Car Accident Severity Prediction

Gaurav Amrutkar
Sep 20, 2020

1. Introduction

The objective is to reduce the frequency of car collisions in a community, a model must be developed to predict the severity of an accident given the current weather, road and visibility conditions. When conditions are bad, this model will alert drivers to remind them to be more careful.

Based on the available historical data of the accidents including whether condition and severity of accident occurred we have developed a model, which based on the current conditions will make the people aware to be careful during certain instances which will ultimately reduce the number and severity of accidents.

2. Data acquisition and cleaning

The data is open source in which the target variable is the severity of accident. The target variable has two values which indicates the high and low severity of accidents. There are around 37independant variables in the data which can be used for the prediction of severity.

Out of the 37 independent variables around eight variables are the key or report variables which are of no use for the prediction. From the remaining 30 variables are the combination of the numeric and categorical variables with two datetime variables.

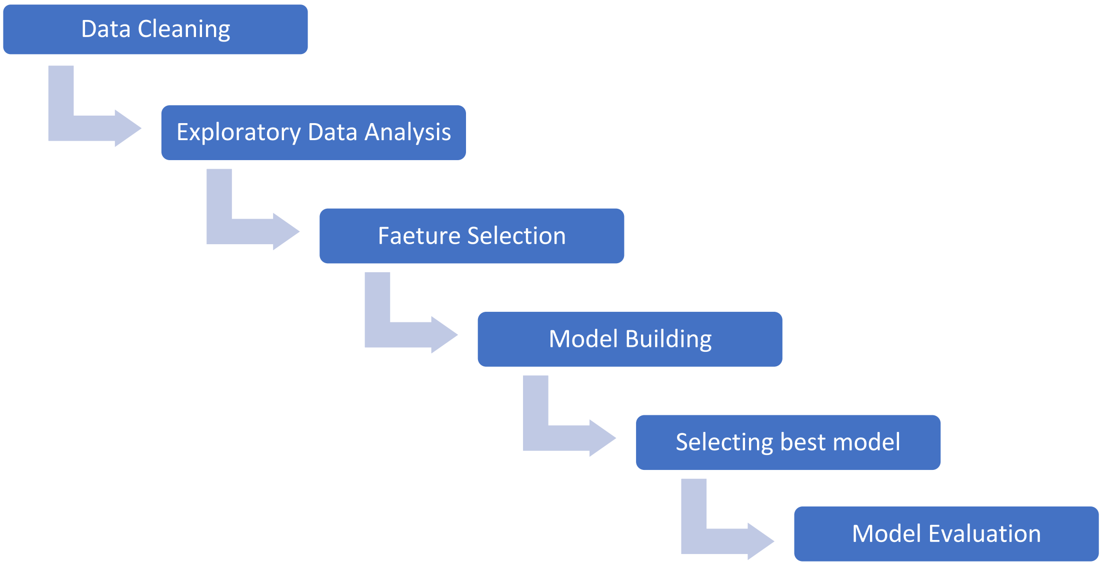
The variable 'SEVERITYDESC' is 100% correlated with the target variable severity of accident and from its name it means that it is the result of severity of accident hence it cannot be the cause rather effect. Hence it was not considered in the model.

For the categorical independent variables, we have seen its distribution with the severity of the accident and grouped the values of high cardinality categorical variables. The details of this distribution are mentioned in the next section of Exploratory data analysis.

The box plot of the numeric variables was created and the values of this variables were restricted to 1% to 99% of the total distribution in order to deal with the outliers.

