Predicting the impact of an accident on traffic

Yoshita Buthalapalli

Abstract

The primary focus of this project is to predict the severity i.e the effect that an accident has on traffic. While predicting this, we will also get to analyse the factors that play a role in an accident severity. Understanding this information, like the weather conditions or the time of the day that are more prone to accidents, will aid in planning and allocating financial and human resources in a better manner.

1 Introduction

For training and testing the models, we will be using a dataset obtained from Kaggle named US Accidents - A Countrywide Traffic Accident Dataset (2016 - 2020). This is a countrywide car accident dataset, which covers 49 states of the USA. The accident data has been collected from February 2016 to Dec 2020. This dataset covers a wide range of features from accident location, time, weather conditions to points of interest indicating the presence of traffic signal or bus station nearby.

The next step after choosing a dataset is data cleaning and data preprocessing (described briefly in section 2) followed by training the models. For predicting the severity of an accident, we tried training four classification algorithms and compared their performance on the test data. The methods used are support vector machines, multinomial naive bayes, decision trees and random forest (described in section 3).

2 Data cleaning and preprocessing

- Few unwanted features like source of data, description of accident, street number,Country etc were dropped.
- Duplicate rows if any have been removed.
- If a column had Value errors like NaN or missing data and if it's ratio to the total number of rows is less then that row was dropped.
- One-hot representation eg: timezone or wind direction

- Aggregation. Eg: replace missing values with mean
- New attributes. Eg: convert start_time to month, day, weekday
- Abstraction. Eg: weather condition like snow,sleet or ice to snow
- Normalizing continuous values like temperature.

3 Methods used

3.1 Support vector machine

SVM is fundamentally a two class classifier. If there are K classes then an approach known as one versus one can be used. In this approach, K(K-1)/2 different two class SVMs are trained on all possible pairs of classes. For predicting a class, the one with highest number of votes is chosen. This approach is computation heavy and hence takes some time when the data is large. For this project, we used svm from sklearn library in python. By default this uses one vs one approach for multiclass classification. But still the accuracy was very low (0.48) and we had to move on to try another approach.

3.2 Multinomial Naive Bayes

This method is based on the assumption that there is conditional independence between every pair of features given the class. MultinomialNB from sklearn library in python was used which learns class prior probabilities by default. The likelihood of a class is calculated using,

$$p(\mathbf{x} \mid C_k) = \frac{(\sum_i x_i)!}{\prod_i x_i!} \prod_i {p_{ki}}^{x_i}$$

The accuracy was 0.561 which is better than previous svm result but still low.

3.3 Decision Tree

A decision tree is formed by simple conditions on individual features that recursively split the input space. The splitting is done until no improvement is possible. To measure this improvement we use impurity measure. There are two types of impurity measures namely entropy and gini. Using gini, we got an accuracy of 0.697 and using entropy it was 0.72.

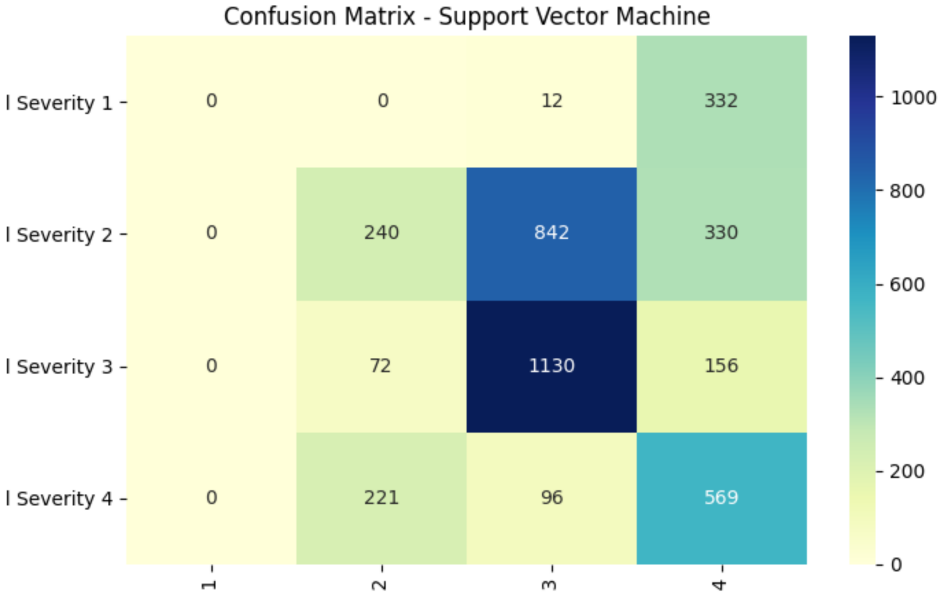
3.4 Random Forest

This consists of a set of decision trees. Each decision tree is built using a subset of features available. The predicted class with majority voting is then chosen.The parameters used for this model was 500 trees with maximum depth

