UNIVERSITY OF MANNHEIM

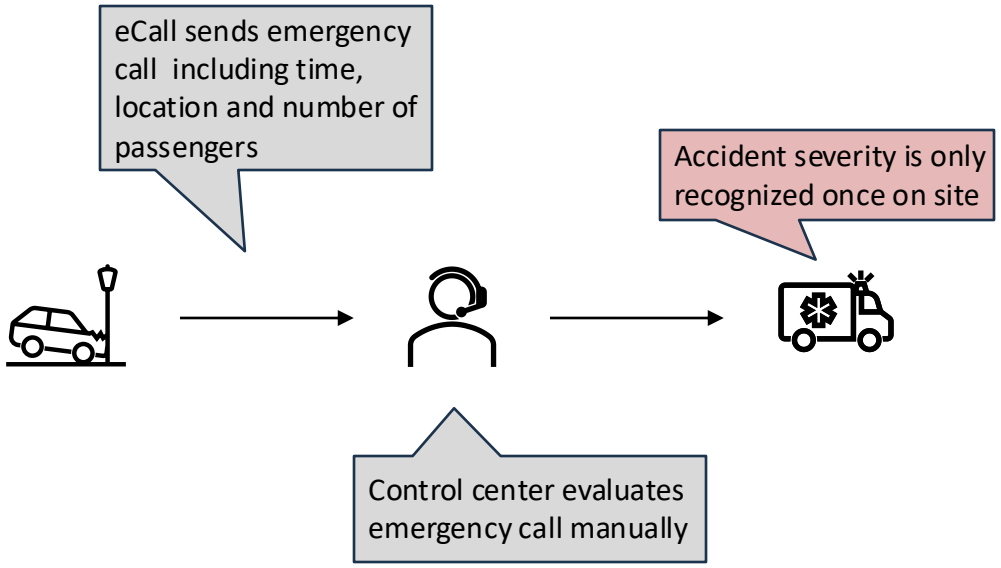
Project Report Group 6

Predicting Injury Severity in Road Accidents – A Real-Time Classification Approach

David Cebulla, Gabriel Himmelein, Lukas Ott, Artur Loreit, Aaron Niemesch

Application Area and Goals

Predict severity. Speed up rescue. Save lives.

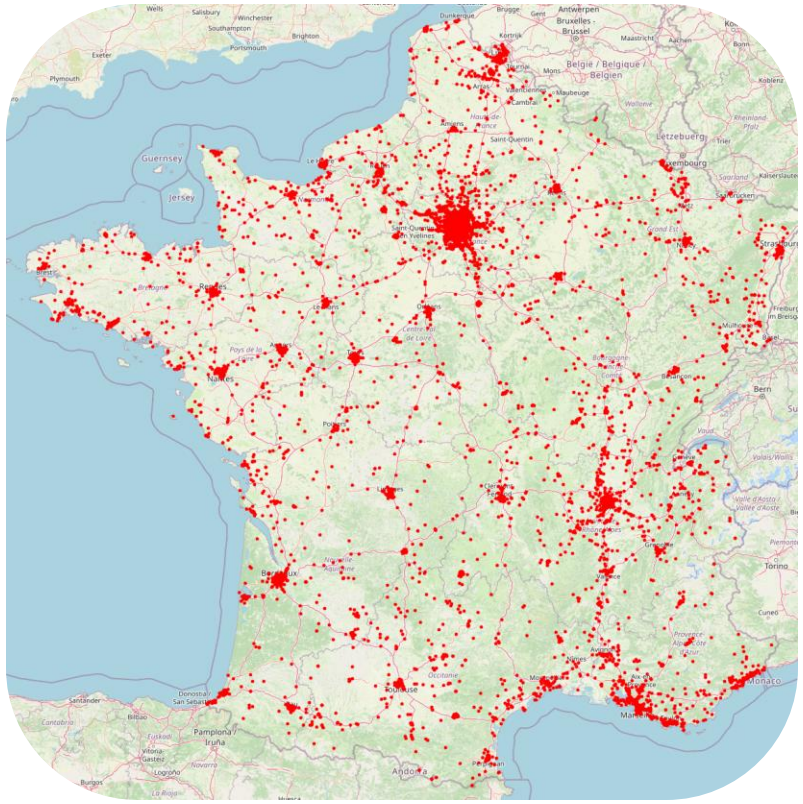


Problem definition:

- Delay due to manual accident classification
- Errors in assessing the severity of an accident
- ⇒ Saving time and improving accuracy can save lives

Structure and Size of Data Set

„Annual Road Traffic Injury Database“ by ONISR

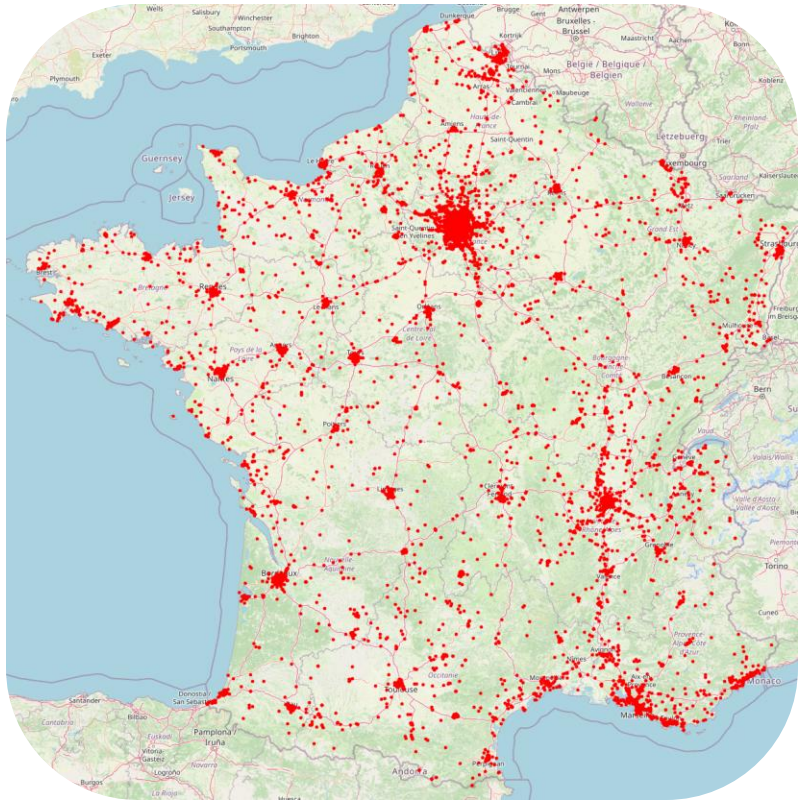


Dataset evaluation:

Files	**4 per year (characteristics, locations, users, vehicles)**
Target	**Individual level of severity**
Train set	**493.214 (2019-22)**
Test set	**125.505 (2023)**
Features	**71 original features**
Used Features	**23 original, 34 engineered features**

Structure and Size of Data Set

Included Aspects and Features in Final Data Set



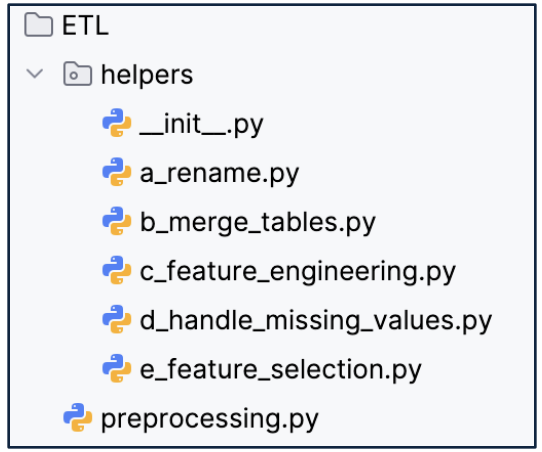
Feature Descriptions:

Time Context	Includes time of day categories
Geospatial	GPS coordinates
Environment	Physical site attributes with risk-based ordinals
Crash dynamics	Collision mechanics
Vehicle attributes	Specifies primary vehicle type
Personal attributes	Demographic data
Safety Equipment	Use of protective gear
Clustering	Accident personas
Target	Ordinal target (0,1,2)

