Street accidents UK 2018

PREDICTING SEVERITY LEVEL WITH MACHINE LEARNING

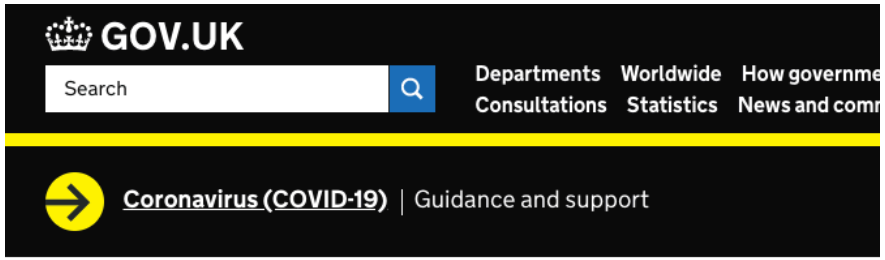
Introduction

Consequences:
- Economical
- Social

Trend is increasing (UK)
Budget cuts for safety road interventions

Field of primary intervention?

Data

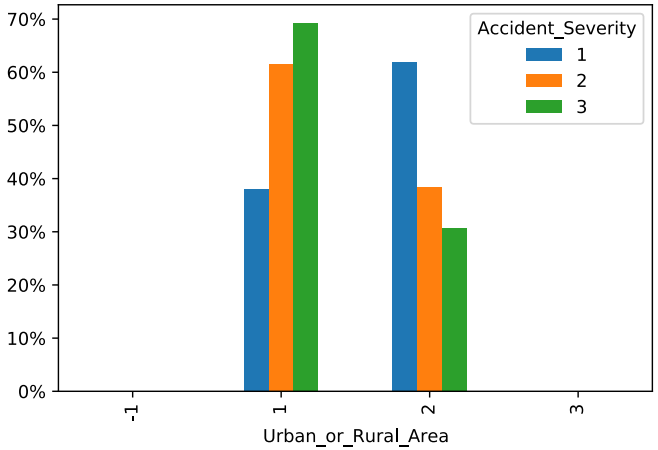


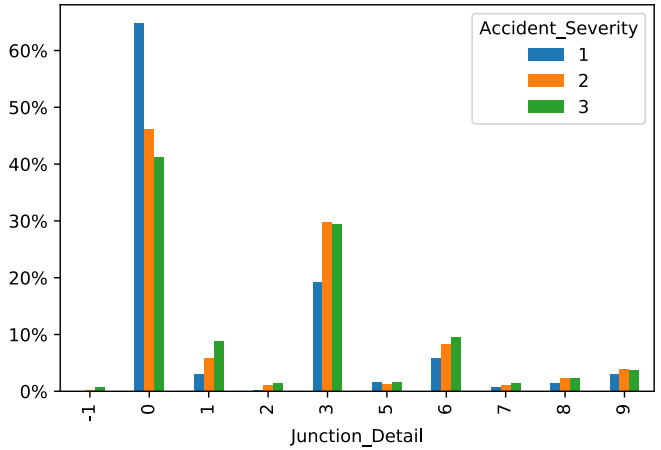
Sources

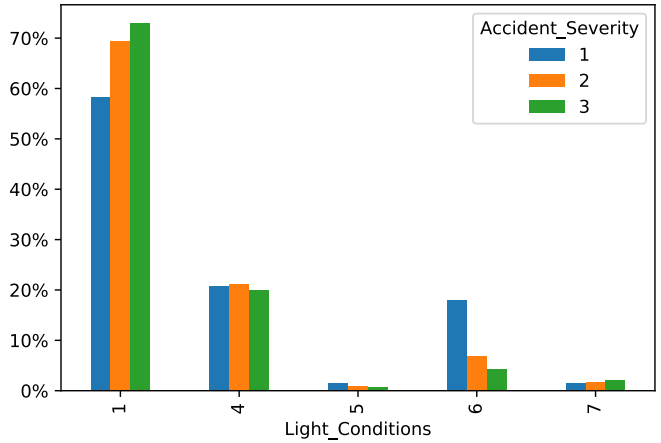
Database name	URL	Type of data	Source
`dftRoadSafetyData_Accidents2018`	http://data.dft.gov.uk.s3.amazonaws.com/road-accidents-safety-data/dftRoadSafetyData_Accidents_2018.csv	general informations regarding the site of the accident, such as "road conditions", "light conditions" and "weather conditions"	Department for Transport (GB)
`dftRoadSafetyData_Vehicles_2018`	http://data.dft.gov.uk.s3.amazonaws.com/road-accidents-safety-data/dftRoadSafetyData_Vehicles_2018.csv	technical info of every vehicle involved in the accident (such as "age", "type of vehicle", "engine capacity"...), but also of the driver (like band of age and IMD decile)	Department for Transport (GB)
`dftRoadSafetyData_Casualties_2018`	http://data.dft.gov.uk.s3.amazonaws.com/road-accidents-safety-data/dftRoadSafetyData_Casualties_2018.csv	data corresponding to individual informations of the casualties such as "sex", "severity", "age", "role at the moment of the accident" (driver, passenger or pedestrian)	Department for Transport (GB)

