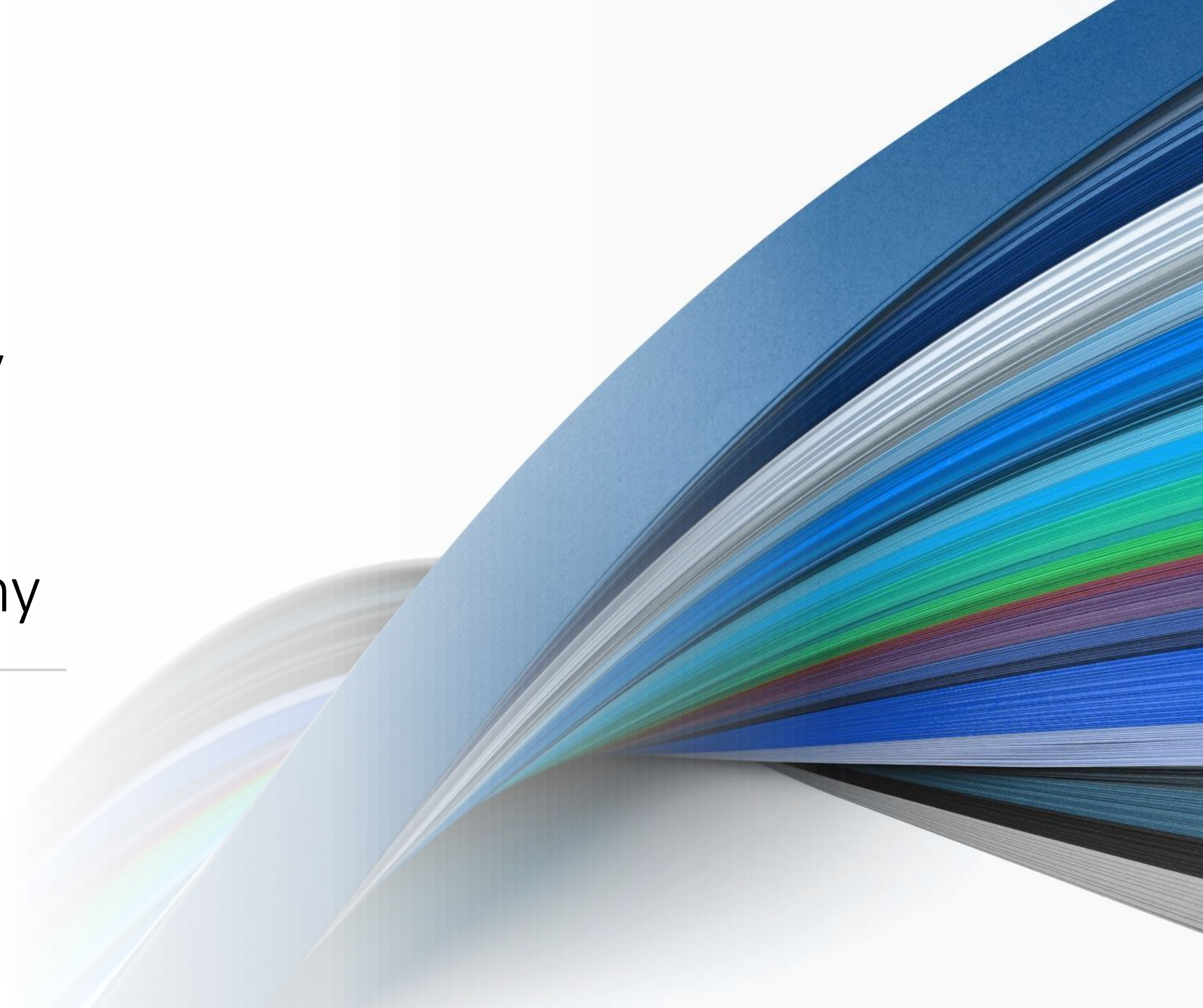
Predicting severity of car collisions in Seattle for Insurance Company

IBM Coursera Capstone Project



Importance of severity index for Insurance company

- The company will highly benefit from the information as if they know and segment their policies depending on the classification of the car accident, they could deduce clauses that would help them reduce costs due to affected customers.
- Customers acquiring the company service can be benefited as they'll be more aware of the type of accidents where they'll get coverage on, and how they could use the insurance service more effectively.

