High-resolution traffic accident prediction in Berlin

Utilizing road segment, time, and weather features to predict hourly traffic accident risk for road segments in Berlin.

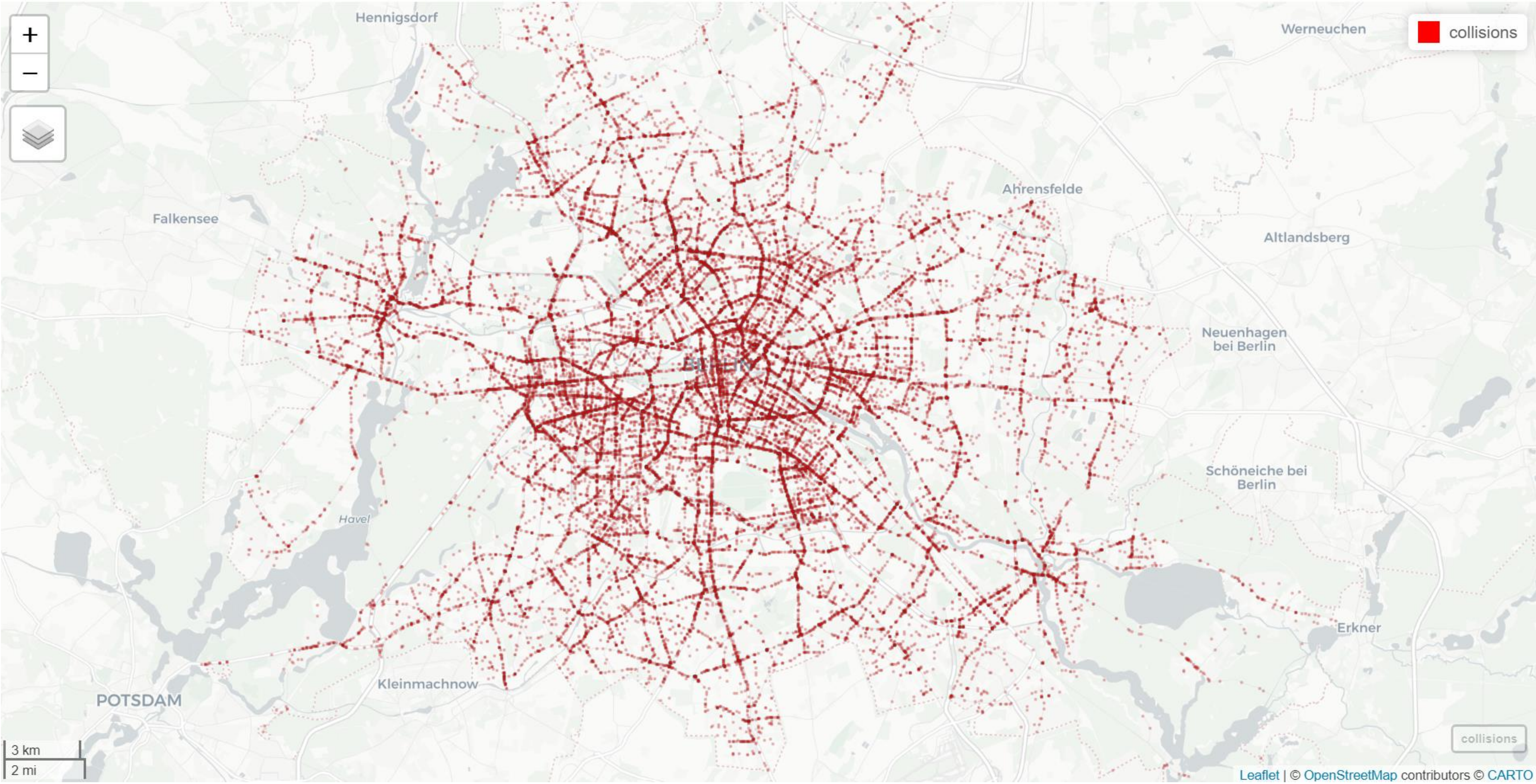
Machine Learning – Group F – May 12th, 2022