Methodology – exploratory statistics -

Relevant features (not transformed)

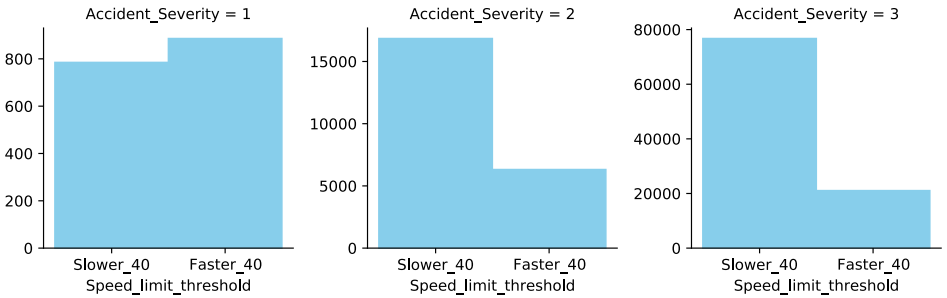


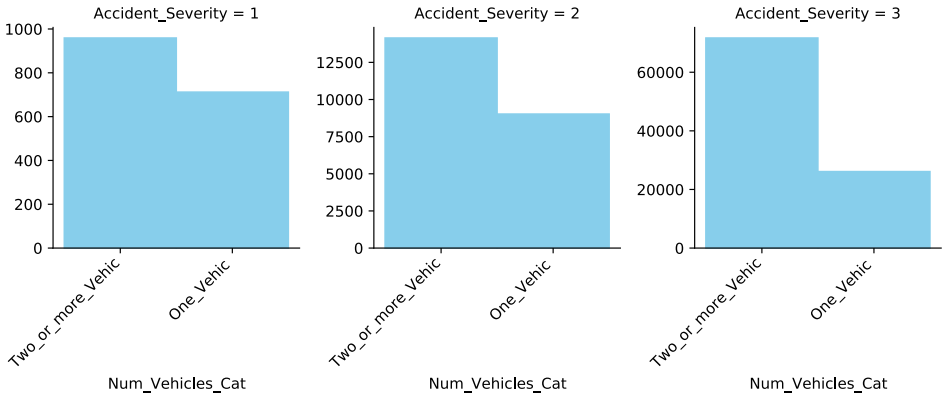


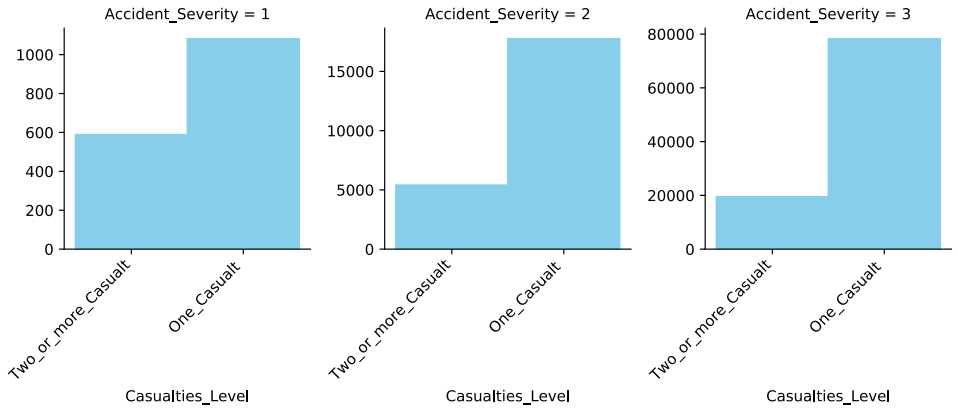


Methodology – exploratory statistics -

Relevant features (transformed)







