Predicting Accident Severity for Seattle City

Shilpa Banerjee

Predicting Accident Severity is valuable for Seattle city

Introduction:

- The Seattle community has expressed concern over the serious road accidents taking place in certain road junctions
- The increasing number of serious accidents will eventually bring down the road safety score of this suburb there being leading to unpleasant consequences.
- It would be great if there is something in place that could warn people, given the weather and the road conditions about the possibility of you getting into a car accident and how severe it would be.

Data acquisition and cleaning

Data Sources:

- This is the Seattle accidents database which discusses the fatality of accidents based on a set of conditions. The fatality is measured using attribute SEVERITYCODE has been given values 1 or 2.

Data Cleaning:

- Out of total 37 columns, columns with redundant values, mostly or completely null values and ID columns were removed.

Feature Selection:

- Original dataset comprised of 37 columns and 194673 records. After cleaning up we have 13 columns and 184146 records remaining.

Exploratory Analysis

Target variable:

- The severity of the accident is the target variable that needs to be predicted. It has values 1 or 2 explained below:

SEVERITYCODE	Description
1	Property damage
2	Injury

Exploratory Analysis

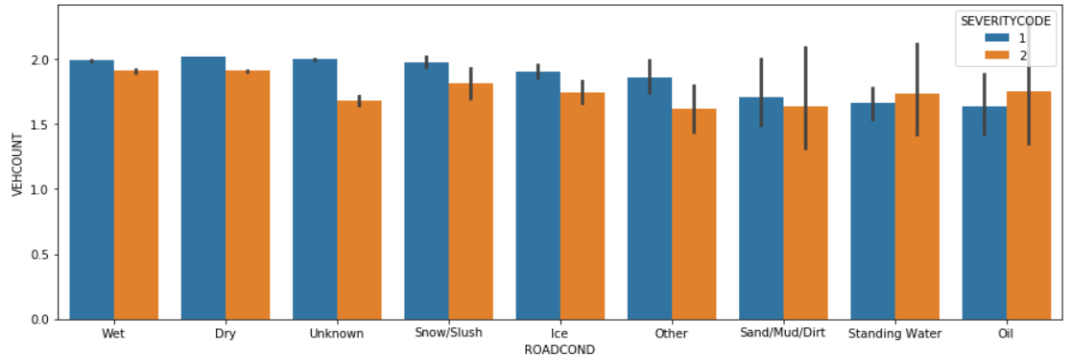
I have analysed the severity of the accident under the following two groups of relationships:

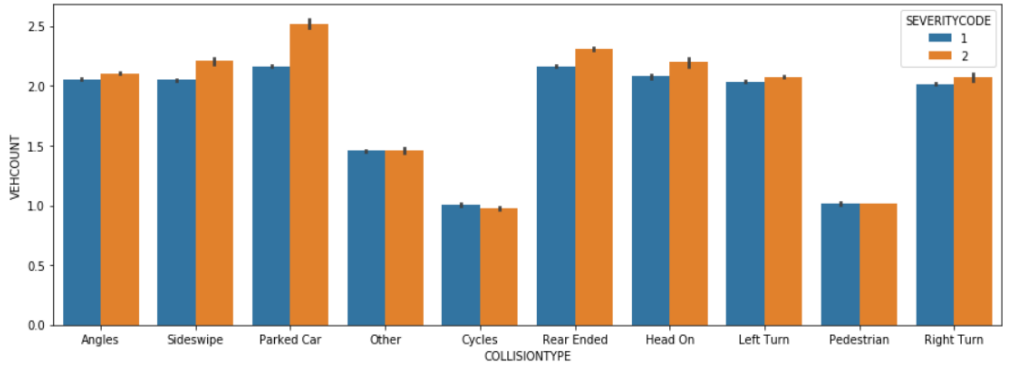
- Total number of vehicles involved in collision(VEHCOUNT) with road condition (ROADCOND) and type of collision(COLLISIONTYPE):
  - VEHCOUNT ----- COLLISIONTYPE ----- SEVERITYCODE
  - VEHCOUNT ----- ADDRTYPE ----- SEVERITYCODE
  - VEHCOUNT ----- ROADCOND ----- SEVERITYCODE