Data source and Cleaning

- **Source**
  - Data from SDOT Traffic management division and Traffic records group.
    - All collisions have been provided by SPD and recorded by Traffic Records. The data set contains all type of collisions produced within the limits of the city of Seattle from year 2004 to the present year.
- **Cleaning**
  - In total 194378 rows and 38 columns in the raw data set.
  - Columns with lots of null values were converted to dummies and originals were dropped.
  - Duplicate columns and ID columns were dropped.
  - Rows with missing or duplicate information was dropped from the data set.
  - Converted data types
  - Final working data set had 184986 rows and 12 columns

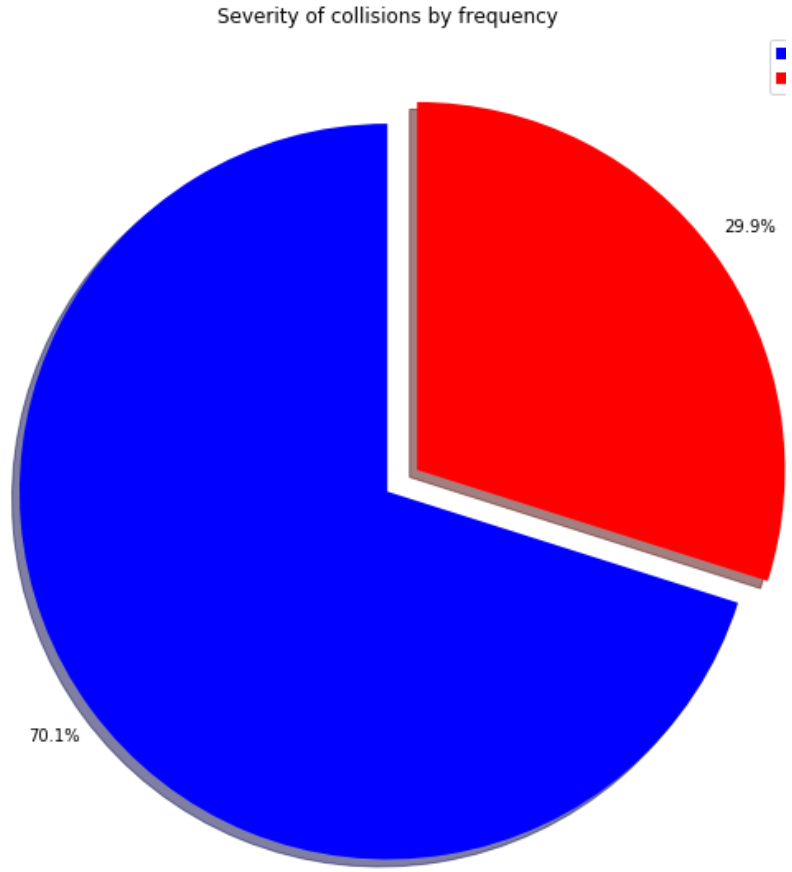
