Car fleet synthesis for agent-based mobility models: a comparison of machine learning and discrete choice methods

Marjolaine Lannes, Nicolas Coulombel, Yelva Roustan

Laboratoire Ville Mobilité Transport

Réseau de recherche Qualité de l'air en Ile-de-France

îledeFrance

École des Ponts ParisTech

Université Gustave Eiffel

Literature review

ML vs DCM comparison in car ownership models

Table 1. Decision types and approches in car ownership modeling literature

Reference	Decision types					Approach		Model type		
	Vehicle ownership	Car ownership	Car size	Fuel type	Car age	Actual	Forecast	DCM	ML	NN
Brownstone et al. (2000)				x		x	x	x		
Mohammadian and Miller (2002)			x				x	x		x
Whelan (2007)			x				x	x		x
Potoglou and Kanaroglou (2008b)		x				x		x		
Paredes et al. (2017)		x				x		x	x	
Kaewwichian et al. (2019)		x				x			x	x
Basu and Ferreira (2020)	x					x		x	x	
Dixon et al. (2021)		x				x	x		x	x
Zambang et al. (2021)	x					x		x	x	

Note. DCM: discrete choice model, ML: machine learning, NN: neural network

Methodology
Problem statement

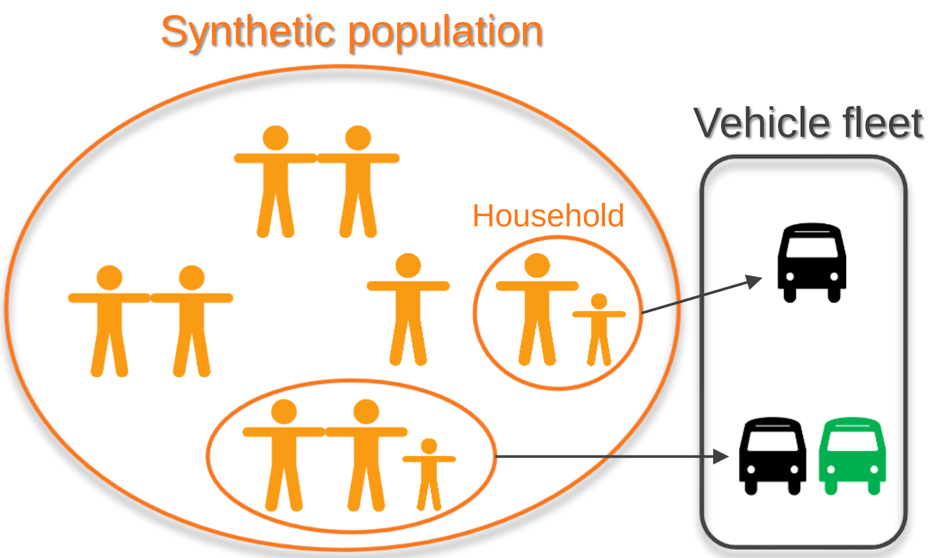


Figure : Vehicle fleet micro-representation based on households characteristics

Main objective : a microscopic and spatialized representation of **vehicle fleet** based on households characteristics of the synthetic population

Problem : which **models** and which **characteristics associated with households** would optimize the prediction of the vehicle fleet of a synthetic population ?

AI vs discrete choice methods

Methodology

Model types

Discrete choice method

Logistic regression

Supervised learning classification methods

Boosting methods

- Gradient boosting
- Ada Boost
- Light gradient boosting machine

Discriminant analysis

- Linear discriminant analysis
- Quadratic discriminant analysis

Decision trees methods

- Decision tree classifier
- Extra Tree Classifier
- Random Forest Classifier

Naive Bayes

Support vector machines

Nearest neighbors

Dummy classifier

Ridge classifier

Methodology

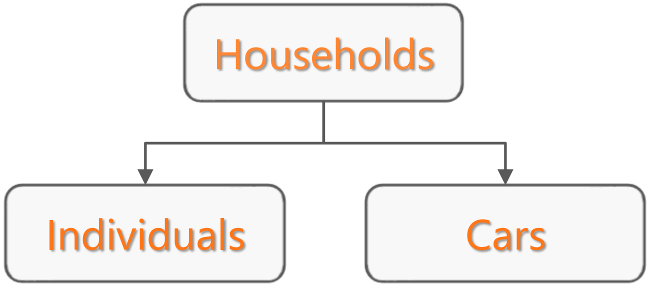
Evaluation metrics

Indicators of performance	Formula	Interpretation
Accuracy score	$\text{Acc} = \dfrac{\text{true predictions}}{\text{number of predictions}}$	Percentage of accurate predictions in the test sample, easy to interpret
Area under the curve (AUC) of the receiver operating characteristic (ROC)	$MAUC = \dfrac{2}{C(C-1)} \sum_{i<j} A(i,j)$	AUC converts ROC curve to a value in the range of [0.5, 1], where 1 means perfect classifier and 0.5 means no better than random classification. Multi-class AUC is average AUC of all pairs of classes.
F1-score	$\text{F1} = \dfrac{2}{\frac{1}{precision} + \frac{1}{recall}}$	F1 is the harmonic mean of the precision and recall, value in range [0,1]. Preferable to accuracy in case of a large class imbalance.
Cohen's kappa (κ)	$\mathcal{K} = \dfrac{\text{Acc} - P_e}{1 - P_e},$	Measure of agreement between observed and predicted or inferred classes for cases in a testing dataset, included in [-1,1]. If negative, a random classification is better.
Matthews Correlation Coefficient (MCC)	$MCC = \dfrac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$	For multi-class : average MCC of all pairs of classes Measure of agreement, included in [-1,1]. If negative, a random classification is better

Laboratoire Ville Mobilité Transport
Qi² Réseau de recherche Qualité de l'air en Ile-de-France
îledeFrance
École des Ponts ParisTech
Université Gustave Eiffel

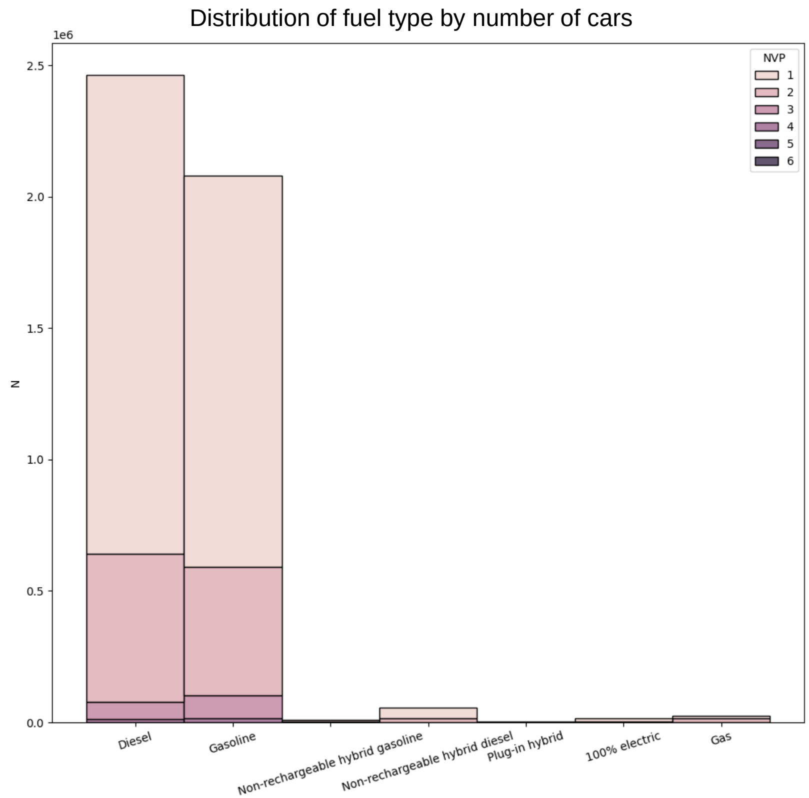
Data – GTS 2018

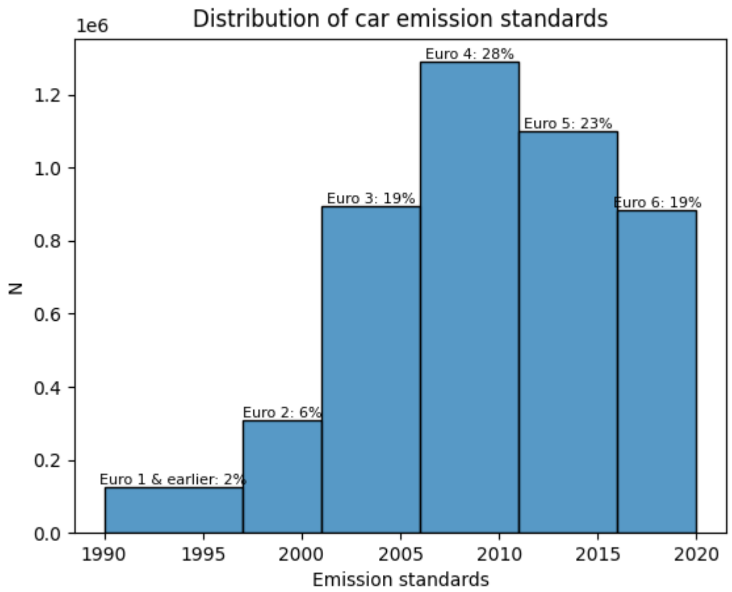
- Global Transport Survey, GTS (*Enquête globale de Transport, EGT*)
- Greater Paris metropolitan area (Île-de-France)
- Results for 2018 of the GTS 2018-2022



