

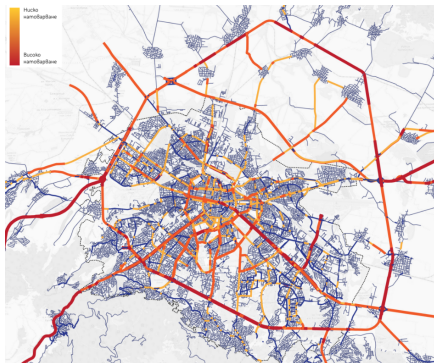
Ensemble learning models for data imputation and forecasting

Danail Brezov

Department of Mathematics, UACEG

Физика и химия на Земята, атмосферата и океана
25-27 септември, 2023

Why do we want to study the city traffic?







- 1 prompt for infrastructural upgrades
- 2 urban planning & incentives policies
- 3 correlated with noise, pollution, etc.
- 4 major cause of stress & health issues
- 5 strong predictor for the quality of life
- 6 weather & climate anomalies in DPA
- 7 a fun toy model for complex systems

Major traffic arteries of Sofia city.

Related publications:



-  Za Zemiata. Spatially Based Scenarios for Introduction of Low Emission Zones in Stolichna Municipality (2023) <https://www.zazemiata.org/resources/report-transport-lez-sofia/>.
-  Burov, A., Brezov, D. Transport Emissions from Sofia's Streets - Inventory, Scenarios, and Exposure Setting. In *Environmental Protection and Disaster Risks*; Dobrinkova, N., Nikolov, O., Eds.; EnviroRISKS 2022: Lecture Notes in Networks and Systems; Springer: Cham, Switzerland, Volume **638** (2023).
-  Dzhambov A., Dimitrova V., Germanova N., Burov A., Brezov D., Hlebarov I., Dimitrova R., *Joint associations and pathways from greenspace, traffic-related air pollution, and noise to poor self-rated general health: A population-based study in Sofia, Bulgaria*, Environmental Research **116087** (2023).
-  Brezov D. and Burov A., Ensemble Learning Traffic Model for Sofia: A Case Study, Applied Sciences **13**(8):4678 (2023)

What data can we rely on?



Table 1. Features used for training the ML model.

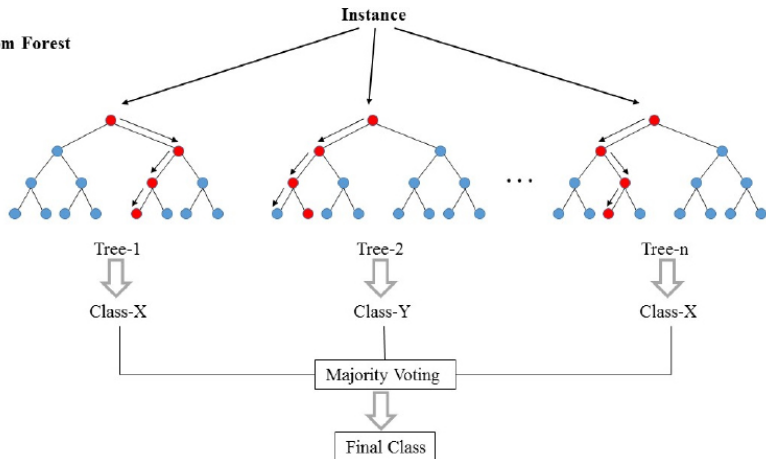
Column Name	Data Content	Values
baseTYPE	street type classification	40,736
IMMIS_RT	traffic situation typology	7637
C_KAPAZ	estimated road capacity	7637
StrDMNDRCT	number of directions	40,736
EMIT_SPEED	speed limit	7637
EMIT_GRDNT	street slope (gradient)	7637
SSINTr_In	space syntax (integration)	40,736
SSChr_In	space syntax (choice)	40,736
X0, Y0	coordinates of the centroids	40,736
OTMsurface	Open Transport Map (OTM) street surface type	5710
OTMtraffic	OTM traffic model	7637
TT200mHMc	TomTom traffic count data heatmap $r = 200$ m	7637
ACTUSEmean	heatmap $r = 200$ m estimated motorized users POI and cadastral data based	40,651
ACTLIVmean	heatmap $r = 200$ m estimated motorized inhabitants census based	40,497
exIDWmean	IDW-interpolated point-based RF traffic model ¹	7637

¹ We used clusterization into segments according to traffic and consecutive data imputation with RF-regression.

