Predicting the Severity of Traffic Collisions in Los Angeles

Jitong Yang

August 30, 2020

Outline

1

PART ONE

INTORDUCTION

Purposes, Background, Brief Introduction of Models

Background

- There are increasing number of people start to buy a vehicle for their daily use
- However, the likelihood of traffic accident also becomes increasingly higher along with more vehicles appear on the road.
- Collisions result in unnecessary injuries to people and cause unexpected delays.
- This is especially the case in metropolises.

Purpose

- ❖ Find appropriate determinants of traffic collisions severity in Los Angeles
- ❖ Use features to predict severity under different machine learning models
- ❖ Select a model that provides the most accurate predictions

Brief Introduction of Models

1. K-Nearest Neighbors
2. Decision Tree
3. Logistic Regression
4. Support Vector Machine
5. Random Forest

2

PART TWO

Data

Properties of Data, Data Resources, Target & Potential Features

Data Information

- Data Source: Kaggle.com

- Observations: around 3.5 million records of traffic accidents in the United States from 2016 to 2020. I choose those in *Los Angeles* only

- Contain 49 attributes, such as severity, latitudes & longitudes, weather conditions, presence of pedestrian crossing, the period of day, etc.

Target & Features

- Target: Severity (integers from 1 to 4)

- Potential Features： Temperature, Humidity, Visibility, wind speed, pressure, weather conditions, road locations, period of day, weekdays

PART THREE

Methodology

Exploratory Data Analysis, Machine Learning Algorithms

Summary Statistics

	Severity	Temperature (F)	Humidity (%)	Pressure (in)	Visibility (mi)	Wind_Speed (mph)
mean	2.372	66.691	61.244	29.900	9.115	4.868
std	0.502	9.003	20.440	0.158	1.964	3.205
min	1	37.9	3	28.83	0	0
1st quartile	2	60.1	50	29.81	10	3.5
median	2	66	64	29.91	10	4.868
3rd quartile	3	72	77	30	10	5.8
max	4	106	100	30.5	10	36.8
mode	2	64	78	29.91	10	4.868
skewness	0.737	0.412	-0.630	-1.042	-2.418	0.908
kurtosis	-0.932	0.156	-0.198	4.391	5.120	2.949
count	79169	79169	79169	79169	79169	79169

Group Statistics: Weather Condition

Weather Condition	Severity	Weather Condition	Severity	Weather Condition	Severity
Blowing Dust	2	Haze	2.358	Light Thunderstorms and Rain	2
Clear	2.456	Heavy Rain	2.239	Mist	2.5
Cloudy	2.263	Heavy T-Storm	2	Mostly Cloudy	2.363
Drizzle	2	Light Drizzle	2.25	Mostly Cloudy / Windy	2
Fair	2.201	Light Rain	2.315	Overcast	2.502
Fair / Windy	2.273	Light Rain / Windy	2.5	Partly Cloudy	2.314
Fog	2.412	Light Rain with Thunder	3	Partly Cloudy / Windy	2
Patches of Fog	2.5	Scattered Clouds	2.5	Thunder	2
Rain	2.327	Shallow Fog	2.333	Thunderstorm	2.5
Rain / Windy	2	Smoke	2.53		

❖ Weather conditions could be **useful** to predict severity

Group Statistics: Period of Day

Period	Severity
Day	2.355
Night	2.403

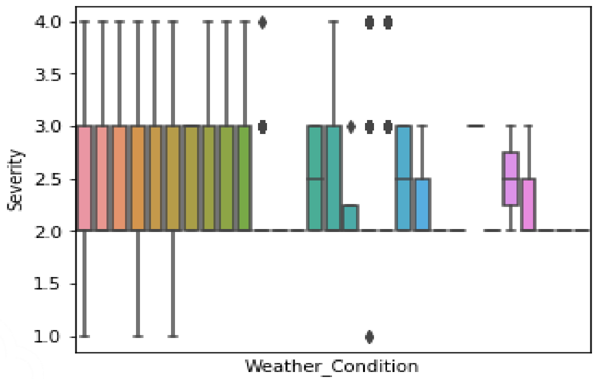
- Period of day could **not** be useful to predict severity