35. The accuracy for this model was better than all the previous models, it was 0.76.

4 Experimental Results

4.1 Support vector machine

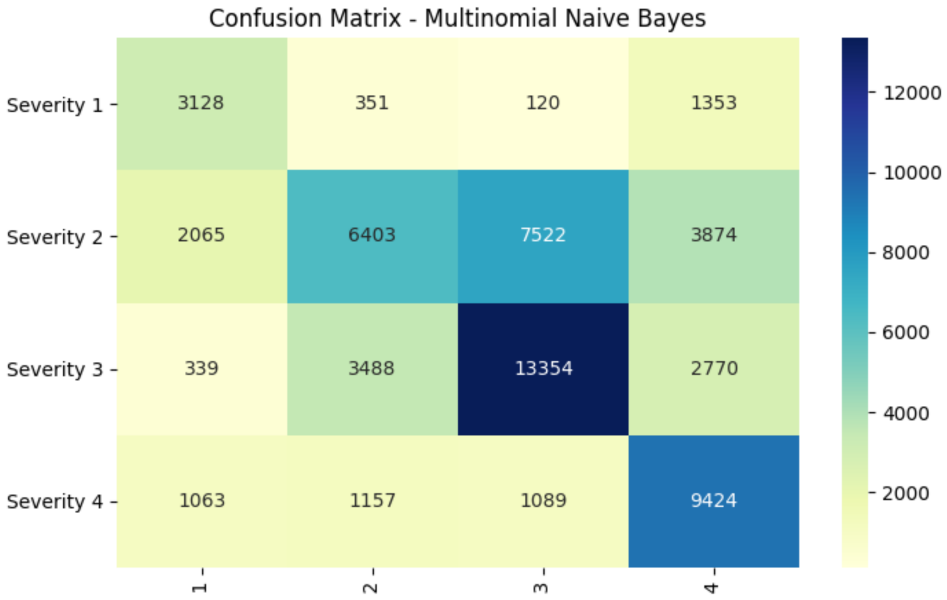


rows represent actual class and the columns represent predicted class values

Average precision for test set : 0.35
Average recall for test set : 0.41
Accuracy: 0.484
F1 score: 0.351

The parameters that gave a reasonable accuracy were polynomial kernal of degree 2 and penalty value 1. however, the accuracy is very low and computation time is high for this model.