Key statistics

- 3,156 households
- 6,928 individuals
- 3,204 cars





Laboratoire Ville Mobilité Transport

Réseau de recherche Qualité de l'air en Ile-de-France

îledeFrance

École des Ponts ParisTech

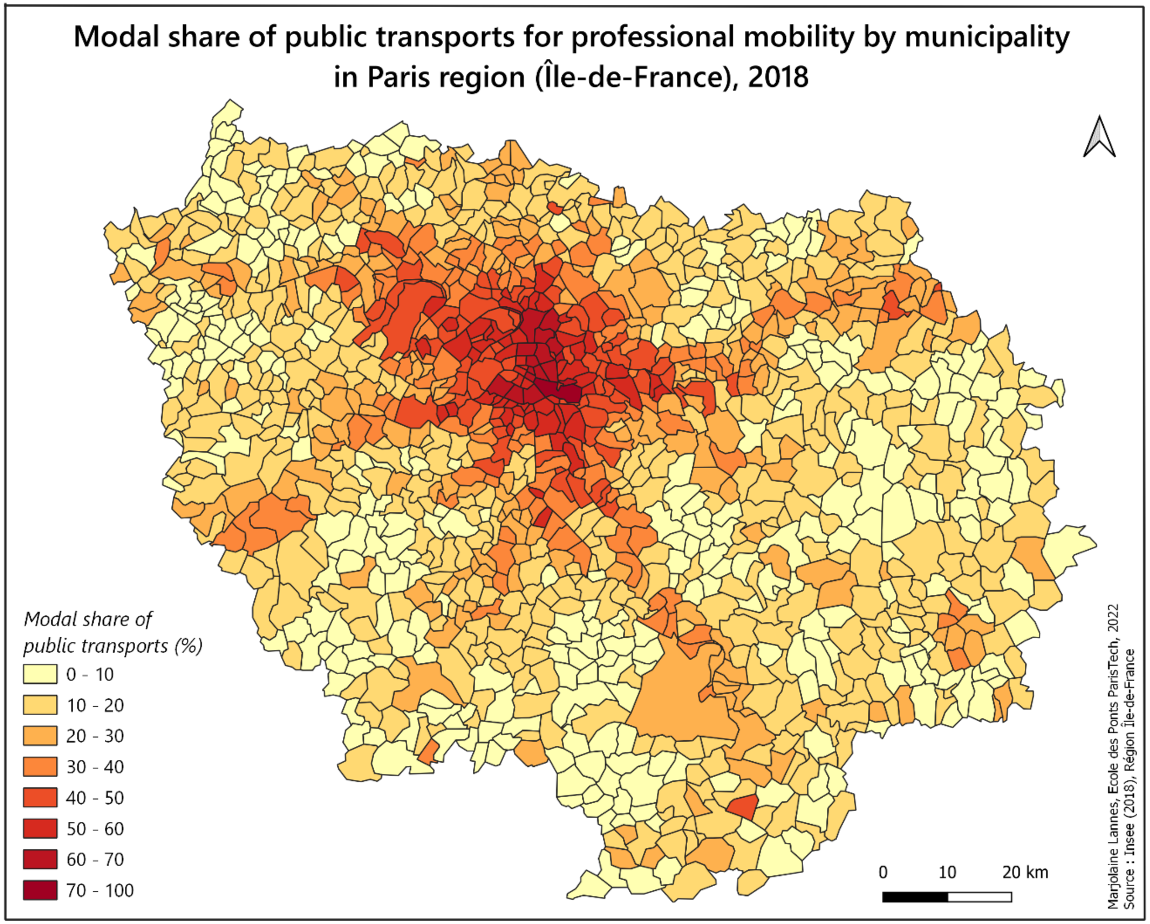
Université Gustave Eiffel

Data – *Public transport modal share*

An indicator of regional accessibility

Public transport modal share : share of residents in a city taking public transports to go to their workplace

Modal share calculated from national census for professional mobility i.e. MobPro survey (INSEE)



Laboratoire Ville Mobilité Transport

Data – Variables

Variable	Description	Variable types
Age	Class of age of the oldest individual of the household	Socio-economic
Income *	Logarithm of the household's income	Socio-economic
N_workers	Number of employed actives in the household	Socio-economic
Household_type	Type of household (couple with / without children, single man/woman, monoparental family mother/father)	Socio-economic
Housing_type	Type of housing (flat, house, others)	Socio-economic
PT_share	Share of home city residents taking public transports to go to their workplace	Build environment
PT_share_work	Share of workplace city residents taking public transports to go to their workplace	Build environment
Commuting_distance *	Maximum commuting distance within household	Build environment
Parking	Presence of a private parking at home	Build environment
Parking_at_workplace	At least one person in the household has parking at their workplace	Build environment
N_cars	Number of cars owned by the household	Predicted variable
Fuel_type	Fuel type of the vehicle	Predicted variable
Euro_norm	European emission standard of the car, depending on the year the car was first put on the road	Predicted variable

* : with the indicator variable which values 1 if the household responded to the question, 0 otherwise