Example: from count of the number of car per accident to two levels: “one car” and “more than one car”

Methodology – data cleaning and transformation -

Unbalanced data
- SMOTE technique

imblearn.over_sampling.SMOTE

class **imblearn.over_sampling.SMOTE**(*sampling_strategy='auto', random_state=None, k_neighbors=5, m_neighbors='deprecated', out_step='deprecated', kind='deprecated', svm_estimator='deprecated', n_jobs=1, ratio=None*) [source]

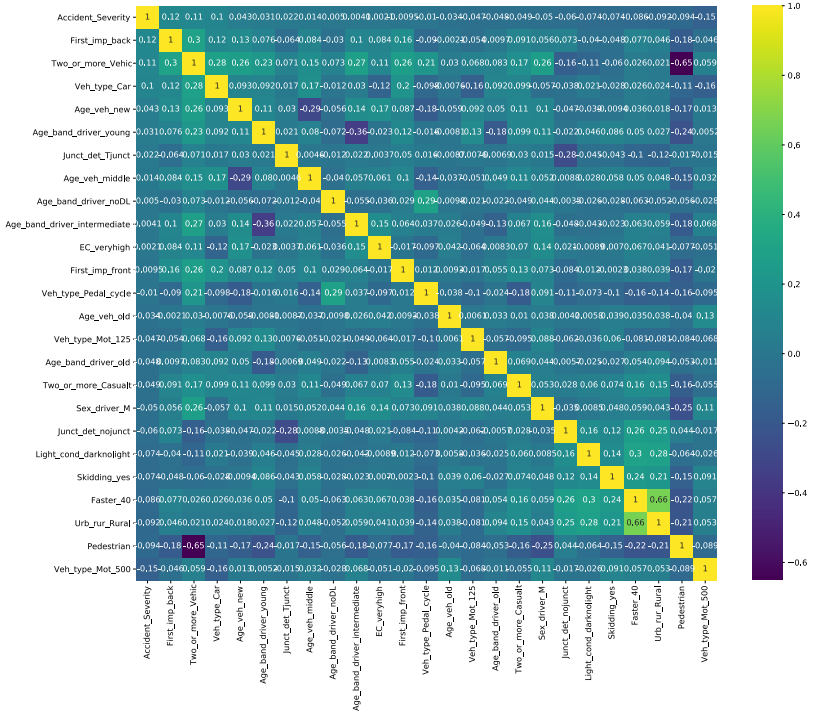
Counts for each category: 1 (fatal), 2 (severe) and 3 (slight)
Before SMOTE

```
Counter({3: 97799, 2: 23165, 1: 1671})
```