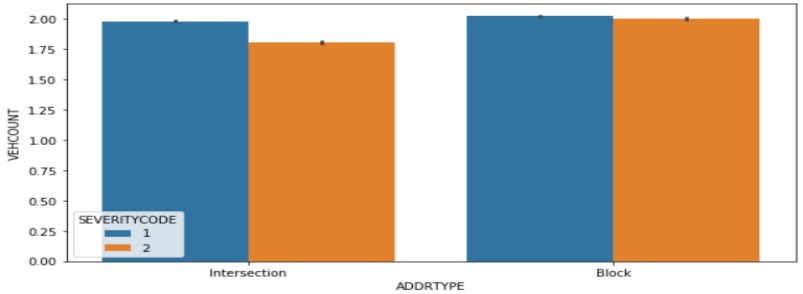
- Total number of people involved in collision(PERSONCOUNT) with road condition (ROADCOND) and type of collision(COLLISIONTYPE):
  - PERSONCOUNT ----- COLLISIONTYPE ----- SEVERITYCODE
  - PERSONCOUNT ----- ADDRTYPE ----- SEVERITYCODE
  - PERSONCOUNT ----- ROADCOND ----- SEVERITYCODE

Exploratory Analysis

Total number of vehicles involved in collision(VEHCOUNT) with road condition (ROADCOND) and type of collision(COLLISIONTYPE). An even spread of values shows that these features are suitable to be used in building the model