Fuel type results

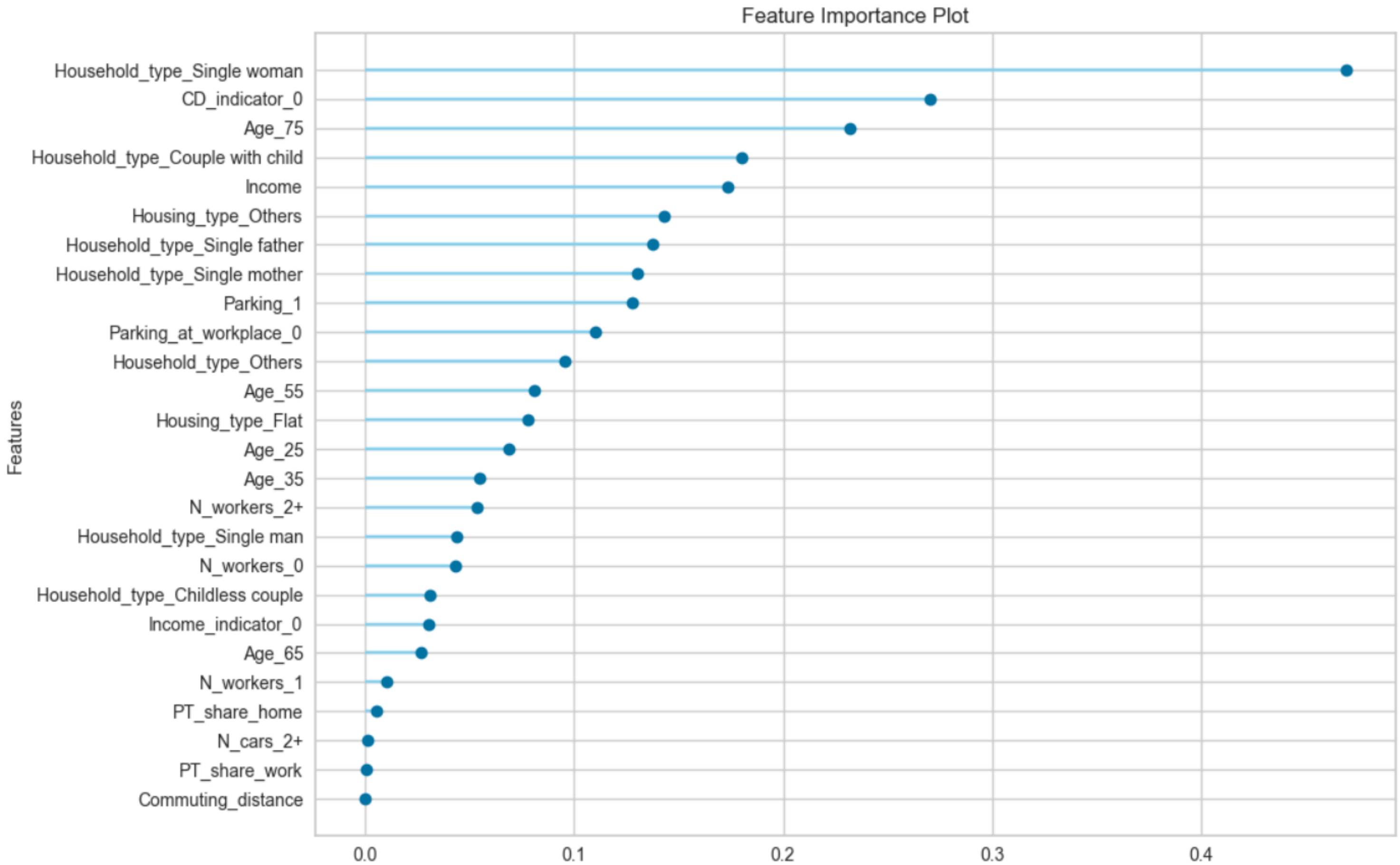
AI vs discrete choice model performance

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
Ridge Classifier	0.587	0.000	0.325	0.582	0.582	0.184	0.185	0.004
Linear Discriminant Analysis	0.586	0.450	0.325	0.582	0.582	0.184	0.185	0.007
Logistic Regression	0.586	0.451	0.325	0.581	0.582	0.182	0.183	0.189
Gradient Boosting Classifier	0.568	0.432	0.335	0.568	0.568	0.154	0.154	0.271
Light Gradient Boosting Machine	0.562	0.414	0.341	0.560	0.560	0.139	0.139	0.086
Random Forest Classifier	0.562	0.405	0.341	0.559	0.560	0.137	0.138	0.097
Extra Trees Classifier	0.558	0.393	0.340	0.556	0.556	0.131	0.131	0.082
Decision Tree Classifier	0.538	0.381	0.354	0.539	0.537	0.097	0.097	0.007
K Neighbors Classifier	0.511	0.377	0.284	0.506	0.508	0.033	0.033	0.011
Dummy Classifier	0.504	0.350	0.275	0.254	0.338	0.000	0.000	0.004
Ada Boost Classifier	0.480	0.357	0.292	0.569	0.505	0.104	0.113	0.032
SVM - Linear Kernel	0.479	0.000	0.279	0.570	0.417	0.057	0.071	0.01
Naive Bayes	0.158	0.369	0.141	0.631	0.247	0.036	0.054	0.008
Quadratic Discriminant Analysis	0.082	0.354	0.214	0.461	0.127	0.007	0.012	0.004

- **Ridge classifer** slightly outperforms linear discriminant and logistic regression
- **F1-score** is closer to 1 than to 0 for most classifiers, indicating a quite satisfying prediction
- **Cohen's kappa** indicates a slight, nearly fair for gradient boosting classifier, agreement : $\kappa \in [0 ; 0{,}20]$
- **Matthews Correlation Coefficient** (MCC) also reaches +0,185 > 0 for gradient boosting, attesting a slight agreement

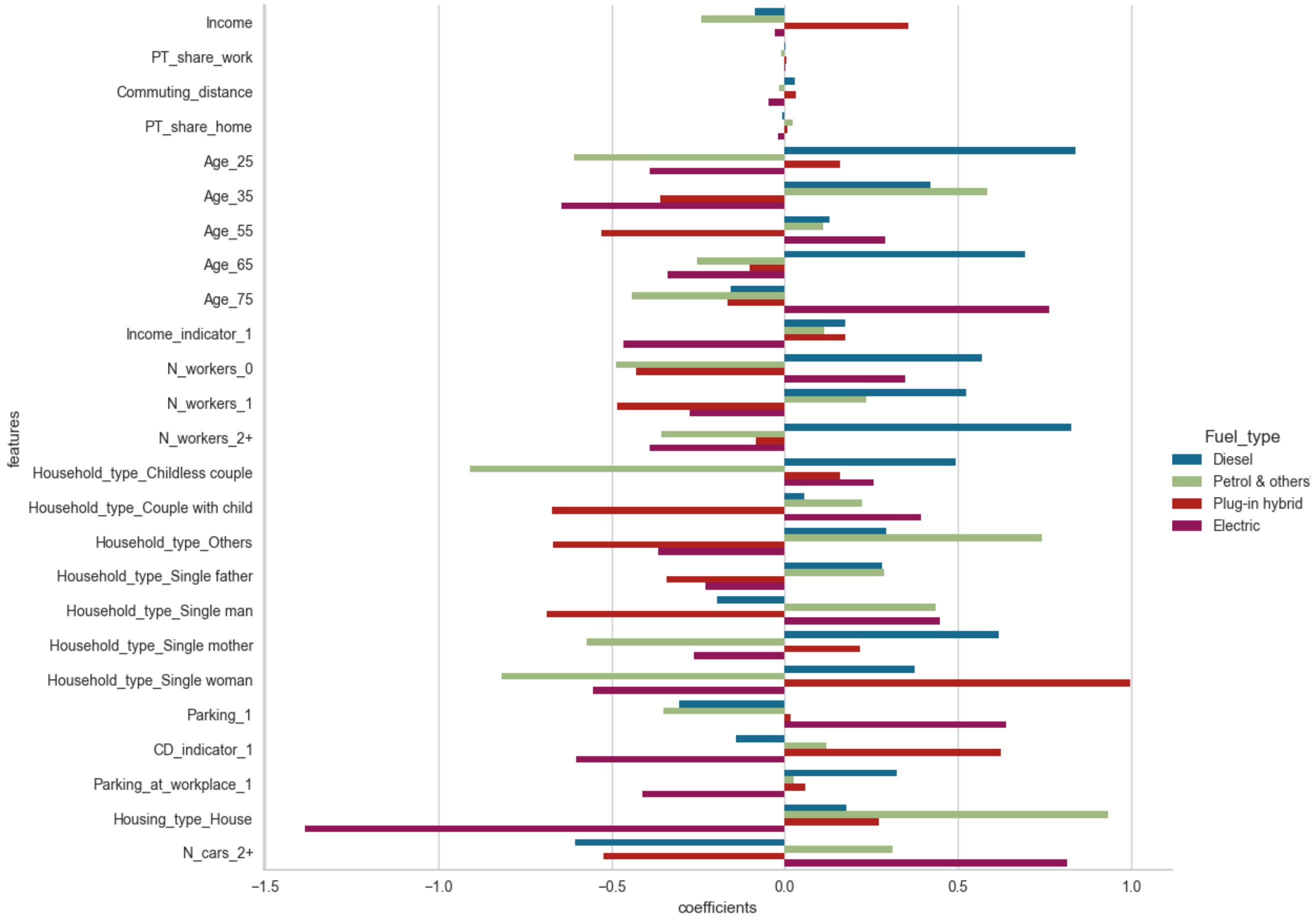
Fuel type results

Feature importance (ridge classifier)



Fuel type results

Logistic regression coefficients



Laboratoire Ville Mobilité Transport

Réseau de recherche Qualité de l'air en Ile-de-France

îledeFrance

École des Ponts ParisTech

Université Gustave Eiffel

Results synthesis

Decision type	Best model type	Model performance	Most important variables
Households car ownership	Gradient boosting	F1-score : 0,763 Cohen kappa : substantial, agreement ($\kappa = 0{,}629$) $\boldsymbol{MCC = +0,630}$ **: strong positive relationship**	Absence of parking at home Housing : flat Public transport share home Income Commuting distance Household : couple with child
Cars fuel type	Ridge classifier	F1-score : 0,582 Cohen kappa : slight agreement ($\kappa = 0{,}184$) $\boldsymbol{MCC = +0,185}$ **: slight agreement**	Household : single woman Commuting distance indicator Age (>75) Household : couple with child Income
Cars emission standard	Linear discriminant analysis	F1-score : 0,282 Cohen kappa : slight agreement ($\kappa = 0{,}053$) $\boldsymbol{MCC = +0,054}$ **: slight, negligible relationship**	Income No worker Commuting distance indicator One worker Household : single woman/man

Laboratoire Ville Mobilité Transport

Discussion

Results analysis

- Car ownership: results consistent with litterature (accessibility and built environment variables)
- Fuel type: sociodemographic variables (energy cost)
- Car age: income and commuting distance

Contribution of machine learning

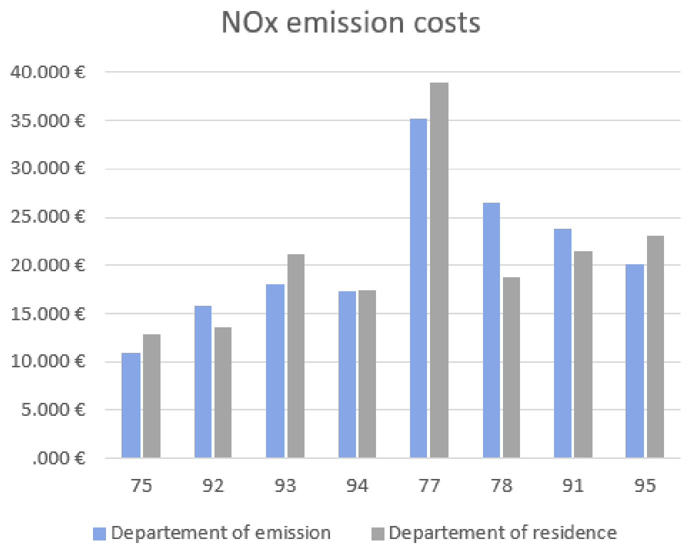
- For all decision levels, machine learning outperforms DCM
- MCC as evaluation metric: imbalanced dataset