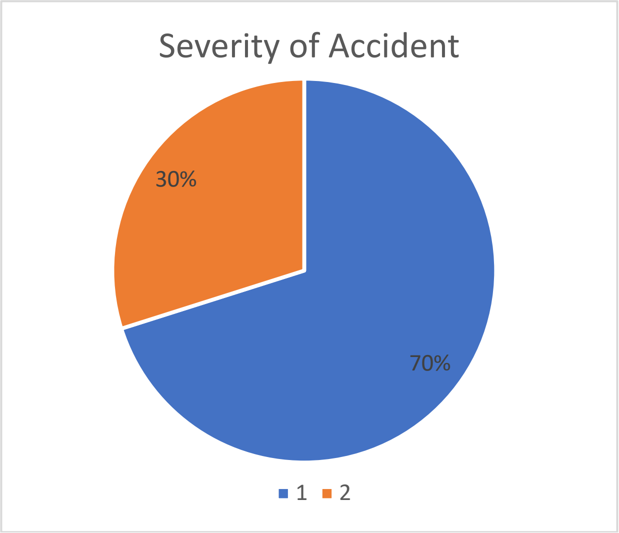
3. Methodology

The detailed step-by-step methodology followed for the model building is explained in the below figure.



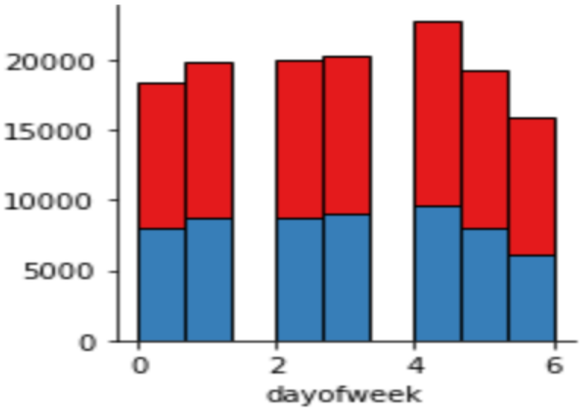
4. Exploratory Data Analysis

The distribution of the target variable is as follows:

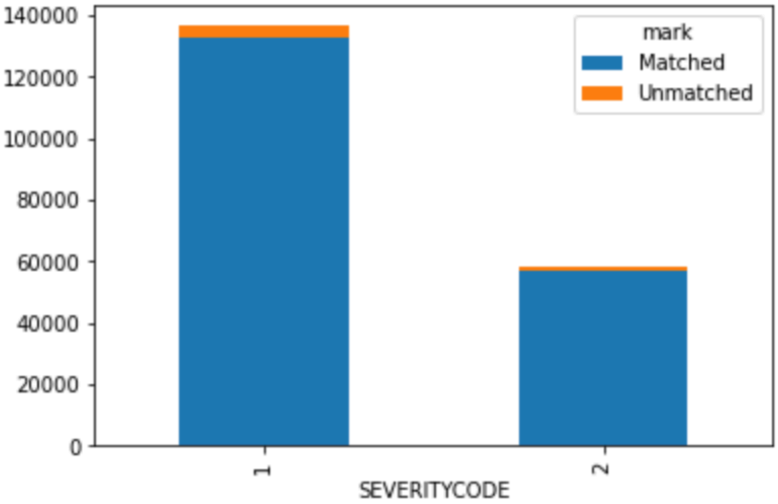


The distribution of other independent categorical variables with respect to the target variables is as follows-

