Capstone

Predict Car Accident Severity in Seattle.

Stepan Rakhimov, September 17 2020

The Plan

What do we do?

1. Introduction
2. Data acquisition and cleaning
3. Exploratory Data Analysis
4. Predictive Modeling
5. Conclusions
6. Future directions

Introduction

Why?

- Car accidents is still a problem
- We have navigation and helpers already
- Roads safety varies
- Take historical accidents severity into account
- More advanced navigation

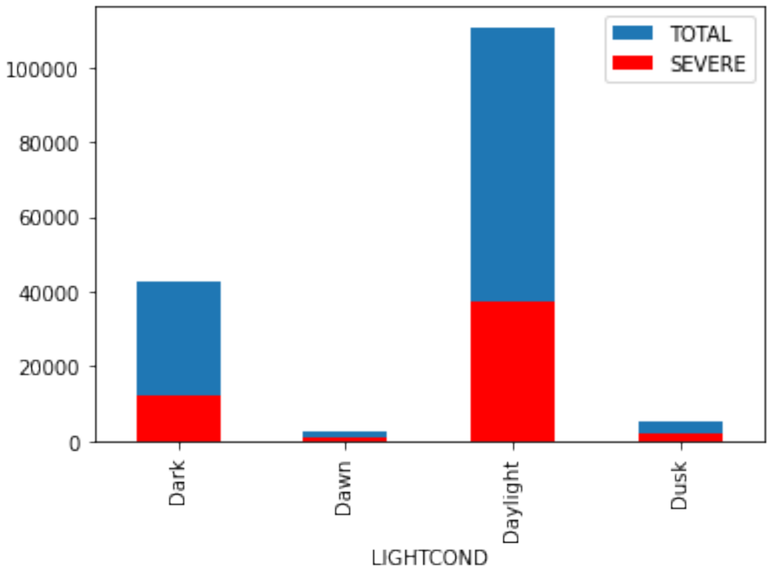
Data acquisition and cleaning

Seattle car accidents 2004-2011

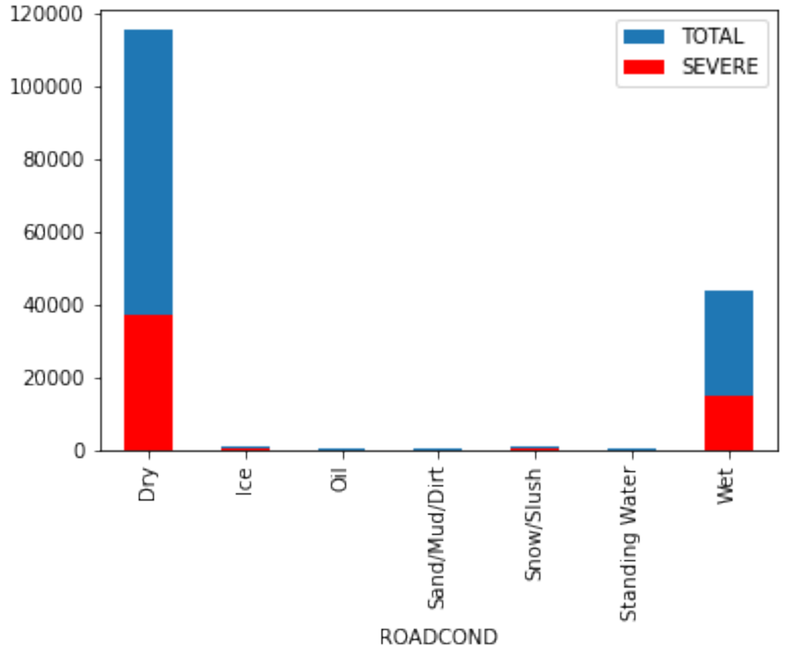
- Accidents dataset from Traffic Management of the city of Seattle, CSV formatted file and PDF formatted metadata
- Severity, weather, road, geo point, number of vehicles and persons involved etc
- 39 features and 194673 rows
- Dropped irrelevant data, null, “Other” and “Unknown”
- One-hot encoded, severity to boolean, 28 features in final set