After SMOTE

```
Counter({3: 13000, 1: 9000, 2: 9000})
```

Very high correlation between features absent, no multicollinearity problem



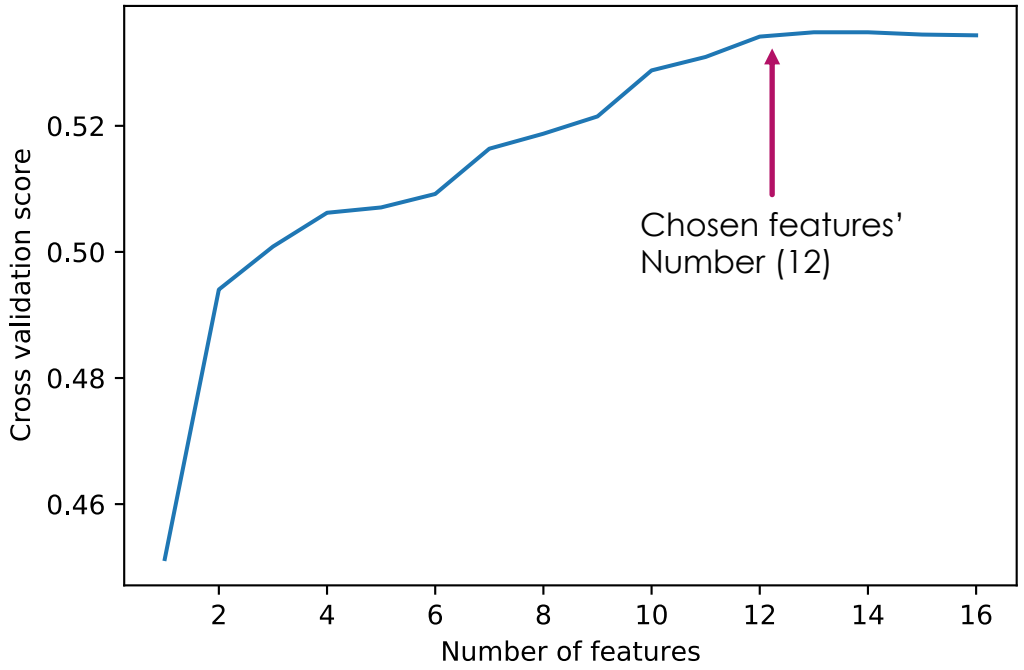
Methodology – features selection -

Feature selection (Recursive Feature Elimination + Cross Validation)

sklearn.feature_selection.RFE

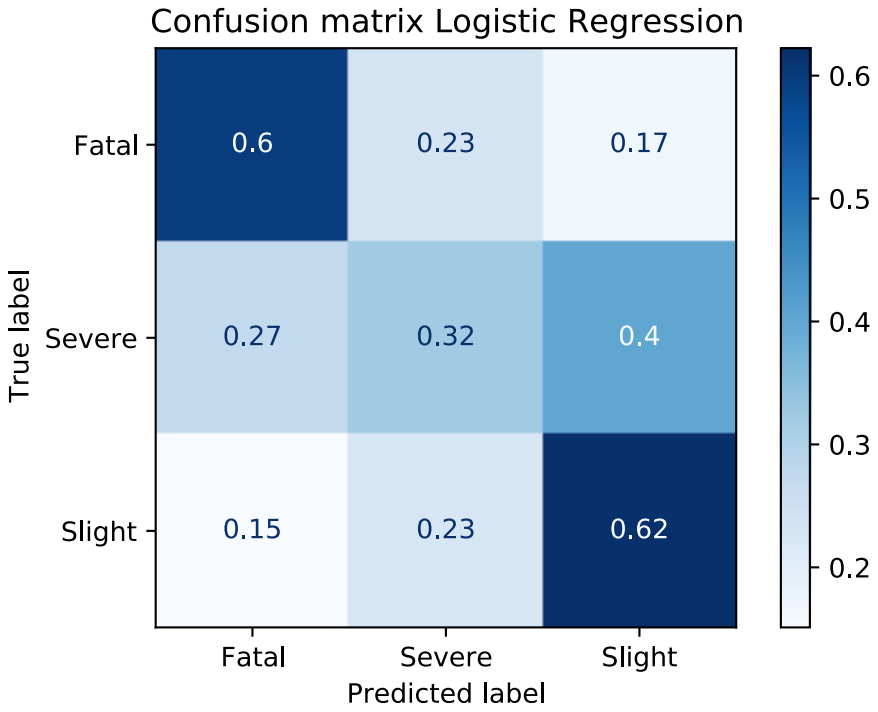
class sklearn.feature_selection. **RFE**(*estimator*, *, *n_features_to_select=None*, *step=1*, *verbose=0*)
[source]

```
No of features to keep: 12
Keep following features: Index(['Two_or_more_Casualt', 'Faster_40', 'Junct_det_nojunct',
       'Light_cond_darknolight', 'Urb_rur_Rural', 'Veh_type_Car',
       'Veh_type_Pedal_cycle', 'Veh_type_Mot_500', 'Skidding_yes',
       'Age_band_driver_old', 'Pedestrian', 'EC_veryhigh'],
      dtype='object')
[0.48358871 0.48733871 0.49439516 0.49237903 0.49931452 0.50229839
 0.50354839 0.51076613 0.52032258 0.52314516 0.52524194 0.52806452
 0.52774194 0.52754032 0.52798387 0.5275    ]
```