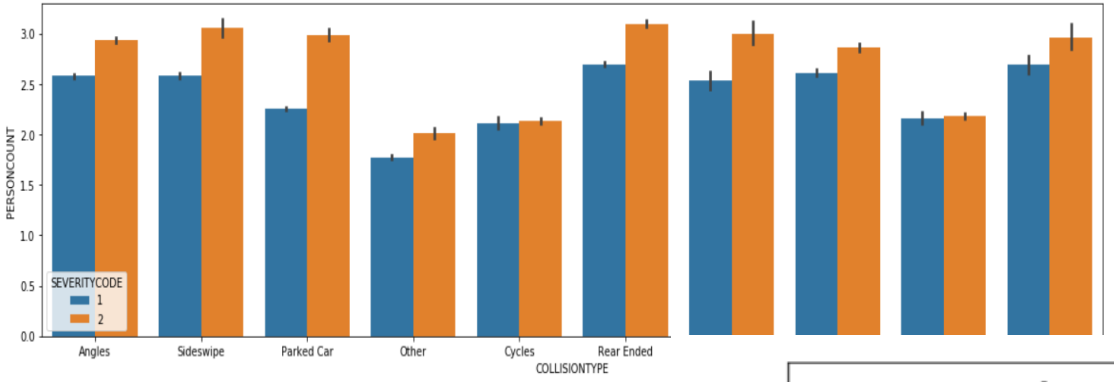


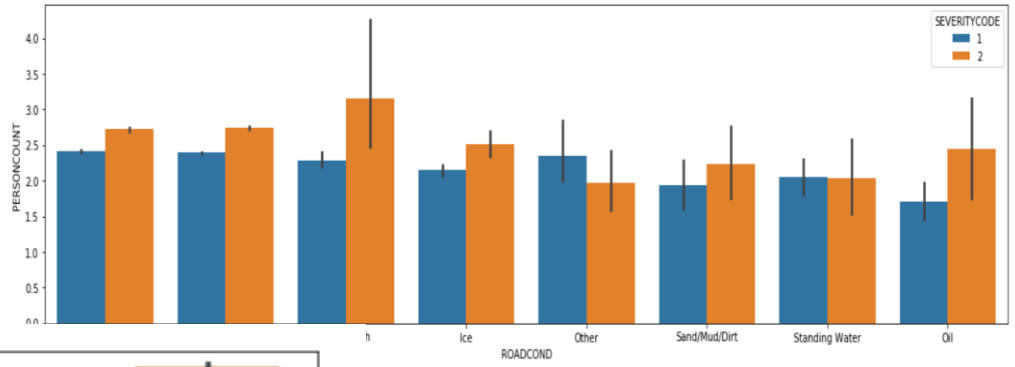


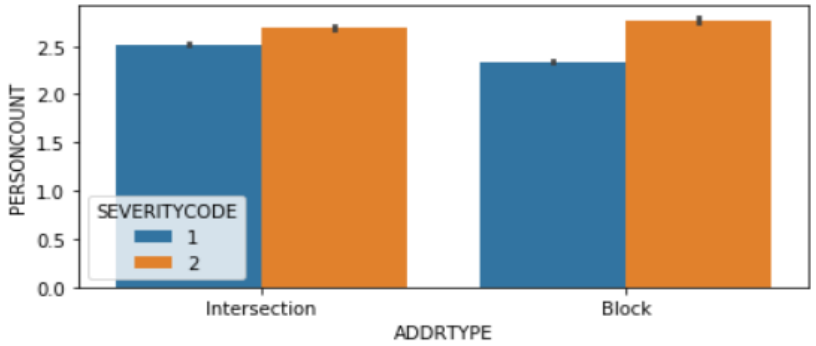


Exploratory Analysis

Total number of people involved in collision(PERSONCOUNT) with road condition (ROADCOND) and type of collision(COLLISIONTYPE). An even spread of values shows that these features are suitable to be used in building the model







Predictive Modelling

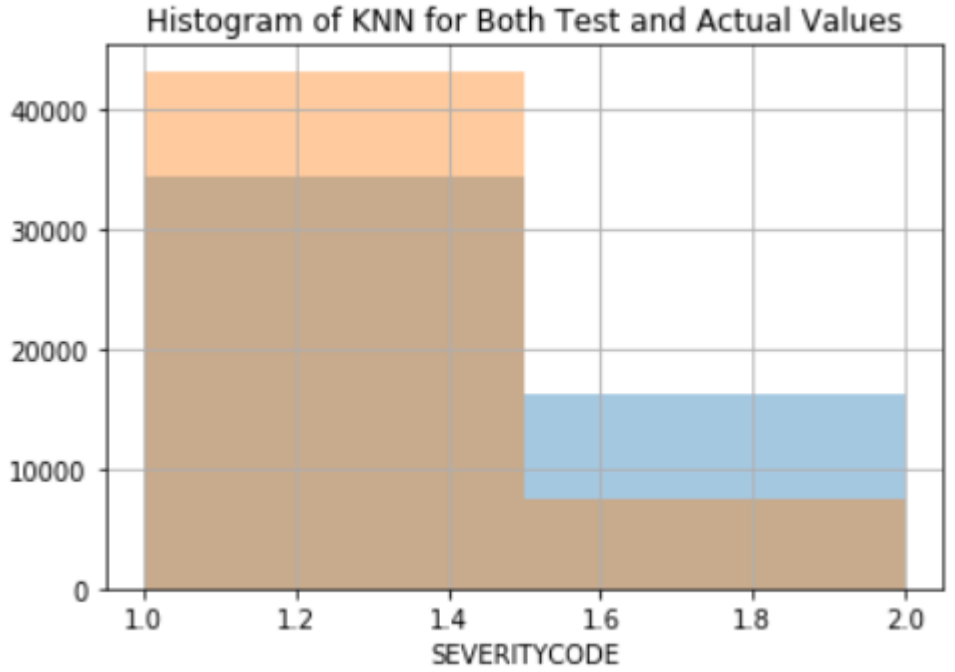
As we have to predict categorical variable, I am using classification algorithms :

