Romanian Journal of Regional Science
Vol. 17, No. 2, Winter 2023

DOI: https://doi.org/10.61225/rjrs.2023.06

A NON-LINEAR AND INTERACTION EFFECT ANALYSIS OF DISTANCE AND TRANSPORT ACCESSIBILITY ON BICYCLE USE: THE EXAMPLE OF THE UNIVERSITY STAFF IN LYON (FRANCE)

Mehmet Güney Celbiş[a], Nathalie Havet[b], Louafi Bouzouina[c,]*

[a] Urban Planning Economics and Transport Laboratory – LAET-ENTPE, CNRS, University of Lyon, France and UNU-MERIT, Maastricht, The Netherlands
[b] Urban Planning Economics and Transport Laboratory – LAET-ENTPE, CNRS, LSAF, University of Lyon, France
[c] Urban Planning Economics and Transport Laboratory – LAET-ENTPE, CNRS, University of Lyon, France
* Corresponding author
Address: LAET-ENTPE, 3 Maurice Audin, 69518 Vaux-en-Velin Cedex, France
E-mail: louafi.bouzouina@entpe.fr

Biographical Notes

Mehmet Güney CELBİŞ is a research associate professor of sustainable development at LAET-Transportation Urban Planning Economics Laboratory, the Graduate school of Civil, Environmental and Urban Engineering (ENTPE), University of Lyon, France, and an affiliated researcher at UNU-MERIT, Maastricht, the Netherlands. His research is centered on investigating urban mobility and spatial relationships through the use of interpretable machine learning tools. ORCID: https://orcid.org/0000-0002-2790-6035

Nathalie HAVET is a professor of economics at LAET-Transportation Urban Planning Economics Laboratory, the Graduate school of Civil, Environmental and Urban Engineering (ENTPE), University of Lyon, France. Her research interests revolve around topics of applied microeconometrics and public policies. They include daily mobility, particularly focusing on linking labour market. ORCID ID: https://orcid.org/0000-0001-7454-8771

Louafi BOUZOUINA is a research professor of sustainable development, LAET-Transportation Urban Planning Economics Laboratory, the Graduate school of Civil, Environmental and Urban Engineering (ENTPE), University of Lyon, France. His research work focuses on metropolitan dynamics, daily mobility and urban segregation, dealing with interactions between the locations of households and businesses, land use, accessibility and uses of transport with an aim to model and evaluate urban policies with the constraint of sustainability. ORCID ID: https://orcid.org/0000-0001-8975-5124

Abstract

This study explores the individual and spatial level determinants of the adoption cycling as a commuting mode by university staff members using data from Lyon, France (the MobiCampus-UdL survey). The empirical approach of the study is centered on the use of a gradient boosting machine prediction implemented using the XGBOOST framework, followed by the use of an interpretable machine learning method, namely Shapley Additive exPlanations (SHAP). We uncover various complex interactive and nonlinear relationships among model features and a binary outcome of being or not being a bike user for commuting. Our main findings suggest that policies designed towards broadening individual access to bicycles through ownership or sharing, in addition to the provision of shared cycle networks within 7 km of major employment centres can increase the adoption of cycling by commuters. Furthermore, among other results, we also observe that promoting regular teleworking among university staff, particularly for those who live at a distance more than 5 km of their place of work, could encourage commuting by bike. We also observe that cycling and public transport become complementary modes when home-work distances are greater that about 7 km.

Keywords: home-campus mobility; bicycle use; university staff; non-linear effects; bike-sharing accessibility; Machine learning
JEL Classification: R40, R58, C14.

1. Introduction

In November 2021, the city of Lyon, France, made 10,000 shared bikes available for free of charge to socially disadvantaged young people aged 18 to 25 (Dimitrova, 2021). Given the relatively low level of bicycle ownership in France within the EU, the move marks an increased commitment by policymakers to support sustainable mobility in the post-Covid framework. The increased willingness to use bicycles as an urban mobility mode during the pandemic offers new opportunities to expand now bicycle use among urban populations (Azevedo et al., 2023; Nikiforiadis et al., 2020). In particular, university environments seem to be privileged places to promote cycling, due to their spatial arrangements, demographic profiles, and cultural settings (Mateo-Babiano et al., 2020). Moreover, university communities are subpopulations that are more likely than average to favour cycling for commuting (Balsas, 2003; Van den Berg and Russo, 2017). In this respect, the bicycle friendliness of university environments can be harnessed and used as a catalyst and driver of change towards more sustainable travel behaviour and commuting modes in wider urban areas such as Lyon (Balsas, 2003; Kelarestaghi et al., 2019).

Given the benefits of daily cycling that have been highlighted in the literature in terms of health, reduction of CO2 emissions, air pollution, road and public transport congestion and cognitive performance (e.g. Prince et al., 2021; Neves and Brand, 2019; EDF, 2016; Martens, 2004; Andersen et al., 2000), the real challenge for public policy makers is to identify ways to increase its use. For example, would it make more sense to encourage households to own a bicycle, to improve the accessibility to bike-share stations where people live and/or work, to reduce the supply of other modes of transport or, on the contrary, to strengthen and adapt intermodality? These questions arise for the general population, but especially for those populations that are the most likely to change their transport habits, such as the university community (Wilson et al., 2018). Understanding the factors that currently influence cycling is necessary in order to design effective pro-cycling transport policies for the future.

