Predicting Traffic Accident Severity

Jialiang Zhang

October 07, 2020

1. Introduction

1.1 Background

Nowadays, with the car counts increasing rapidly in global area, there are more and more traffic accidents occurring every day. The current situation impacts everyone's daily life, as well as the city management problem for the governments.

Hence, it is important to analyze the severity of traffic accident factors with select the appropriate data label within weather conditions, special events, roadworks, traffic jams, etc.

1.2 Problem

This project aims to predict the severity of accidents with historical data that could provide the emergency services for involved citizens.

1.3 Interest

The benefits of data insight: the corresponding government could allocate the resources feasibly. Besides, the drivers could be notified to change the planned route for avoiding the traffic accident.

The transportation department should be interested in the data analysis and predicate which can save time and budget. In the meantime, the citizens will save time and will not put themself into the risk.

2. Data

2.1 Data Acquisition

The original data comes from the following Kaggle dataset which is divided into 4 parts including: *characteristics*, *places*, *users*, *holiday*.

2.2 Feature Selection

The dataset that resulted from the feature selection consisted in 839,985 samples, each one describing an accident and 29 different features.

In the *characteristics* dataset, features: "lighting", "localisation"(agg), "type of intersection", "atmospheric conditions", "type of collisions", "department", "address", "time" and the coordinates.

In the *places* dataset, features: "road categories", "traffic regime", "number of traffic lanes", "road profile", "road shape", "surface condition", "situation", "school nearby" and "infrastructure".

By the *users* dataset, created the following features:

- num_us: total number of users involved in the accident.
- ped: Weather there are pedestrians involved or not.
- critic_age: If there is any user in between 17 and 31 years old.
- sev : maximum gravity suffered by any user involved in the accident:
    - 0 = Unscathed or Light injury
    - 1 = Hospitalized wounded or Death

In the *holiday* dataset to craft a new feature indicating the accident accuracy during a holiday.

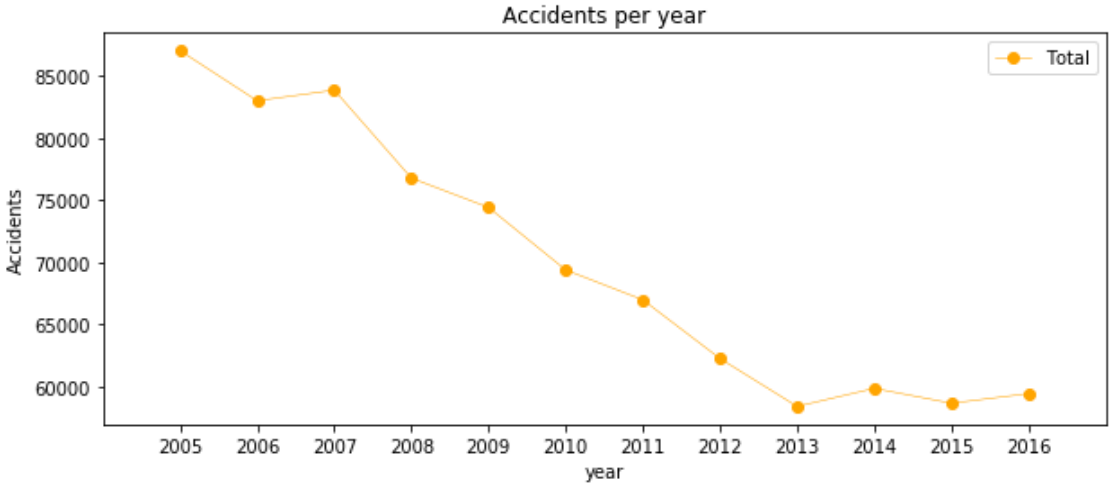
2.3 Data Wrangling

The first step was to deal with missing values and outliers. Initially the latitude, longitude and road number were dropped form the data. Then keeping with replacing the missing values, the analysis was divided in two groups of features. Finally with the rest of the features with missing values, the transportation department, the number of lanes, the road profile and shape and the situation at the time of the accident, the NaN and outliers were replaced with the feature's most popular value.