K- Nearest Neighbours (KNN)

Using KNN with k = 6 we get best accuracy of 73.41%

Jaccard Similarity Score = 73.28%

F1 Score = 69.61%



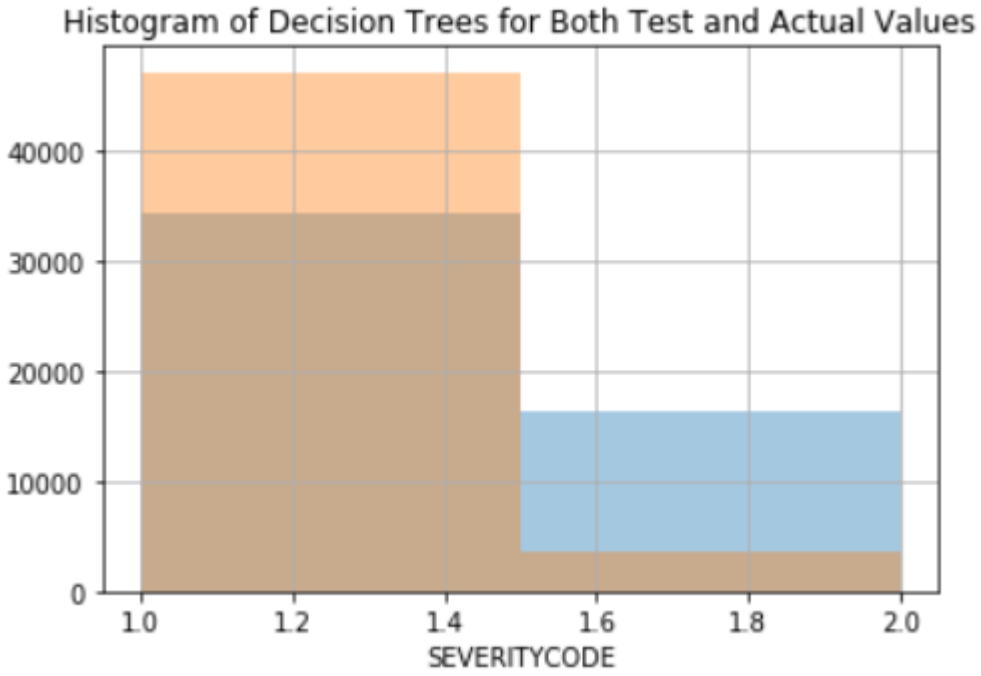
Predictive Modelling

Decision Trees

With entropy = 4 we get best accuracy of 73.08%

Jaccard Similarity Score = 73.08%

F1 Score = 66.42%



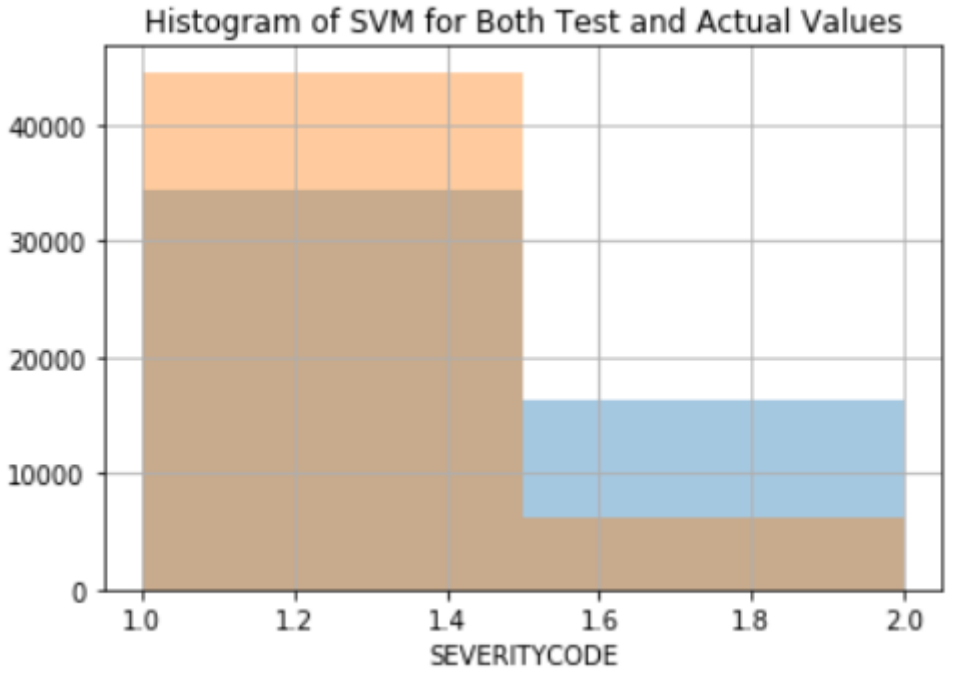
Predictive Modelling

Support Vector Machines (SVM)

SVM has following accuracy :

Jaccard Similarity Score = 74.01%

F1 Score = 69.81%



Predictive Modelling