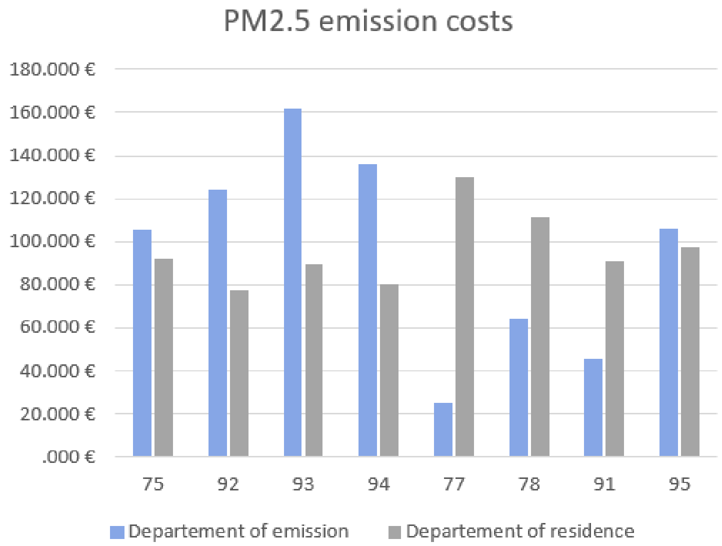
Outlook

- More data for better performance? (especially for underrepresented classes)
- Comparison with *Parc Auto* survey

Outlook

- **Integration of the model in MATSim for the calculation of emissions and exposures at the household level**
- PhD topic : *Modeling the exposure to air pollution in Île-de-France region: uncertainty analysis with a multi-agent approach*





Laboratoire Ville Mobilité Transport

Qi2 Réseau de recherche Qualité de l'air en Ile-de-France

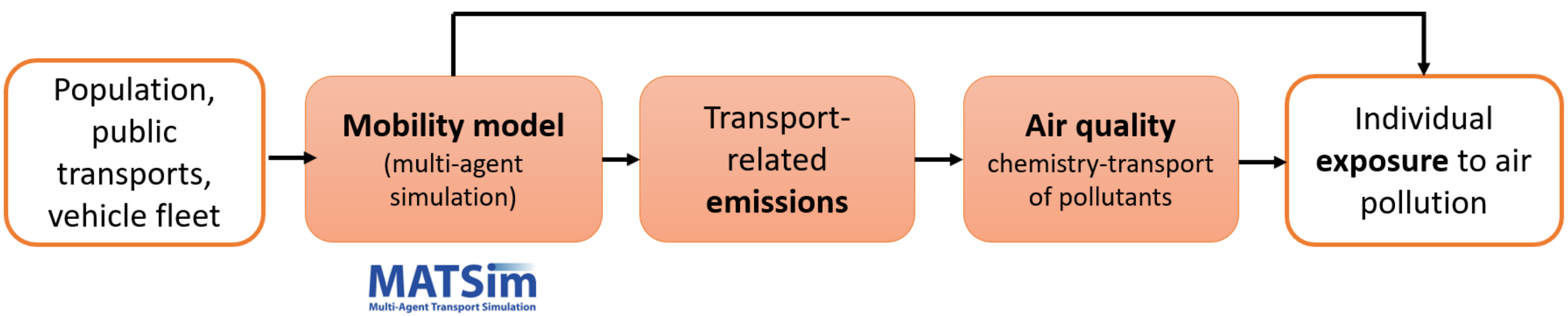
îledeFrance

École des Ponts ParisTech

Université Gustave Eiffel

Outlook

- Integration of the model in MATSim for the calculation of emissions and exposures at the household level
- **PhD topic :** ***Modeling the exposure to air pollution in Île-de-France region: uncertainty analysis with a multi-agent approach***



Thank you for your attention

Bibliography

Basu, R. and Ferreira, J.: Understanding household vehicle ownership in Singapore through a comparison of econometric and machine learning models, Transportation Research Procedia, 48, 1674–1693, https://doi.org/10.1016/j.trpro.2020.08.207, 2020.

Brownstone, D., Bunch, D. S., and Train, K.: Joint mixed logit models of stated and revealed preferences for alternative-fuel vehicles, Transportation Research Part B: Methodological, 34, 315–338, 2000.

Dixon, J., Koukoura, S., Brand, C., Morgan, M., and Bell, K.: Spatially Disaggregated Car Ownership Prediction Using Deep Neural Networks, Future Transportation, 1, 113–133, https://doi.org/10.3390/futuretransp1010008, number: 1 Publisher: Multidisciplinary Digital Publishing Institute, 2021.

Kaewwichian, P., Tanwanichkul, L., and Pitaksringkarn, J.: Car ownership demand modeling using machine learning: decision trees and neural networks, International Journal of GEOMATE, 17, 219–230, https://doi.org/https://doi.org/10.21660/2019.62.94618, 2019.

Mohammadian, A. and Miller, E. J.: Nested Logit Models and Artificial Neural Networks for Predicting Household Automobile Choices: Comparison of Performance, Transportation Research Record, 1807, 92–100, https://doi.org/10.3141/1807-12, publisher: SAGE Publications Inc, 2002.

Paredes, M., Hemberg, E., O'Reilly, U.-M., and Zegras, C.: Machine learning or discrete choice models for car ownership demand estimation and prediction?, in: 2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS), pp. 780–785, https://doi.org/10.1109/MTITS.2017.8005618, 2017.

Potoglou, D. and Kanaroglou, P. S.: Modelling car ownership in urban areas: a case study of Hamilton, Canada, Journal of Transport Geography, 16, 42–54, https://doi.org/10.1016/j.jtrangeo.2007.01.006, 2008.

Whelan, G.: Modelling car ownership in Great Britain, Transportation Research Part A: Policy and Practice, 41, 205–219, https://doi.org/10.1016/j.tra.2006.09.013, 2007.

Zambang, M. A. M., Jiang, H., and Wahab, L.: Modeling vehicle ownership with machine learning techniques in the Greater Tamale Area, Ghana, PLOS ONE, 16, e0246 044, https://doi.org/10.1371/journal.pone.0246044, publisher: Public Library of Science, 2021.

Appendix

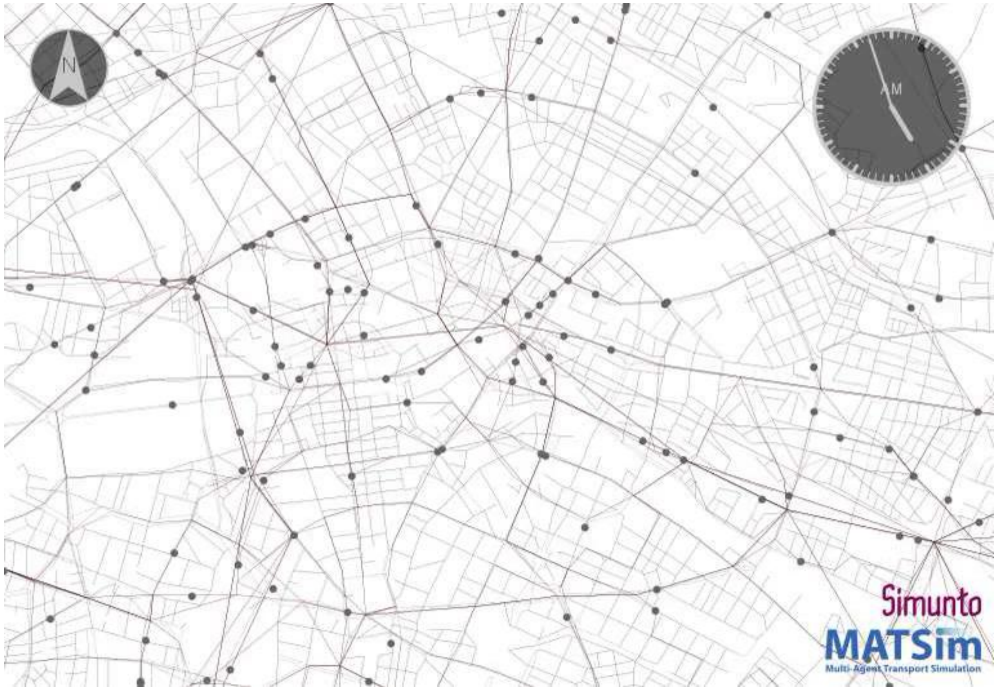
Context

Air quality challenge : mobility & emissions model

Context

- Public policies to **improve air quality** focus on vehicle fleet regulation, low emission zones
- Need to represent **vehicle types in mobility models** to precise and spacialize emissions by road trafic
- **Agent-based models** (ABM) for mobility modelling : require a synthetic population