4.2 Multinomial Naive Bayes

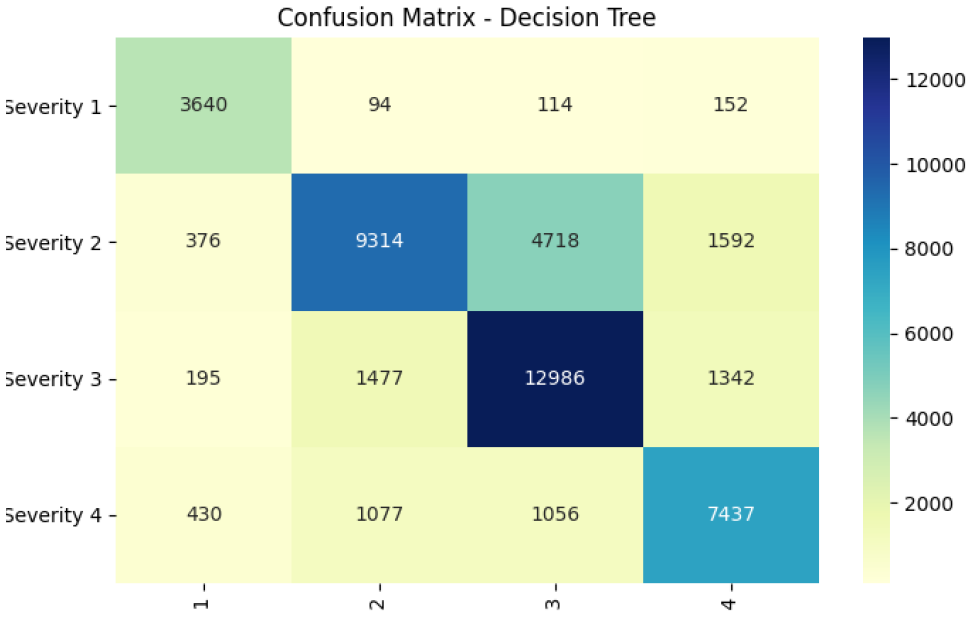


rows represent actual class and the columns represent predicted class values

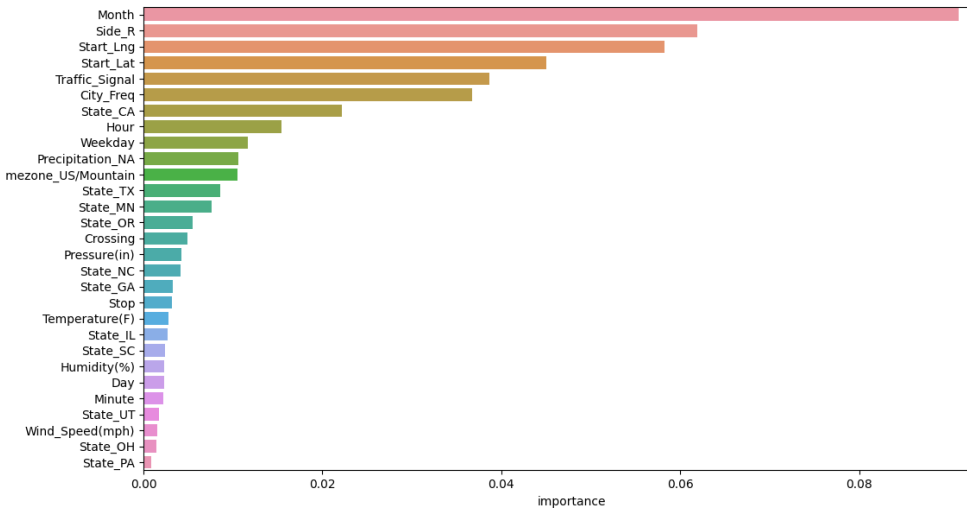
Average precision for test set : 0.55
Average recall for test set : 0.59
Accuracy: 0.561
F1 score: 0.552

This method couldn't perform well because multinomial Naive Bayes classifier is suitable for classification with discrete features like integer counts.

4.3 Decision Tree



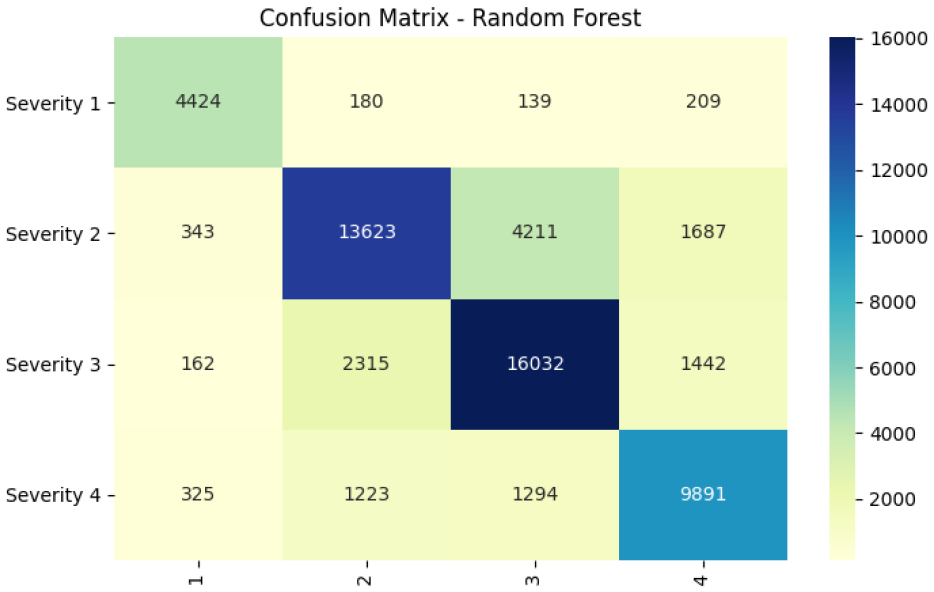
Average precision for test set : 0.72
Average recall for test set : 0.72
Accuracy: 0.703
F1 score: 0.703