Our study focuses on the bicycle use for commuting by staff members of the university community in the Lyon metropolitan area. We chose this target population because university staff members are demographically similar to non-university urban commuters, but also work in a bicycle-friendly setting. Furthermore, transport habits are generally observed more in the context of commuting rather than for other purposes (Rondinella, 2015). In order to explore and identify the determinants of the bicycle use for commuting for this particular group, we carried out three surveys in 2017, 2018, and 2019, as part of the MobiCampus-UdL project. The sample consists of the staff members from around the twenty campuses of the University of Lyon (UdL). We implement randomized and sequential machine learning (ML) ensemble models based on weak learners performing recursive binary partitioning (i.e. individual regularized classification trees). Through the use of interpretable ML tools based on the computations of Shapley Additive Explanations (SHAP) values, we investigate – among other results – the interactive and non-linear relationships algorithmically learned from the MobiCampus-UdL surveys. Indeed, machine learning is useful for finding complex non-linear relationships and capturing the hidden patterns of variables because, unlike traditional statistical models, machine learning has no prior assumption that data must follow a particular distribution (Ji et al., 2022). Furthermore, while many studies have concluded that the longer the trip distance, the lower the share cycling in the mode choice for commuting, the interaction effects between distance and other variables on cycling behaviour have rarely been evaluated. Consequently, we explore in particular the complex interactions between distance and accessibility to different modes close to home or on campus. Our findings can provide valuable insights for promoting and improving the development of urban cycling in France. They also complement the scarce quantitative research on cycling in France (see, for example, Papon, 2003; Jensen et al., 2010; Héran, 2015; Raux et al., 2017).

The remainder of the paper is organised as follows. Section 2 describes the MobiCampus-UdL travel surveys and methods used in the research, and the main results are discussed in Section 3. Section 4 concludes the paper with policy and research implications.

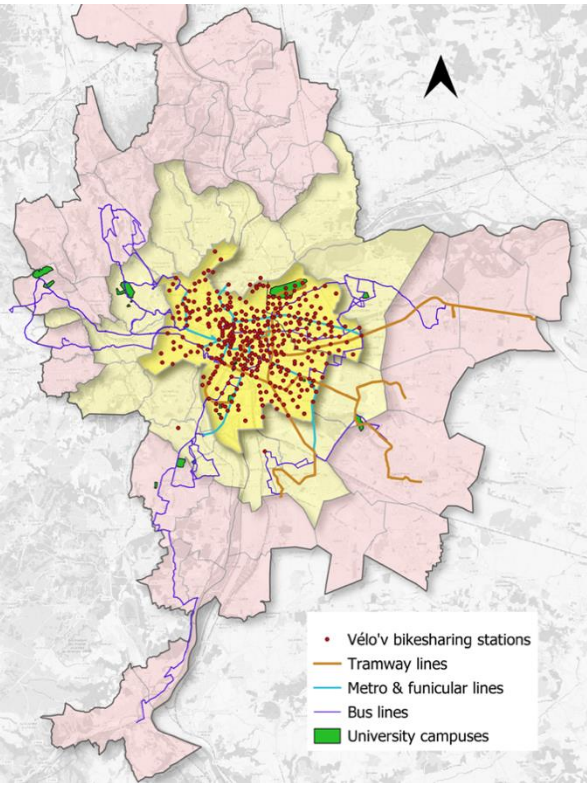
2. Data and methodology

2.1 The MobiCampus-UdL Travel Surveys

We use data from the three web-based MobiCampus-UdL travel surveys conducted in March-April 2017, 2018 and 2019 among the staff of the "University of Lyon" (UdL), an academic community of 26 institutions dedicated to education and research in the Greater Lyon area (Metropole de Lyon, France). The UdL has 7,000 researchers and 3,000 administrative and technical staff spread over 21 campuses and university sites. Like many universities in France and around the world, the UdL authorities have been working for several years on Mobility Plans to improve the accessibility of the campuses and university sites in a metropolitan area of more than 1.4 million inhabitants, while reducing the use of cars, which are more polluting and take up more space, in favour of alternative modes of transport. A recent UdL assessment showed that, on average, home-to-work travel accounts for 14% of the carbon footprint of its institutions (Université de Lyon, 2022). In this context, the MobiCampus-UdL project, carried out by the Transport Urban Planning Economics Laboratory with the financial support of the UdL and the Lyon Urban Community, aims to understand the daily travel behavior of the university community in order to help campus managers to plan coherent mobility management policies.

Of the 26 UdL institutions, 17 participated in the three waves of the survey (2/3 of the UdL population). The questionnaires were sent to their staff via institutional e-mail addresses and included different categories of questions: i) *their occupation* (type of institution (university vs. business and engineering school), type of job (researcher vs. administrative-technical staff), usual working patterns and hours; teleworking habits); ii) *usual mode(s) travel to and from the campus and transport membership and ownership* (bicycle/car ownership, public transport membership, bike-sharing membership); iii) *general socio-economic and demographic characteristics* such as age, gender, marital status (single, couple with/without children), household income class, and place of residence (city centre, inner suburbs, outer suburbs, out of agglomeration; distance from home to campus). A fourth category of information on *individual accessibility* to different modes of transport for home and work (campus) was constructed from the respondents' precise locations on interactive maps and the infrastructure network information. The Lyon agglomeration has more than 120 bus lines, 4 metro lines (32km), 7 tram lines (66km), 2 funiculars, 6 railway stations and more than 400 bike-sharing stations with about 5,000 bicycles in operation. However, the level of public transport is uneven across the 21 campuses; most bike and public transport stations are located in the city centre (Figure 1). We therefore constructed dummy variables indicating whether the respondent's home (or workplace) was less than one kilometre from a train station, close to a public transport station (i.e. less than 500 metres from a metro station, less than 400 metres from a tram stop or less than 300 metres from a bus stop), and less than 300 metres from a bike-share station.

Figure 1. Campuses of the University of Lyon (UdL) and public transport lines and self-service bike stations in Lyon agglomeration



Source: own processing.