Ma. Adelle Gia Arbo, MDS 2023

Helena Bakic, MDS 2023

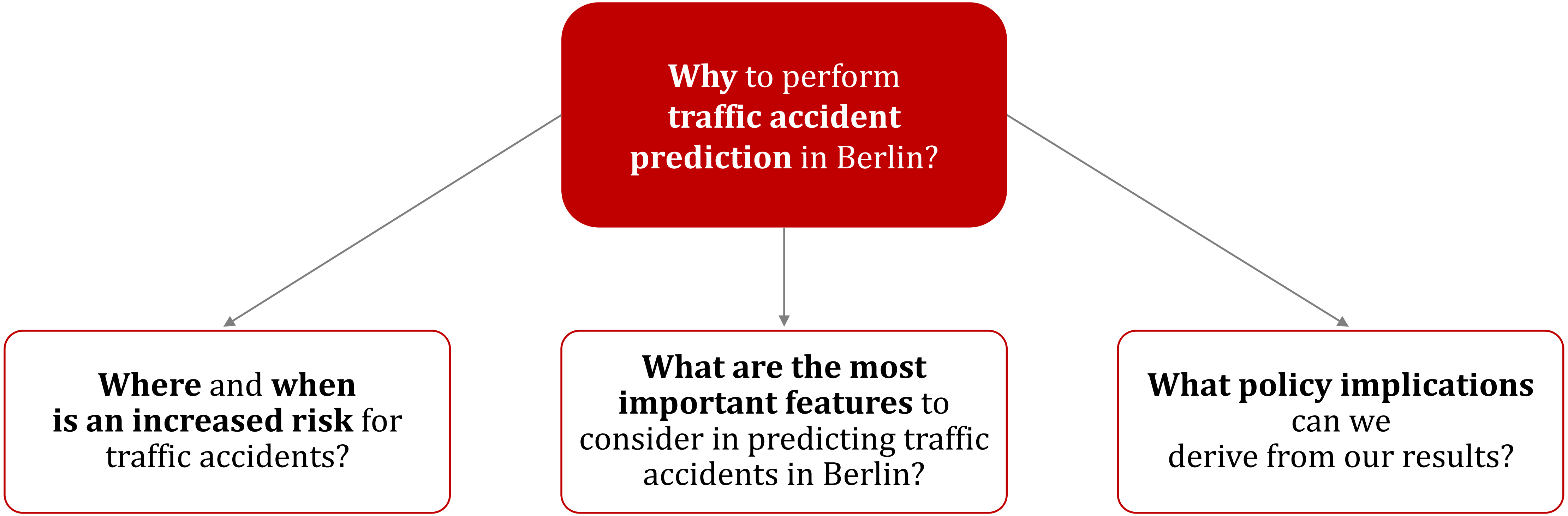
Benedikt Ströbl, MDS 2023

Hertie School



Hertie School

A binary classification problem prone of imbalanced data in rare-event prediction



Why to perform **traffic accident prediction** in Berlin?

Where and **when is an increased risk** for traffic accidents?

What are the most important features to consider in predicting traffic accidents in Berlin?

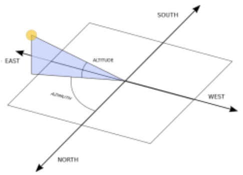
What policy implications can we derive from our results?

→ **For a given date, hour, and road segment**: Does an accident happen or not?

Our project was structured along the machine learning workflow

Data collection, Pre-processing

Acquisition of collision, road network, weather, and time data. Extensive feature engineering and the generation of negative example combination—road segment time pairs where no accident has occurred.



Modelling

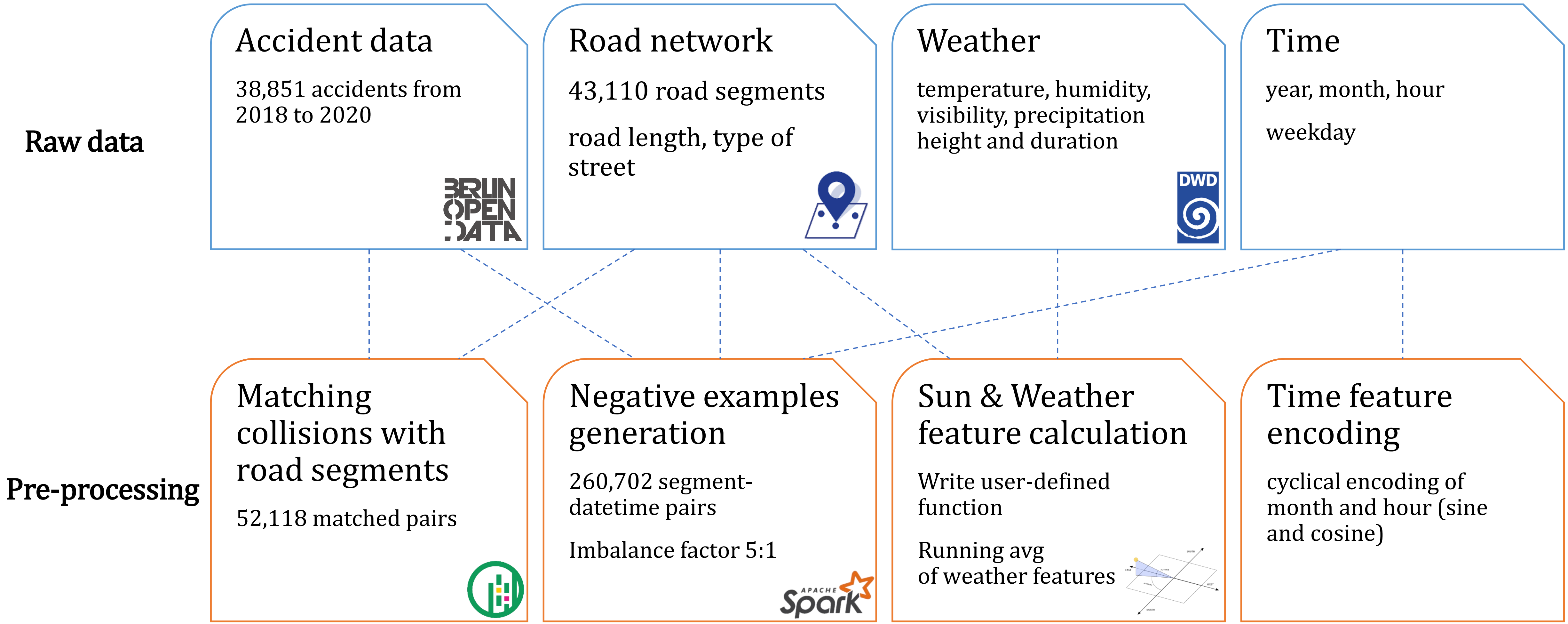
Design and implementation of models that fitted our binary classification problem and mitigated with the imbalanced data issue. Different re-sampling techniques have been designed to achieve better performance.