Dropped columns in raw data set

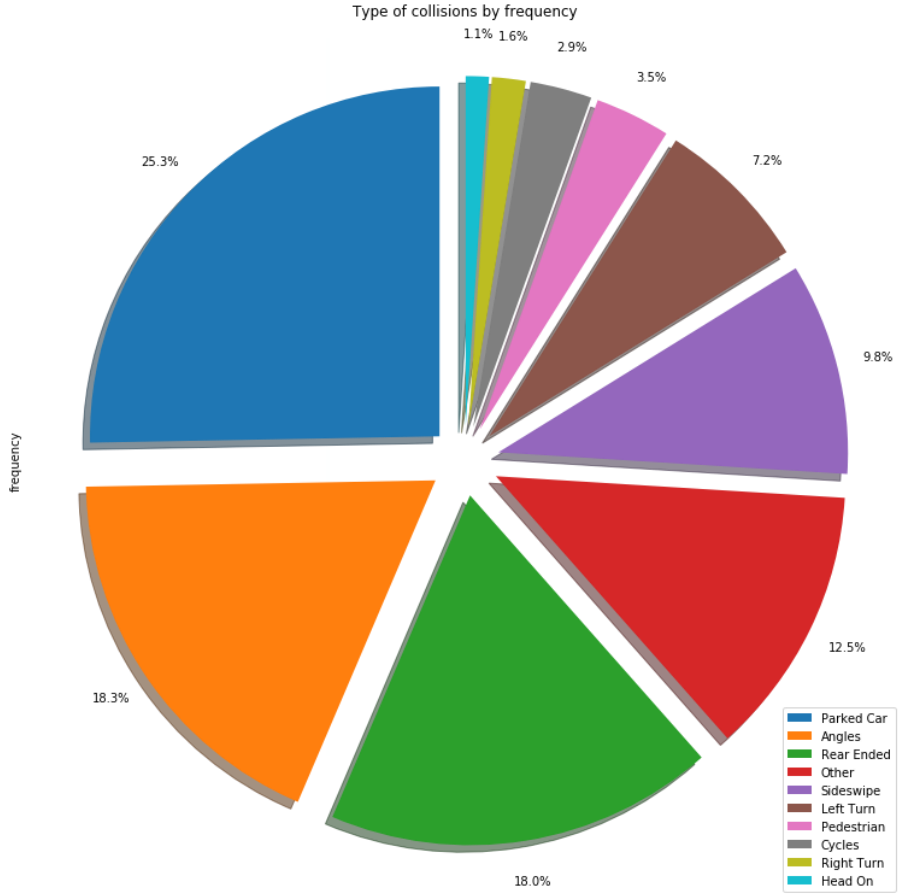
Kept attributes	Dropped attributes	Reason
ADDRTYPE	X, Y, LOCATION	Location of the accident was defined as unimportant. More important is where the accident took place
ST_COLCODE, ST_COLDESC	'OBJECTID','INCKEY','COLDETKEY', 'REPORTNO','STATUS','INTKEY'	Dropped all the id relevant columns as they are used just for indexing the collision
SEVERITYCODE	'SEVERITYDESC','SDOT_COLCODE', 'SDOT_COLDESC','SDOTCOLNUM', 'SEVERITYCODE.1'	Similar features, and more indexing
PEDCYCLCOUNT, INCDTTM, PEDROWNOTGRNT	'PERSONCOUNT','PEDCOUNT', 'VEHCOUNT','INCDATE',	Redundant date column, redundant involvement of pedestrians, and numbers of vehicles involved seemed irrelevant for our analysis

