Predicting the Severity Level in Traffic Collisions

A Inan

October 11, 2020

Background and Project Objective

- Collisions affect both people and institutions.
- They concern human lives, lead to financial losses and traffic interruptions.
- Predicting the level of severity would help the local authorities, insurance companies, car producers.
- The objective of the project is to build a classifier model to predict the level of severity in case of a traffic collision.

Data

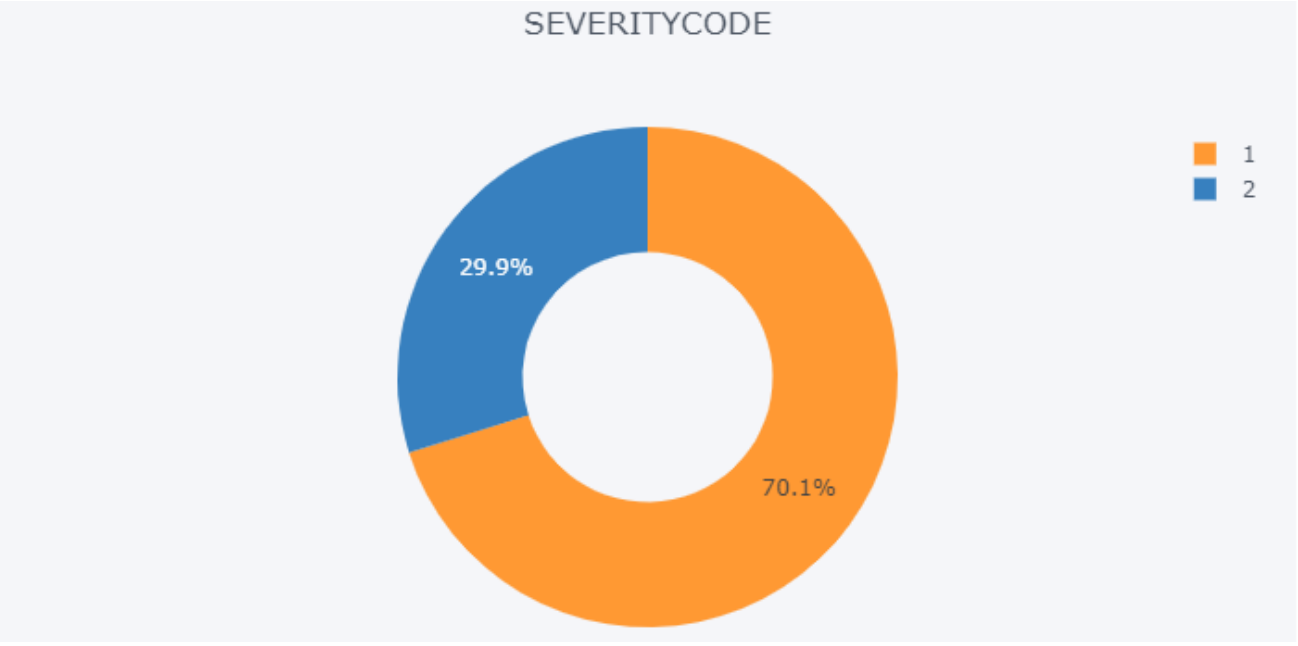
- The data is downloaded from the website of Cursera.
- The original data is taken from Seattle Police Department and Seattle Traffic Department records.
- It provides information on the collisions between 2004-2020. It has 38 columns and 194,672 rows.
- It includes different types of IDs, location, address, date and time, severity, collision type, number of persons and vehicles involved, weather, condition of the road etc.

Data Manipulation

- Columns considered unnecessary, duplicate and with more than 95% missing values were dropped.
- Categories in some of the variables were rearranged.
- New variables from time stamp were created.