Tuning & Evaluation

Conduction of random and grid search with cross validation to find the best parameter specification for our models and select the best performing one according to key metrics. Sampling strategies have also been optimized.

Note: There are, of course, more packages we used throughout our project workflow, but these illustrate the key tools that were necessary.

Data pre-processing was required across various data sources and engineering steps



Raw data

Accident data

38,851 accidents from 2018 to 2020

Road network

43,110 road segments

road length, type of street

Weather

temperature, humidity, visibility, precipitation height and duration

Time

year, month, hour

weekday

Pre-processing

Matching collisions with road segments

52,118 matched pairs

Negative examples generation

260,702 segment-datetime pairs

Imbalance factor 5:1

Sun & Weather feature calculation

Write user-defined function

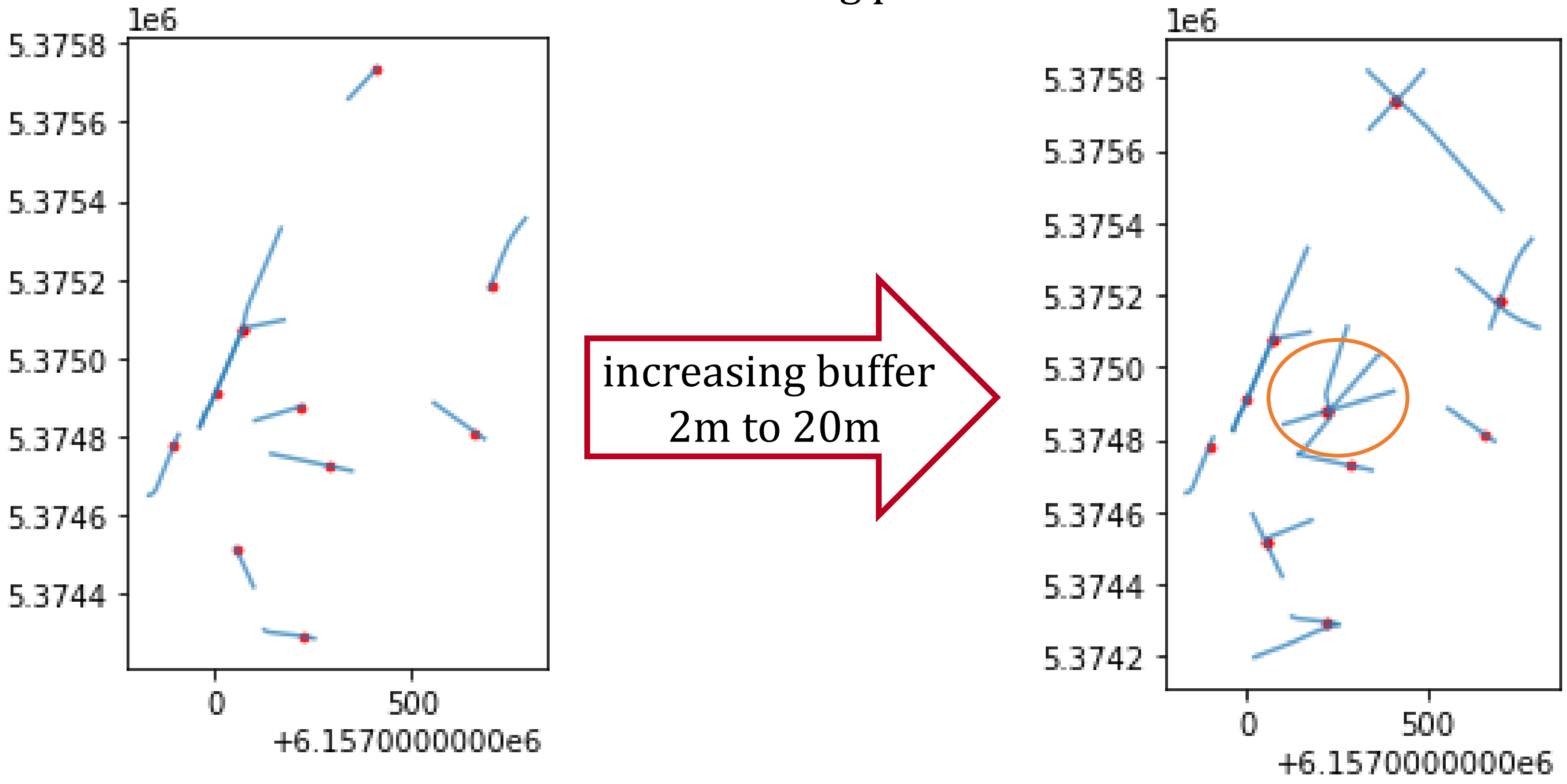
Running avg of weather features

Time feature encoding

cyclical encoding of month and hour (sine and cosine)