Preprocessing

Data Merging Strategy

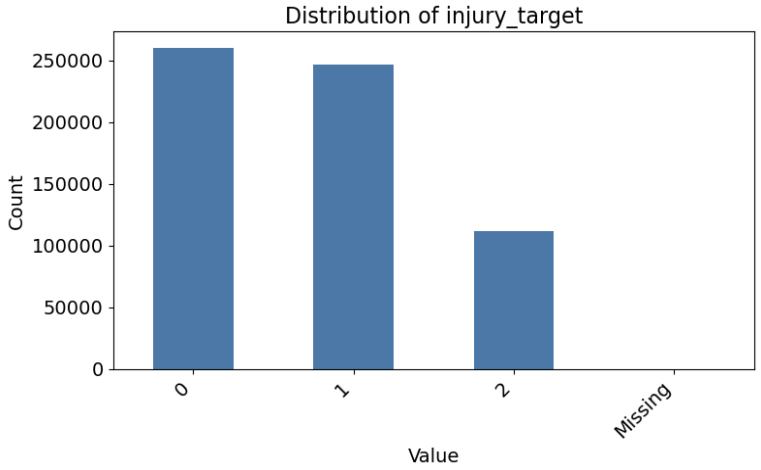
Data standardization strategy	Engineered features:	
Creation of synthetic *user_id* for the years 2019 and 2020	(1) Vehicle antagonist resolution: • Impact score to vehicle categories based on mass and risk (e.g., HGV Truck = 6, Bicycle = 2) • Highest scoring antagonist vehicle encoded into record	(2) Location deduplication: • Weighted completeness score assessment for most data-rich description of scene • (Road Category: 2.0, Speed Limit: 2.0, others: 1.0)



Preprocessing

Data Merging Strategy

Feature selection and cleaning	Engineered features	Encoding of target value
Imputation of structural missingness • Actual missing values: 0 or 'Unknown' • Pedestrian and antagonist cases: -1 or 'none' Columns were dropped with at least 5% missing values	• Temporal features • Road complexity index • Bucketing of vehicle and user attributes	Severe injury and death were put in the same category



Data Mining

Categories of algorithms considered

*Rebalancing Required

Boosting	Bagging	Regression
• **CatBoost**	• **BalancedRandomForest**	• **RidgeClassifier**
• HistGradientBoosting	• BalancedBaggingClassifier	• RidgeRegression
• GradientBoosting	• EasyEnsembleClassifier	• LassoRegression
• AdaBoost		• CatBoostRegression
• LightLGM		

Ordinal
- Ordinal-IT
- Ordinal-AT

SVM
- SVC
- LinearSVC
- SDGClassifier + Nystroem

ANN
- MLPClassifier

Data Mining

Model Selection

3-fold Cross Validation

- Reason: Dataset Large (>430k), Resampling (130k)
- Tradeoff between Precision and Computation Time
- Alternative: Separate **Validation Set**

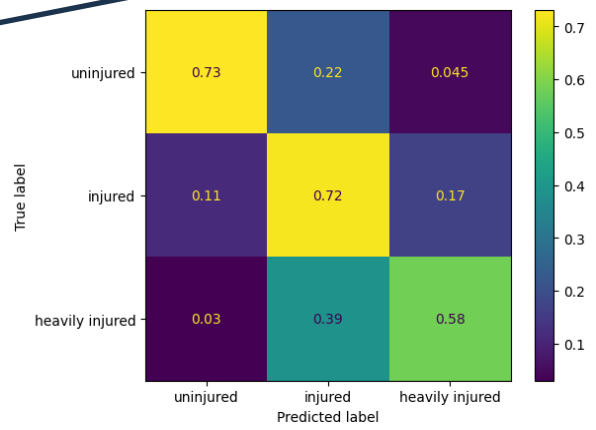
F1-Macro (Optimization Target)