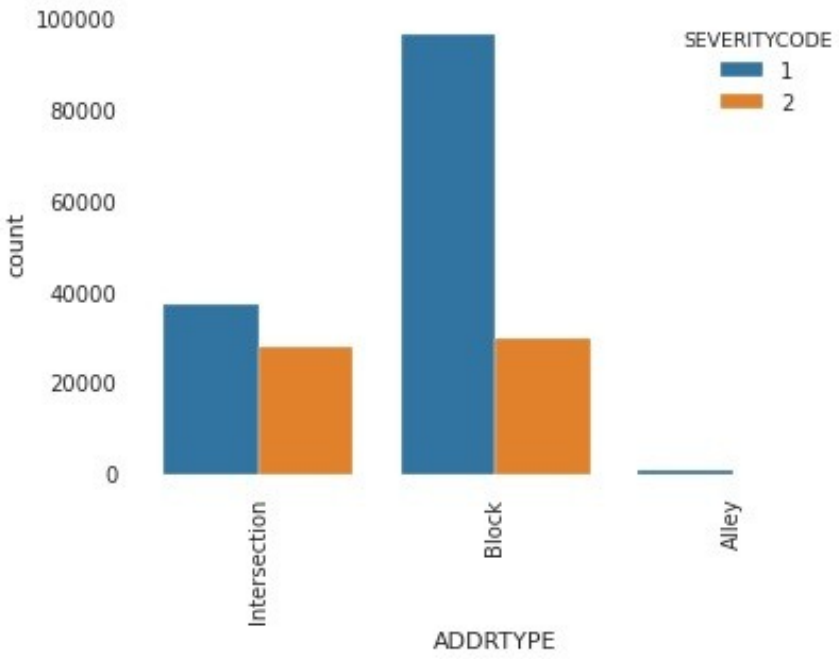
Exploratory Data Analysis

- 70% of the collisions in the data has severity level of 1.



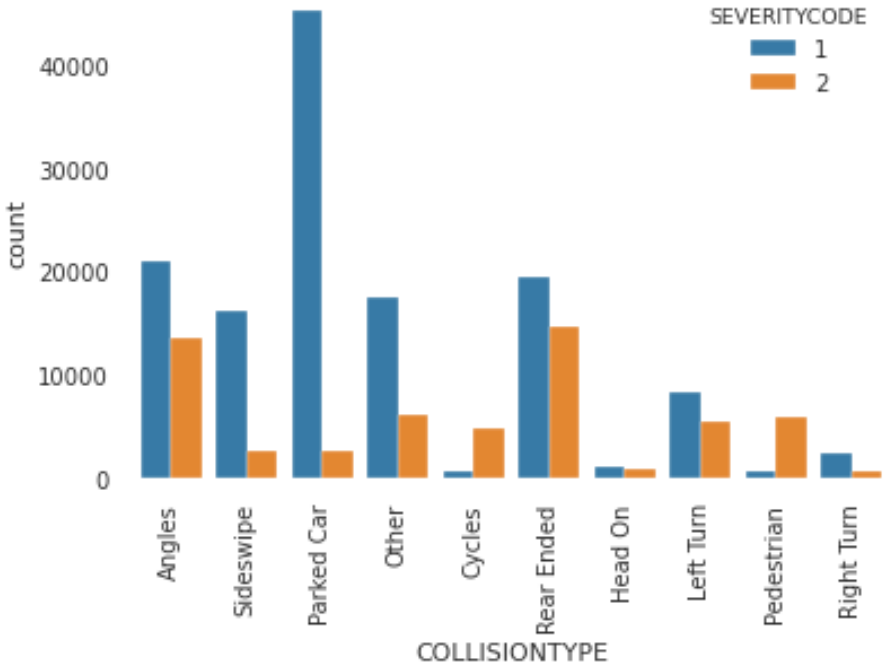
Exploratory Data Analysis

- Collisions at the intersections have higher severity.