Geolocation-matching of road segments and collisions with GeoPandas

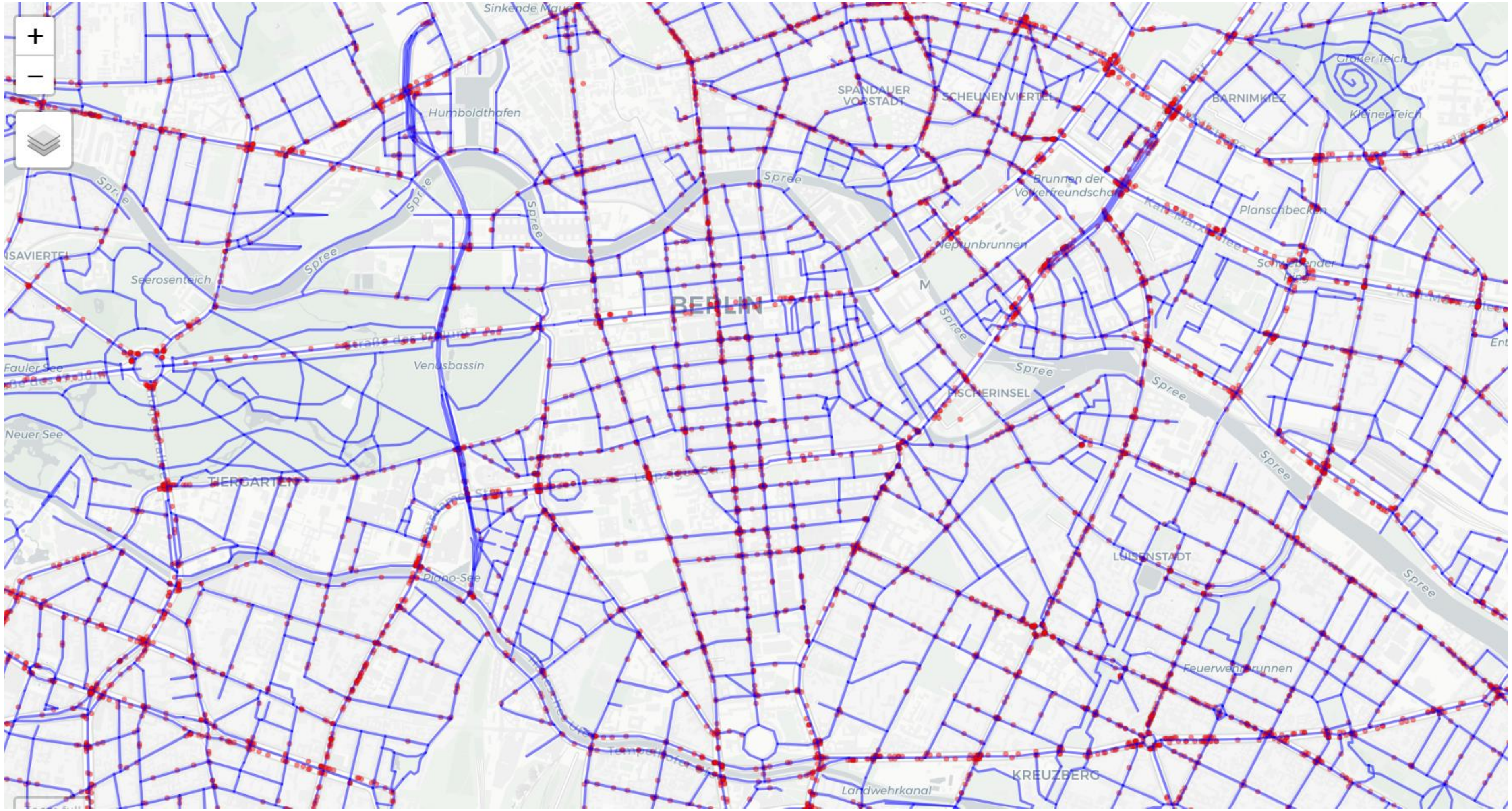
buffer as a tuning parameter



increasing buffer 2m to 20m

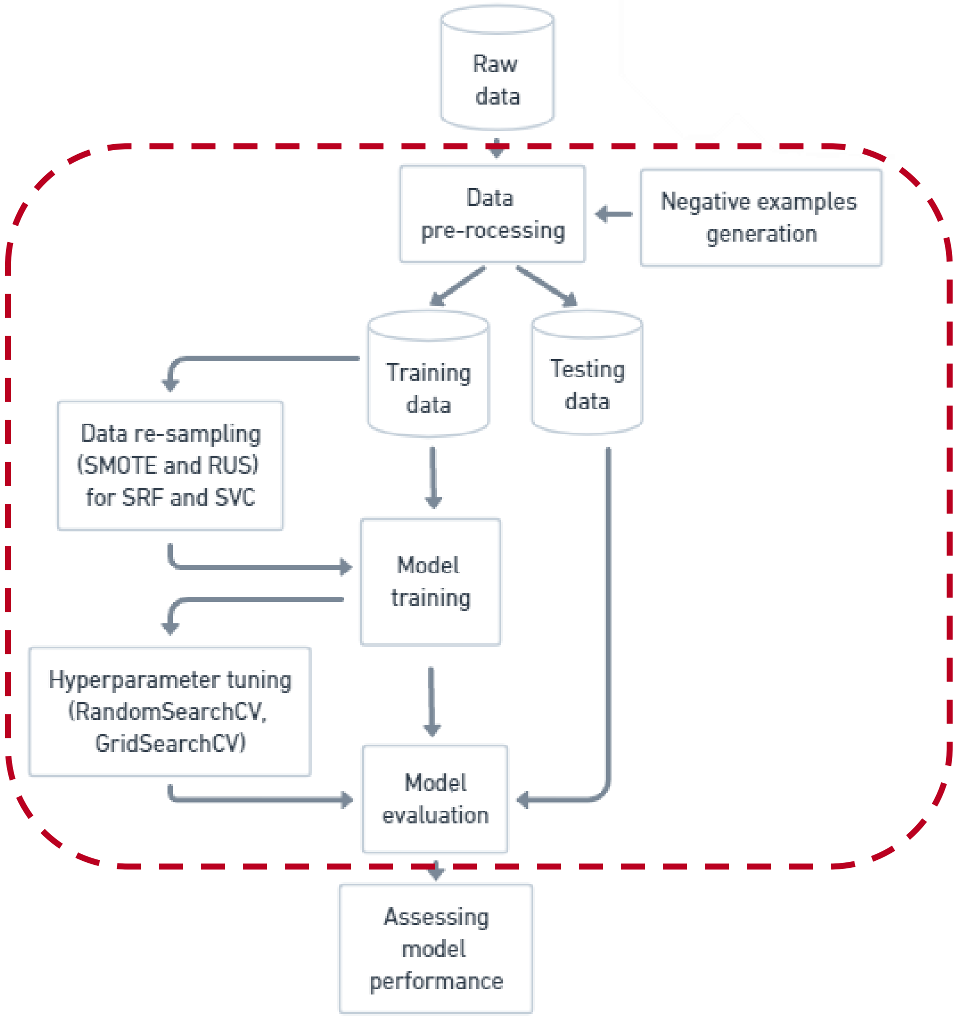
Note: Sample of accidents that occurred in the Alt-Kaulsdorf district shown.

...let us end up with 52,118 pairs of collisions and road segments in Berlin



Note: Created using Leaflet and OpenstreetMap.

We applied various types of models to minimize the imbalanced data issue



Raw data
Data pre-rocessing
Negative examples generation
Training data
Testing data
Data re-sampling (SMOTE and RUS) for SRF and SVC
Model training
Hyperparameter tuning (RandomSearchCV, GridSearchCV)
Model evaluation
Assessing model performance

