Predicting the Severity Level in Traffic Collisions

A Inan

October 11, 2020

1. Introduction

1.1 Background

Traffic accidents and collisions affect both people and institutions since they have impact on human lives. They also lead to financial losses and traffic interruptions. In this regard, predicting the severity level of a collision in advance would help the local authorities such as police, emergency medical service and fire department to allocate necessary staff and equipment and to handle more delicate incidents that concern human lives in a better way. Insurance companies would also use these predictions to better deal with financial costs arising from collisions. Car producers would analyse the models, the most important factors and the predictions to produce safer vehicles. Besides, traffic interruptions lead to time losses and this means further financial losses and social distress. Traffic authorities would use these predictions to handle interruptions due to traffic accidents.

1.2 Problem

The objective of the project is to build a classifier model to predict the level of severity in case of a traffic collision based on the given details of the collision.

1.3. Interest

As mentioned above, local authorities, insurance companies and car producers would be very interested in a robust model to predict the level of severity in case of a collision.

2. Data acquisition

The data is downloaded from the website of Cursera. The original data is taken from Seattle Police Department and Seattle Traffic Department records. It is provided by Cursera within the framework of the Machine Learning Capstone project. The data provides information on the collisions between 2004-2020. It has 38 columns and 194,672 rows. 16 columns have numerical and the rest have object type of data.

The data includes details of a traffic collision including different types of IDs, location, address, date and time, severity, collision type, number of persons and vehicles involved, weather, condition of the road etc.

3. Data Manipulation

There were columns with duplicate information such as IDs given by different authorities, location, type of collision and codes for the collisions. Also, two columns had more than 95% missing values. In this framework the following columns were dropped from the data.

- X
- Y
- INCKEY

- COLDETKEY
- REPORTNO
- INTKEY
- LOCATION
- EXCEPTRSNCODE
- EXCEPTRSNDESC
- SEVERITYCODE1
- SEVERITYDESC
- INCDATE
- SDOT_COLCODE
- SDOTCOLNUM
- ST_COLCODE
- SEGLANEKEY
- CROSSWALKKEY

The variables named as SPEEDING, PEDROWNOTGRNT, INATTENTIONIND had only one category as Y and the rest was NaN values. It was assumed that the NaN values were rather N and therefore NaN values were replaced by N.

The variable named UNDERINFL had 4 categories, namely 0, 1, N, Y. 0 and 1 were replaced by N and Y respectively.

SDOT_COLDESC and ST_COLDESC had too many categories and it would not be reasonable to use all the categories in these variables. Thus, the categories with lower frequency were accumulated under the category of OTHER to decrease the number of categories.

INCDTTM was the variable of time stamp. The hour, day and month were derived from the time stamp as three different variables to see whether these time units make a difference in the level of severity.

4. Exploratory Data Analysis (EDA)

Pandas-profiling library and visualization libraries of Seaborn and Plotly were used for EDA.

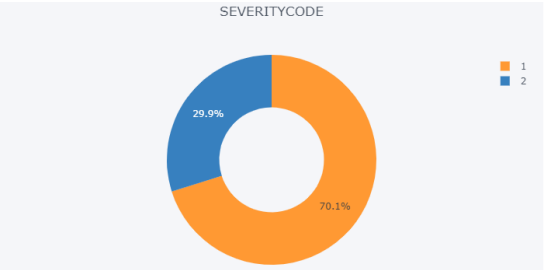
Pandas-profiling was useful to see the summary info of each variable in many aspects.

Since there were no continuous variables in the data, countplots piechart and kdeplot were used to visualize the variables.

Important findings from EDA are presented below.

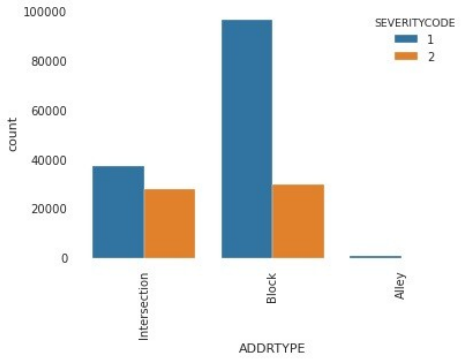
4.1 Severity Level

There are two levels of severity: 1 and 2. 70% of the collisions in the data has severity level of 1.



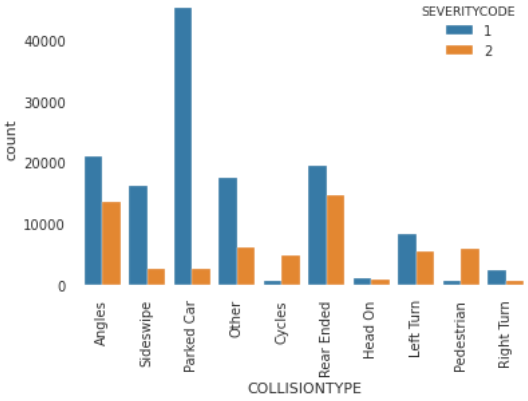
4.2 Collisions at the intersections have higher severity

The collisions at the intersection have a higher severity level compared to block and alley with respect to the address type of the collision.