3,670 people from the 17 participating institutions completed the questionnaire, giving a response rate of 30%. Table 1 presents the sample characteristics. Half of the university staff respondents are teachers and researchers and half are administrative and technical staff. A significant proportion telework at least occasionally, with 17% doing so regularly, despite our pre-Covid study period. Of these, 47% live in the city centre, 14% in the inner suburbs, 11% in the outer suburbs and almost 28% outside the agglomeration. 29.5% of them live less than 5km away from their place of work and more than 15% live more than 30km away. The respondents' places of residence are well served by public transport and bike-sharing networks: 71% live less than 300m from a public transport station and 44% live less than 1km from a bike-sharing station. The accessibility of these modes of transport is even higher for their place of work: 92% and 70% respectively work close to a public transport station and a bike-sharing station. Accessibility to the train is much lower (28% for home and 13% for work). In terms of means of transport, a large majority of university staff have a driving licence and own a car (76%). In comparison, only 29% own a private bicycle and 11% have a subscription to the shared bicycle network. However, owning a mode of transport does not necessarily mean using it for commuting. In our sample, 15.7% of employees commute by bicycle[1], which is a much higher cycling rate than in the general French working and student population. In 2021, only 4.5% of French people who commute to work or study on a daily basis report that they have used a bicycle for all or part of their journey. [2]

Table 1. Descriptive statistics

Variable Code	Variable Description	Frequency
Bikeuse	Use of bicycle for commuting	15.7%
	Gender:	
Male	Men	52.8%
	Women	47.2%
Age	Age:	
	under 35 years old	19.3%
	35 to 44 years old	28.3%
	45 to 54 years old	32.2%
	55 years old and over	20.2%
Maritalstatus	Marital status:	
	Couple with children	50.4%
	Couple without children	23.6%
	Single	19.3%
	Single with children	6.7%

[1] This figure includes not only the exclusive use of the bicycle to get to work, but also use in combination with one or more other modes of transport.

[2] https://www.notre-environnement.gouv.fr/themes/amenagement/transport-et-mobilite-ressources/article/les-francais-et-le-velo-en-2022#Part-du-velo-comme-mode-de-deplacement-principal-pour-se-rendre-au-nbsp

Income_cat	Household income per month: Less than 2,000 euros [2,000 , 3,000[euros [3,000 , 4,000[euros [4,000 , 6,000[euros 6,000 euros and above	 15.6% 18.6% 19.0% 26.4% 11.1%
Type_etab	Type of institutions: University Business and engineering school	 48.6% 51.5%
Profession	Occupation: Lecturer-Researcher Technical and administrative staff	 48.0% 52.0%
Teleworking	Teleworking: Yes, sometimes Yes, regularly No	 29.0% 17.1% 53.9%
Zone_resi	Home location: City Centre (Lyon, Villeurbanne) Inner suburbs Outer suburbs Outside the agglomeration	 47.1% 13.7% 11.4% 27.7%
Distance	Home-campus distance: Less than 1 km 1 to 3 km 3 to 5 km 5 to 15 km 15 to 30 km 30 to 50 km 50 km and more	 2.9% 12.4% 14.2% 38.5% 16.6% 8.0% 7.4%
Drivingcarown	Driving licence and car ownership	76.1%
Bikeown	Own a bike	28.6%
BikeShM	Bike sharing membership	11.4%
PubTransM	Public transport membership	37.2%

PCarsh	Practicing car-sharing	1.9%
PCarpool	Practicing car-pooling	19.6%
AccSSB_res	Accessibility to self-service bicycle station (residence)	44.4%
AccPubT_res	Accessibility to public transport station (residence)	70.6%
AccTrain_res	Accessibility to train station (residence)	28.1%
AccSSB_campus	Accessibility to self-service bicycle station (campus)	69.6%
AccPubT_campus	Accessibility to public transport station (campus)	91.6%
AccTrain_campus	Accessibility to train station (campus)	13.0%
Arrivaltime	Usual arrival time (campus)	
	Before 7.30 am	12.2%
	Between 7.30 and 8.00 am	21.4%
	Between 8.00 and 8.30 am	20.6%
	Between 8.30 and 9.00 am	26.6%
	Between 9.00 and 9.30 am	13.7%
	Between 9.30 and 10.0 am	3.2%
	After 10.0 am	2.4%
Departuretime	Usual departure time (campus)	
	Before 5.00 pm	16.4%
	Between 5.00 and 5.30 pm	19.9%
	Between 5.30 and 6.00 pm	20.7%
	Between 6.00 and 6.30 pm	19.4%
	After 6.30 pm	23.7%
	Number of observations	3,670

Source: MobiCampus-UdL Travel surveys (2017, 2018, 2019) among the university staff, own calculations.

2.2 Methods

We investigate the impact of individual characteristics, distance from home to campus, and accessibility to different modes of transport on the choice of cycling to work. We employ a gradient boosting machine (GBM), developed by Friedman (2001, 2002) through the application of the Extreme Gradient Boosting (XGBoost) algorithm by Chen and Guestrin (2016) due to its advantages in highlighting the most important factors affecting individuals' choices and the overall impact on the outcome, and in accurately recognizing non-linear relationships (James et al., 2013; Varian, 2014). Furthermore, tree-based machine learning algorithms are often difficult to interpret due to their non-parametric and prediction-oriented structures. In order to ensure informative statistical inference, our empirical approach is particularly centered on interpretable ML techniques implemented ex-post on the predictions generated by the tree-based approaches. More specifically, in terms of model interpretation, we take advantage of the SHapley Additive exPlanation (SHAP) technique (Ribeiro et al., 2016; Strumbelj and Kononenko, 2014), developed by Lundberg and Lee (2017), which is based on the Shapley values introduced by Shapley (1953) in the framework of cooperative game theory. These techniques are model agnostic and focus on retrospectively explaining the results of the original predictive models through a series of operations using different randomisation procedures in an iterative manner for the purpose of computing feature value effect sizes and directions. Thus, the empirical analysis consists of two steps: prediction and interpretation.