Standard Random Forest

Balanced Random Forest

Standard Random Forest w/ SMOTE and RUS*

XGBoost

Support Vector Classifier

Support Vector Classifier w/ SMOTE and RUS*

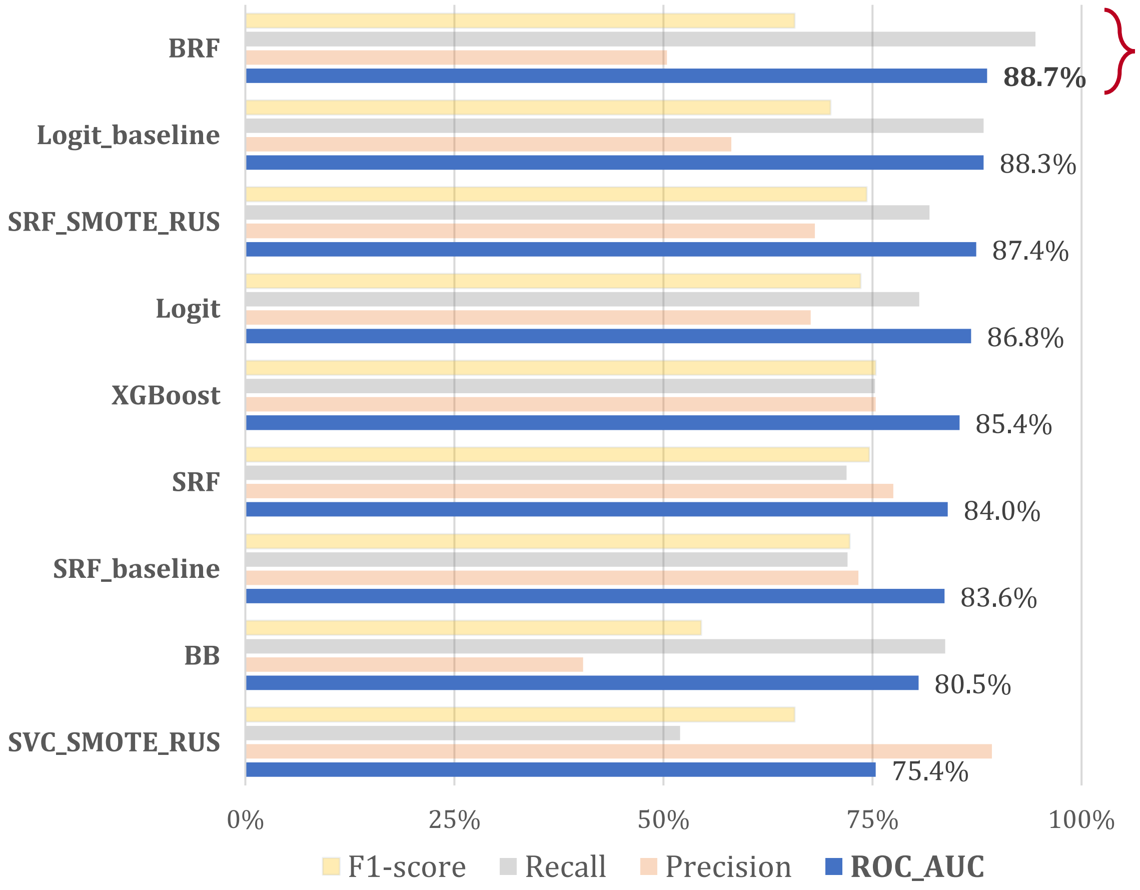
Logistic Regression

*Our self-implemented models using combined re-sampling with SMOTE and RUS

Four key metrics were selected to choose the best-performing models

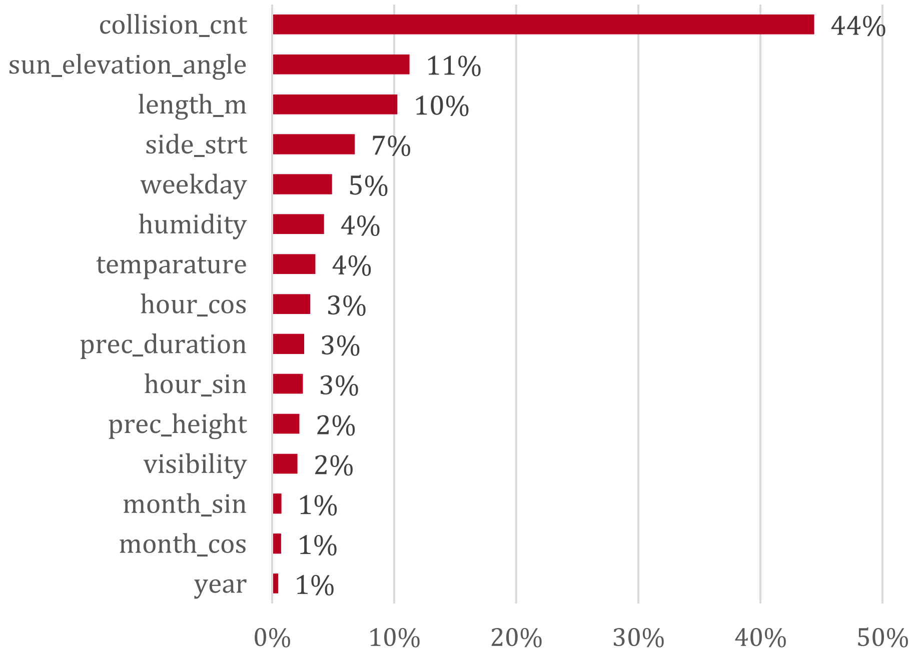
Performance across our different classification models on testing set

Models are compared across key metrics—roc_auc, recall, precision, and F1-score



Feature importances of BRF

Overall, individual feature importances sum to 1



Thank you!