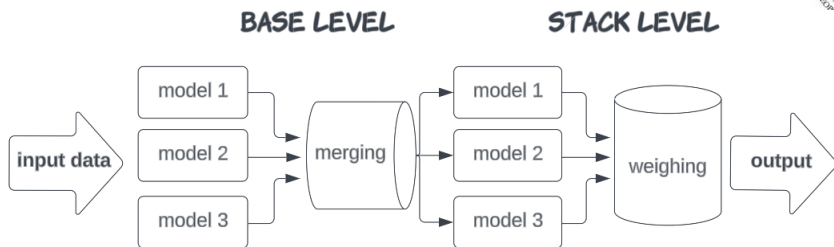
Ensemble learning algorithms



Random Forest

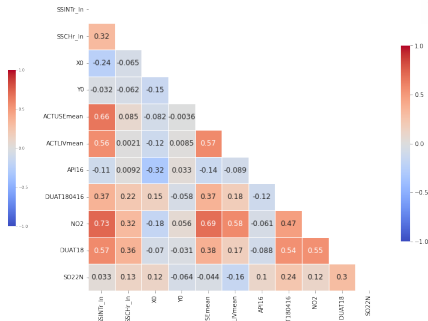
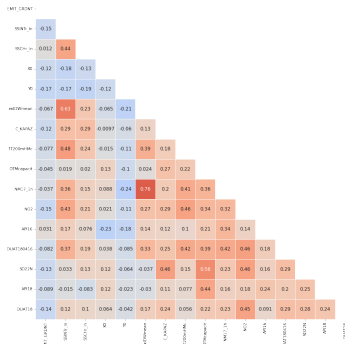


AutoML advantages



- many automated features in just a few lines of code
- rich libraries and powerful stacking algorithms
- multi-modal learning (tabular data, text and images)

So, let's start cooking..



Correlation heat map of the numerical features: primary vs. total SN.

Assessing the performance: primary street network



Table 2. Features with missing values predicted by the corresponding model ¹.

Feature	Content	% Missing	Model	Accuracy	r ²
OTMcapacit	capacity	25.23%	XGB/RSO	82.62%	0.83
NM17_1h	hourly 2017 traffic count	93.70%	WE_L2	64.29%	0.94
NO ₂	NO ₂ levels	99.21%	WE_L2	90.04%	0.74
API16	road traffic	92.61%	XGB/RSO	92.60%	0.91
DUAT180416	street traffic	60.05%	WE_L2	80.02%	0.43
SO22N	street traffic	99.36%	WE_L2	78.84%	0.33
API18	bidirectional road traffic	99.63%	WE_L2	81.33%	0.97
MIX18	mixed traffic	97.81%	WE_L2	71.68%	0.45
DUAT18	street traffic	98.18%	WE_L2	67.95%	0.25

¹ Ranked by accuracy based on MAPE (mean absolute percentage error) estimates. We also use the abbreviations: RSO (Random Search Optimization), WE_L2 (Weighted Ensemble_L2), XGB (Extreme Gradient Boosting).

...

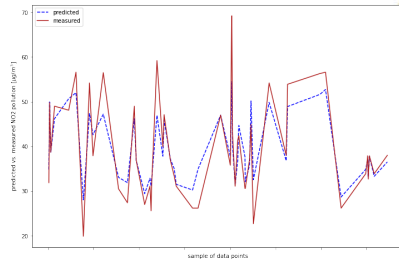
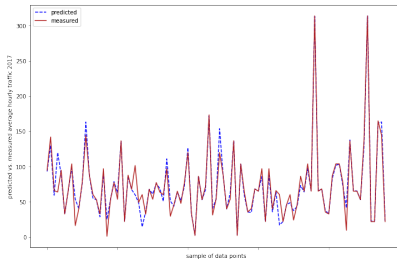


Table 3. Best performing models for the NM17_1h data imputation task according to AutoGluon ¹.