Extreme Gradient Boosting (XGBoost) and prediction

The prediction step is based on the XGBoost algorithm, an efficient implementation of gradient boosted decision trees, developed by Chen and Guestrin (2016). In practice, the individual predictions for the test sample – randomly drawn from the full dataset – are done through building boosted trees based on the recursive binary partitioning algorithm of Breinan et al. (1984). However, the principle of the gradient boosted decision trees is to generate sequential regression trees, where each decision tree learns from the previous tree and affects the next tree to improve the model (by reducing the prediction errors of the previous trees) and build a strong learner (Friedman, 2001; Friedman, 2022). The superiority of XGBoost lies in several innovations over gradient boosting, including a regularised learning objective, optimization in storage and computation, and randomisation. For example, the XGBoost algorithm allows for cross-validation for regularisation and determination of the optimal model parameters such as the learning rate, i.e. the step size for revising the predictions for each observation.

In XGBoost, a suitable parameter set needs to be selected to maximise model performance: parameter tuning is important to prevent overfitting and to improve generalization ability. In particular, overfitting occurs when a model starts to learn noise and random fluctuations and eventually treats them as meaningful facts or concepts. The *learning rate* makes the model more robust by shrinking the weights at each step; the *maximum depth* of the tree represents the maximum number of splits and its higher value could cause overfitting; the *subsample* is the random fraction of the training data prior to growing trees, and its lower value makes the model more conservative and reduces overfitting, but too small values could lead to underfitting; *the number of trees* denotes the number of iterations of the model. In this study, XGBoost is trained on 70% of our initial sample that is randomly selected, and the remaining 30% is used to test the model. We apply a hyperparameter procedure to determine the optimal combination of these parameters. The number in brackets are the values used to construct the hypergrid, while the numbers in bold on the left-hand side are the parameters corresponding to the minimum mean log-loss of the gradient boosting machine's prediction of the validation sets, that are put apart within a 10-fold cross-validation procedure[3]:

- Learning rate: (0.005, .001, .01, .05, .1), **0.005**
- Maximum tree depth: (5, 10, 15), **5**
- Minimum number of observations in terminal nodes: (5, 10, 15), **5**
- Subsample ratio in each iteration: (.5, .7, .9, 1), **1** (i.e. the gradient boosting model is non-stochastic)
- Feature subsample ratio in each iteration: (.5, .7, .9, 1), **0.9**
- Number of trees: 1,247

Model interpretation

As the non-parametric XGBoost models do not generate predictions and effect size estimates simultaneously, we perform additional retrospective analyses on their output. SHapley Additive exPlanations (SHAP), as proposed by Lundberg and Lee (2017), is used to interpret the outputs of

[3] The R package for XGBoost and the cross-validation procedure were developed by Chen et al. (2019). A 10-fold cross-validation procedure is used on the training data to test the stability of the model performance, i.e. the training data is randomly divided into ten subsamples, and ten models are trained in such a way that each time nine subsamples are used to train a model and one subsample is used to test a model.

the models. The calculation and observation-specific representation of SHAP values provides a means to explore the strength of the relationships among model variables (i.e., to estimate the contribution of each variable) as well as their directions (Lundberg and Lee, 2017; Molnar, 2019).

A major advantage of SHAP values over ceteris paribus effect sizes (i.e. estimated elasticities) is that they are computed by replacing randomly selected subsets of model features for each individual in the training data through a random sampling of feature values from within the same dataset (Lundberg and Lee, 2017; Molnar, 2019). Therefore, for each person in our training dataset the effect of a given feature value on the outcome predicted for that person is calculated as ‘mutatis mutandis’ as opposed to ceteris paribus, because the remaining feature values for that individual are randomly replaced multiple times (for combinations of random subsets of variables to be replaced) and the effect is recalculated. In other words, the effect is calculated by allowing everything else to change, rather than keeping the variables at their average value. Strumbelj and Kononenko (2014) presented an approximation of the SHAP values, given the computational inefficiency of the procedure described above, which we implement in the present study. [4,5]

To provide more detail on some of the relationships between features and cycling to work, we also use SHAP dependence plots, which describe the feature value on the x-axis and the corresponding marginal effect calculated by the Shapley value on the y-axis. The positive or negative Shapley value indicates how much a feature positively or negatively affects the prediction of an instance. In addition to examining the main effect of each characteristic on cycling, we also explore the possible interaction effects between the main characteristics and home-to-work distance to better explain cycling behaviour. Interaction effects occur when the effect of one variable depends on the value of another variable, and we check for their existence using the interaction values of feature pairs. If the interaction value of a feature pair is zero, there is no interaction between the feature pair (Lundberg et al., 2019; Lundberg et al., 2020). Therefore, we construct the SHAP dependence plots of feature pairs with interaction values.

[4] For recent illustrations and the mathematical documentations of this process in the context of cycling, traffic accidents, mobility, and employment, see e.g. Ji et al. (2022); Parsa et al. (2020); Celbiş (2022); Celbiş et al. (2023).

[5] We compute and plot the SHAP values using the R package SHAPforxgboost developed by Liu and Just (2020).

3. Results and discussion

3.1 XGBoost model performance

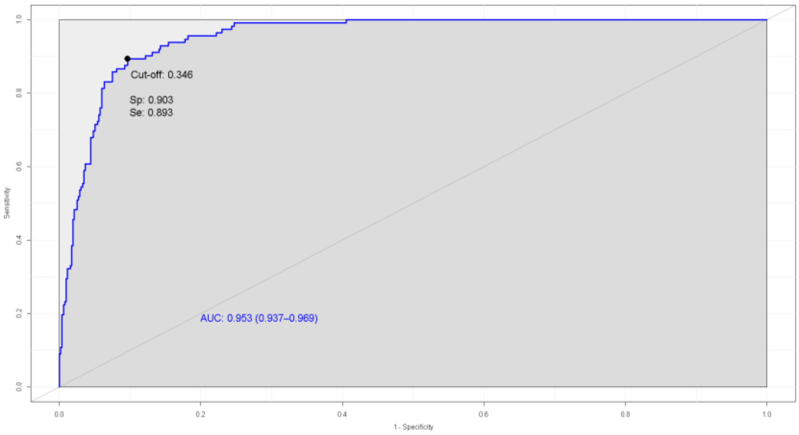
Our main prediction model, the XGBoost, predicts the bicycle use of the persons in the test dataset with 90.8% accuracy whereas the persons who do not use bicycles in this portion of the data account for 82.2% of the sample. The prediction performance of the algorithmic model is assessed more accurately using a receiver operating characteristic (ROC) curve and the associated area under the curve (AUC) value of 0.953 in Figure 2.