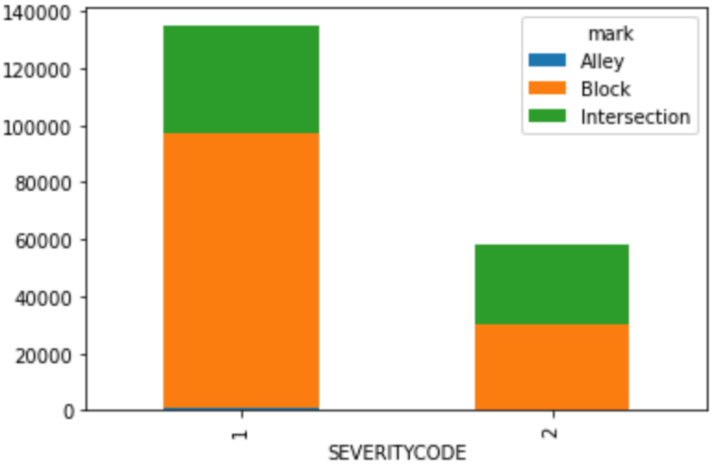
Day of the week



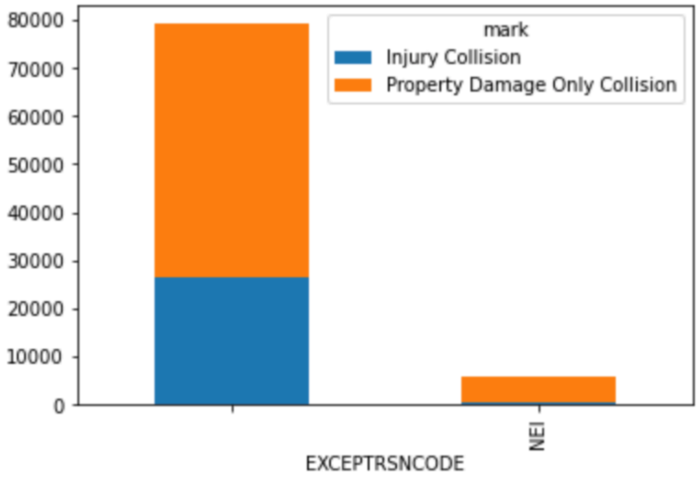
STATUS



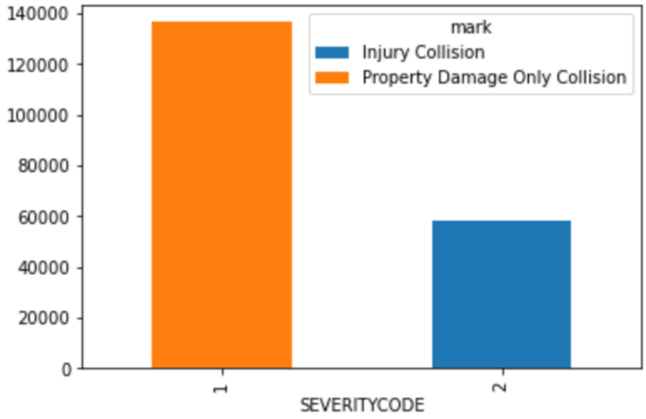
ADDRTYPE



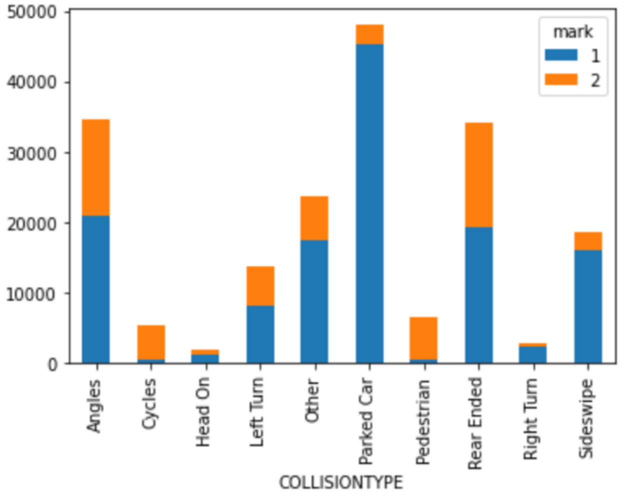
EXCEPTRSNCODE



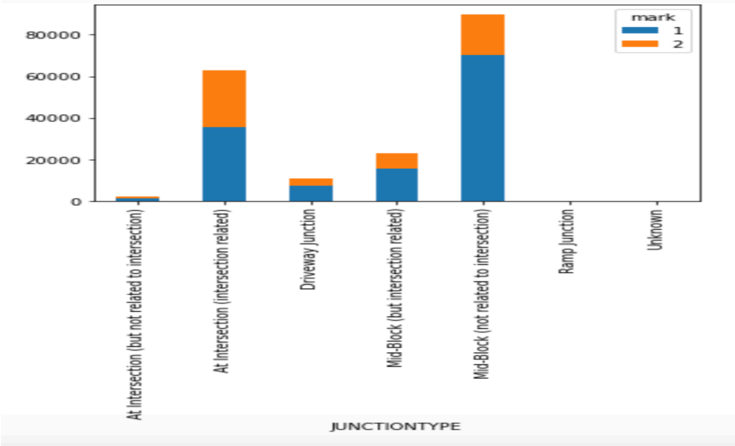
SEVERITYDESC



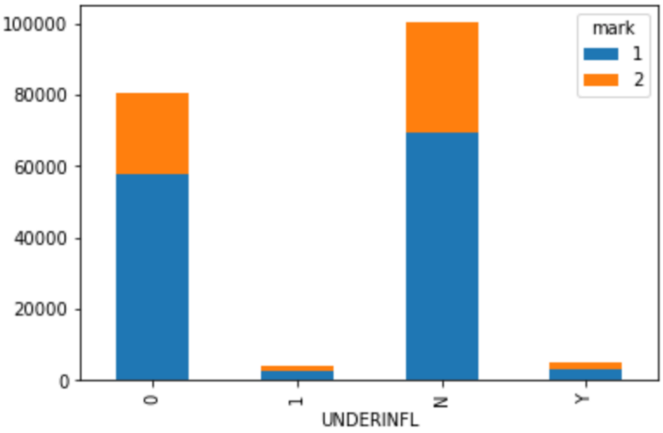
COLLISIONTYPE



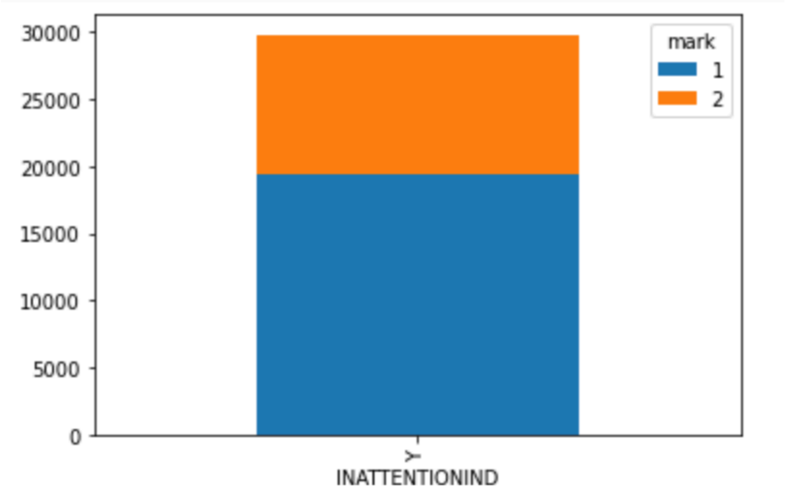
JUNCTIONTYPE



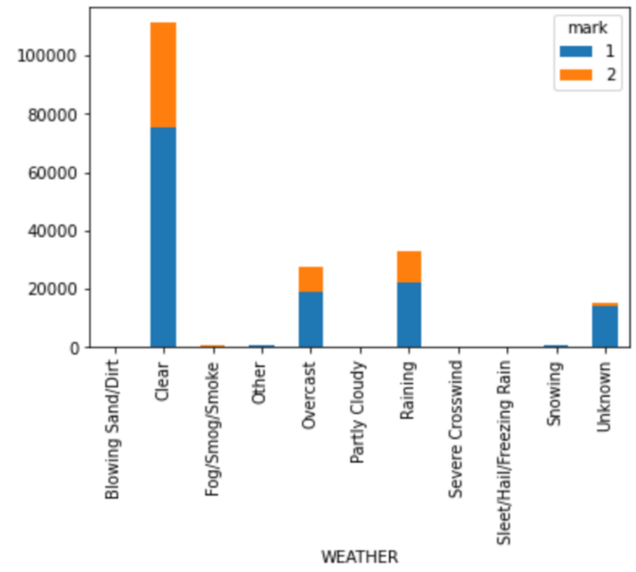
UNDERINFL



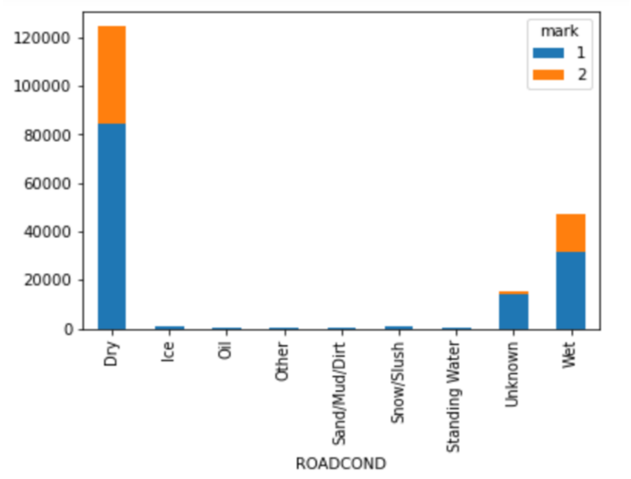
INATTENTIONIND



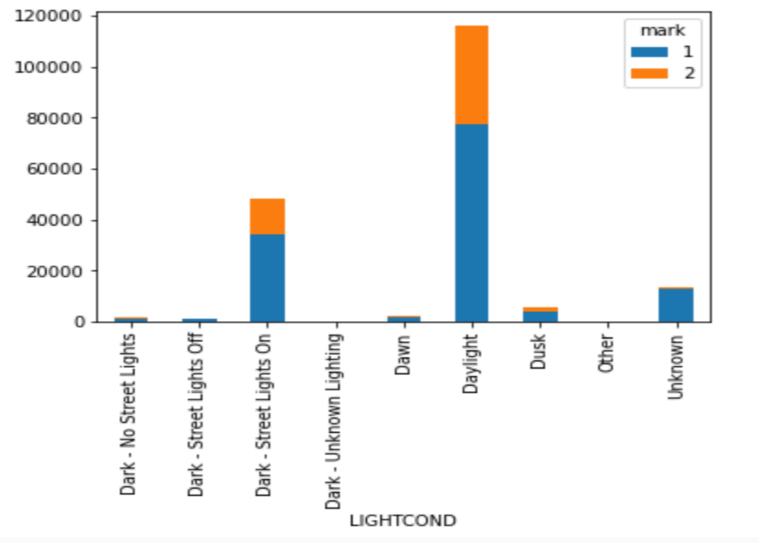
WEATHER



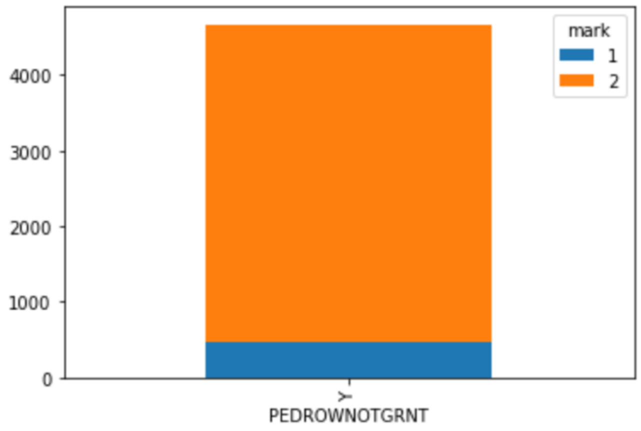
ROADCOND



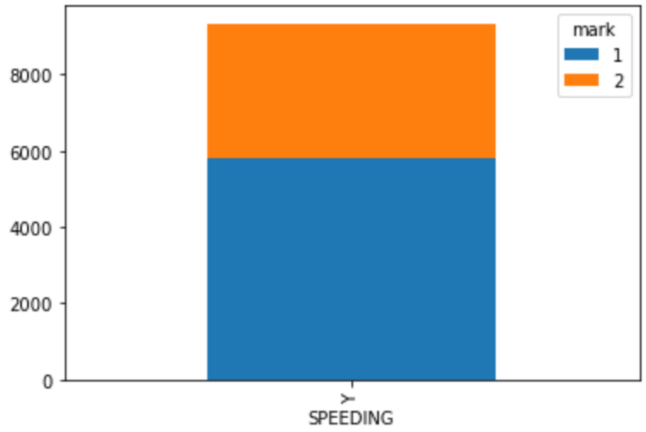