Model	score_val	fit_time	fit_order
WeightedEnsemble_L2	0.894	23.48	9
RandomForestMSE	0.881	0.715	3
XGBoost	0.881	1.639	7
NeuralNetTorch	0.863	20.33	8
ExtraTreesMSE	0.848	0.637	5
CatBoost	0.763	3.560	4
NeuralNetFastAI	0.576	1.354	6

¹ The coefficient of determination r^2 calculated via cross-validation is used as evaluation metric (score_val).

Assessing the performance: secondary street network



Table 4. Features with missing values predicted for the whole city network (we refer to [19], [21–25]).

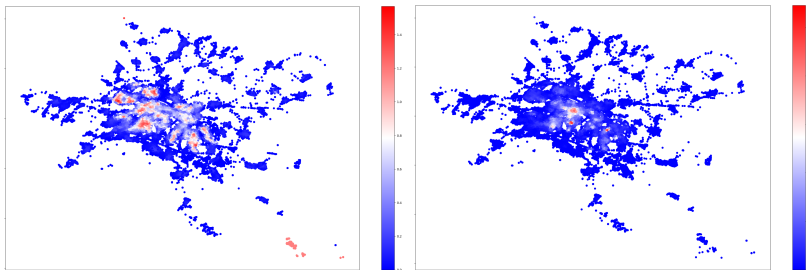
Column ¹	Content	Missing Data	Best Model	Accuracy	r ²
ACTUSEmean	active users of functions	0.21%	RF	97.36%	0.97
ACTLIVmean	density of inhabitants	0.59%	RF	97.88%	0.98
API16	road traffic	98.62%	RF	93.57%	0.90
DUAT180416	street traffic	92.51%	WE_L2	88.83%	0.69
NM17_1h	hourly noise map traffic	99.82%	WE_L2	84.10%	0.91
NO ₂	NO ₂ levels	99.85%	WE_L2	91.90%	0.80
API18	road traffic	99.93%	XGB/RSO	90.92%	0.61
DUAT18	street traffic	81.25%	WE_L2	77.78%	0.24
SO22CC	street traffic	99.79%	WE_L2	69.12%	0.67
DUAT18/WE_L2	modeled	81.25%	WE_L2	95.06%	0.65
SO22CC/WE_L2	modeled	81.16%	WE_L2	94.22%	0.92

¹ The last two rows refer to already modeled data (Table 2), while 'SO22CC' includes also some small streets.

Feature importance

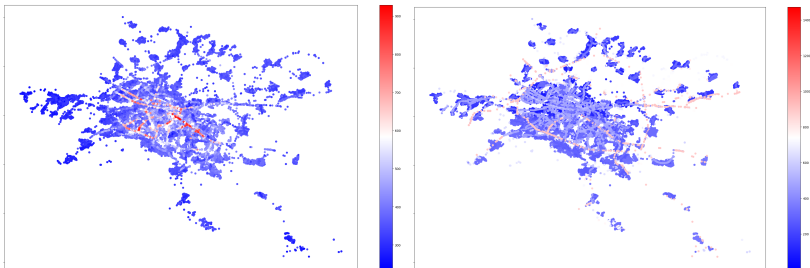


- 1 primary model: street capacity, speed limit, slope, surface type
- 2 secondary street network: functional analysis and space syntax



Functional analysis: motorized residents and users of daily activity POI

Data visualization



The daily average traffic for 2018 and 2022, according to our model.

