

CAR CRASH ANALYSIS

INSIGHTS & ISSUES

PIETER-JAN DROUILLON

Data Science Leuven Meetup - November 14, 2019

SETTING THE SCENE

BACKGROUND

 Computer science

 MaNaMa AI

 Economics

 MaNaMa Statistics

PROFESSIONAL



Academic background

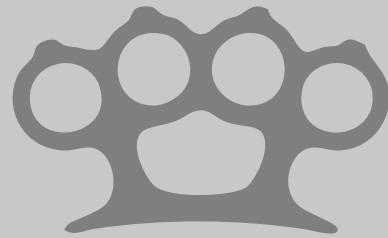


Short-lived career in Supply Chain




Full-stack developer @ Design Is Dead

DESIGN IS DEAD



PG BIG DATA & ANALYTICS IN BUSINESS AND MANAGEMENT

8 Fridays


1 Project


THE PROJECT

- Choose a topic
- Write a proposal
- Execute

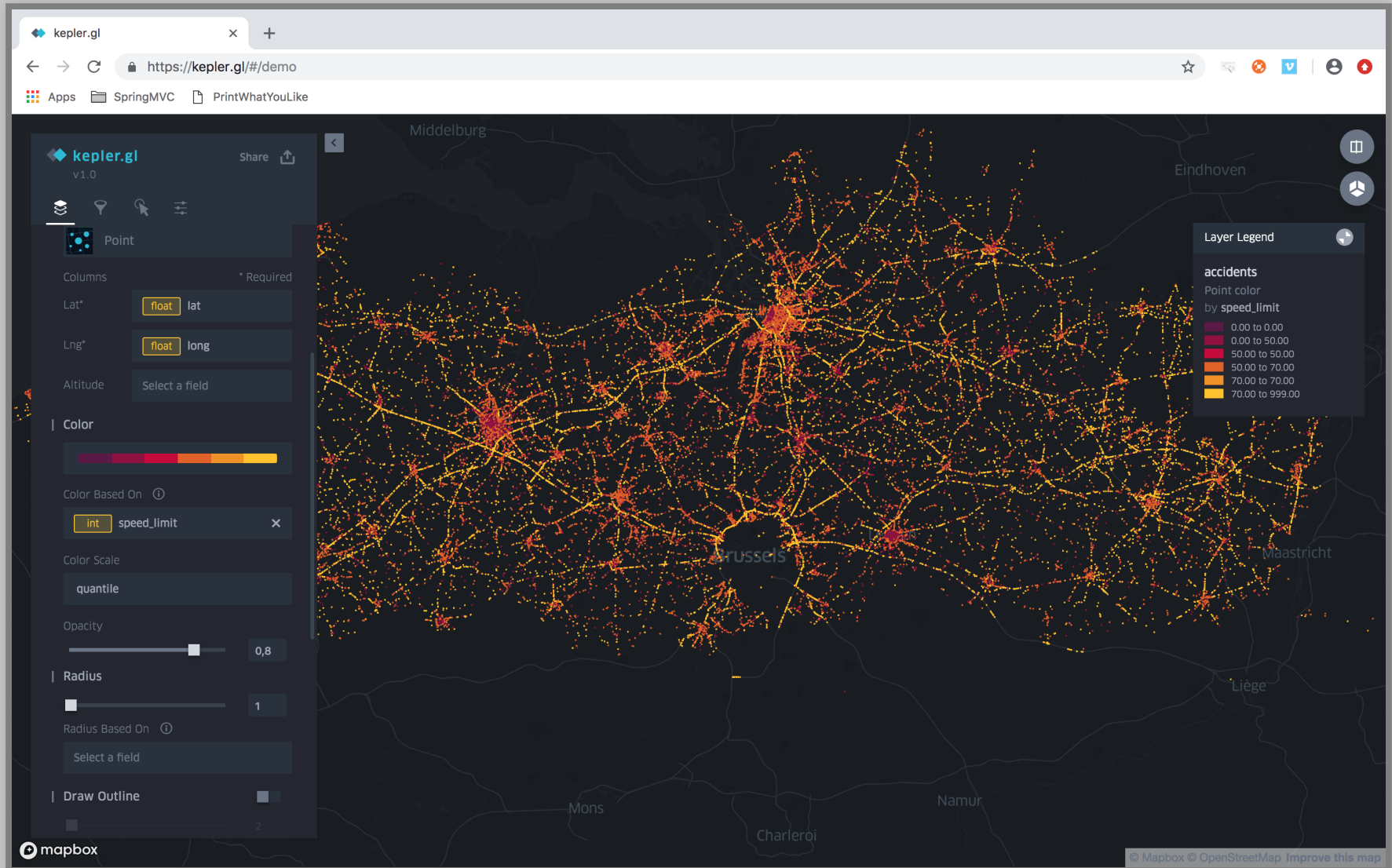
DATASET

- 52K car crashes
- Federal police & Informatie Vlaanderen
- What factors influence crash severity?