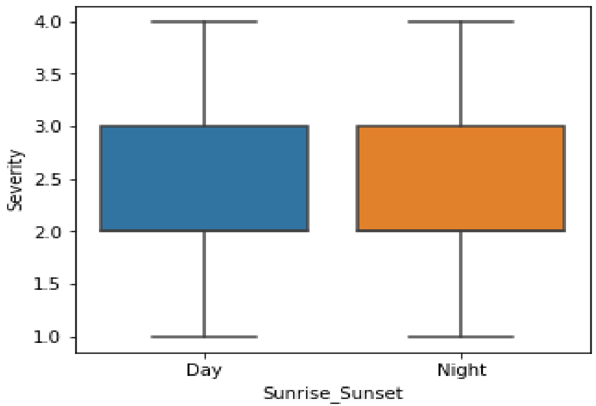
Group Statistics: Weekday

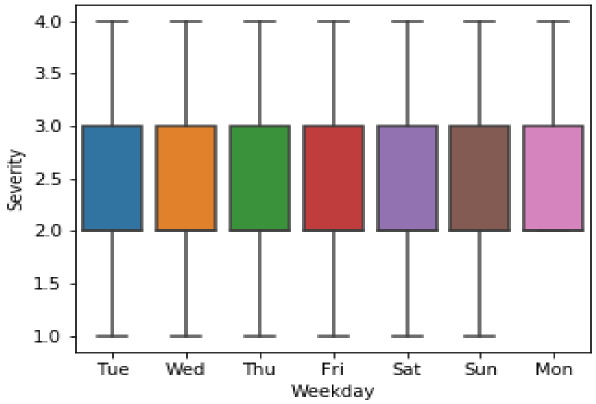
Weekday	Severity
Sun	2.459
Mon	2.352
Tue	2.332
Wed	2.348
Thu	2.363
Fri	2.351
Sat	2.465