Imagine if life was always that good to you..



```
▶ train_data = X.dropna(axis = 'rows')

label = dat[-1].name

from autogluon.tabular import FeatureMetadata
feature_metadata = FeatureMetadata.from_df(train_data)

from autogluon.tabular.configs.hyperparameter_configs import get_hyperparameter_config
hyperparameters = get_hyperparameter_config('default')

hyperparameters

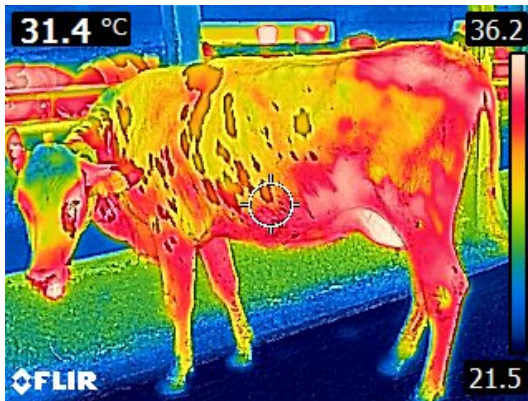
# root_mean_squared_error mean_squared_error mean_absolute_error median_absolute_error mean_absolute_percentage_error r2

from autogluon.tabular import TabularPredictor
predictor = TabularPredictor(label=label, problem_type = 'regression', eval_metric = 'mean_absolute_percentage_error' ).fit(
    train_data=train_data,
    hyperparameters=hyperparameters,
    time_limit=900,
    # presets="best_quality"
)

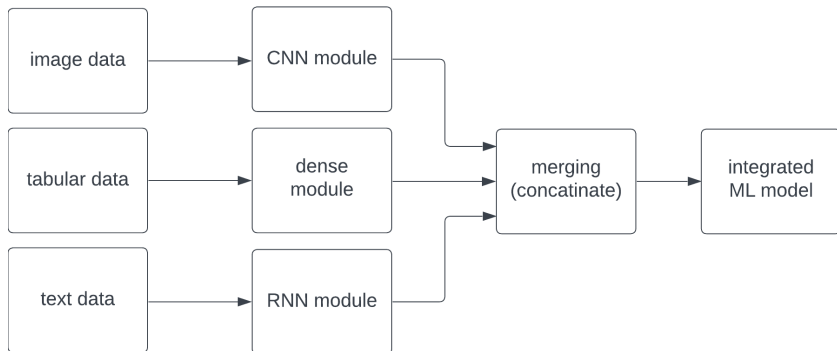
leaderboard = predictor.leaderboard()
leaderboard.sort_values(by='score_val', ascending=False)

pd.DataFrame(leaderboard)
```

This is a picture of a cow..



Multimodal ML with AutoGluon



Why choose between books and movies when you can have both?

Applications in analysis and forecasting

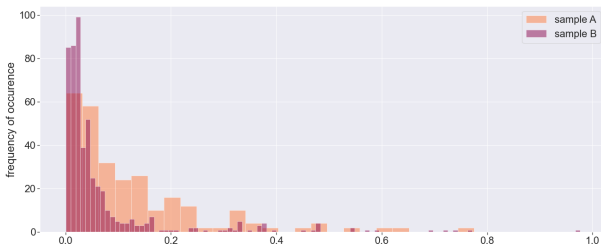
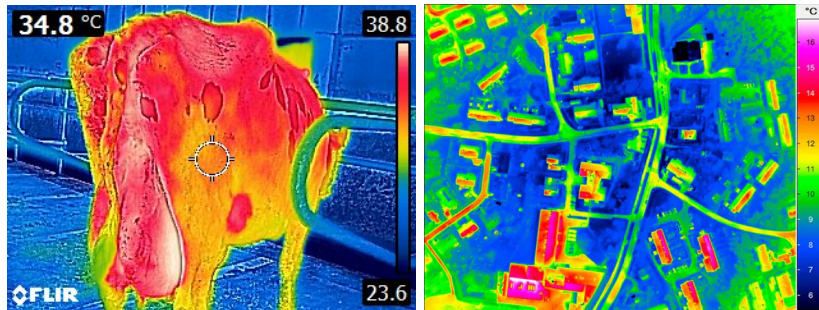


Table 2. Best performing models for the regression task for sample B, according to AutoGluon.

model	R^2	MAE	MedAE
WeightedEnsemble_L2	0.728	0.11°C	0.005°C
NeuralNetTorch	0.701	0.13°C	0.029°C
XGBoost	0.697	0.11°C	0.005°C
CatBoost	0.588	0.19°C	0.112°C

Thank you for your attention!

So, maybe that's the way to go from here..



Acknowledgements:

This work has been carried out in the framework of the grant № КП-06-Н54/2 'Development of a methodology for air quality and human health risk assessment in urban areas'

supported by the Research Fund at the Bulgarian Ministry of Education and Science.