And please do not forget to look at the appendix y’all

Link to our project on GitHub:

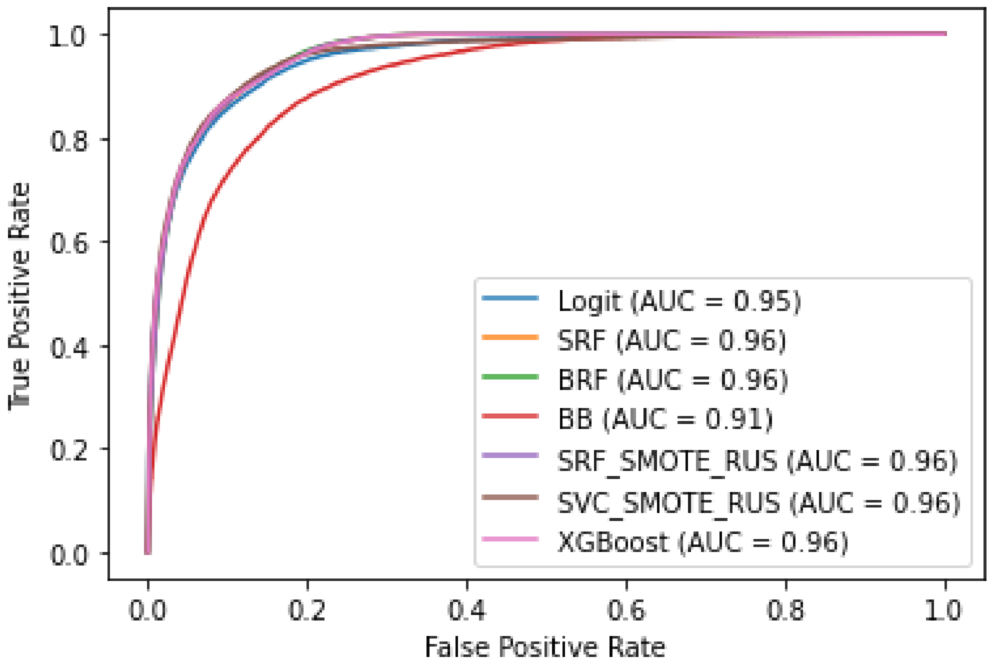
Hertie School

Appendix

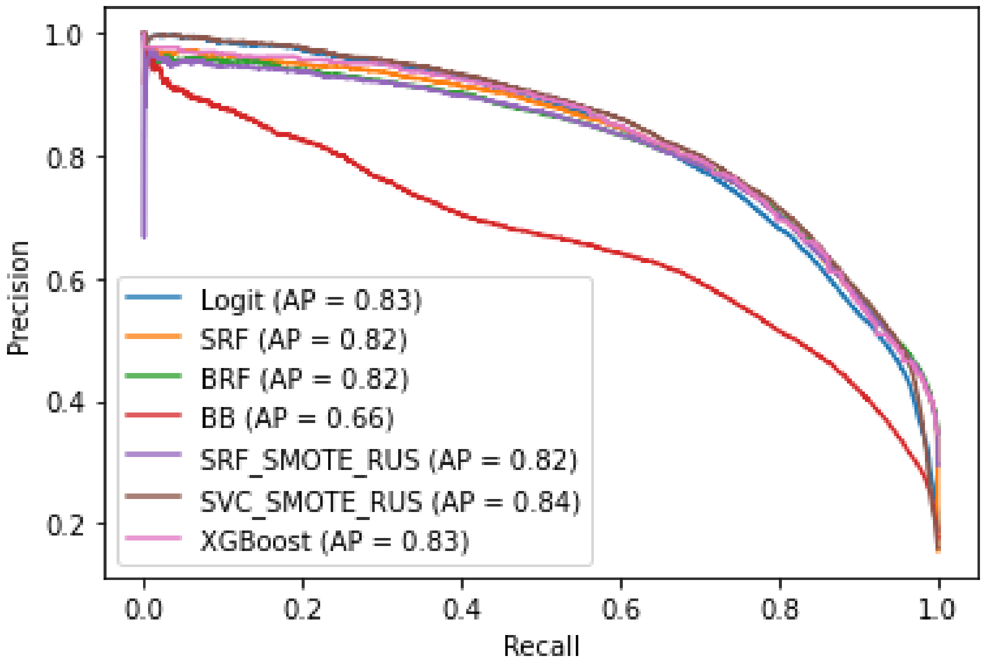
In the following you can find interesting links to our data sets and some additional slides about our models' performances, feature importances, and the team behind this project