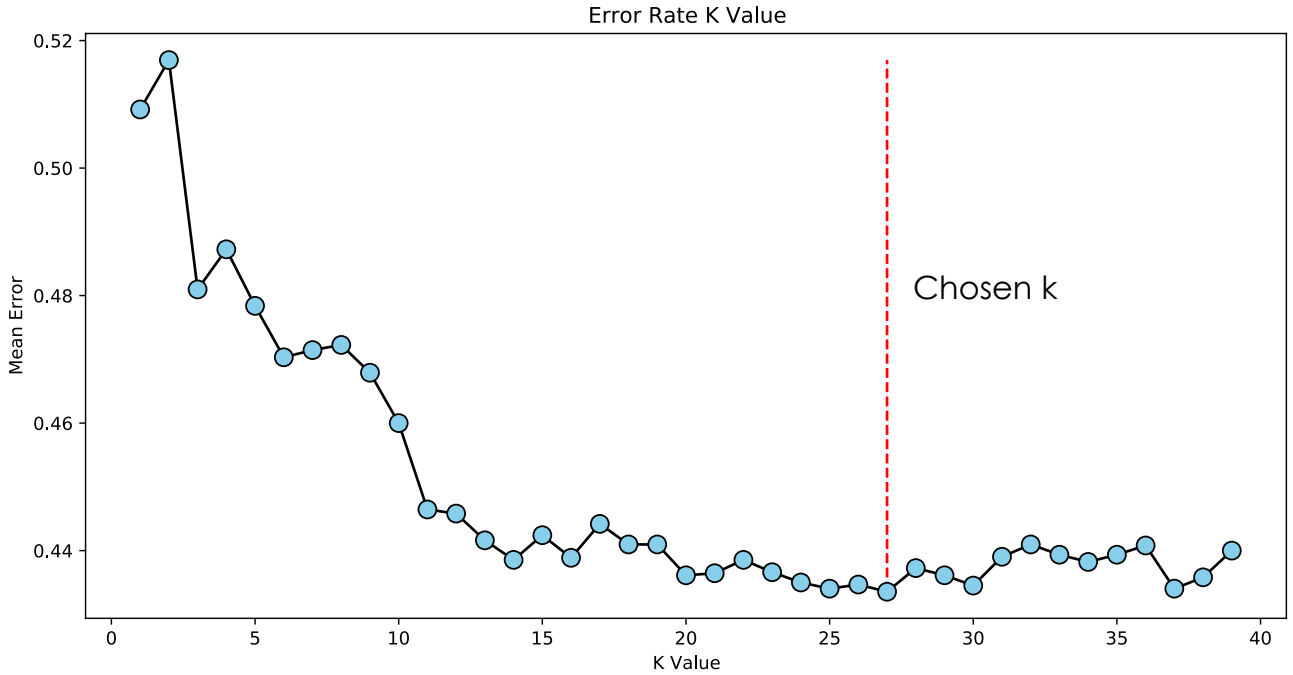
Results – Modeling -

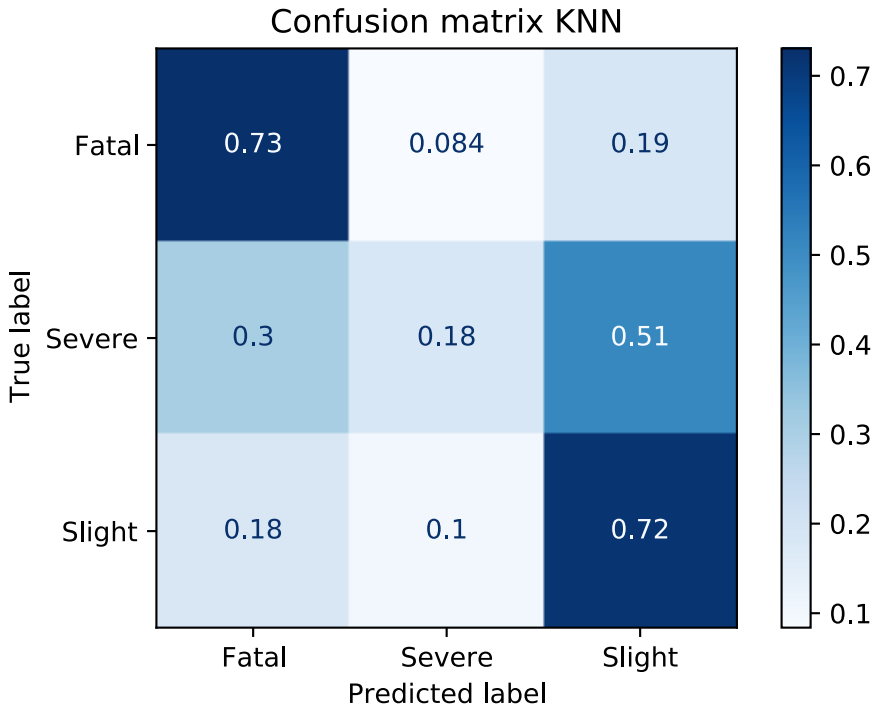
Logistic Regression



Results – Modeling -

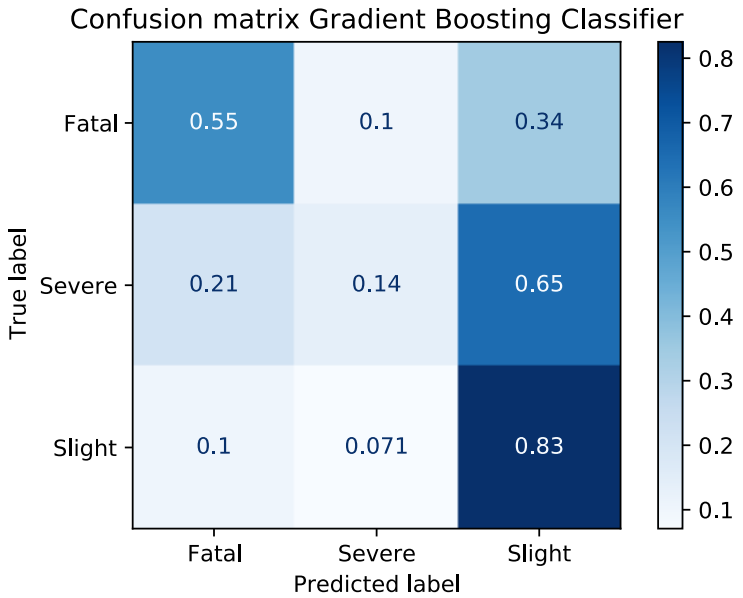
KNN



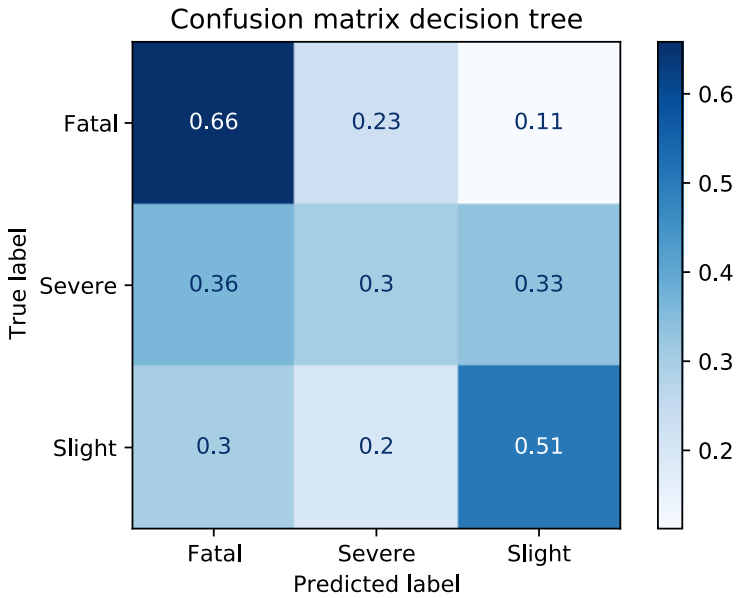


Results – Modeling -

Decision Tree Classifier

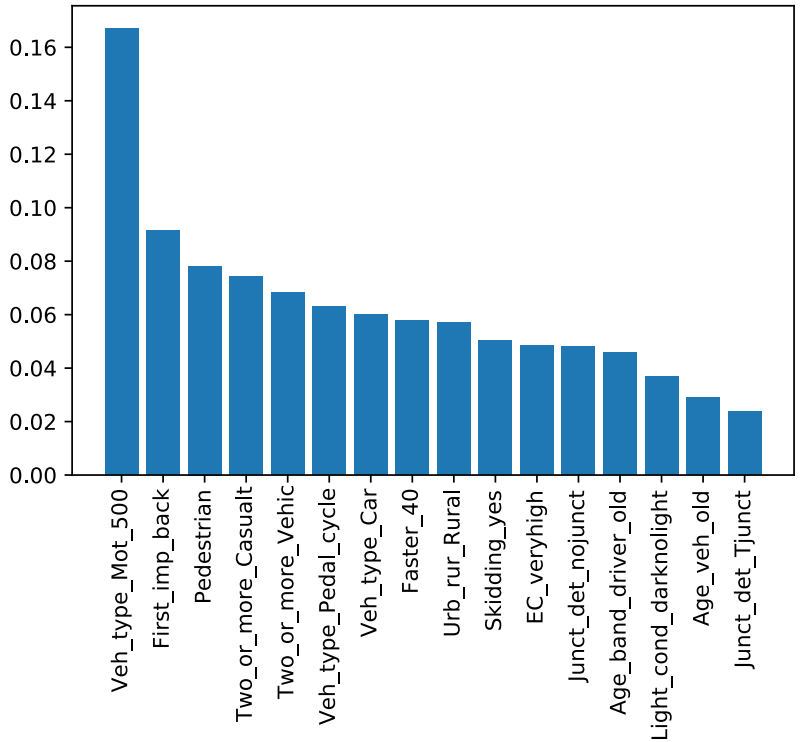


Gradient Boost Classifier