Multi-Agent Transport Simulation (MATSim) : Berlin simulation



Main objective : a microscopic and spacialized representation of vehicle fleet based on households characteristics of the synthetic population

Laboratoire Ville Mobilité Transport

Qi 2 Réseau de recherche Qualité de l'air en Ile-de-France

îledeFrance

École des Ponts ParisTech

Université Gustave Eiffel

Literature review
Disaggregated car ownership choice modeling

<u>Discrete choice modeling (DCM) in transportation research :</u>

1973 : development of conditional multinomial logit (MNL) by McFadden

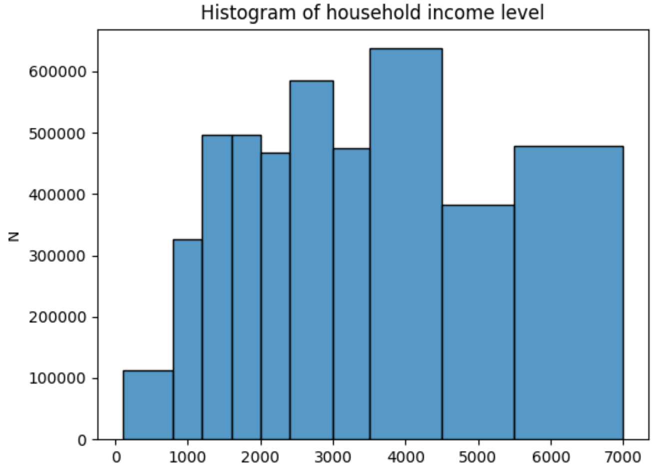
1975 : first application of MNL to car ownership by Lerman and Ben-Akiva

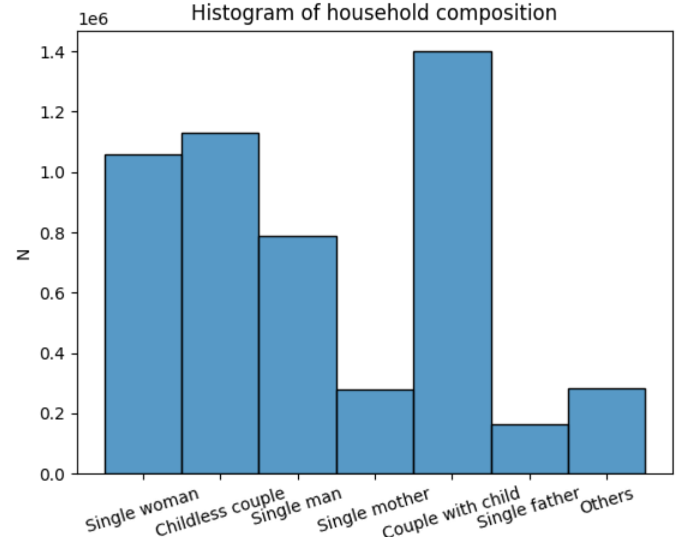
1980 : joint car ownership and mode choice DCM by Train

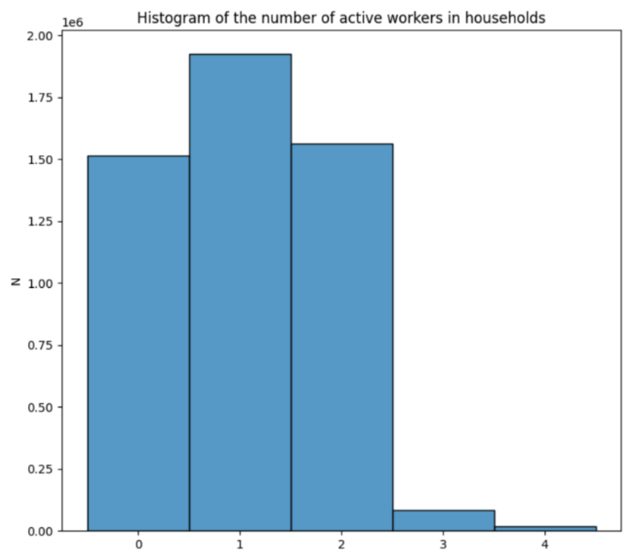
1985 : publication of « Discrete Choice Analysis: Theory and Application to Travel Demand », written by Lerman et. al.

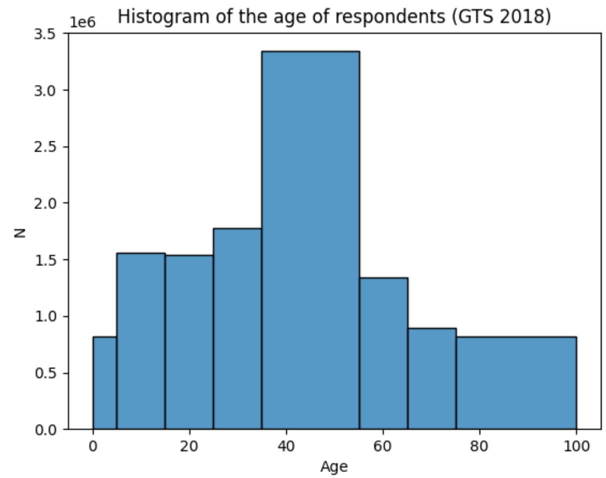
1994 : first study comparing mobility surveys and public census data in car ownership modeling by Purvis

Data – GTS 2018









Laboratoire
Ville
Mobilité
Transport

Qi 2
Réseau de recherche
Qualité de l'air
en Ile-de-France

îledeFrance

École des Ponts
ParisTech

Université
Gustave Eiffel

Car ownership results

AI vs discrete choice model performance

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
Gradient Boosting Classifier	0.766	0.897	0.766	0.763	0.763	0.629	0.630	0.201
Logistic Regression	0.755	0.893	0.759	0.753	0.752	0.614	0.616	0.450
Random Forest Classifier	0.752	0.892	0.747	0.749	0.749	0.605	0.606	0.077
Light Gradient Boosting Machine	0.751	0.888	0.749	0.750	0.749	0.605	0.605	0.049
Linear Discriminant Analysis	0.747	0.887	0.768	0.753	0.741	0.610	0.618	0.005
Extra Trees Classifier	0.746	0.879	0.745	0.743	0.743	0.597	0.598	0.082
Ridge Classifier	0.732	0.000	0.736	0.734	0.722	0.580	0.588	0.003
Decision Tree Classifier	0.688	0.743	0.685	0.690	0.688	0.504	0.505	0.014
Ada Boost Classifier	0.663	0.840	0.722	0.708	0.656	0.500	0.525	0.034
Naive Bayes	0.649	0.831	0.715	0.754	0.647	0.491	0.538	0.003
SVM - Linear Kernel	0.646	0.000	0.657	0.680	0.616	0.452	0.481	0.010
K Neighbors Classifier	0.554	0.693	0.521	0.549	0.546	0.271	0.274	0.013
Dummy Classifier	0.491	0.500	0.333	0.241	0.323	0.000	0.000	0.003
Quadratic Discriminant Analysis	0.381	0.557	0.430	0.502	0.349	0.113	0.148	0.006

- **Gradient boosting** slightly outperforms logistic regression and other artificial intelligence models
- **F1-score** is largely closer to 1 than to 0, indicating a satisfying prediction
- **Cohen's kappa** indicates a substantial agreement for gradient boosting: $\kappa \in [0{,}61 ; 0{,}80]$
- **Matthews Correlation Coefficient** (MCC) reaches +0,606 for gradient boosting indicating a strong positive relationship