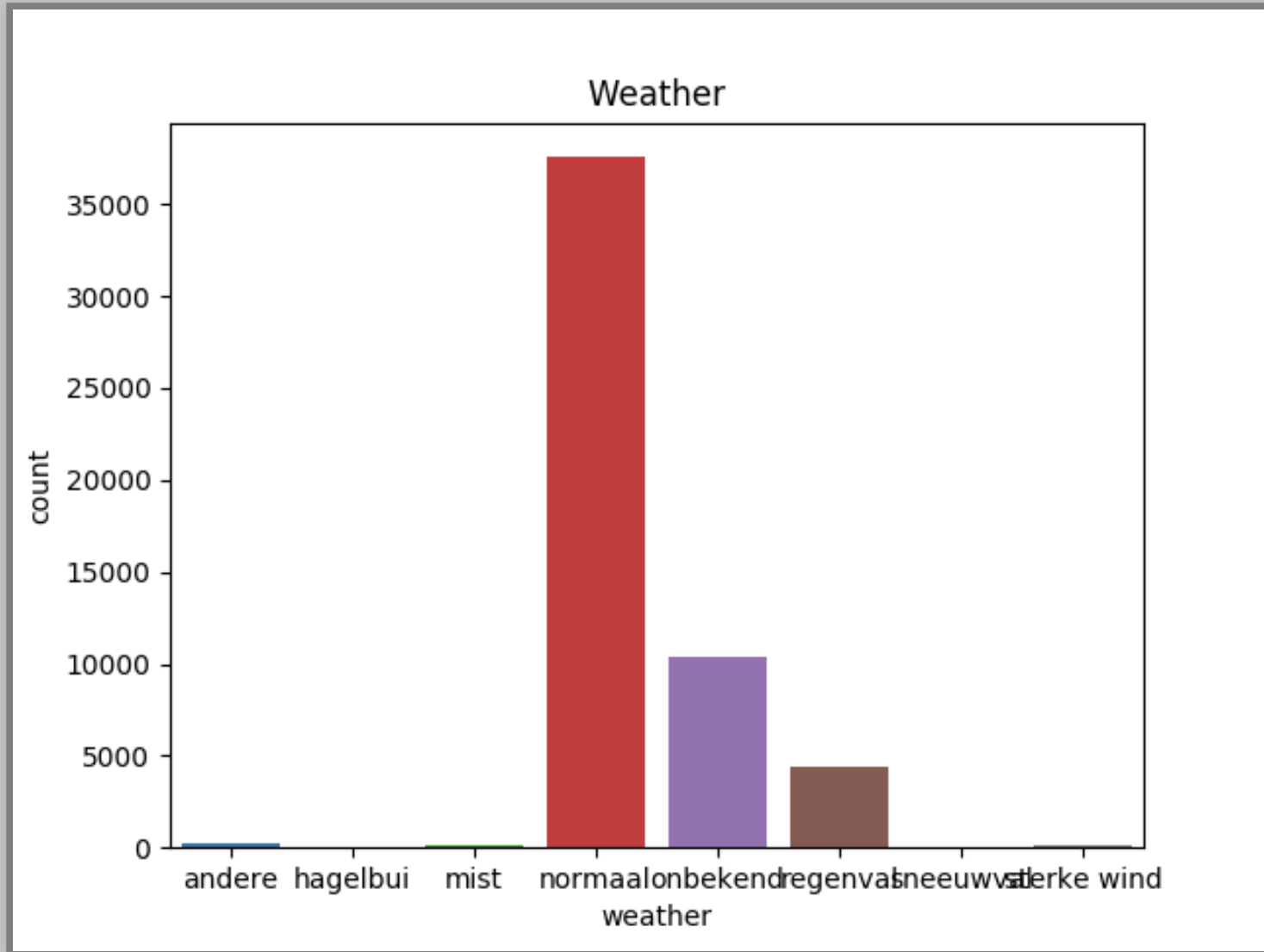
EXPLORATION

- Pixel coordinates
- Weather condition
- Road state
- Speed limit
- Timestamp
- Severity