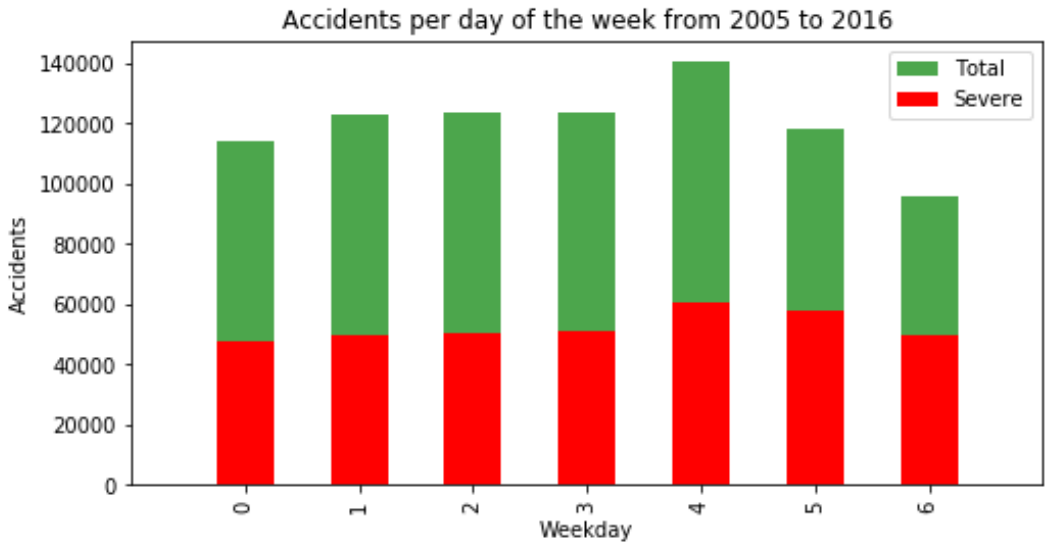
3. Methodology

3.1 Exploratory data Analysis

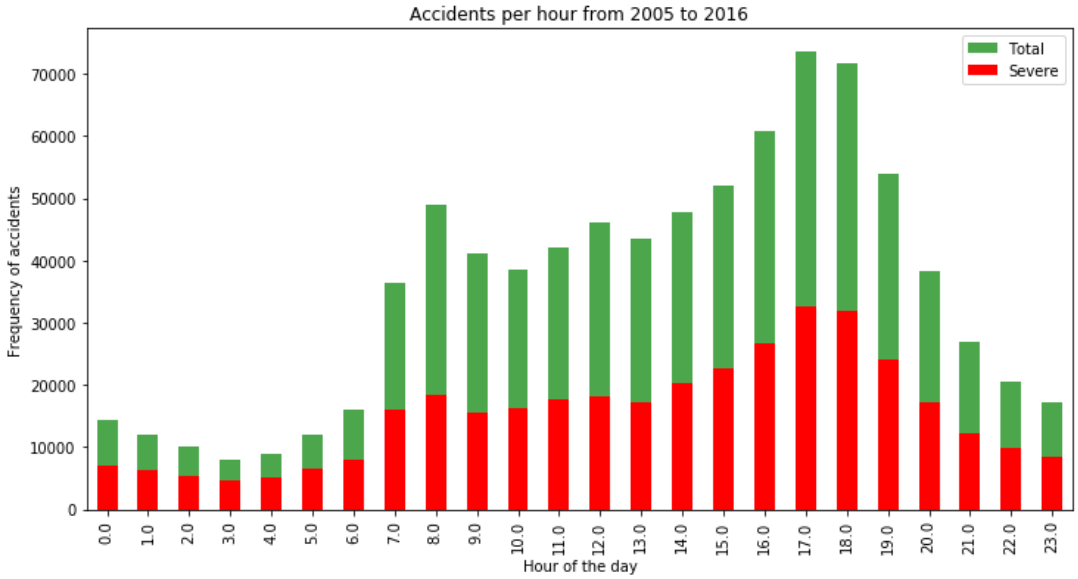
The accident rate (per year) decrease to stable after 2013.



The accident rate (per year) will reach to PEAK on Friday.



The accident rate (per year) will reach to PEAK at 17:00—18:00.



3.2 Model development

These algorithms provided a supervised learning approach predicting with certain accuracy and computational time. Algorithm in the Model development:

Random Forest, 10 decision trees

Logistic Regression, C=0.001

K-Nearest Neighbor, K=16

Support Vector Machine, 7500 sample data

4. Results

This table reports the results of the evaluation of each model. According the evaluation, the best performance model is Random Forest.

Algorithm	Jaccard	f1-score	Precision	Recall	Time(s)
Random Forest	0.722	0.72	0.724	0.591	6.588
Logistic Regression	0.661	0.65	0.667	0.456	6.530
KNN	0.664	0.66	0.652	0.506	200.58
SVM	0.659	0.65	0.630	0.528	403.92

5. Discussion

These features had been that the target of this classification problem was simplified to two different classes, low and high severity. there was still significant variance that could not be predicted by the models in this study.

6. Conclusion

The project built and compared 4 different classification models to predict whether an accident will have a high or low severity. These models can have multiple application in real life.