Laboratoire Ville Mobilité Transport

Réseau de recherche Qualité de l'air en Ile-de-France

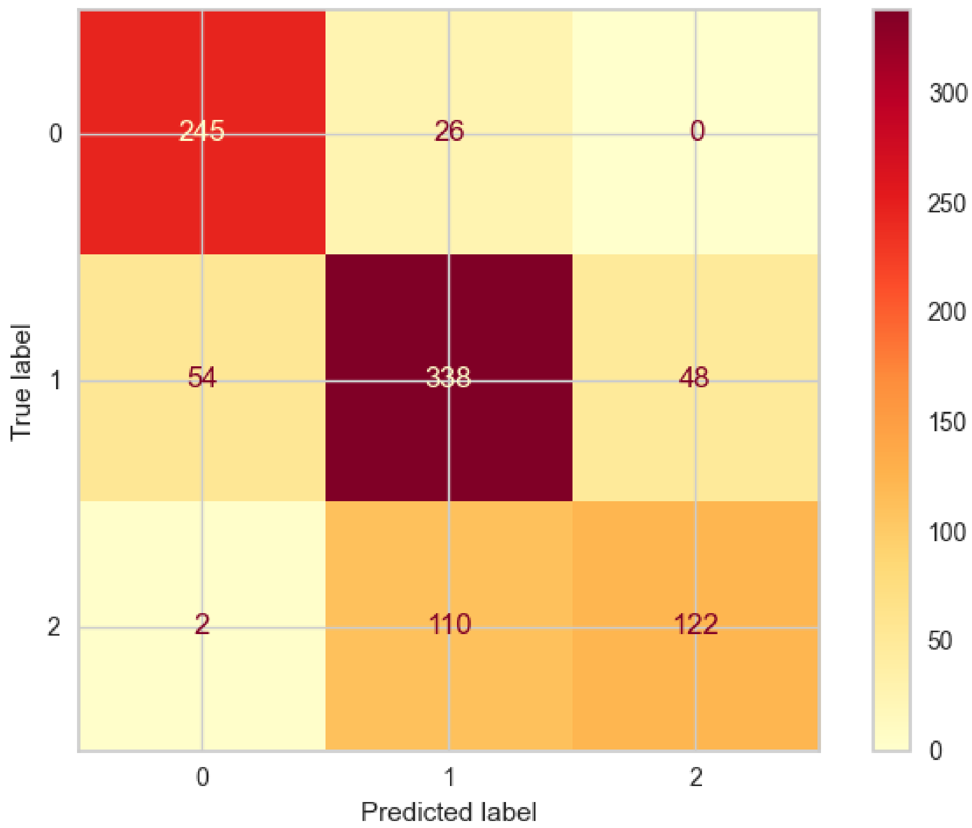
îledeFrance

École des Ponts ParisTech

Université Gustave Eiffel

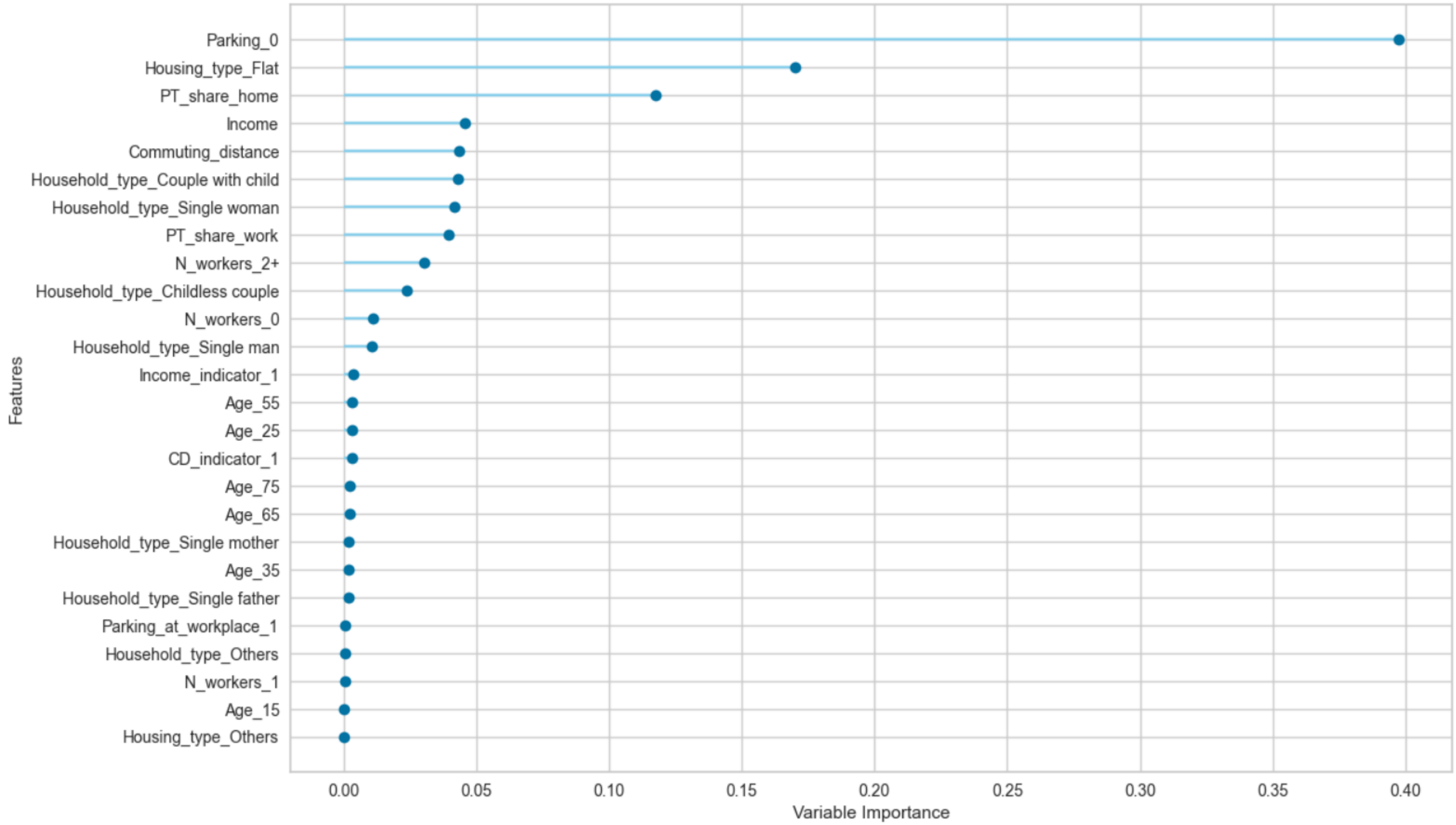
Car ownership results

Confusion matrix for gradient boosting classifier



Car ownership results

Feature importance (gradient boosting classifier)