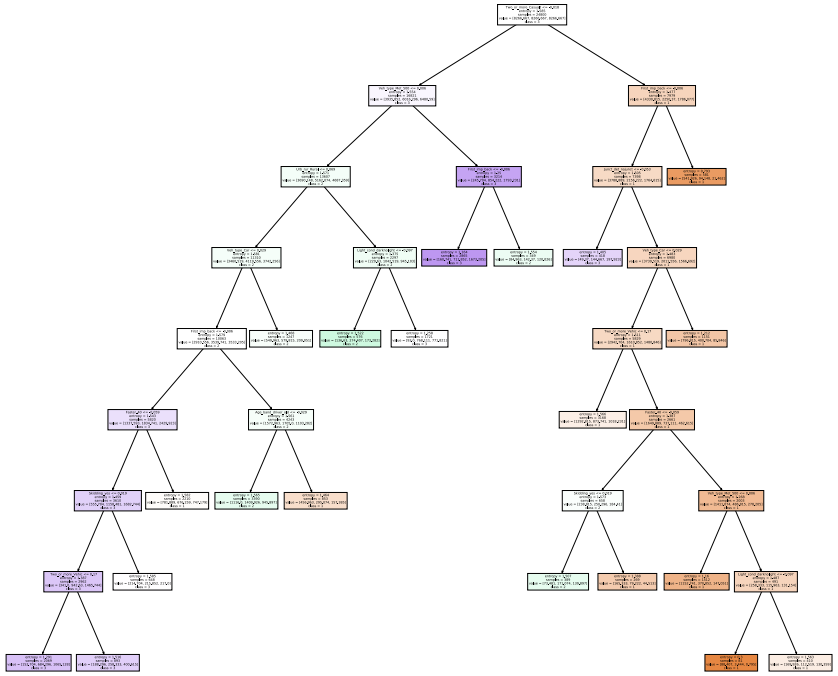


Results – Evaluation -

Feature Importance



Tree evaluation



Conclusion

Model accuracy

- Not sufficiently accurate to be deployed (accuracy with training data nerver reaches 60%, but overfitting is well corrected – around 50% accuracy with test data)
- “Slight” and ”Severe” categories are not weel discriminated by the majority of models; more data and tailored feature selection must be done

Insights from the model

- Increment of traffic stops (especially for motorbikes with more than 125 cc)
- Measures aimed to increase visibility and better regulate traffic flow in correspondence of areas with high pedestrian density and road junctions.