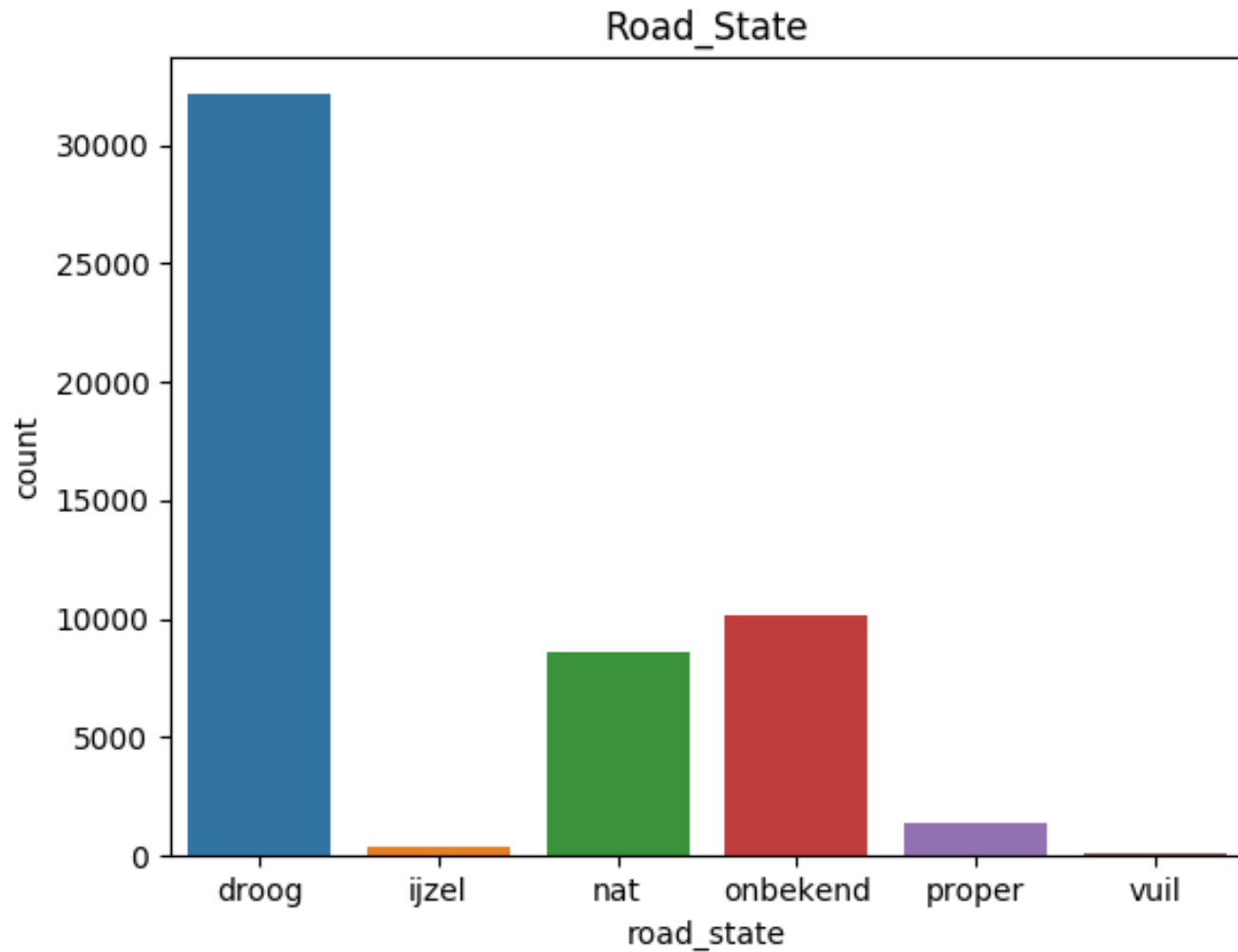
PIXEL COORDINATES



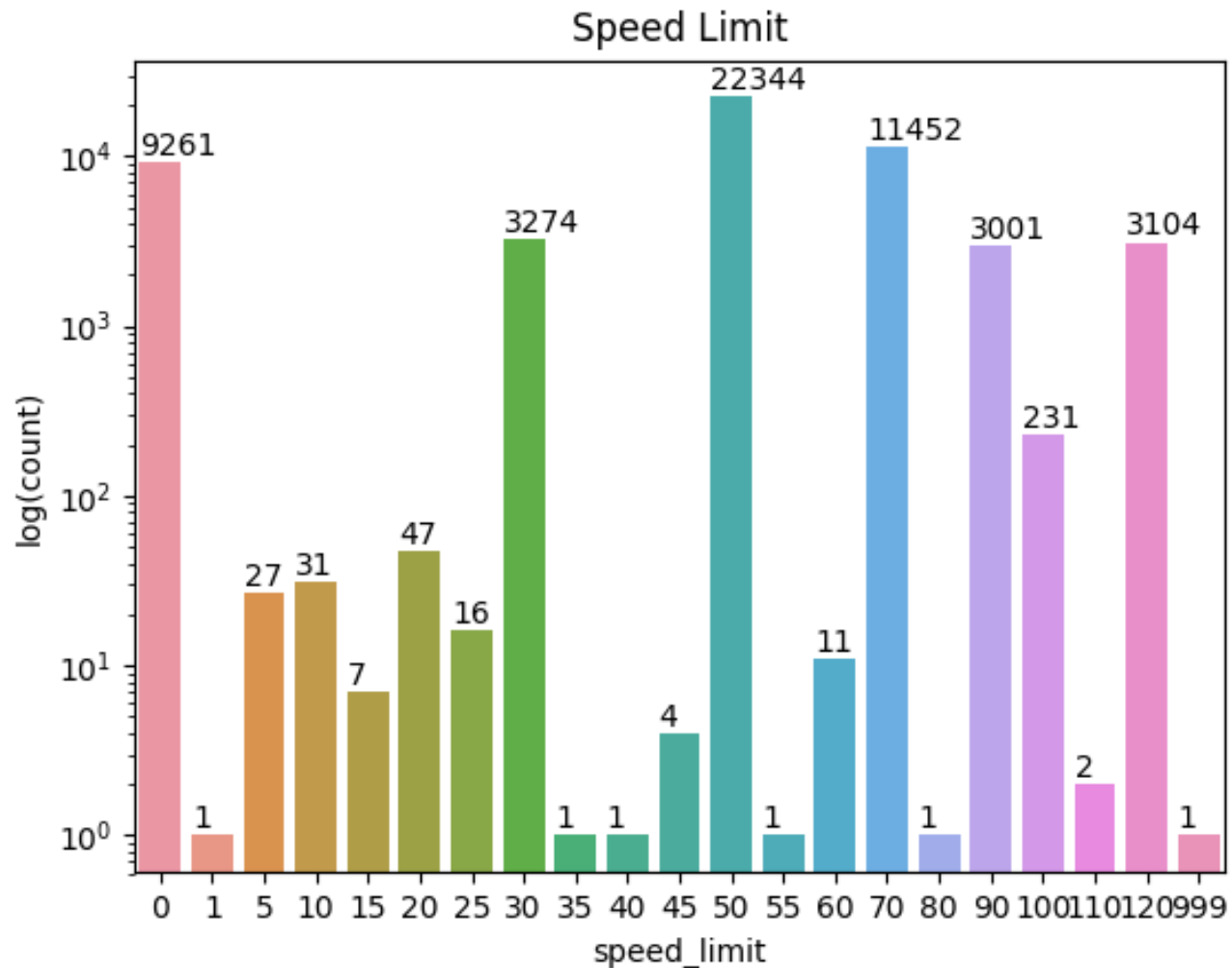