Laboratoire Ville Mobilité Transport

Qi 2 Réseau de recherche Qualité de l'air en Ile-de-France

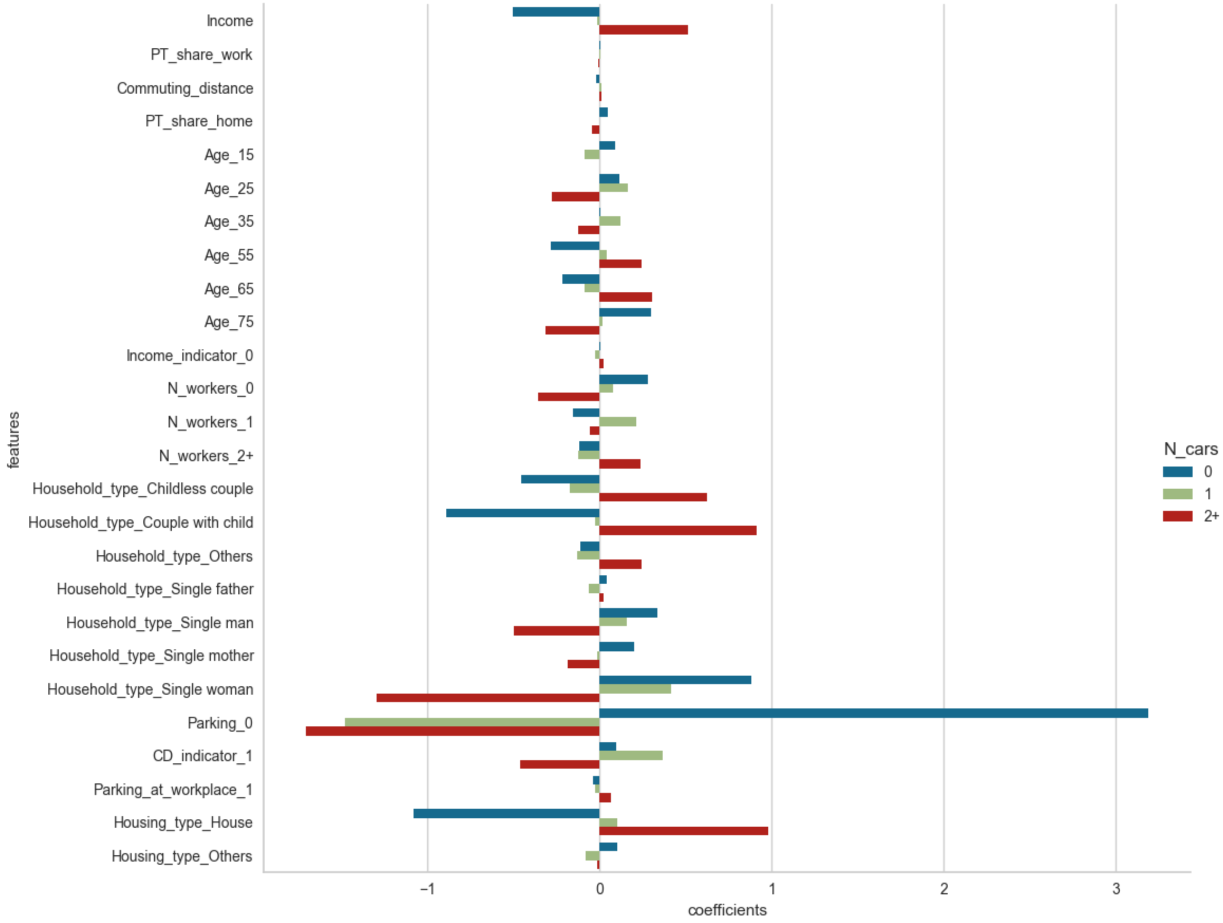
îledeFrance

École des Ponts ParisTech

Université Gustave Eiffel

Car ownership results

Logistic regression coefficients



Laboratoire Ville Mobilité Transport

Réseau de recherche Qualité de l'air en Ile-de-France

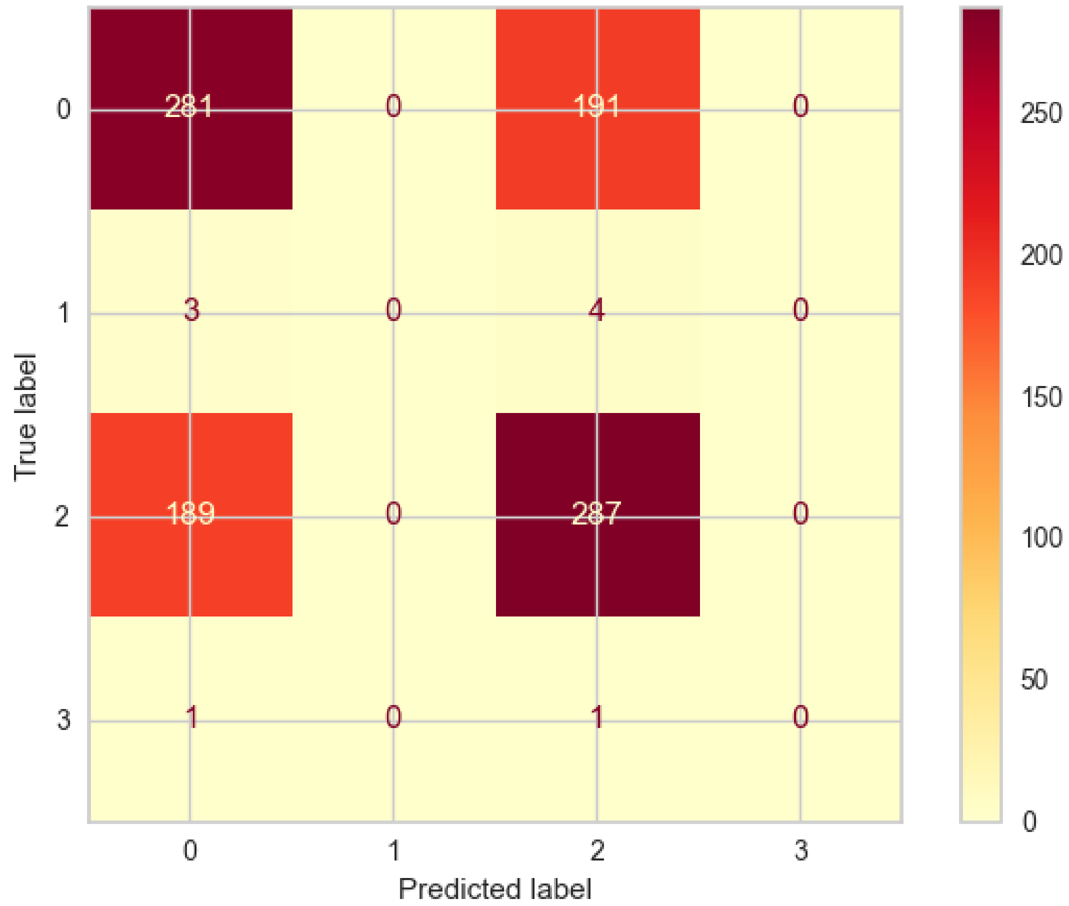
îledeFrance

École des Ponts ParisTech

Université Gustave Eiffel

Fuel type results

Confusion matrix for ridge classifier



Emission standards results

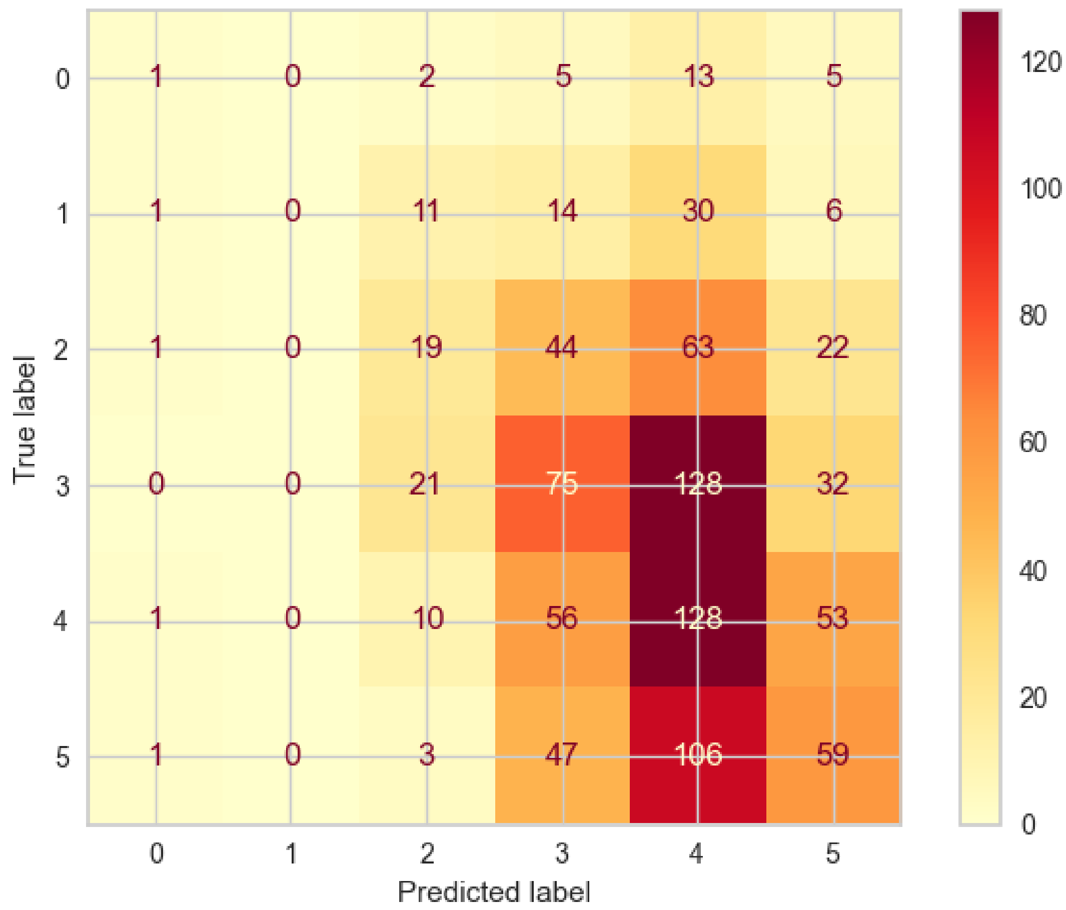
AI vs discrete choice model performance

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
Linear Discriminant Analysis	0.282	0.553	0.204	0.268	0.265	0.053	0.054	0.007
Logistic Regression	0.281	0.548	0.197	0.260	0.258	0.047	0.048	0.234
Ridge Classifier	0.282	0.000	0.193	0.253	0.258	0.046	0.047	0.004
Ada Boost Classifier	0.274	0.541	0.200	0.259	0.251	0.042	0.044	0.034
Light Gradient Boosting Machine	0.256	0.550	0.190	0.242	0.247	0.036	0.036	0.121
Gradient Boosting Classifier	0.261	0.538	0.183	0.238	0.244	0.031	0.032	0.405
Extra Trees Classifier	0.245	0.526	0.184	0.239	0.240	0.027	0.027	0.086
Random Forest Classifier	0.251	0.537	0.179	0.236	0.242	0.027	0.027	0.099
K Neighbors Classifier	0.240	0.513	0.181	0.238	0.236	0.023	0.023	0.014
Naive Bayes	0.099	0.523	0.186	0.244	0.107	0.014	0.023	0.008
Decision Tree Classifier	0.227	0.511	0.177	0.232	0.228	0.018	0.018	0.006
Quadratic Discriminant Analysis	0.174	0.503	0.179	0.239	0.147	0.006	0.006	0.004
SVM - Linear Kernel	0.216	0.000	0.178	0.146	0.120	0.003	0.004	0.015
Dummy Classifier	0.265	0.500	0.167	0.070	0.111	0.000	0.000	0.003