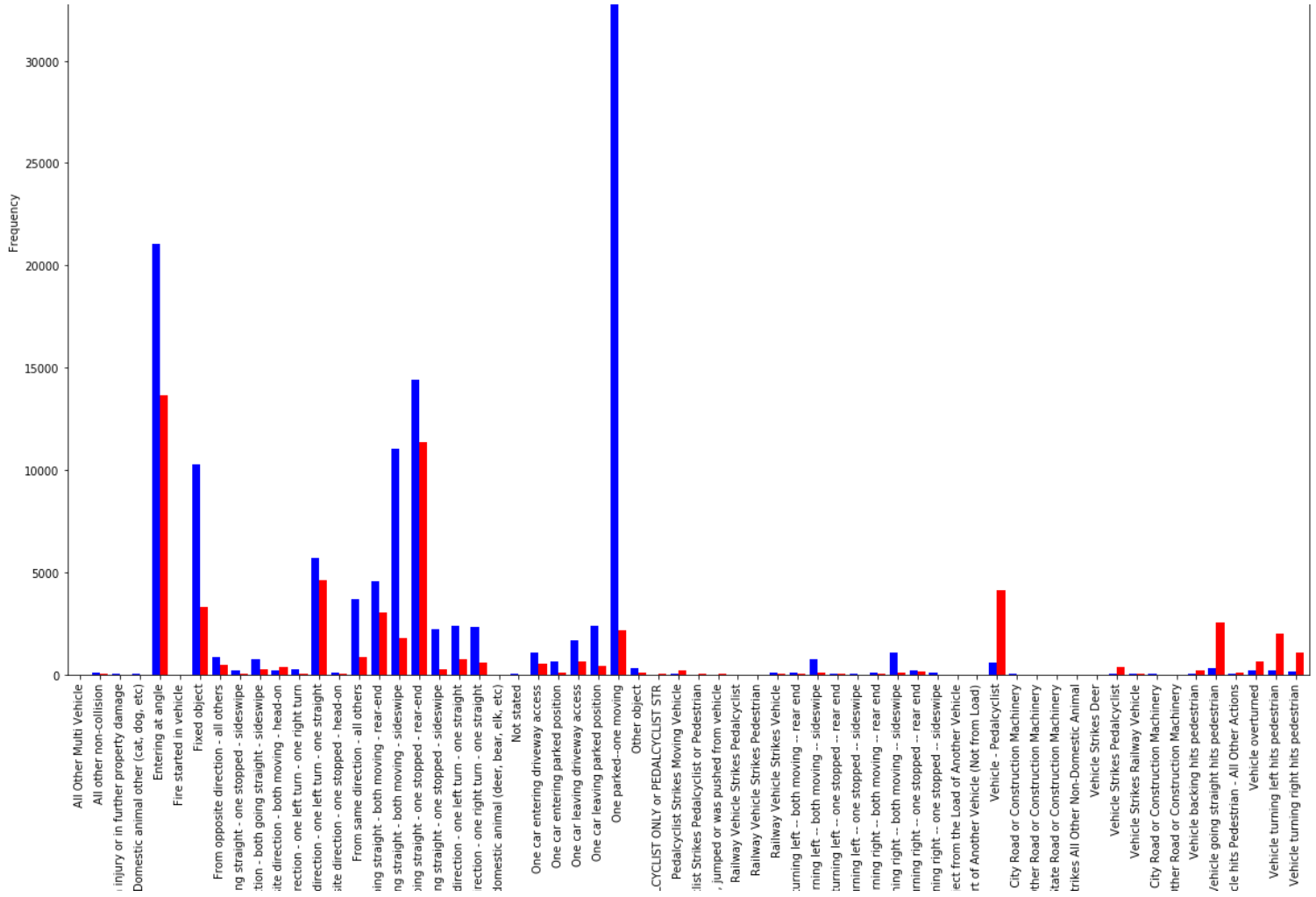
Exploratory Data Analysis

- **Determining reasons that caused the most accidents**
  - Relationship between accidents and weather = Negative
  - Relationship between accidents and road condition = negative
  - Relationship between accidents and distraction = negative
  - Relationship between accidents and alcohol = negative
  - Relationship between accidents and lighting = negative
  - Relationship between accidents and type of junction= negative
  - Relationship between accidents and type of collision= High correlation

Types of collision and Severity index frequency distribution







Relationship between severity index and type of collision

Feature set building for classification model

Attribute selected	Reason
Alcohol_No, Alcohol_Yes	High correlation with severity index
PEDCYCLCOUNT	High correlation with severity index, easily categorized as injury
ST_COLCODE	Main distribution of cases with a lot of variables to analyze
HitParkedYes, HitParkedNo	High correlation, it can be easily categorized as property damage
Inattention	High correlation with severity index
Speeding	High correlation with severity index, easily categorized as injury
Pedestrians	High correlation with severity index, easily categorized as injury

Predictive Modeling

- The model wants to determine the class or type of severity associated with the attributes selected, I decided to base the predictive model in a classification model, in which the result could be either 1 or 2. As such I decided to test the model using two different approaches, using K-Nearest Neighbors, and Logistic Regression.
- Each model certainty was measured using Jacquard index score, F1, and Log loss.
- Also a Confusion Matrix was built in order to analyze the results of the predictions the model made.