WEATHER CONDITIONS



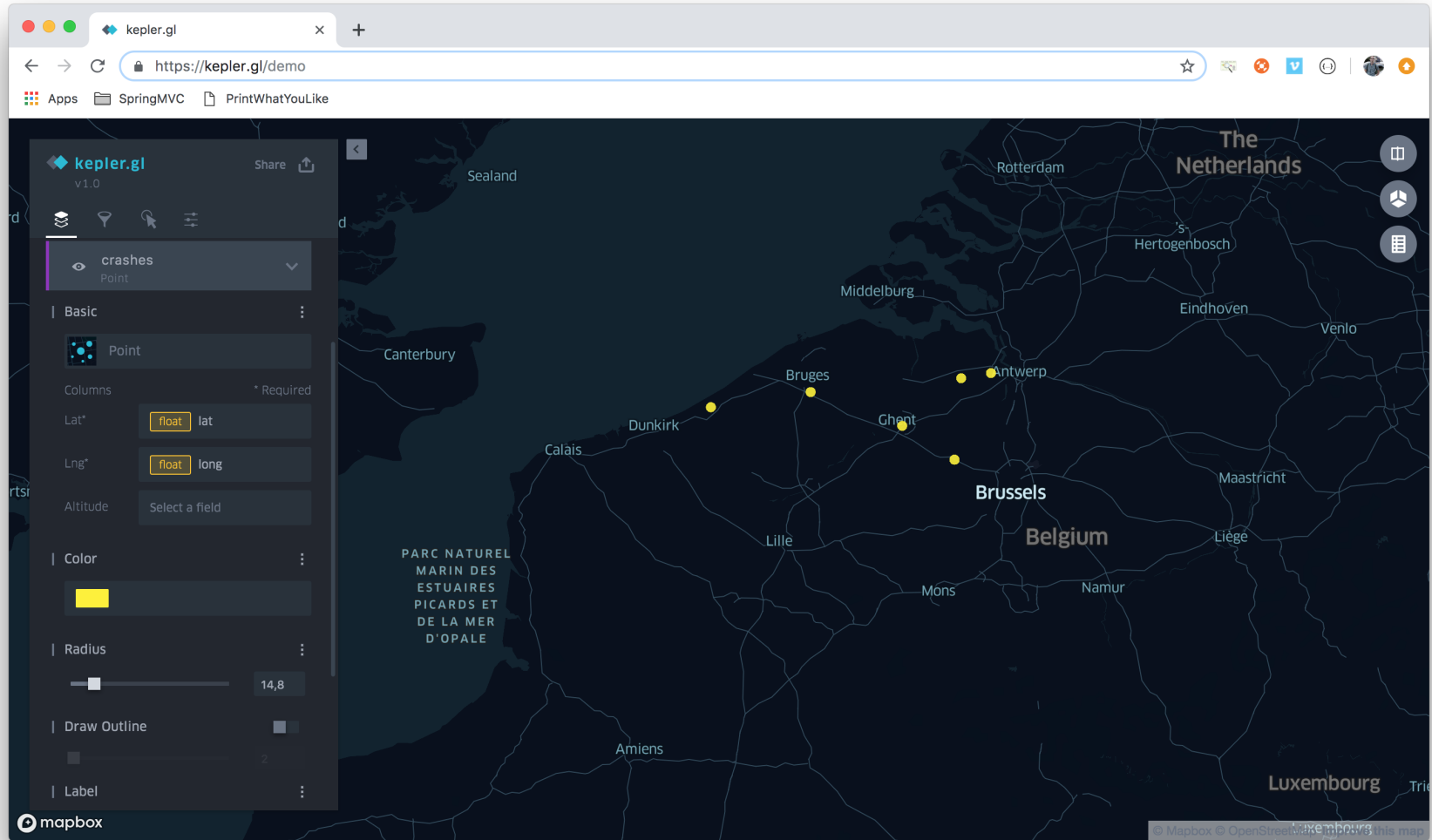
ROADSTATE



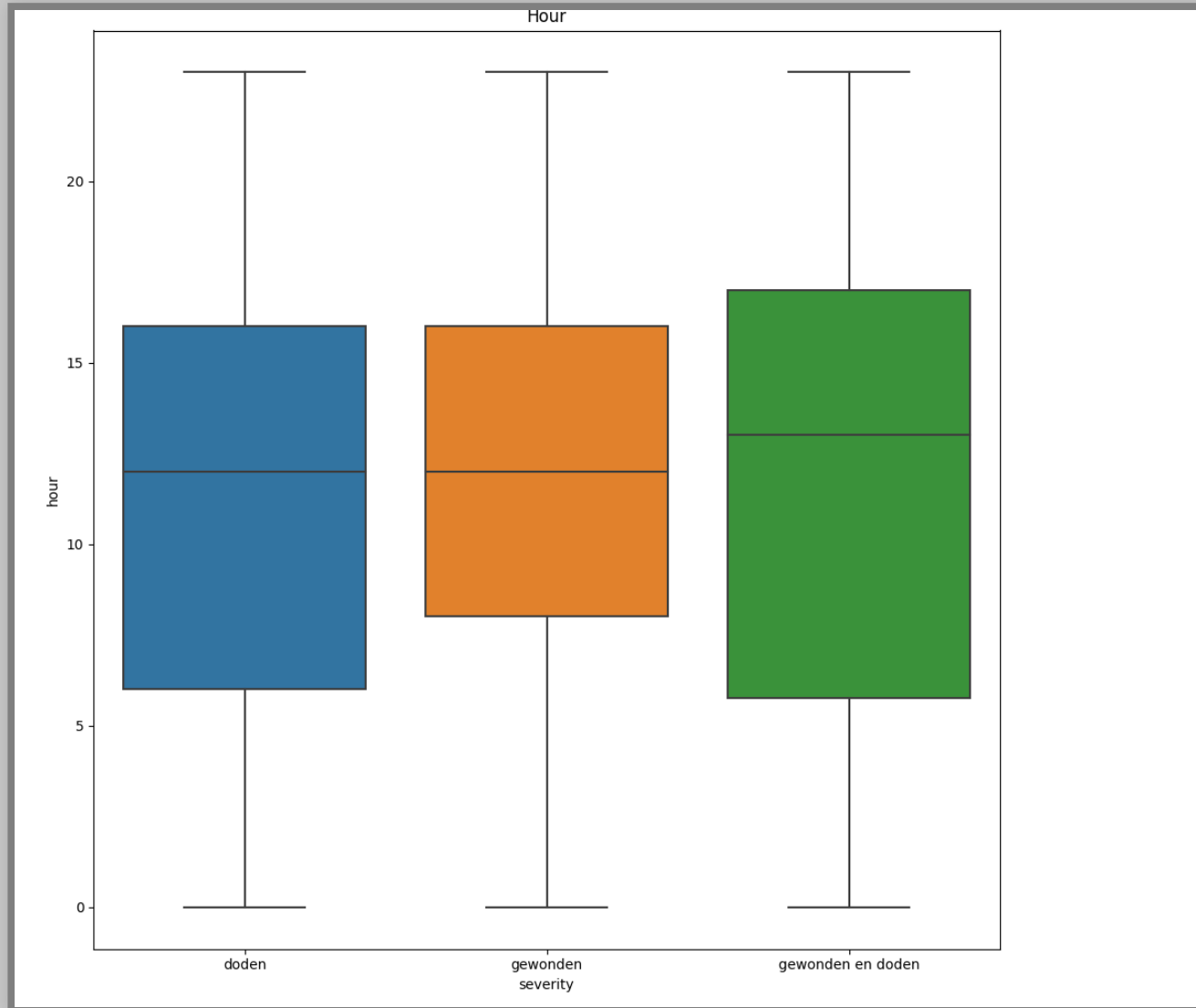