- Weekday could **not** be useful to predict severity

Visualization: Categorical Variable



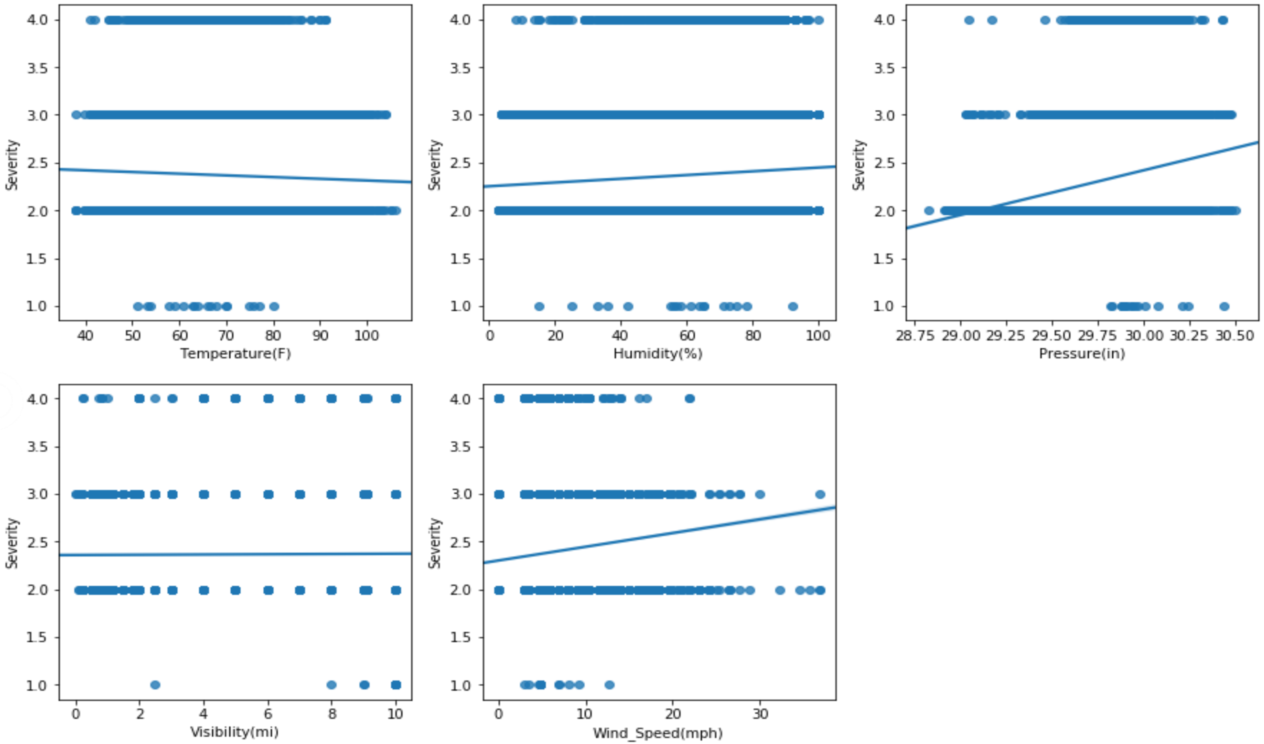




Correlation Analysis

	Severity	Temperature (F)	Humidity (%)	Pressure (in)	Visibility (mi)	Wind_Speed (mph)
Severity	1*					
Temperature (F)	-0.032*	1				
Humidity (%)	0.08*	-0.477	1			
Pressure (in)	0.147*	-0.245	-0.028	1		
Visibility (mi)	0.005*	0.177	-0.361	0.007	1	
Wind_Speed (mph)	0.092*	0.11	-0.036	0.106	0.085	1

Correlation Analysis



❖ Visibility could **not** be useful to predict severity

Methods

- ❖ Classification algorithms: K-Nearest Neighborhoods (KNN), Decision Tree, Logistic Regression, Support Vector Machine (SVM), Random Forest

- ❖ I split the dataset into training set (75% observations) and test set (25%).

- ❖ Training set: train the model under each algorithm

Methods

- Test set: predict traffic accident severity and assess the quality and accuracy of each model.

- Find the optimal number of nearest neighbors for KNN

- Compute Jaccard index and F1-score and visualize distributions of predicted values and actual values