A further look into the structure of the data prior to proceeding to the main XGBoost results are done by examining the existence of anomalies in the complete dataset using the isolation forest method developed by Liu et al. (2008) and clustering the respondents of the survey through the use of a random forest proximity plot resulting from a random forest (Breiman, 2001) prediction with 500 trees using the same training data. The resulting random forest proximity matrix (based on the proximities of the out-of-bag observations) is reduced to three dimensions through metric multidimensional scaling.[6] Each sphere in the three dimensional random forest proximity plot presented in Figure 3a corresponds to a respondent in the training data where the orange color corresponds to bicycle users and the grey color represents the non-bike users. The sphere sizes represent the home-campus distance for the corresponding individual where larger spheres represent longer distances. Clusters of persons traveling short and long distances are visible in the random forest proximity plot. On the other hand, the bike users in orange are densely clustered. An enlarged view of the bikers cluster is presented in Figure 3b which shows that this cluster is mostly made up of persons with lower home-campus distances.

However, the analysis of the above mentioned potential anomalies shows that we do not have highly anomalous observations in our data set (see Appendix).

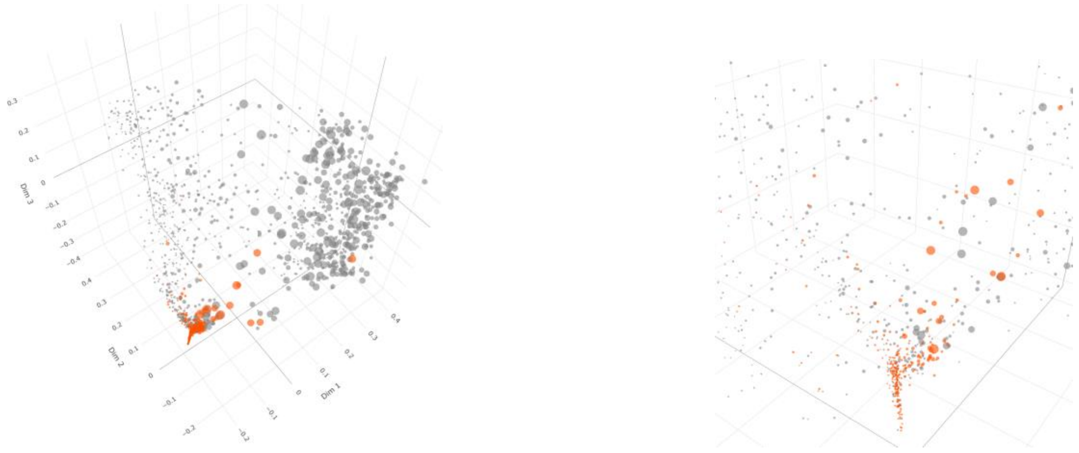
[6] See Aldrich and Auret (2013); Breiman and Cutler (2020); Friedman (2001) for the details of this technique.

Figure 2. Receiver Operating Characteristic Curve



Source: own processing.

Figure 3. Random Forest Proximities



a) Random Forest Proximity Plot (Full)

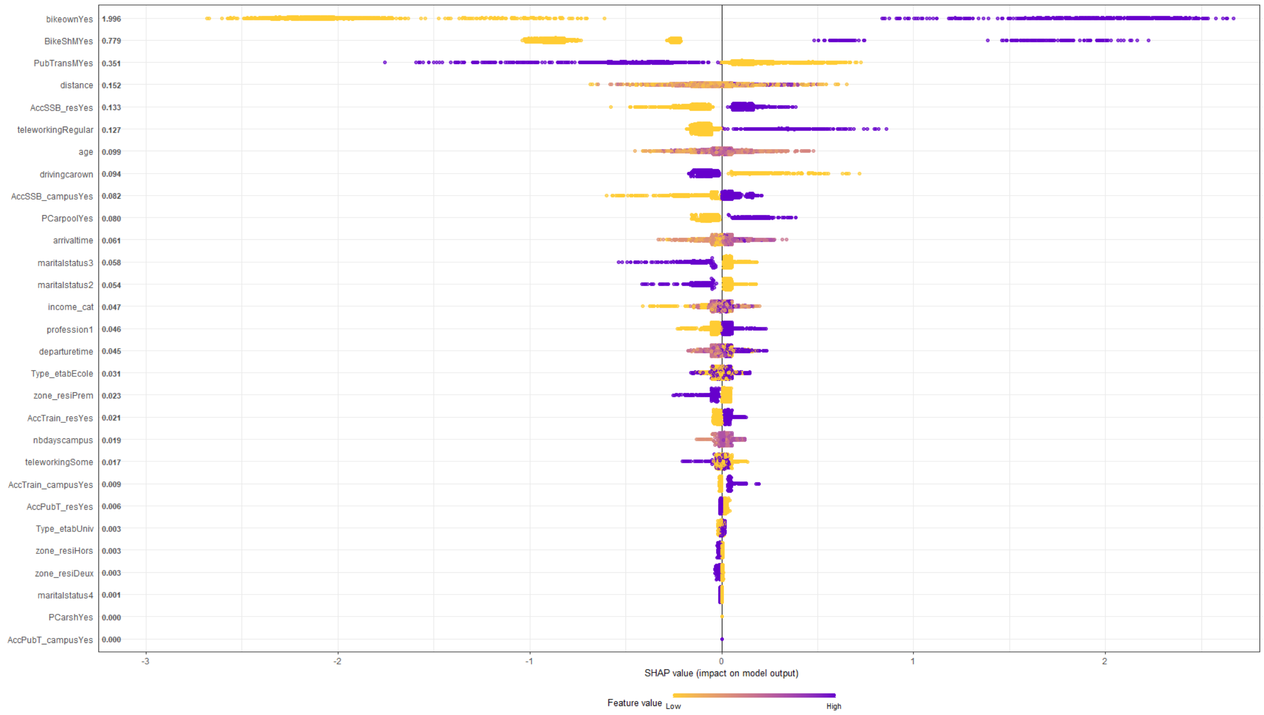
b) Random Forest Proximity Plot (Close-up on the Cluster of Bikers)

Source: own processing.