SPEED LIMIT



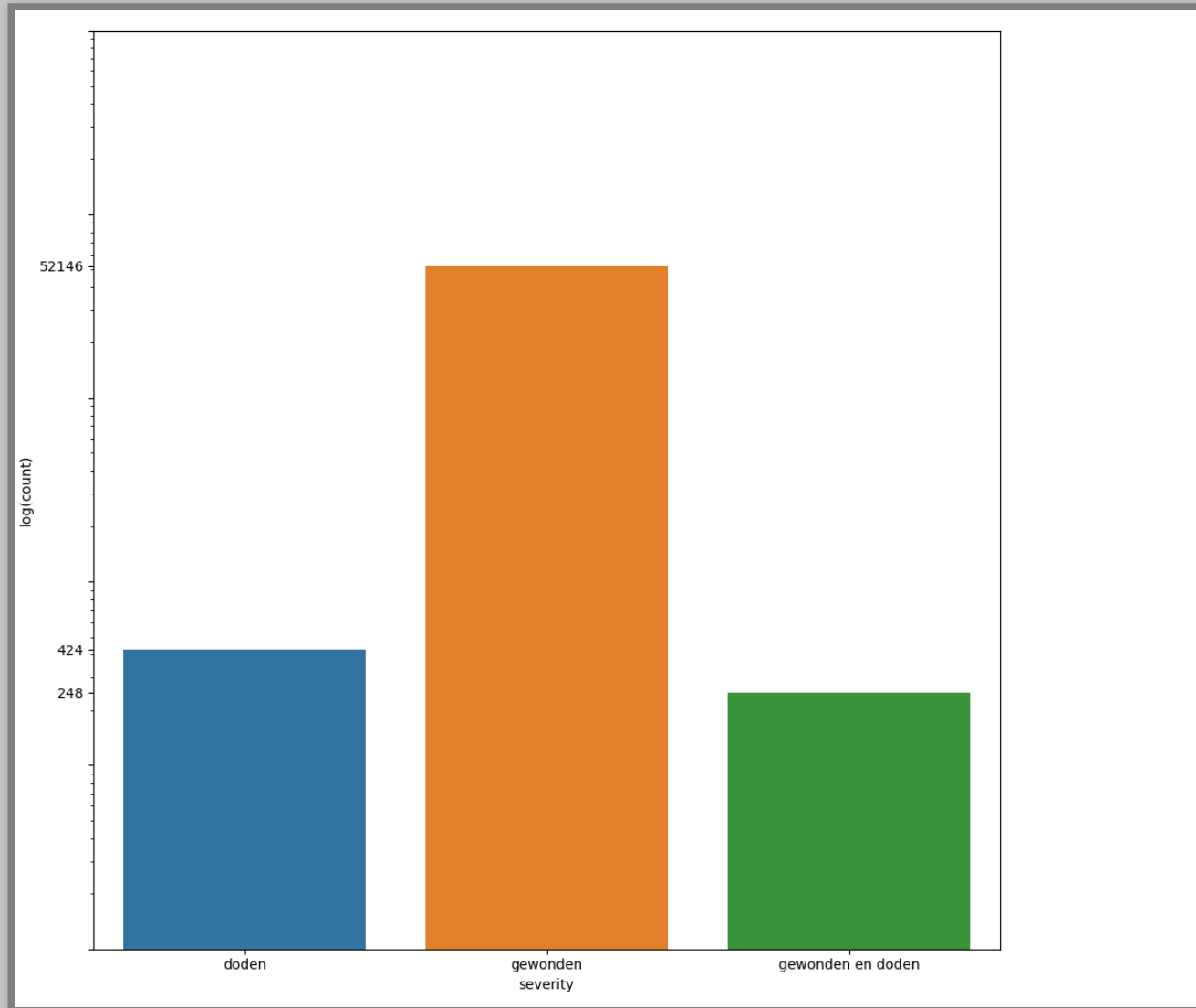
TIMESTAMP



TIMESTAMP / HOUR



SEVERITY



FEATURIZING