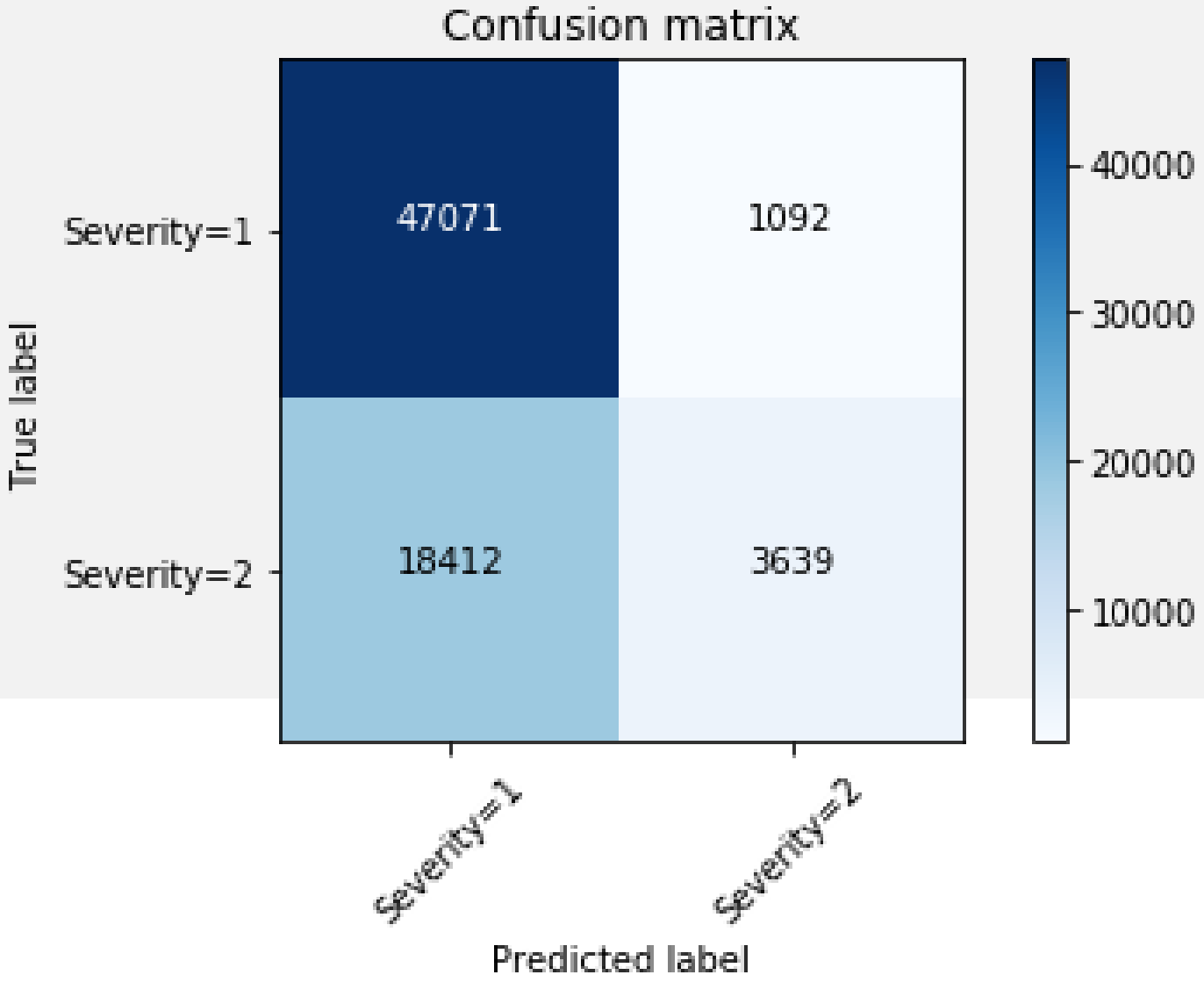
Accuracy scores for both models

- K-Nearest Neighbors

	Accuracy score	F1 score	Jacquard Similarity score
KNN	0.7407	0.6863	0.7407

- Logistic Regression

	Acc Score	F1 score	Jacquard Similarity score	Log loss
Logistic Reg	0.7222	0.6536	0.7222	0.5625



Confusion Matrix

Discussion of Results

- From the model we built, it is clear that the data used to predict the severity of the accidents still suffers from uncertainty as we have a high accuracy rate for accidents related to property damage, but we have a low accuracy score when determining accidents that are related to injuries of the persons involved. Still the overall prediction score using KNN model and Logistic Regression has an accuracy rate of 73%, so we can still consider our model as accurate to determine the severity of an accident.
- As the most common type of collisions are against parked cars, this parameter may have influenced the model to be more biased toward property damage as we determined that this type of collisions was associated to a lower severity index. To build a more accurate model we will need further information or attributes that could help us determine a better relationship between the severity index and the attribute selected.

Conclusion

- From the recollection of the data, we have determined that the accidents that involve a higher severity index are those related to accidents involving pedestrians and cyclist with almost 100% certainty. So, we recommend to the insurance company that it should modify policies of car accidents that are related to hitting pedestrians or cyclists, and to some extent motorcycles. This way, the insurance company can decide if they reduce the coverage of the person that caused the accident and covering most of the medical expenses of the pedestrian or cyclist involved in the accident.