3.2 Feature analysis and SHAP interaction values

Figure 4 shows the SHAP summary plot resulting from the XGBoost prediction that orders variables, on the y-axis, based on their importance to use bicycle for commuting. Each dot represents a person in the training data. The x-axis represents the contribution of each variable value (higher values are shown in darker colors) on the deviation of a respondent's predicted log-odds from the mean predicted value arising from the corresponding feature value for all persons in the training data.

Figure 4. SHAP Values – Classification of Cyclists and non-cyclists for commuting

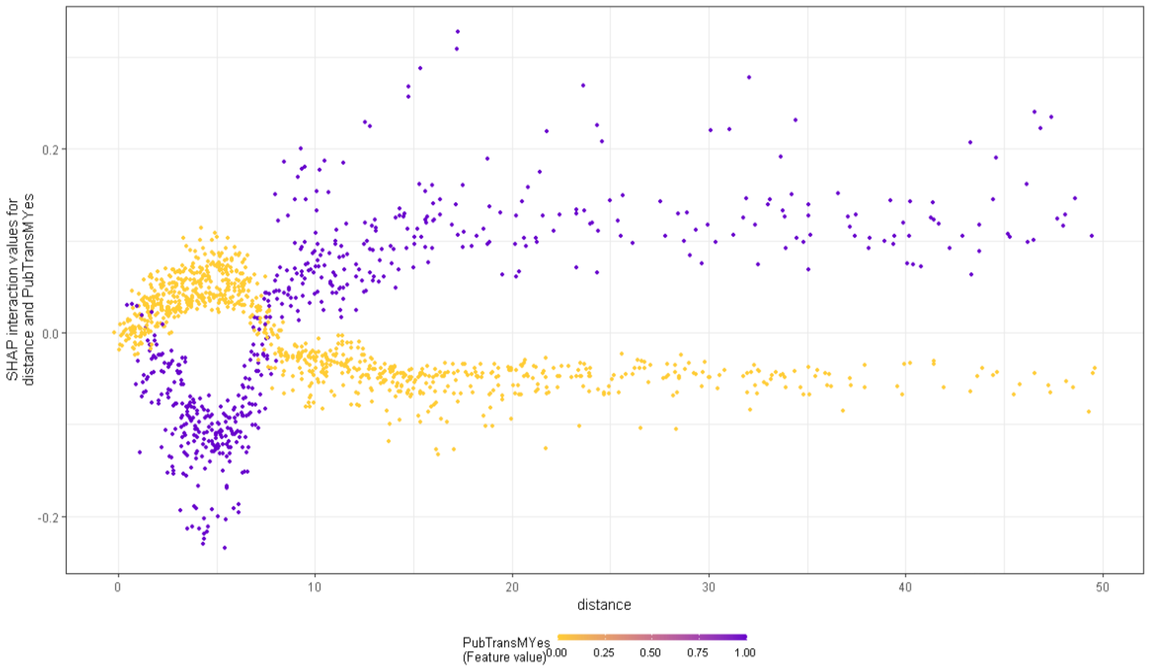


Source: own processing.

In line with previous studies (Dill and Voros, 2007; Munoz Lopez, 2016), we can see from the SHAP summary plot that all types of access to a bicycle are the most important features in the model, and also the most important in explaining bike commuting among university staff. In particular, bike users are much more likely to own a private bicycle. In addition, having a bike-sharing membership is also a major factor in profiling cyclists, with having a membership is potentially more positively associated with being a bike user, but not having a membership not being associated with a drastic reduction in the likelihood of using a bicycle for commuting. Therefore, policies that encourage personal purchase of a bicycle might be more likely to increase the use of bicycles for commuting than policies that encourage membership of a bike-sharing service. Public transport season ticket ownership is associated with a lower relative likelihood of commuting by bicycle. This finding would suggest that, overall, cycling and the use of public transport are substitutable rather complementary modes of transport among university staff. However, this effect may be non-linear and in particular may vary according with the distance between home and work. We explore the possible interaction effects of these two variables on the bicycle use, using interaction values of pairs of characteristics: the combination of the SHAP dependence plot and interaction values is shown in in Figure 5. It shows that people with a public transport pass (purple dots) are less likely to cycle when the distance from home to work is relatively short (less than about 7 km), but more likely to cycle for longer distances. The opposite pattern is observed for people without a public transport season ticket. Therefore, cycling and public transport would become complementary modes beyond a certain distance threshold. Our result supports the thrust of the recent French *Cycling and Walking Plan* 2023-2027, which aims to make these two modes of transport attractive alternatives to the private car by combining them with public transport for long journeys.

Private car ownership is negatively related to bicycle choice (Figure 4), as expected from the literature (Parkin et al., 2007; Barberan et al., 2017). However, its contribution to explaining bicycle use for commuting is much less important than bicycle ownership or public transport membership.