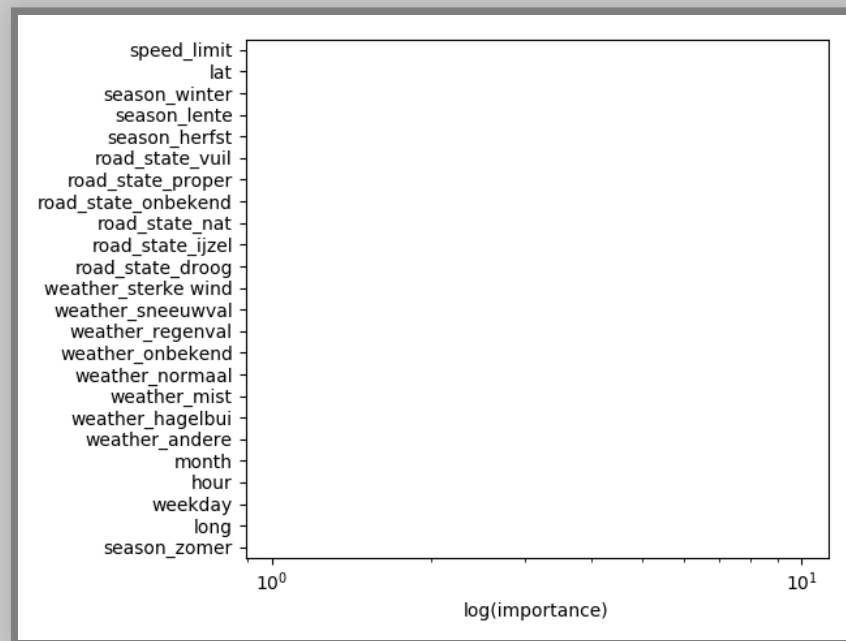
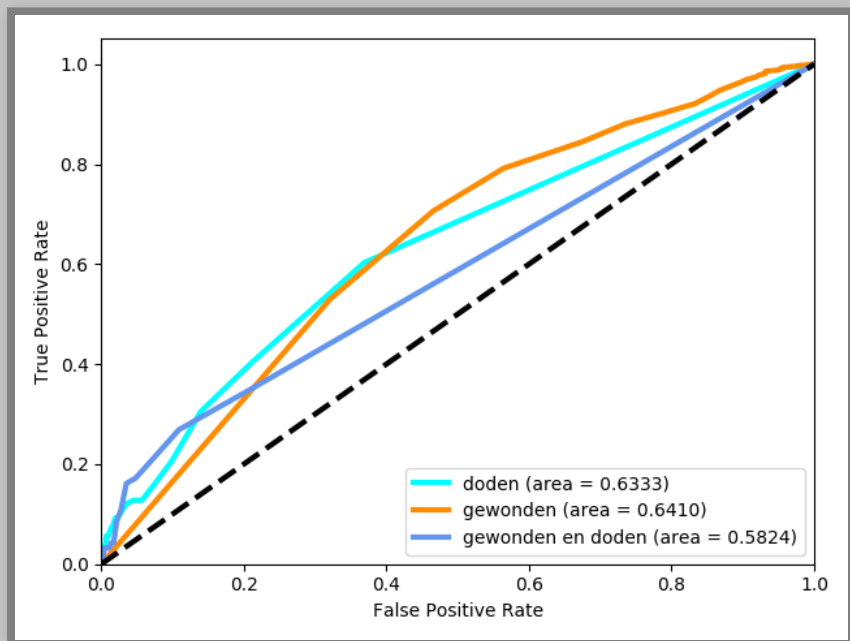
- Timestamp
- Dummies