- **Linear discriminant analysis** slightly outperforms logistic regression
- **F1-score** is closer to 0, indicating an unsatisfying prediction
- **Cohen's kappa** indicates a slight agreement : $\kappa \in [0{,}10\ ; 0{,}20]$, except for SVM, dummy and naive Bayes models (to exclude)
- **Matthews Correlation Coefficient** (MCC) also indicates a negligible relationship, slightly better than random

Emission standards results

Confusion matrix for linear discriminent analysis



<u>Class equivalent : construction year</u>

0 : before 1997
1 : 1997-2000
2 : 2001-2005
3 : 2006-2010
4 : 2011-2015
5 : since 2015

Laboratoire Ville Mobilité Transport

Réseau de recherche Qualité de l'air en Ile-de-France

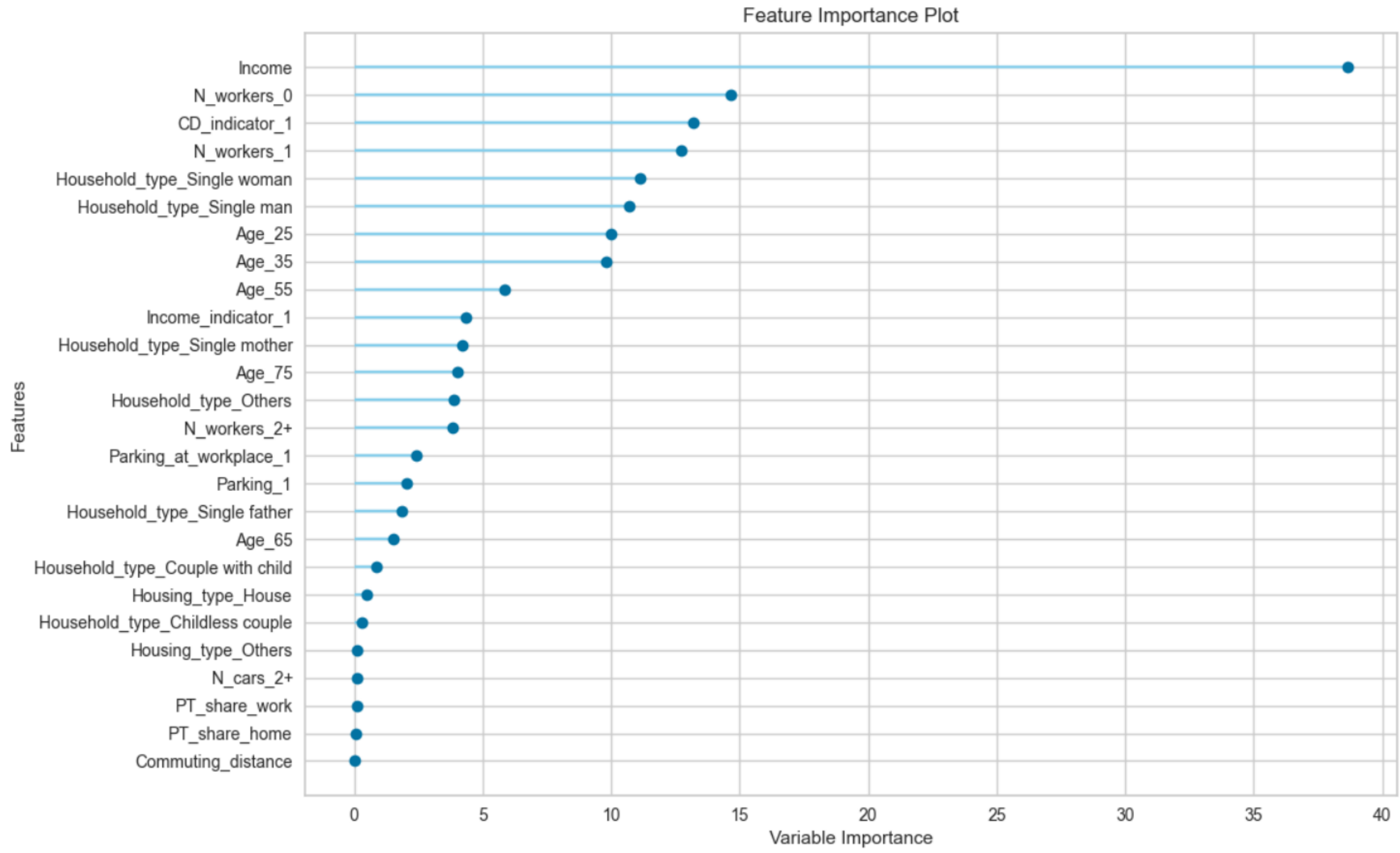
îledeFrance

École des Ponts ParisTech

Université Gustave Eiffel

Emission standards results

Feature importance (linear discriminent analysis)



Laboratoire Ville Mobilité Transport

Réseau de recherche Qualité de l'air en Ile-de-France

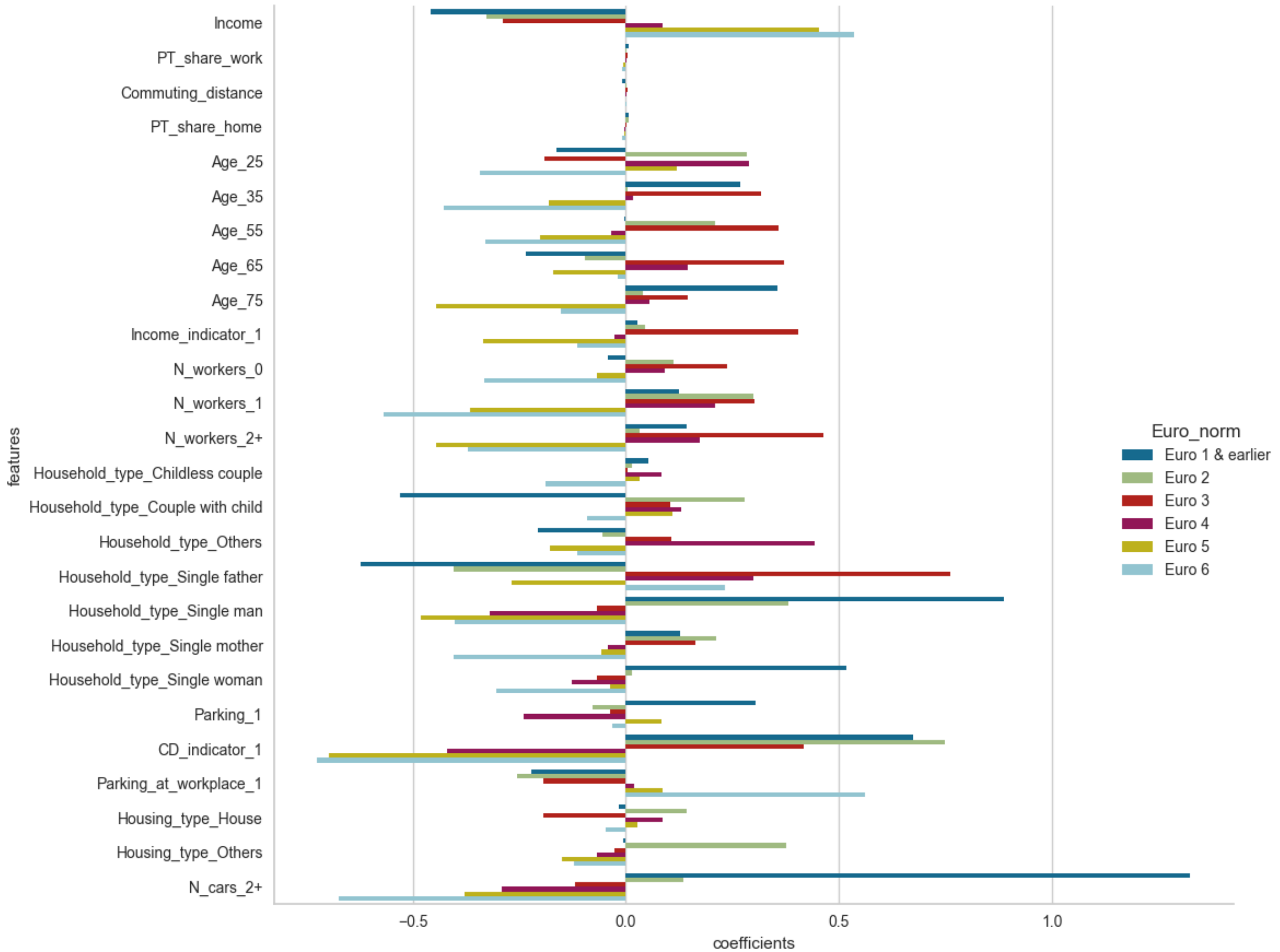
îledeFrance

École des Ponts ParisTech

Université Gustave Eiffel

Emission standards results

Logistic regression coefficients



Laboratoire Ville Mobilité Transport

Qi2 Réseau de recherche Qualité de l'air en Île-de-France

îledeFrance

École des Ponts ParisTech

Université Gustave Eiffel