LIGHTCOND



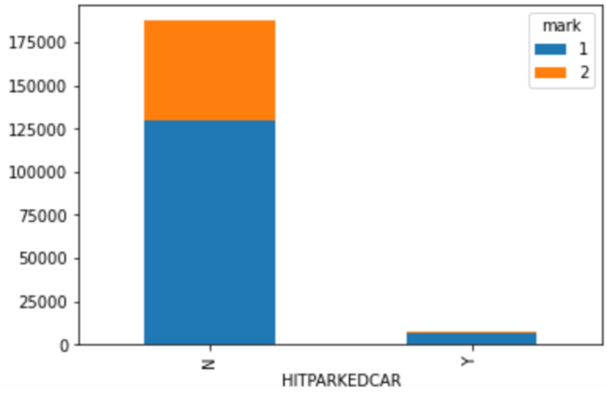
PEDROWNOTGRNT



SPEEDING



HITPARKEDCAR



Based on the above distribution of the categorical variables the ones with high cardinality were groped into small number of values in order to reduce the overfitting and complexity of model.

5. Predictive Modeling

In order to start with predictive modeling, we need to first select the important features. The dummy variables were created for the categorical variables and then all the variables were normalized.

For the feature selection the random forest model with the default parameters was used and the top 20 features by importance were selected for the modelling purpose. The list of top features with their importance's is as follows-

	Feature Name	Importance
1	SDOT_COLCODE	17%
2	PERSONCOUNT	16%
3	PEDCOUNT	9%
4	VEHCOUNT	7%
5	PEDCYLCOUNT	7%
6	CROSSWALKKEY	4%
7	INATTENTIONIND_Y	3%
8	ADDRTYPE_Intersection	3%
9	JUNCTIONTYPE_At Intersection (intersection related)	3%
10	PEDROWNOTGRNT_Y	3%
11	SEGLANEKEY	2%
12	SPEEDING_Y	2%
13	WEATHER_Other	2%
14	JUNCTIONTYPE_Mid-Block (not related to intersection)	2%
15	ADDRTYPE_Block	2%
16	ROADCOND_Other	2%
17	LIGHTCOND_Daylight	2%
18	UNDERINFL_N	1%
19	UNDERINFL_0	1%
20	WEATHER_Overcast	1%

Further using the above top features, the data was split into train-test sample in order to evaluate the model.

For the new data the various models were created and their performance for the optimal combination of hyperparameters is as follows-

Classifier	F1 Score
KNN	0.70
Decision Tree	0.70
SVC	0.69
Logistic regression	0.70
Random Forest	0.71

Out of all the models random forest has the best performance and hence is finally selected for the prediction.

6. Conclusion

We were able to achieve ~76% AUC which is a quite better performance index. Hence by implementing this model we can reduce the number of accidents and of course significantly reduce their severity.