Explore our data

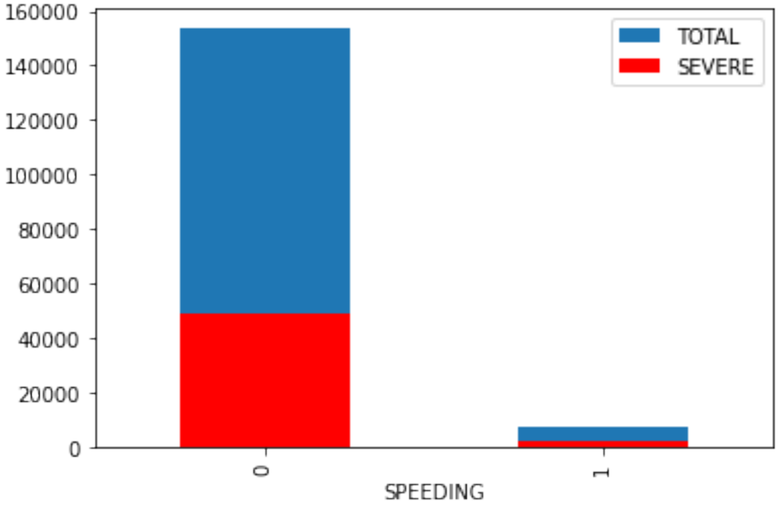
Can we spot relations?



Less severe in darkness



Dry is not safe

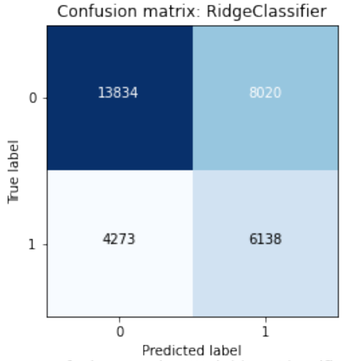


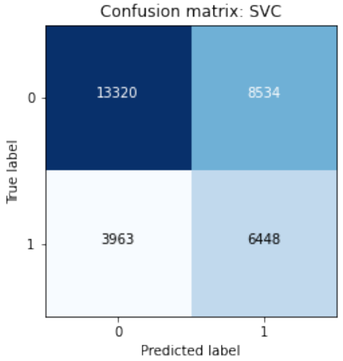
Speeding is not always the cause

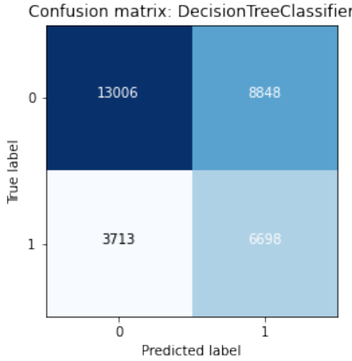
Models

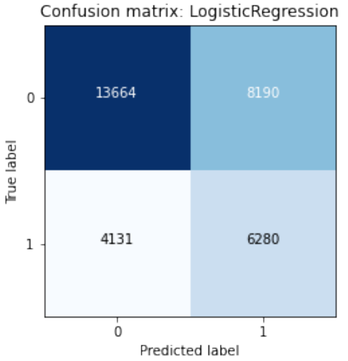
Is there a best algorithm?

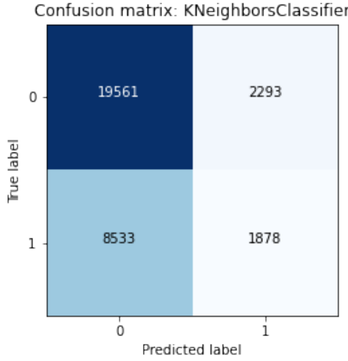
- Decision Tree
- K-Means
- SVM
- Logistic Regression
- Gradient Boost
- Random Forest
- Ridge Regression

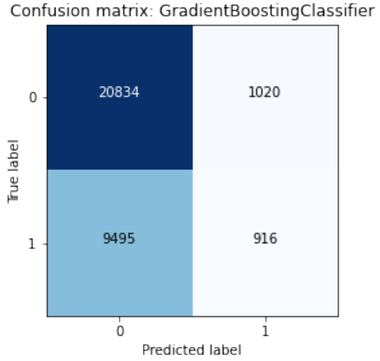












	accuracy	f1	logloss
DecisionTreeClassifier	0.610693	0.623287	0.704222
KNeighborsClassifier	0.664466	0.613634	None
LogisticRegression	0.618131	0.629733	0.662557
SVC	0.612676	0.624919	None
GradientBoostingClassifier	0.674105	0.588722	0.627829
RandomForestClassifier	0.614567	0.626017	0.667945
RidgeClassifier	0.618999	0.630189	None

Results

Depends on application

Do we need probability?

- no - RidgeClassifier with alternative of SVC
- yes - DecisionTreeClassifier with alternative of LogisticRegression and RandomForestClassifier

Future

Can we improve?

- Sure we can
- More experience
- More models
- More feature engineering
- More data
- Other sets of data