RANDOM FOREST

0. Init
1. Train RF
2. Test RF
3. Construct ROC
4. Compute feature importance
5. Construct PDPs

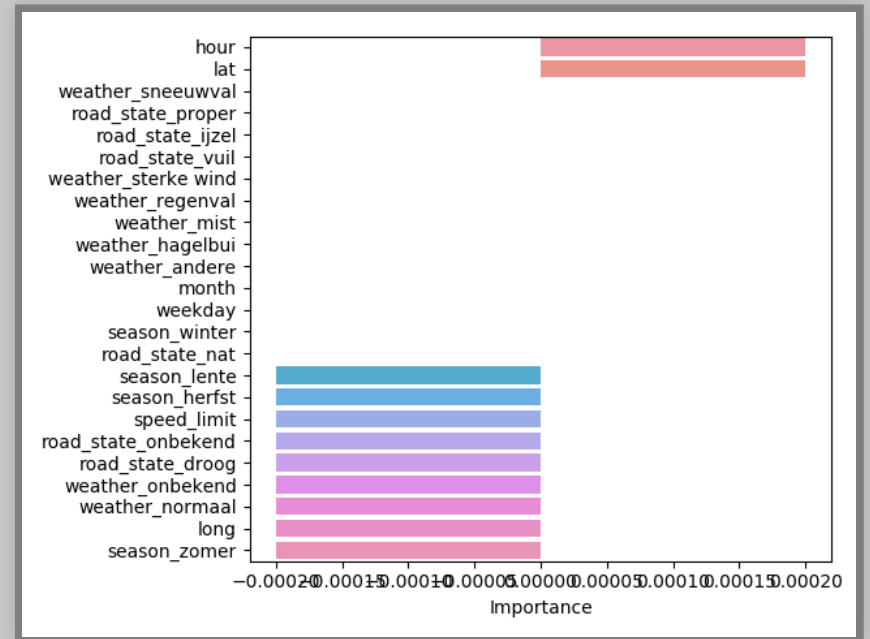
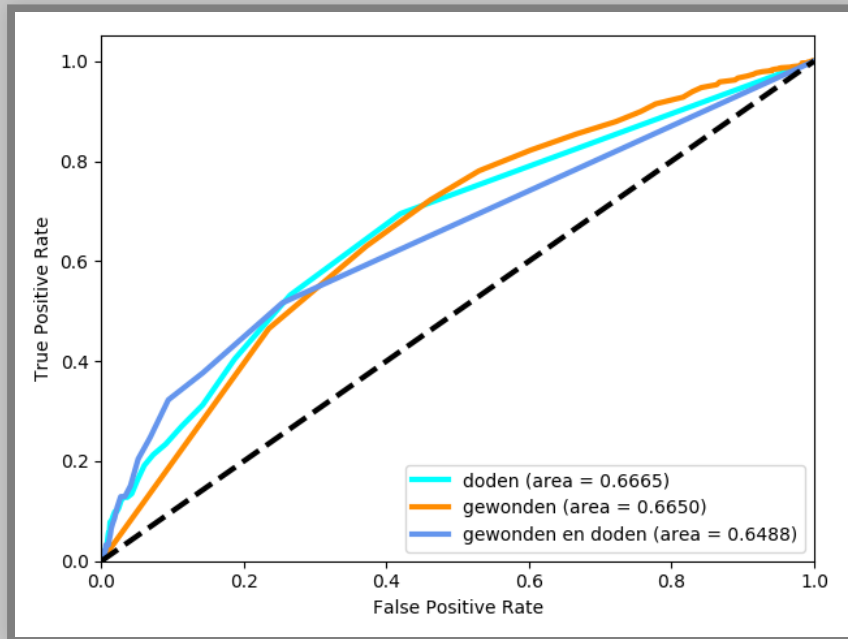
AS-IS