Figure 5. Interaction effects between public transport membership and home-to-work distance

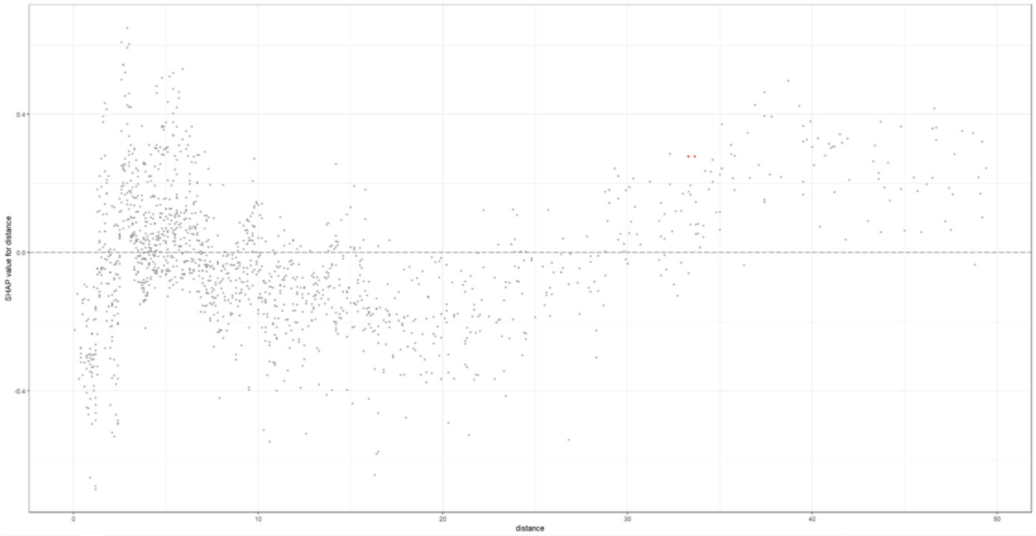


Source: own processing.

Distance to work is the fourth most important factor in our model. However, as its effect is not clearly represented in Figure 4, we need to plot its SHAP dependence (Figure 6). Figure 6 shows the non-linear attribution of home-to-work distance to the use of the bicycle for commuting. When the home-to-work distance is very short (<3 km), the distance is associated with decreased probabilities of using a bicycle. The SHAP values for cycling then increase together with distance, and then turn negative again after about 7 km. They become positive again after about 30km, which may correspond to commuters who use bikes in combination with other modes of transport, as at this distance, the people with an increased likelihood of cycling are mainly those who have a public transport membership.

.

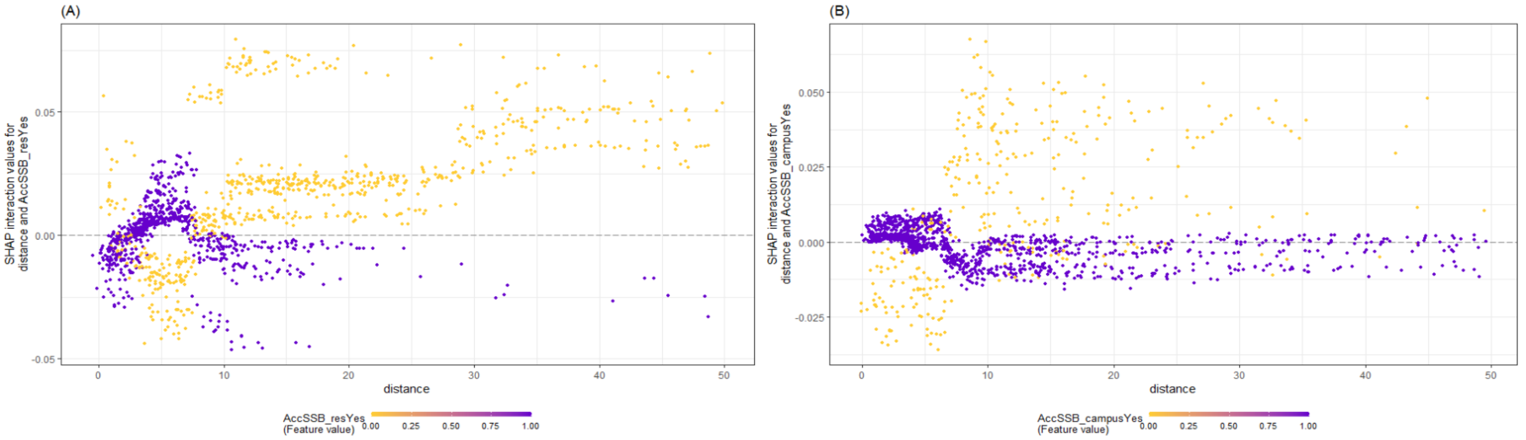
Figure 6. SHAP dependence plot for home-to-work distance variable



Source: own processing.

Not surprisingly, Figure 4 shows that accessibility to a shared bicycle station has a strong influence on the use of the bicycle for commuting. Nevertheless, a number of refinements can be made to this result, which may be important for the effective deployment of such shared cycle networks. Firstly, the proximity of a bike-share station to home (origin accessibility) has a higher contribution to cycling to work than the proximity of a bike-share station to work (destination accessibility) (with SHAP importance values of 0.133 and 0.082 respectively). It is predicted that respondents who have access to a bike-share station near their home are more likely to cycle. Secondly, the influence of accessibility to a bike-share station on campus is asymmetric for cycling to work for university staff. The lack of accessibility to a bike-share station on campus has a stronger negative relationship with the outcome compared to its limited positive effect. Thirdly, there are again non-linear effects with distance behind these two overall effects on the accessibility of shared bike stations, as shown in Figure 7, corresponding to the SHAP dependence plots of the pairs of these features with interaction values. Individuals living close to a bike-share station (purple dots in Figure 7a) are much more likely to cycle to work, but only when the distance from home to work is between 3 and 7 kilometres. For shorter distances, accessibility to the shared network is associated with lower levels of cycling, and for distances over 7 kilometres the effect is also negative, but very small. Similarly, the relationship between accessibility to an on-campus bike-share station and the likelihood of cycling to work changes with distance, particularly for the university employees who do not have these access points on their campus (orange dots in Figure 7b): it is negative until the home-work distance is less than about 7 kilometres, and then becomes positive. Moreover, people working on campuses close to a bike-share station (purple dots in Figure 7b) are more likely to cycle to work if they live less than 7 km from their workplace, and not just between 3 and 7 km, as the accessibility of this shared network to home is. Improving accessibility to a shared cycle network therefore seems to be a lever for increasing cycling, but our results suggest that deployment efforts should focus on residential areas within 7 km of major employment centres (such as campuses), as beyond this the negative effect of distance seems to dominate. Furthermore, this 7 km threshold corresponds to a cycle time of around 30 minutes at an average speed in an urban area (12 to 14 km/h).