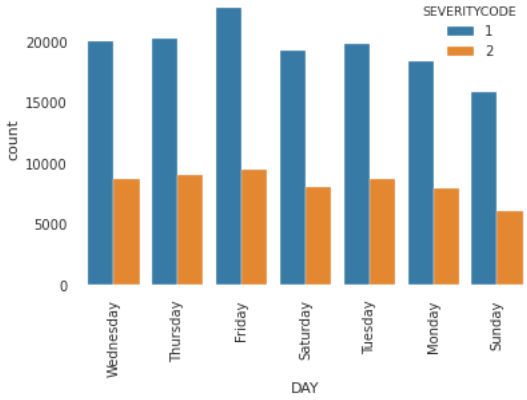
4.3 Collision type is an important differentiating factor regarding the level of severity

Categories of angles, cycles, rear ended, left turn and pedestrian have higher severity level.

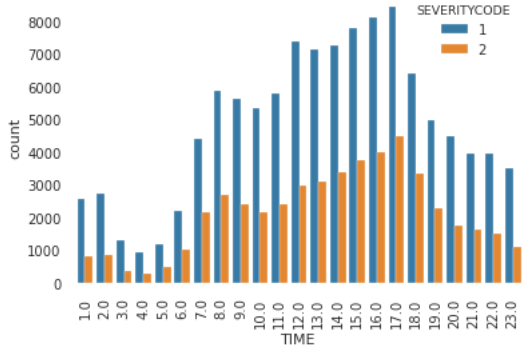


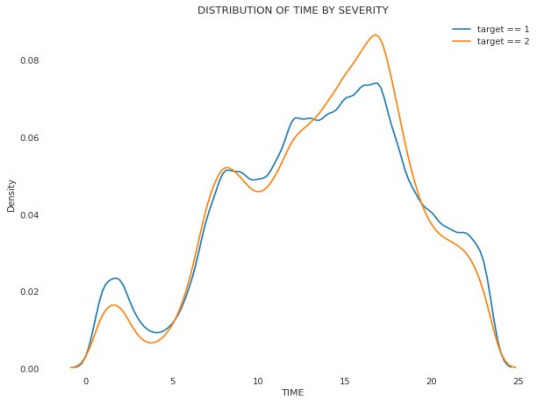
4.4 The day and time of the collision makes a difference

The number of collisions is higher on Friday and lower on Sunday.



The collisions in the afternoon have higher level of severity.





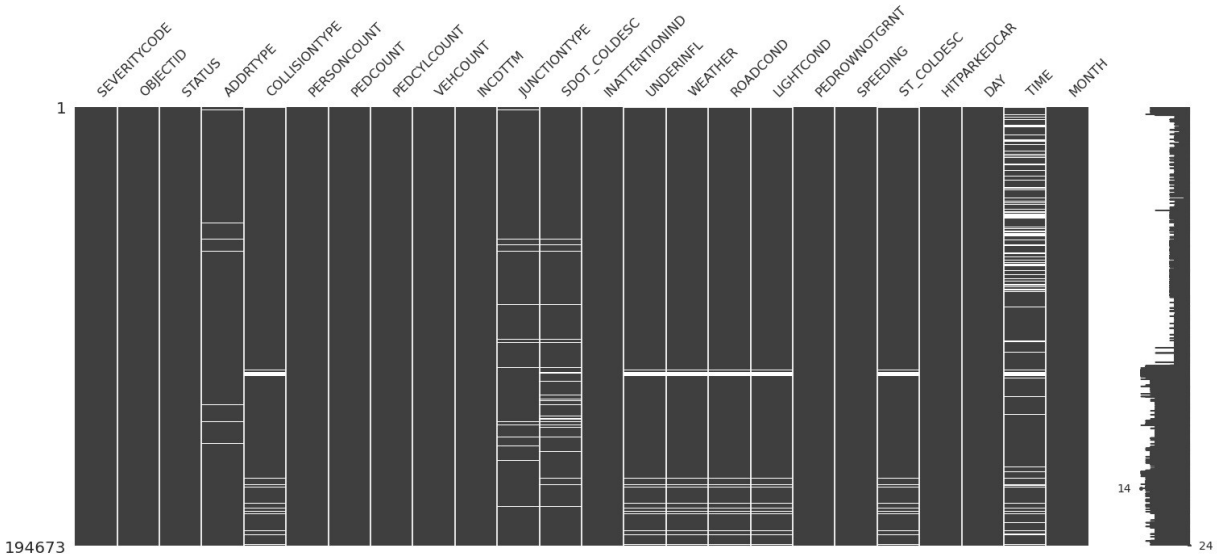
5. Missing Values

The total number and percentage of missing values in the data for each variable are as follows.

