Predicting the severity of car collisions

Maximovich Alexander

September 24, 2020

1. Introduction / Business Problem

As the number of cars is growing and consequently more car accidents occur, prevention of the car accidents has become important in the modern society. Car accident kills or injures people, damages vehicles, buildings, goods, or suffer a material loss. What can we do to decrease its occurrence?

Information technologies have been rapidly developed in recent decades. Large volumes of data are being collected, stored and analyzed to make different processes more efficient. Minimization of quantity of car incidents and their severity is also the area for applying different data science techniques.

It is very important for society to decrease number of collisions and lower damage to people and property. As a result, state organizations responsible for traffic safety are key stakeholders in this business problem. Using results derived from car collisions data analysis, they will be able to make changes in traffic regulation, adding new road signs and crossroads to improve traffic safety.

To solve this business problem, we need to acquire detailed data about car collisions, prepare data for further analysis, choose proper data analysis tool and construct the model which allows to determine key factors responsible for road incidents and produce warning signals when some critical conditions appear.

This predictive model will be deployed in systems used in traffic regulation processes and will estimate current traffic, weather and other conditions around road network to produce recommendations which preventive actions to be made to minimize the quantity of cars collisions and their severity.

2. Data acquisition and cleaning

Seattle Police Department (SPD) has been collecting detailed data about cars collisions. It will be used for developing model that allows to determine locations, weather conditions, days of week, time of day and other factors that are helpful for car collisions prediction. ¶

the dataset is available at https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

File with metadata is available at https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf

Data contains 194,673 records with 38 variables. The target variable is the severity code that classifies each collision according to type of sequences.

SEVERITYCODE	SEVERITYDESK
1	Property Damage Only Collision
2	Injury Collision

All variables have been analyzed if they can be used as factors that allow to predict the severity of car collision. Dependent variables and technical ones (like incidents' keys) were excluded. As well, variables that contain incident's details that are available only after the incident had happened (like number of cars involved and etc.) were excluded. Then all remaining variables were checked for data completeness, those that contained more than 50% of empty records were excluded too.

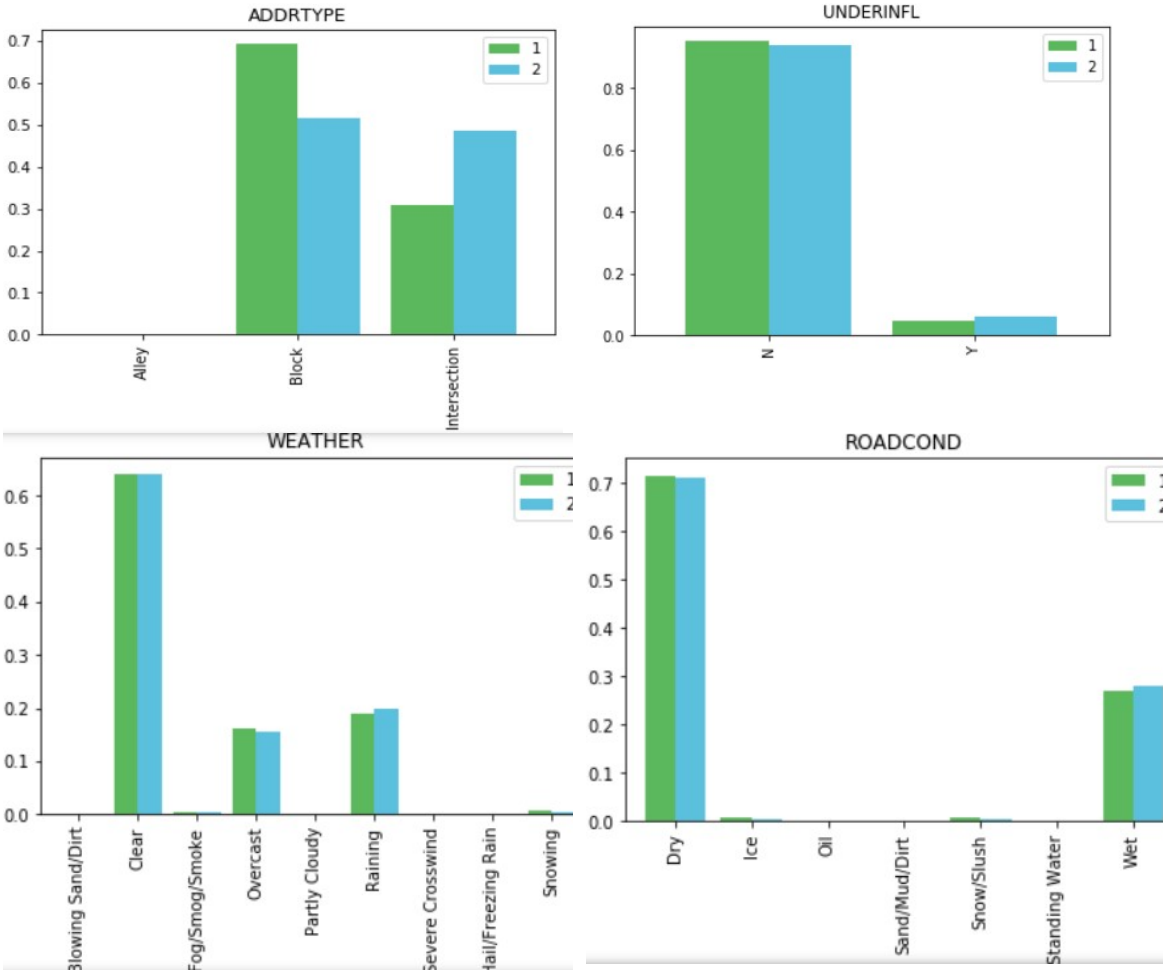
As a result, only 7 variables were selected as independent ones that can be used to predict the target value:

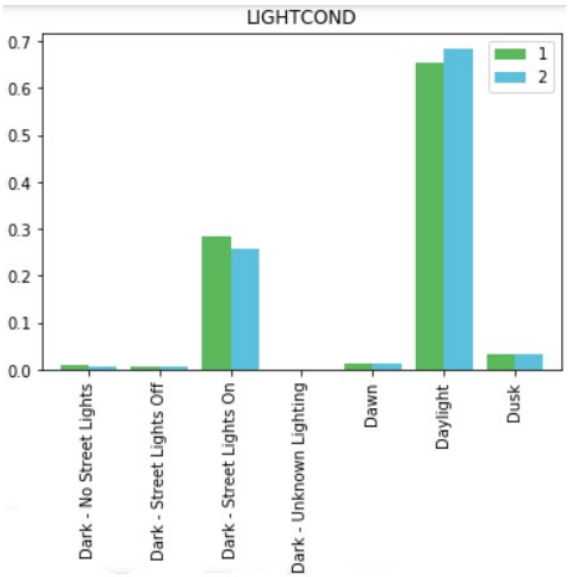
Name of variable	Variable description
ADDRTYPE	Collision address type
INCDATE	The date of the incident.
UNDERINFL	Whether or not a driver involved was under the influence of drugs or alcohol.
WEATHER	A description of the weather conditions during the time of the collision.
ROADCOND	The condition of the road during the collision.
LIGHTCOND	The light conditions during the collision.
JUNCTIONTYPE	Category of junction at which collision took place

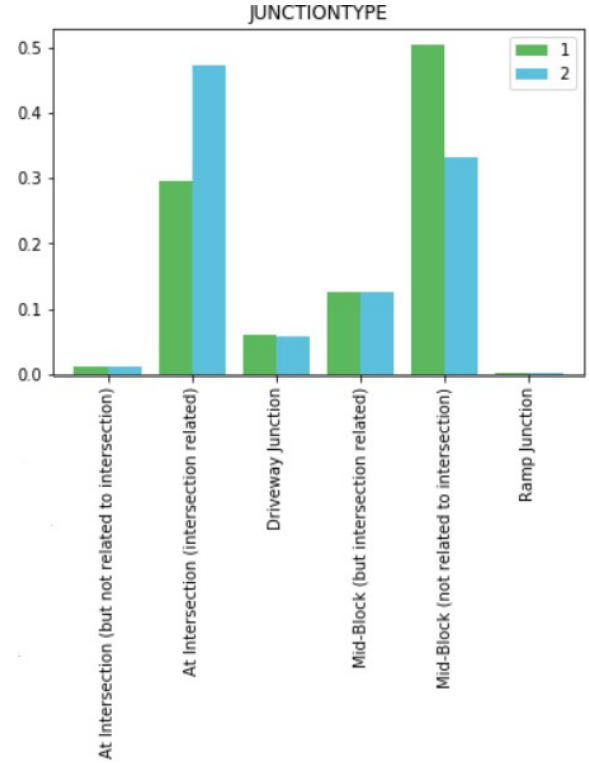
The dataset with 7 independent variables and the target variables was checked for empty and incorrect data. All records with any empty fields and fields containing 'Unknown' or 'Other' values were deleted. As a result, the number of records decreased to 167,335.