RANDOM FOREST



OVERSAMPLING

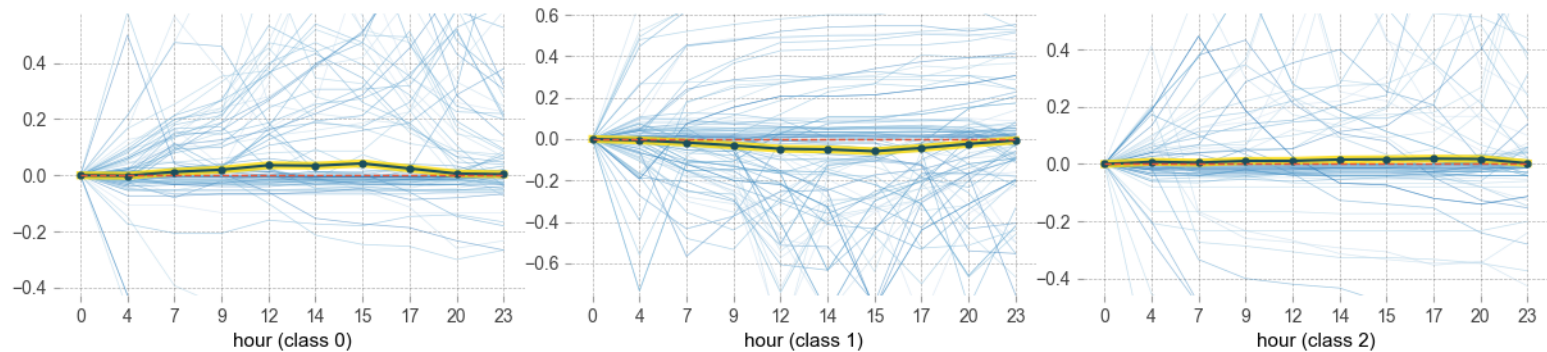
RANDOM FOREST



OVERSAMPLING

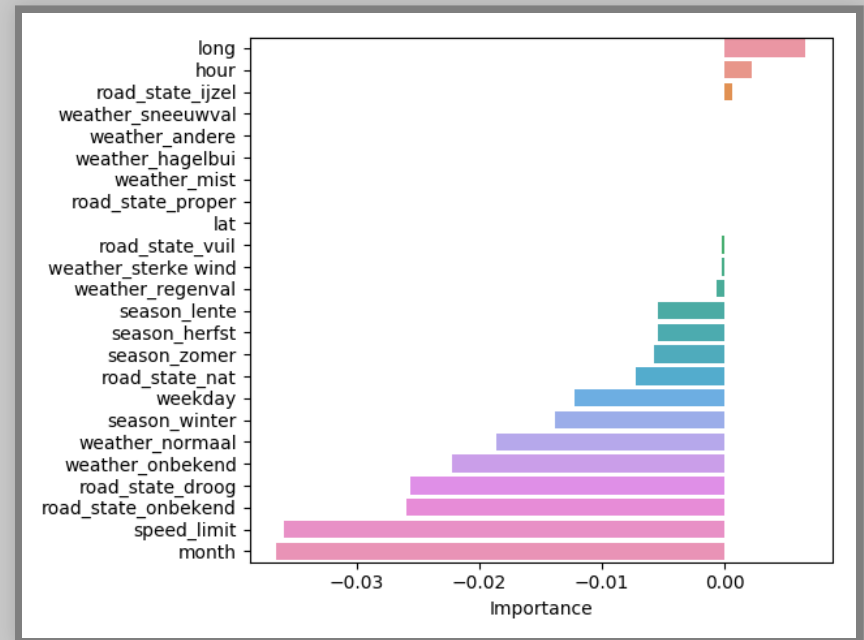
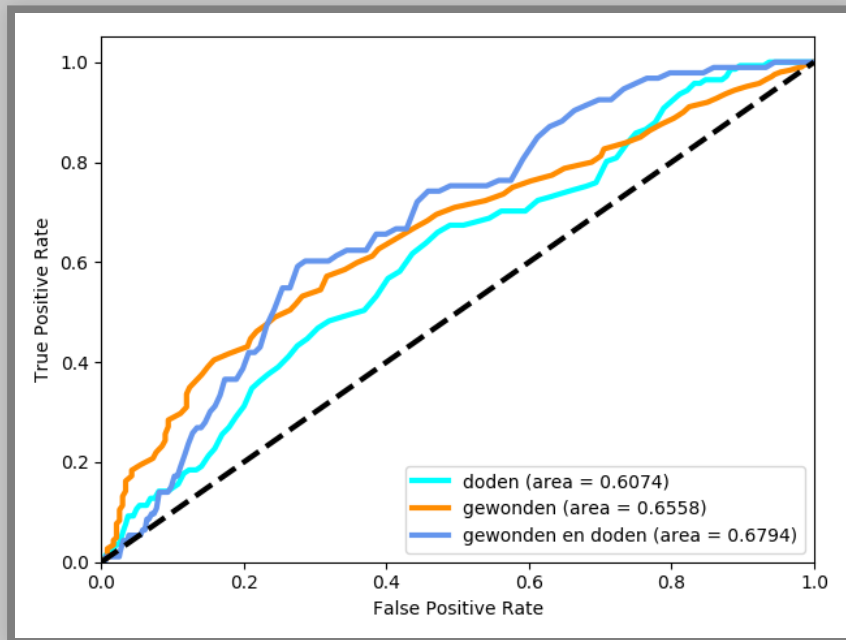
PDP for feature "hour"

Number of unique grid points: 10



UNDERSAMPLING

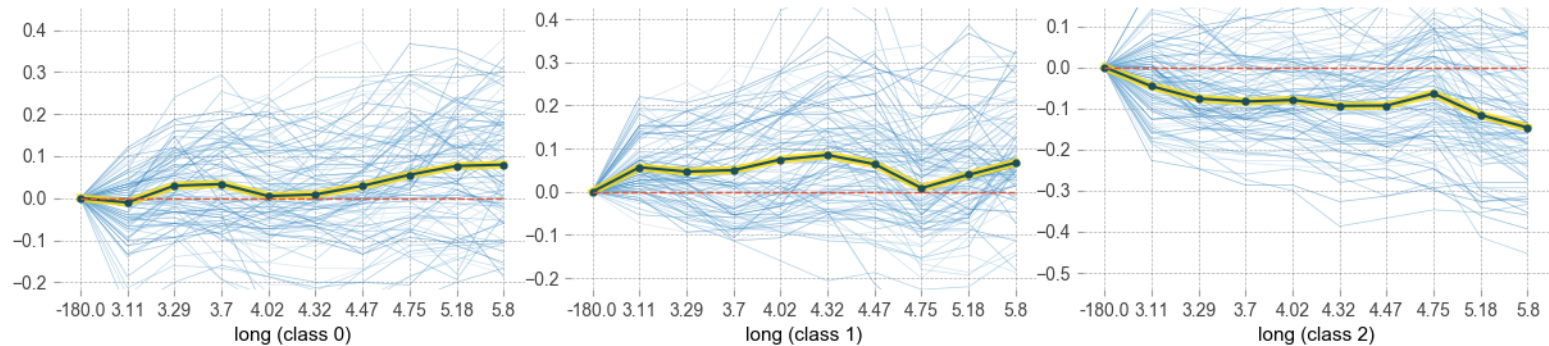
RANDOM FOREST



UNDERSAMPLING

PDP for feature "long"

Number of unique grid points: 10



CONCLUSION

IMPORTANT FACTORS

- Location
- Time
- Weather

PREPARATION

80₂₀

IMPROVEMENTS

- Daylight
- PDP with 2 variables
- Data quality?
- Overlap target values

Q & A

PIETER-JAN DROUILLON | WWW.PIETERJD.BE

Data Science Leuven Meetup - November 14, 2019