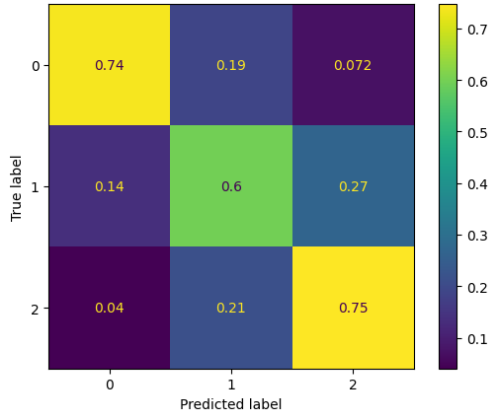
- Equal Weight for all Classes irrespective of Distribution
- **Important**: only used after checking for ordinal performance
  - **Confusion Matrices** used to check for Error Types
  - Check Recall for Severly Injured
  - Generated using **`[cross-val-predict]`**

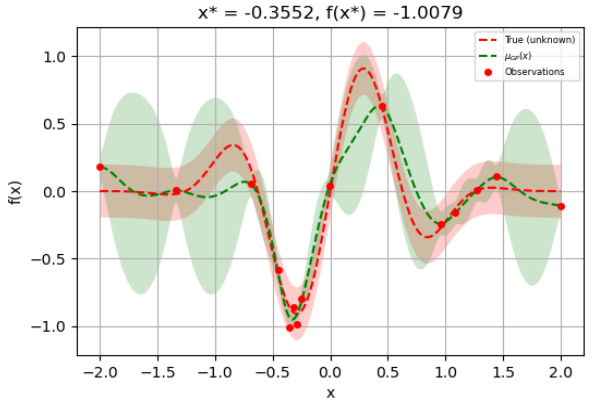
Hyperparameter Tuning: Bayesian Optimization

- Dataset Size: Reduce Computation Time
- E.g.: CatBoost Hyperparameter Training: 1h 40min (on GPU!)

F1-Makro: ≈ 0.7







Data Mining

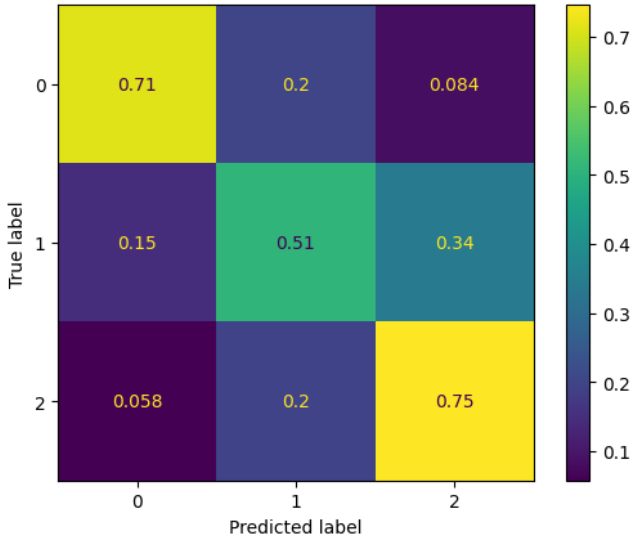
Hyperparameter Tuning

Model	Hyperparameter	Search Space	Best Value	F1-Macro
Ridge	alpha	$\{0.05, 0.10, \ldots, 10\}$	1.5	0.654
BRF	n_estimators	$[50, 400] \cap \mathbb{Z}$	400	0.674
	max_depth	$\{3, \ldots, 20\}$	18	
	min_samples_leaf	$\{1, \ldots, 20\}$	1	
	replacement	{`True`, `False`}	`False`	
	sampling_strategy	{`all`, `not minority`}	`all`	
CatBoost	iterations	$\{1000, \ldots, 5000\}$	4139	0.696
	learning_rate	$[0.01, 0.2]$	0.01	
	depth	$\{4, \ldots, 10\}$	10	
	l2_leaf_reg	$[10^{-2}, 10]$	0.1887	
	border_count	$\{32, \ldots, 255\}$	255	