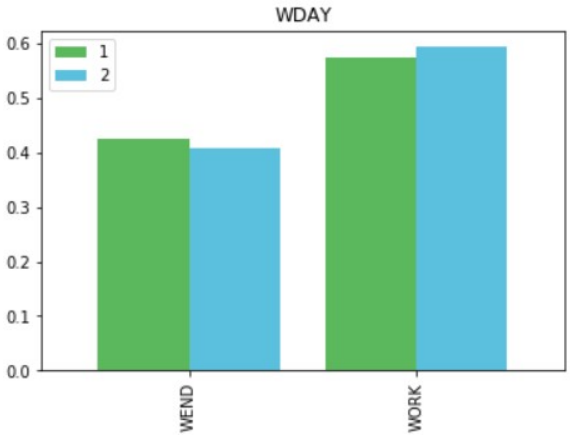
All values in INCDATE was changed to day of week (0 – Monday) and then divided in two subgroups [Monday – Thursday] and [Friday – Sunday].

Bar charts for normalized independent variables were plotted to identify those variables that varies for different severity codes.









There were identified four factors with the most prominent difference for two severity codes: ADDRTYPE, LIGHTCOND, WDAY and JUNCTIONTYPE.

As a result, the dataset which would be used for model's selection and estimation includes these four independent variables and one target variable.

3. Predictive Modeling

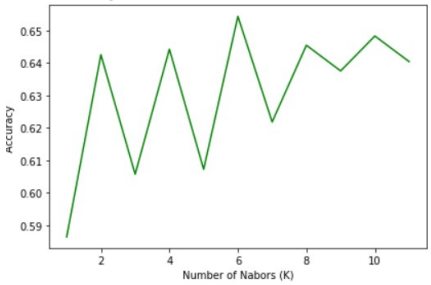
The object of the model is to determine cars collision severity (1 or 2). Classification model is the best choice for such kind of task.

There are three classification algorithms were used to select the algorithm with the highest accuracy: K-Nearest Neighbors (KNN), Logistic Regression (LR) and Support Vector Machines (SVM). The criteria for selection of model with the highest predictive power was average of F1 score and Jaccard similarity score.

Dataset was divided in training set (80%) and test set (20%).

Each of the algorithms was run with different parameters: KNN with different number of neighbors, LR with different solver types, and SVM with different kernels.

KNN (neighbors)



LR (solvers)

```
newton-cg =  0.622
lbfgs =  0.622
liblinear =  0.622
sag =  0.622
saga =  0.622
```

SVM (kernels)

```
linear =  0.668
poly =  0.668
rbf =  0.668
sigmoid =  0.578
```

Then accuracy results of algorithms with selected parameters (number of neighbors = 6, solver = 'liblinear', kernel = 'rbf') was estimated with F1 score and Jaccard similarity score.

Algorithm	Jaccard	F1 score	Average
KNN	0.654	0.573	0.614
SMV	0.578	0.575	0.576
LR	0.622	0.627	0.624

The best average classification power showed Logistic Regression algorithm (62.4% accuracy).

4. Discussion

The model confirmed that severity of car collisions address type, junction type, light condition and day of week (working days of weekends). As a result, additional measures in traffic regulation in intersections on interconnection roads during weekends with daylight time. These actions allow to decrease of collision resulted in injuries of people.

5. Conclusion

In this study, I analyzed key factors that allow to predict the severity of car collisions. I identified that address type, junction type, light condition and day of week (working days of weekends) have the most predictive power. I built classification models to predict the type of car collisions severity and determined that Logistic Regression model had the highest predictive accuracy. This model helps identify places, light conditions and days when additional traffic regulation measures are to be applied that results in lower cases of injuries both pedestrians and drivers.