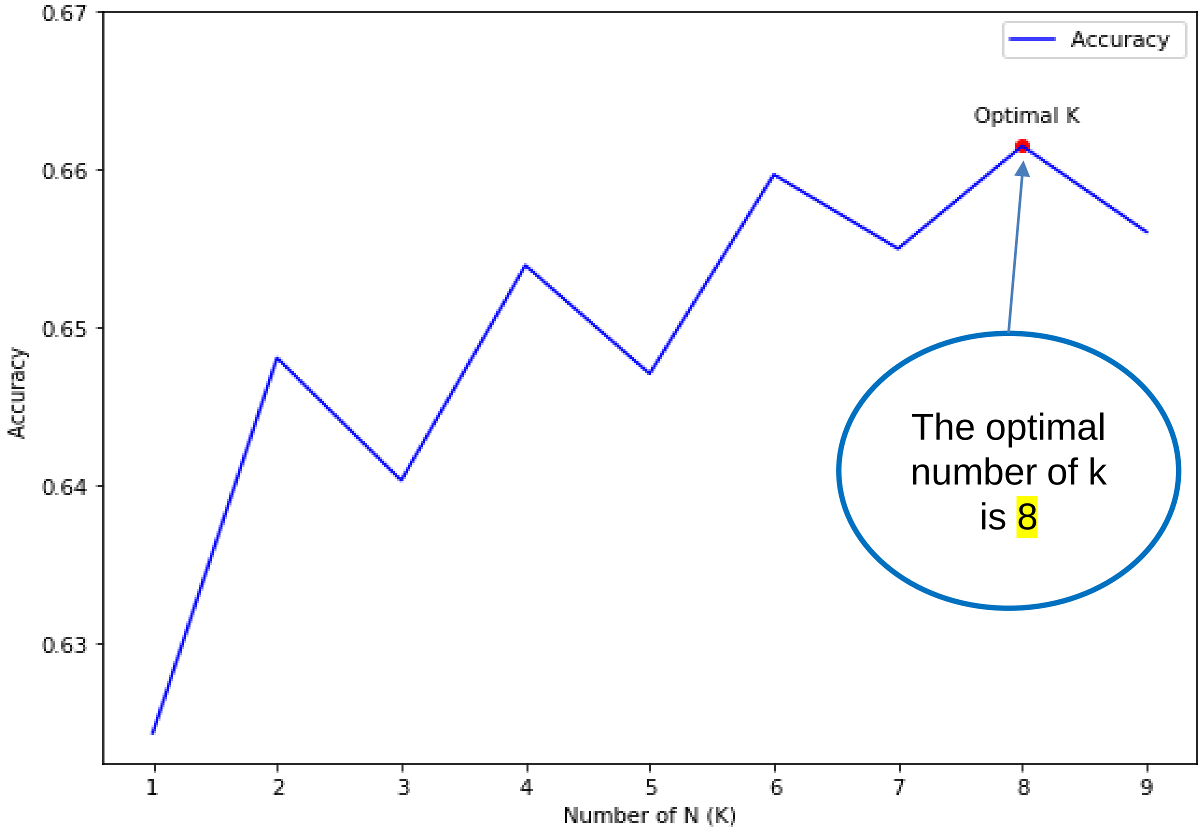
4

PART Four

Results and Discussion

Empirical Results, Model Evaluation, Recommendations

Optimal Number of Nearest Neighbors

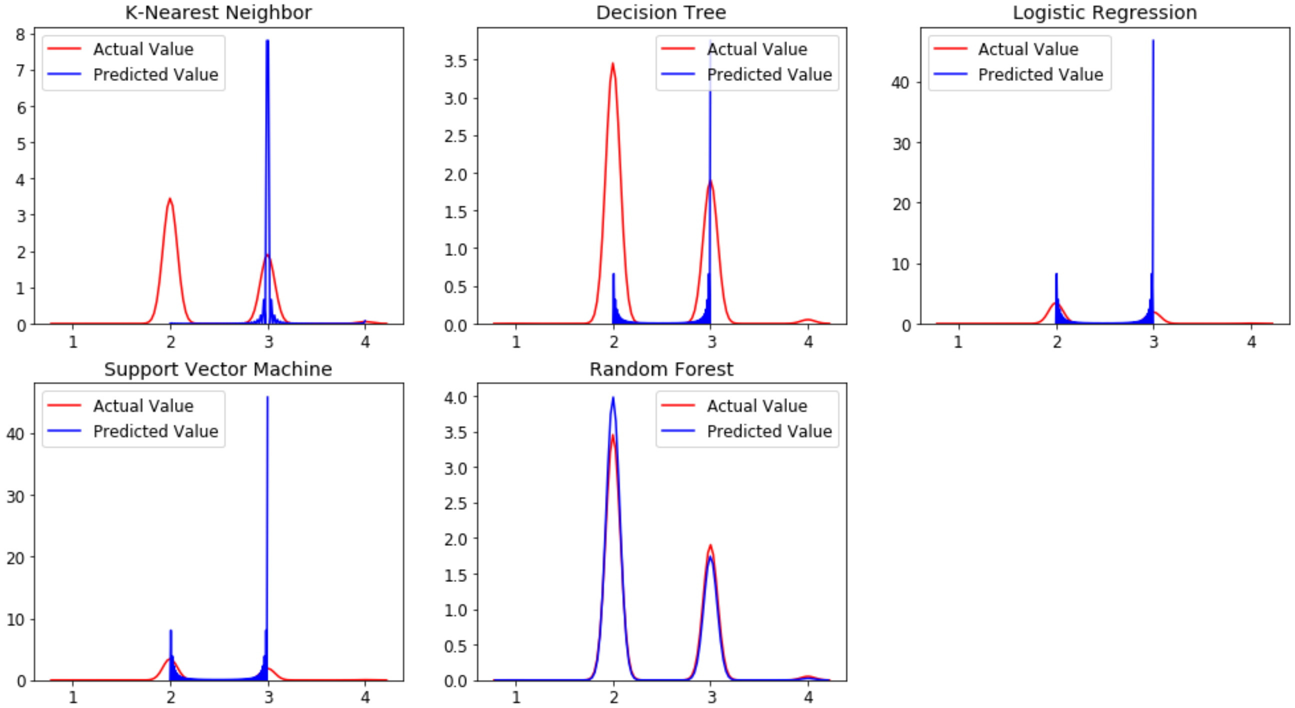


Accuracy Scores

Algorithm	Jaccard index	F1-score
KNN	0.6615	0.6338
Decision Tree	0.6457	0.5227
Logistic Regression	0.658	0.6316
SVM	0.6621*	0.6348
Random Forest	0.6605	0.6525*

❖ SVM and Random Forest are **most** accurate; Decision Tree is **least** accurate

Distribution Plot



❖ Random Forest has the best fit

Recommendations

- Use weather conditions to predict severity of traffic accident instead of weekdays and period of day
- Visibility is not an appropriate predictor
- Use Random Forest for prediction

5

PART Five

Conclusion

Conclusions of the project

Conclusion

- Find determinants of severity and train classification models

- Weather conditions can better predict severity of traffic accident than period of day and weekdays

- Accuracy scores indicate SVM and Random Forest provide most accurate result

- Distribution plot indicates Random Forest can best fit all observations

THANKS FOR YOUR WATCHING