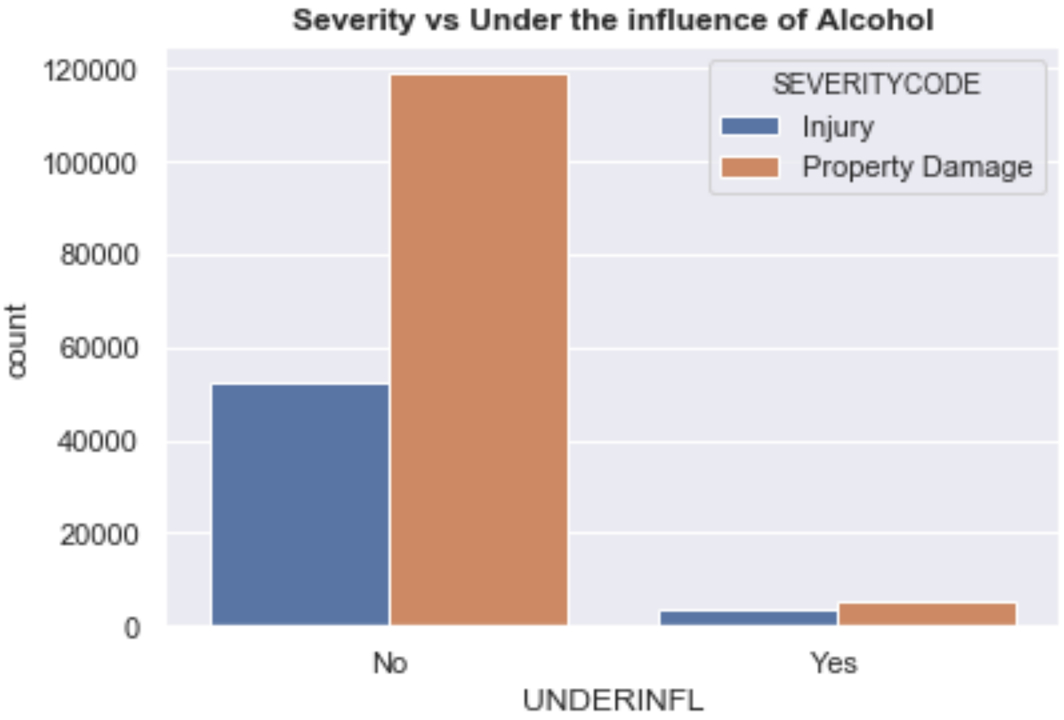
Relationship between address type and severity of accident:



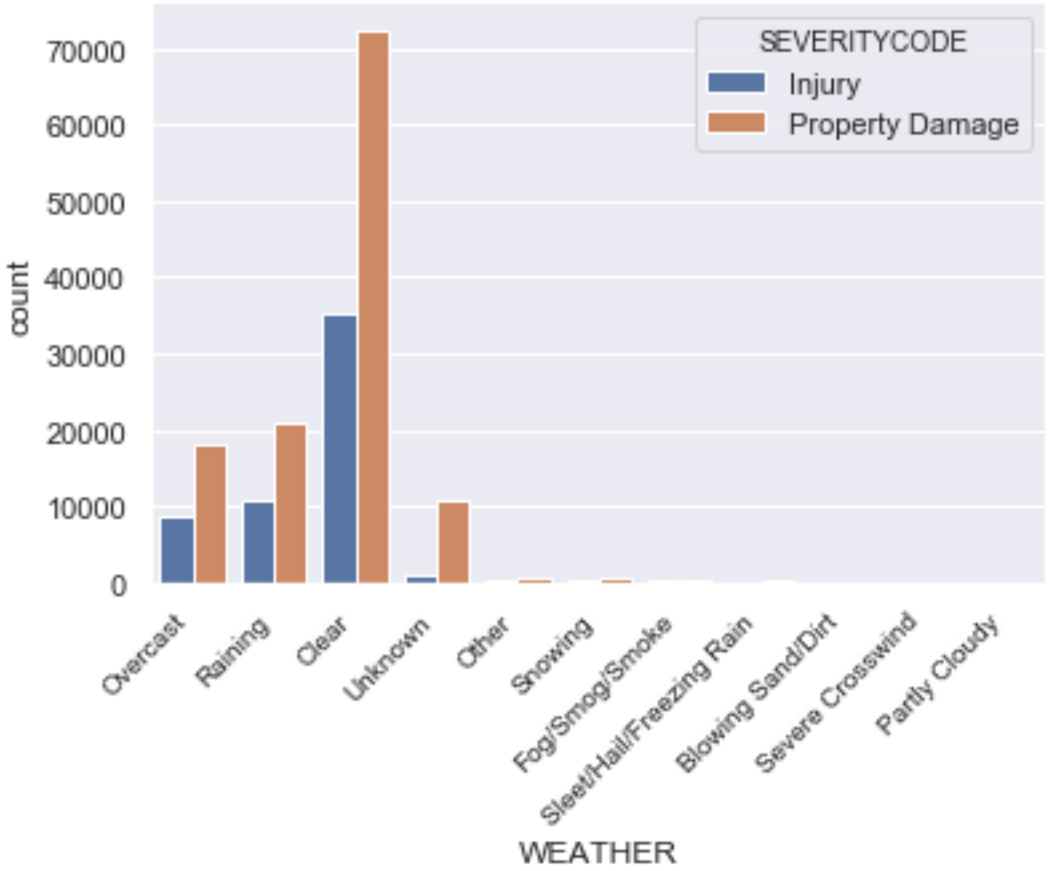
Relationship between Light conditions and severity of accident:



Relationship between Under the influence and severity of accident:



Relationship between Weather conditions and severity of accident:



3.2 Feature Engineering

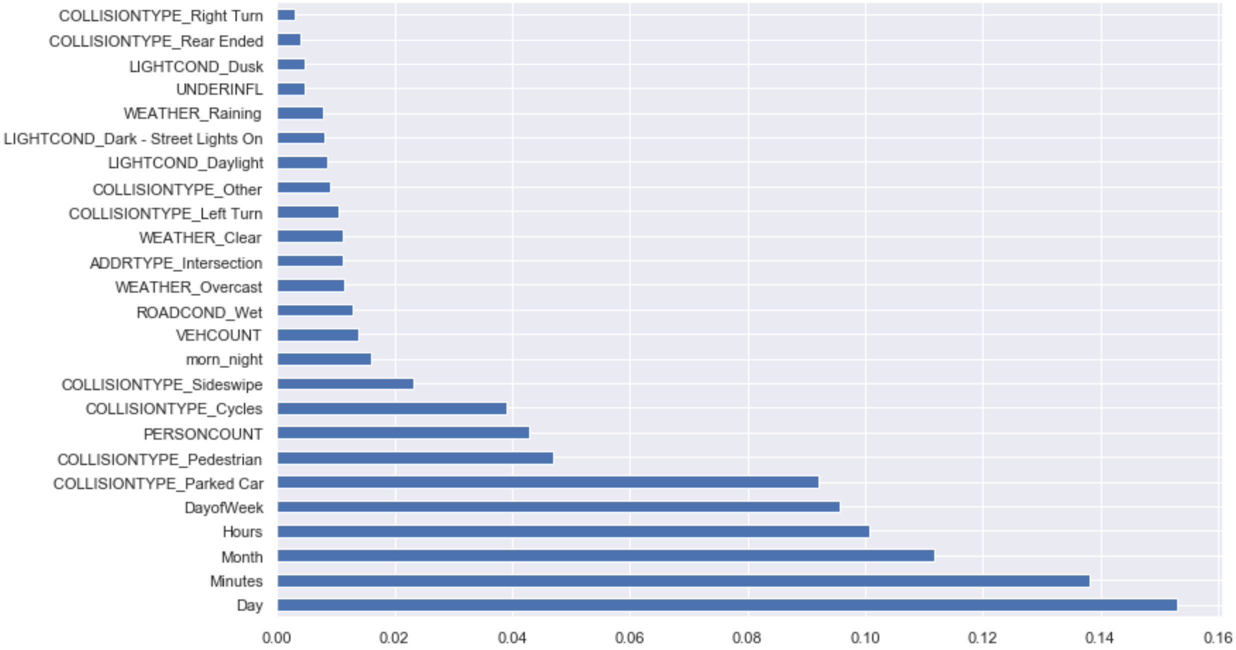
I used One Hot encoding for the Categorical variables, like ADDRTYPE, COLLISIONTYPE, WEATHER, ROADCOND, LIGHTCOND. Out of 5 categorial variables I have created 36 new variables, using one hot encoding technique.
All the Integer variables are kept same, they are SEVERITYCODE, PERSONCOUNT, VEHCOUNT, UNDERINFL.
There was only 1 date variable in the dataset which is INCDTTM(incident date and time). From this variable we created Year, Month, Day, Hours, Minutes, Day of week, Morning or night. Engineered 7 variables from date variable.
The I concat all the categorical, integer and date variables into one dataset for machine learning. So total we have 46 variables and 180067 rows.