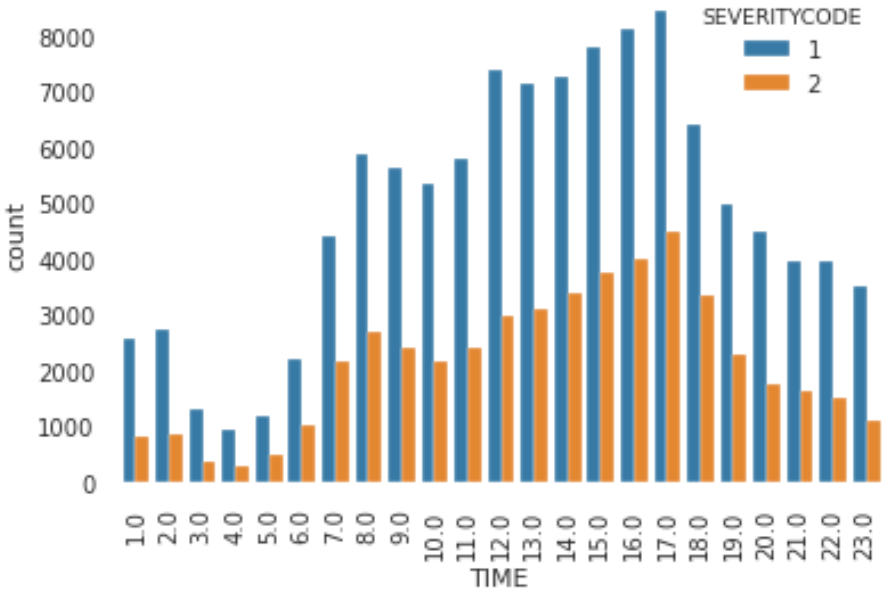
Exploratory Data Analysis

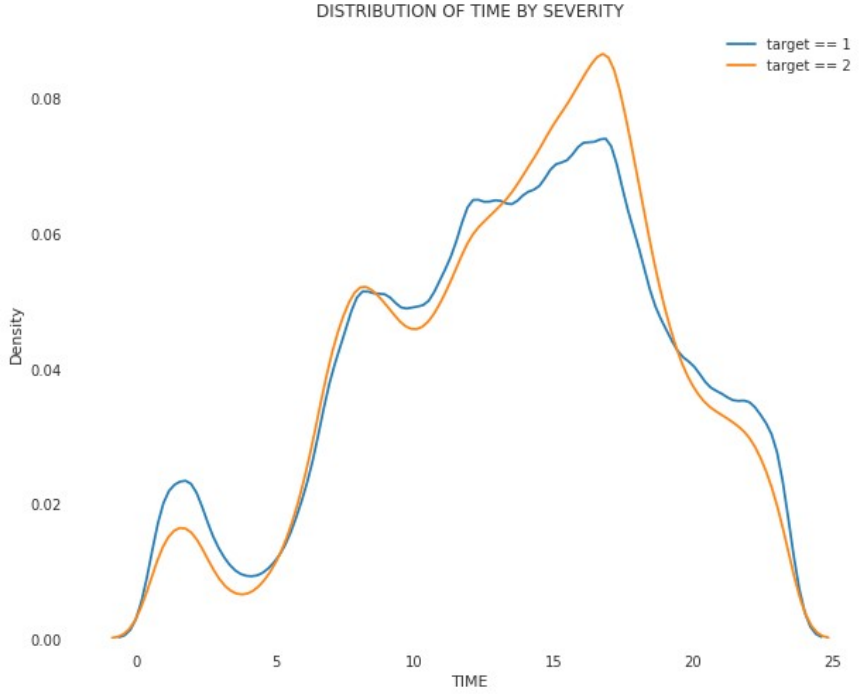
- Collision type is an important differentiating factor.