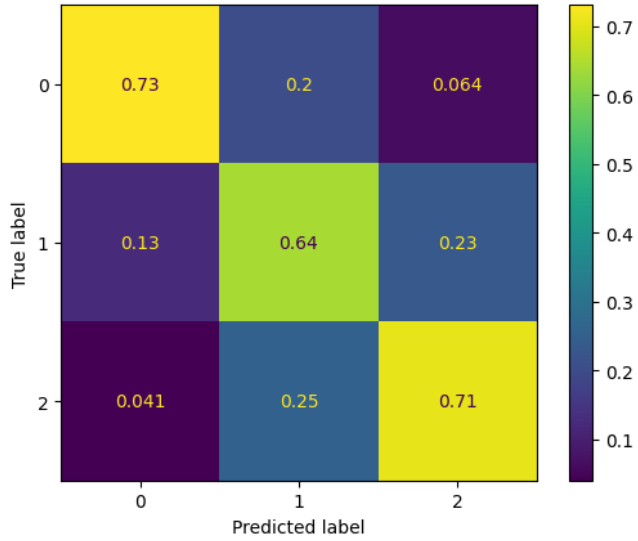
Data Mining

Model Selection: Results

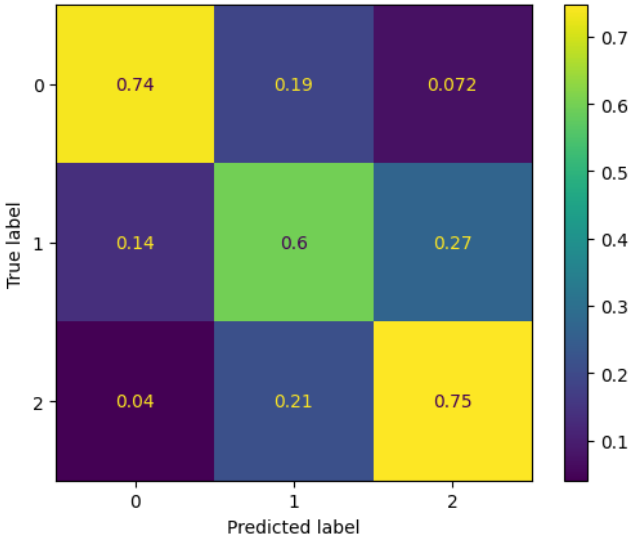
RidgeClassifier
F1-Macro: 0.65



BalancedRandomForest
F1-Macro: 0.66



CatBoost
F1-Macro: 0.7




```
y_pred_cat_cross_val2 = cross_val_predict(cat_classifier_cross_val2, X_train_rebalanced, y_train_rebalanced, cv=cv)

print(classification_report(y_train_rebalanced, y_pred_cat_cross_val2))

ConfusionMatrixDisplay(confusion_matrix(y_train_rebalanced, y_pred_cat_cross_val2, normalize='true')).plot()
```

➡ Only model selection: no generalization performance!

Data Mining - Clustering

Goal: Identification of distinct accident personas

Approaches: K-Prototypes, HDBSCAN, Agglomerative Clustering

Criteria for Profiling:

- Lift filter: A minimum overrepresentation
- Support filter: minimum total representation in the cluster
- Dominance filter: Exclude dominant features

$$\text{lift}(x, c) = \frac{P(X{=}x|C{=}c)}{P(X{=}x)} > 1.5.$$

Evaluation Setup

UNIVERSITY OF MANNHEIM

Test Set: Full **2023 dataset**

107,967 records

Distribution:

- Non-Injured 47%
- Injured 37%
- Severly Injured 16

Domain Baseline

$$\hat{y} = \begin{cases} \text{uninjured}, & \text{if speed\_limit} \leq 50, \\ \text{injured}, & \text{if speed\_limit} < 100, \\ \text{severely injured}, & \text{otherwise.} \end{cases}$$