3.3 Feature Selection

A total of 9 features were selected for this project along with the target variable being severity code.

Feature Variables	Description
ADDRTYPE	Collision address type
COLLISIONTYPE	Collision type
WEATHER	A description of the weather conditions during the time of the collision.
ROADCOND	The condition of the road during the collision.
LIGHTCOND	The light conditions during the collision.
PERSONCOUNT	The total number of people involved in the collision
VEHCOUNT	The number of vehicles involved in the collision.
UNDERINFL	Whether or not a driver involved was under the influence of drugs or alcohol.
INCDTTM	Date and time of the accident

For feature selection I used Extra treesRegressor from sklearn, also plotted the top 25 features. Extremely Randomized Trees Classifier(Extra Trees Classifier) is a type of ensemble learning technique which aggregates the results of multiple de-correlated decision trees collected in a “forest” to output it’s classification result. Out of that all the features top 10 features that has most predictive power is day, minutes, month, hours, days week (all these features engineered from the date variable ‘INCDTTM’) also collisiontype_parkedcar, collisiontype_pedestrion, personcount, collisiontype_cycles, collisiontype_sideswipe.



3.4 Machine Learning and Model Selection

The model is trained on the feature attributes of dataset to predict the severity of the accident. Dataset must be first divided into training set (80%) to train the models and test set (20%) to test the models. The machine learning algorithms used are Logistic Regression, k-Nearest Neighbor, SVM and Random forest. All the all algorithms used are classification, because the problem we are trying to solve is a classification problem.

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable. K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (based on distance). Random forest classifier creates a set of decision trees from randomly selected subset of training set. Model’s accuracy was evaluated, and classification report was generated including precision, recall and f1 scores. Confusion Matrix is a performance measurement for machine learning classification. It is a performance measurement for machine learning classification problem where output can be two or more classes. It is a table with 4 different combinations of predicted and actual values.

4. Results and Evaluations

The final results of the model evaluations are summarized in the following table

	Logistic	SVM	Random Forest	k-Nearest Neighbors
Accuracy	75	74	73	69
No. of True Positives	23764	24711	22139	23670
No. of False Positives	1256	246	2881	1350
No. of False Negatives	7736	8968	6656	9674
No. of True Negatives	3258	2089	4338	1320

Based on the above table, Logistic Regression is the best model to predict car accident severity.

5. Discussion

The dataset is analyzed with the help of visualization and modeling of data.The relationships between the severity code and feature attributes are analyzed. As we can observe the most of the accidents occurred during daytime, on a dry road condition in a clear weather and in block.The collisions which occurred under influence and over speeding are also very less.

6. Conclusion

In this study, I analyzed and visualized the relationship between accident severity (property damage/injury) and under the influence of alcohol, light condition, collision type, road condition, weather condition. I build classification models to predict whether there is change of having property damage or injury during an accident. Out of 4 algorithms I used logistic regression gave me 75% accuracy in predicting type of accident sever, based on light condition, collision type, road condition, weather condition. In future I will use hyper-parameter techniques to increase the accuracy of my model and also create an web app to deploy my model and others to use it for predicting the chance of having accident and how severe it will be.