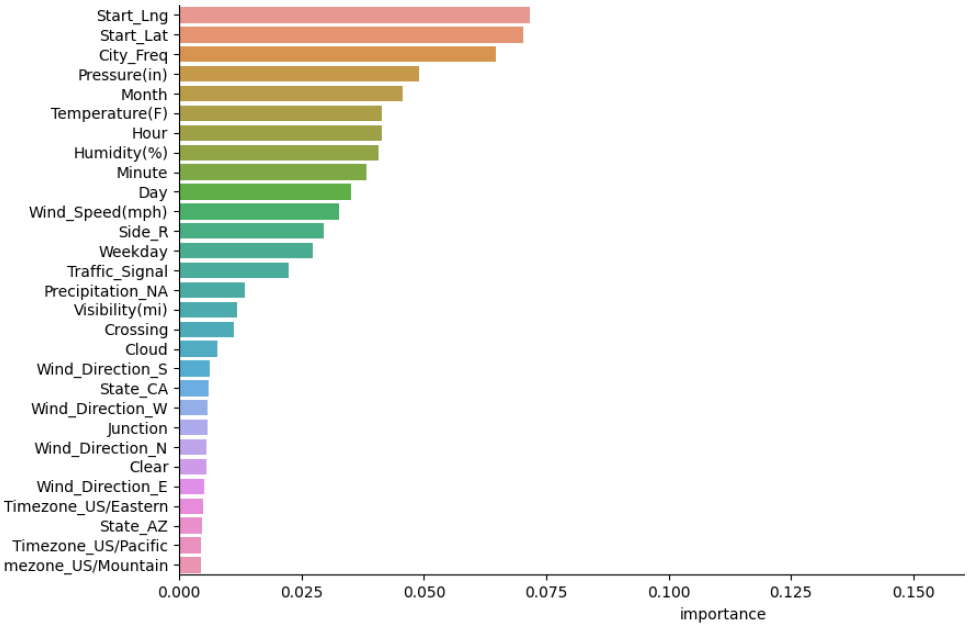
The above graph shows the importance of features for this model. We can infer

that the severity depended mostly on the month when the accident occurred, side of the road, location, city and also on the presence of a traffic signal near the accident location.

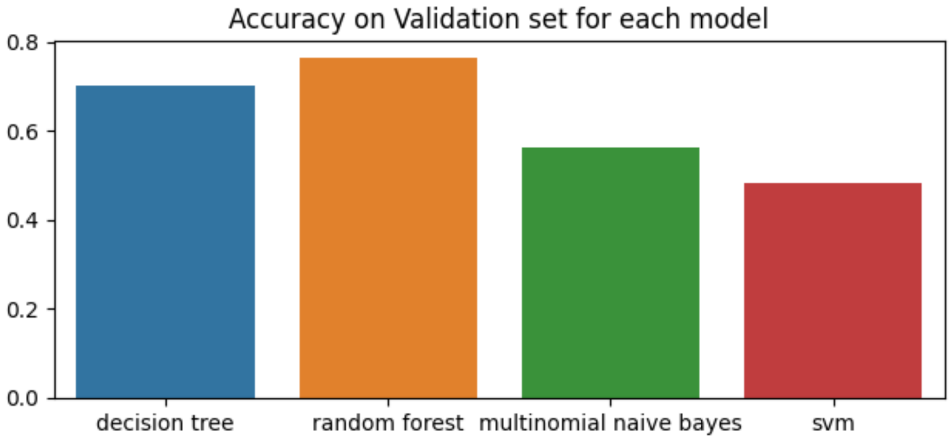
4.4 Random Forest



Average precision for test set : 0.78
Average recall for test set : 0.79
Accuracy: 0.764
F1 score: 0.782



The above graph shows the importance of features for this model. We can infer that the classification of severity depended mostly on the location, city where the accident occurred, pressure, month, temperature and hour of the day. Upon comparing the top 10 feature importance for decision tree and random forest we see that Month, location, city and hour is a common feature in both and they contribute more in the severity prediction.



5 Conclusion

Among the four methods that were used, random forest proved to be better at predicting the severity of an accident followed by decision tree algorithm. An interesting finding from the importance graph for decision tree was that the month and the presence of traffic signal nearby affected the severity of an accident. For future work, the relation between the severity and features of dataset can further be explored and analyzed to get more insights. The accuracy of predicted values could also be improved by maybe using different set of parameters or using other classification algorithms.

6 References

1. Dataset - Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, 2019.
2. Dataset - Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. ”Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights.” In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.
3. For data preprocessing, the workdone on the same dataset by the following notebook was very helpful -
https://www.kaggle.com/jingzongwang/usa-car-accidents-severity-prediction