	Total_nan	Percent_nan
TIME	34381	17.660898
SDOT_COLDESC	9787	5.027405
JUNCTIONTYPE	6329	3.251093
LIGHTCOND	5170	2.655736
WEATHER	5081	2.610018
ROADCOND	5012	2.574574
ST_COLDESC	4904	2.519096
COLLISIONTYPE	4904	2.519096
UNDERINFL	4884	2.508822
ADDRTYPE	1926	0.989351

The number of missing values were reasonable to impute the NaN values.

The missingno library was used to visualize the missing values. It can be seen from the graphic that the NaN values are not random and there are systematic NaN values in different variables.



In this respect, multivariate imputation was preferred compared to univariate imputation. The KNN imputer from the scikitlearn library was used for the imputation of NaN values.

As it is explained later in this report that the data without imputation is also used in Light GBM, for comparison since this model is able to handle NaN values internally.

6. Data preparation

The categorical variables with two categories and more than two categories were converted through the label encoding and one-hot encoding respectively.

Most of the categories had spaces and once they were encoded, the names were corrected to prevent any errors in the models.

The OBJECTID was dropped since it was not needed in the model and might lead to misleading results since it had numerical values. INCDTTM variable with time stamp was also dropped since three variables were already derived from that variable.

The train and test data were split using shuffle and stratified since the data was imbalanced.

7. Modeling and Model Tuning

As mentioned above, this is a classification problem and the following classification models were used to come up with the best model to predict the level of severity.

- Logistic Regression
- Random Forest
- Light GBM (LGBM)

After getting the initial predictions and accuracy scores, 10-fold cross validation (CV) through RandomizedSearchCV tool from the scikitlearn library was used to determine the optimum hyperparameters for these models.

The predictions were made again after model tuning and CV and the final accuracy scores were calculated. The performance of different models within the framework of accuracy score is as follows.

	Accuracy_Score	rank
lgbm_final	0.763375	1
lgbm_final_with_null	0.763375	2
log_reg_final	0.759265	3
rf_final	0.751714	4

As it can be seen from the above that LGBM has a higher accuracy score. Furthermore, the scores of the two LGBM models are the same for the data with and without imputation of NaN values.

8. Feature Importance

In the tree-based algorithms, the feature importance of independent variables can be calculated according to their contribution to the model. LGBM algorithm was used for this purpose. The most important 10 features based on this calculation are presented below.

feature	importance
TIME	2701
PERSONCOUNT	1963
VEHCOUNT	979
SDOT_COLDESC_MOTOR_VEHICLE_STRUCK_MOTOR_VEHICL...	601
ST_COLDESC_One_parked_one_moving	498
INATTENTIONIND	491
UNDERINFL	491
SPEEDING	477
JUNCTIONTYPE_Mid_Block_not_related_to_intersec...	452
PEDCOUNT	399

Three numerical variables, the hour of the collusion, the number of persons and vehicles involved in the collision are the top three variables. The type of collision and the junction type are also important variables. Other noteworthy variables are; whether the collision was due to inattention, whether a driver involved was under the influence of drugs or alcohol, whether or not speeding was a factor in the collision and the number of pedestrians involved in the collision.

The factors above sound reasonable regarding the severity level and also some of these were also noticed during the EDA step.

9. Ensembling the Best Models

The scikit library has a method called VotingClassifier. It tries to ensemble models and come up with a result using the predictions of multiple models. This method was used to check whether this approach would contribute the accuracy score. But, it could not get better scores compared to LGBM. The scores are 0.7631 and 0.7628 for hard voting and soft voting respectively.

10. Conclusion

The traffic collisions concern people, public and private institutions as they concern everybody. They cause loss of human lives, injuries, financial costs and traffic interruptions.

Therefore, the aim of this project was to predict the severity level of a collision given the factors derived from the data of the police department and traffic authority.

Three algorithms were used to build a model and the LGBM algorithm gave the best result with an accuracy score of 76.33%.

The most important factors were the hour of the collusion, the number of persons and vehicles involved in the collision, the type of collision, the junction type, whether the collision was due to inattention, whether a driver involved was under the influence of drugs or alcohol, whether speeding was a factor in the collision and the number of pedestrians involved in the collision.

This model can be useful for local authorities, insurance companies and car producers.

11. Future research

Only three algorithms were used since the data was relatively large and it was time consuming to try additional algorithms. Other algorithms can be used for comparison.

The area and the location of the accidents were not taken into consideration in this analysis, but a cluster analysis can be done to see whether different regions make a difference in the level of severity and added to the model.

In addition, more variables can be brought to the analysis through feature engineering to increase the accuracy score.

Finally, other factors such as the model, type and make of the vehicles could be added to the factor to see whether these would contribute to the model.