Model evaluation: ROC & PR curves

ROC Curves



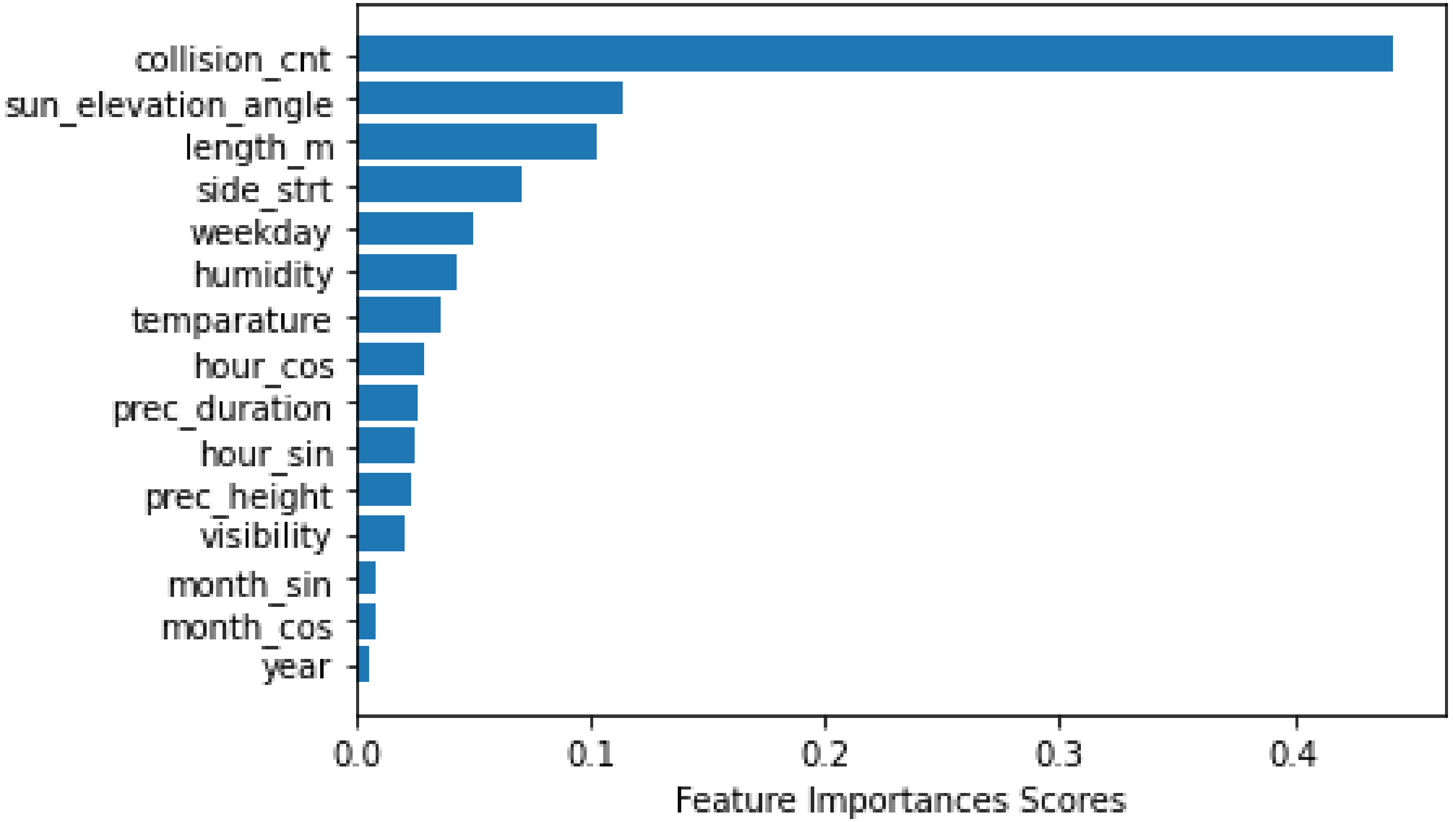
Precision-Recall Curves



Model evaluation: Table of performance metrics

	ROC_AUC	Precision	Recall	F1-score
Logit_baseline	88.3%	58.1%	88.3%	70.0%
SRF_baseline	83.6%	73.3%	72.0%	72.3%
Logit	86.8%	67.6%	80.6%	73.6%
SRF	84.0%	77.5%	71.9%	74.6%
BRF	**88.7%**	**50.4%**	**94.5%**	**65.7%**
BB	80.5%	40.4%	83.7%	54.5%
SRF_SMOTE_RUS	87.4%	68.1%	81.8%	74.3%
SVC_SMOTE_RUS	75.4%	89.3%	52.0%	65.7%
XGBoost	85.4%	75.4%	75.3%	75.4%

BRF Feature Importances



When it rains it pours—but precipitation doesn’t seem to be an important feature?!

Ma. Adelle Gia Arbo
MDS 2023

Helena Bakic
MDS 2023

Benedikt Ströbl
MDS 2023