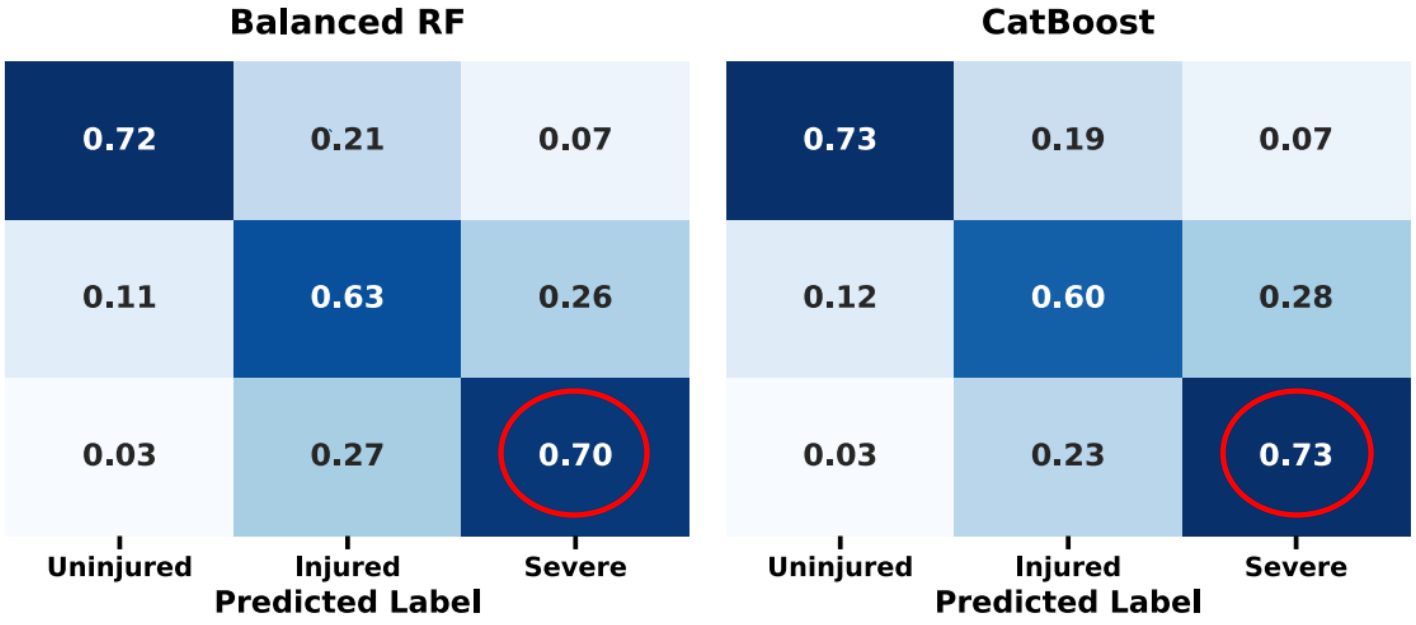
Results
Model Comparison (1)

	RC			BRF			CB			BL		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Uninjured	0.84	0.72	0.78	0.88	0.72	0.79	0.88	0.73	0.80	0.50	0.64	0.56
Injured	0.59	0.46	0.52	0.62	0.63	0.63	0.63	0.60	0.62	0.34	0.30	0.32
Severe	0.40	0.77	0.53	0.47	0.70	0.56	0.46	0.73	0.57	0.16	0.08	0.10
Macro F1		0.61			0.66			0.66			0.33	
Cohen's κ		0.59			0.63			0.6342			0.0910	

Table 4: Test Performance: Ridge Classifier, Balanced Random Forest, CatBoost, Baseline.

➡ CatBoost and Balanced Random Forest perform best!

Results
Model Comparison (2)



- Catboost has the higher recall for the "severe" - class
- This is desirable, since a misclassification of a severe case can be fatal

➡ Therefore, Catboost yields the best results

Results
Feature Importance

Rank	RC	BRF	CB
1	mobile_obstacle_struck_1	mobile_obstacle_struck	type_of_collision
2	vehicle_category_other_none	impact_delta	mobile_obstacle_struck
3	sex_2 (Female)	fixed_obstacle_struck	age_group
4	mobile_obstacle_struck_4	type_of_collision	initial_point_of_impact
5	vehicle_2_3_wheeler	speed_limit	speed_limit

- Nature of the obstacle struck critical for injury prediction
- Variables such as type of collision and initial point of impact highlight the importance of the precise point of impact for prediction
- Speed limit is also a relevant feature, as higher speeds lead to more severe injuries

Results
Accident Personas

Persona 1

- Midday
- Older road users
- Urban areas

Persona 2

- Night-time
- Younger adults
- Low visibility

Persona 3

- Morning rush-hour
- Commuter traffic / Congestions

Persona 4

- High speed
- Pedestrian involvement

Persona 5

- Low-speed
- Urban areas
- No safety belt

Thank You For Your Attention