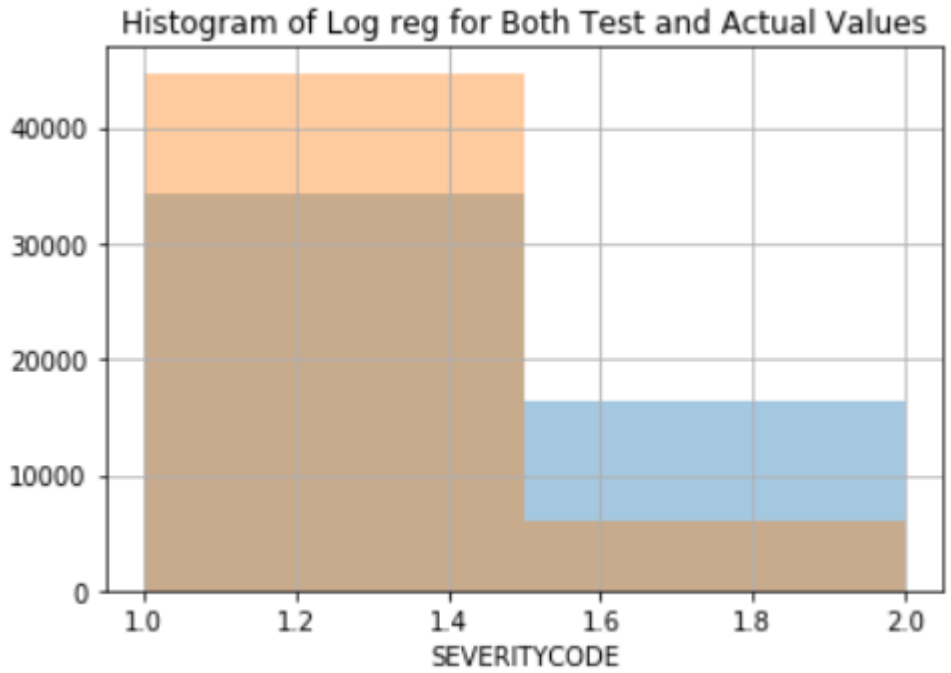
Logistic Regression

We get the following accuracy with Logistic Regression

Jaccard Similarity Score = 73.81%

F1 Score = 69.43%

Log Loss = 51.59%



Conclusion

- The **SVM algorithm** is the most suitable for predicting the categorical target variable.
- It has the best accuracy amongst the three methods as shown by Jaccard score of 74%
- Other algorithms are not too far behind with Jaccard scores of around 73 %
- Thus, using SVM we can make predictions on accident severity based on a set of factors.