Exploratory Data Analysis

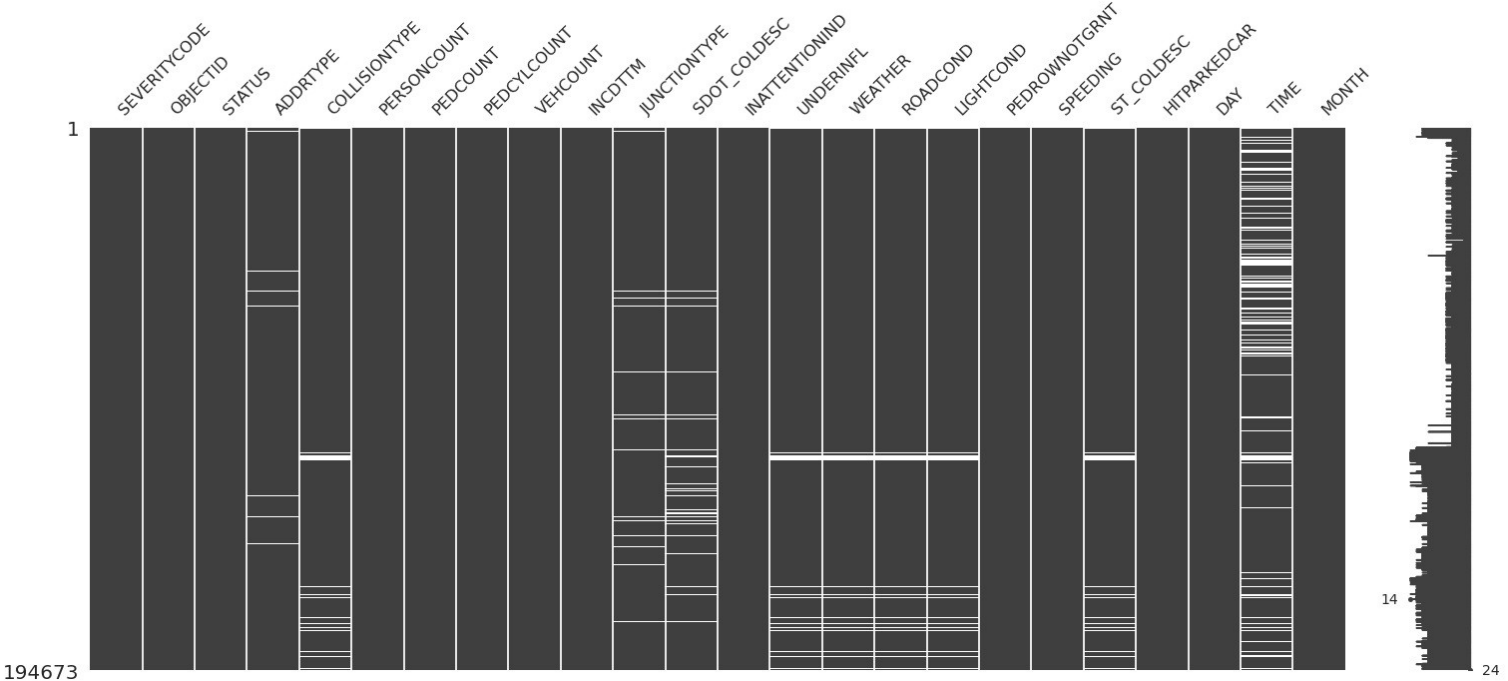
- The time of the collision makes a difference.





Missing Values

- Missing values were not random.
- The KNN imputer from the scikitlearn library was used.



Modeling and Model Tuning

- Logistic Regression, Random Forest and Light GBM (LGBM) models were used to build a model and make predictions.
- 10-fold Stratified Cross validation was used for cross validation since the data was imbalanced.
- RandomizedSearchCV was used for model tuning.

Results

- The accuracy scores of the three models:

	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

Feature Importance

- The top 10 features based on the LGBM model:

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

Conclusion

- 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.

Future Research

- Other aspects need to be considered in the future researches:
  - Different algorithms
  - Area and location
  - Further future engineering
  - The make, model and type of the vehicle