Figure 7. Interaction effects between accessibility to self-service bicycle station (a: to home, b: to campus) and home-to-work distance

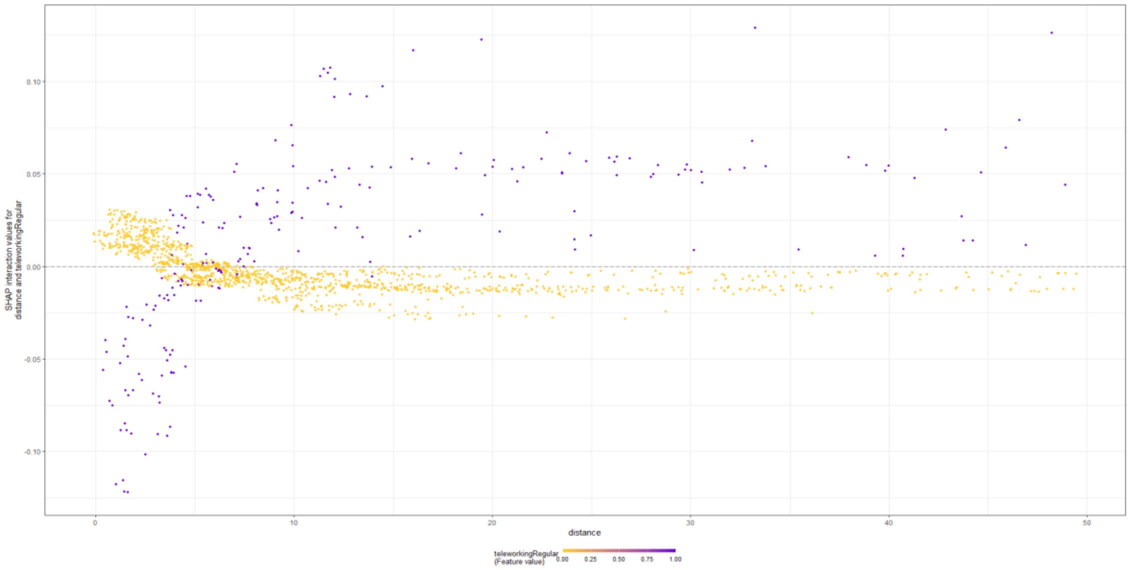


Source: own processing.

For the remaining accessibility-related features, Figure 4 shows a very low SHAP importance score (0.006) for accessibility to public transport close to home and a SHAP importance value of zero for accessibility to public transport on campus. Accessibility to a public transport station is also not a strong determinant of cycling behaviour based in our dataset. Similarly, accessibility to a train station seems to have little influence on the decision to cycle to work (contribution values of 0.021 for origin and 0.009 for destination). Nevertheless, the relationship is positive, confirming a form of complementarity between the two modes, train and bicycle, especially for university staff living far from the campus. Improving the quality of this type of intermodality could therefore be a way of increasing cycling.

In Figure 4, a non-symmetric effect is observed with respect to the telework variable. XGBoost uses one-hot encoding for categorical features which allows us to examine the role of each category in the prediction (Chen et al., 2019). From this perspective, respondents who telework regularly are more likely to cycle, while not teleworking regularly does not reduce the likelihood of cycling as much. Promoting regular teleworking among university staff could therefore encourage them to cycle more on the days when they have to travel to work. Figure 8 suggests that this type of measure could be even more effective for staff who live more than 5 km from their place of work.

Figure 8. Interaction effects between regular teleworking and home-to-work distance



Source: own processing.

4. Conclusion

In addition to its well-documented environmental benefits and contributions to energy efficiency in urban areas, cycling as a commuting mode is often considered as a remedy to the physical inactivity imposed by many modern work environments, resulting in physical health benefits (Raustorp and Koglin, 2019; Nieuwenhuijsen and Khreis, 2019). The physical benefits of cycling have also been shown to occur in conjunction with positive mental health outcomes (Humphreys et al., 2013; Martin and Suhrcke, 2014; Petrunoff et al., 2016; Synek and Koenigstorfer, 2019). For instance, it has been observed that regular biking can reduce psychological distress and increase the life satisfactions of individuals (Ma and Wang, 2021). The positive mental effects of bicycle commuting can be partly attributed to the fact that this mode provides a relaxing and engaging commuting experience in addition to a feeling of greater autonomy due to the avoidance of unpredictability caused by traffic congestion (Wild and Woodward, 2019). Our analysis of the MobiCampus-UdL survey has therefore been focused on explaining the reasons for adopting cycling as a commuting mode, at a personal level, by focusing on university staff.

In order to address our research question, we implemented a sequential ensemble model based on classification trees followed by the usage of interpretable machine learning tools. Our empirical approach allowed us to infer complex relationships present in the MobiCampus-UdL dataset through algorithmic modeling. Upon confirming the existence of a class of bike commuters clustered based on all the features in the training data (i.e. clusters based on random forest predictions), we predicted a binary outcome (bike to work or not) through the application of a gradient boosting machine. Our inferential analysis using SHAP values suggested that access to bicycles, either through ownership or sharing, coupled with spatial proximity to bike-sharing stations, contributed to the likelihood of individuals using bicycles for commuting to varying degrees depending on the distance between home and work. Our findings suggest that the deployment of bike-share stations should be done in residential areas that are, at most, within 30-minute bike ride from major employment centres.

As the machine learning techniques implemented in the present study allowed us to explore interactive and nonlinear relationships between access to cycling and to public transport, our findings uncovered implications for the complementarities and substitutabilities of these modes as a function of distance. More specifically, we observed that cycling and public transport become complementary modes as the distance between home and work increases. We also observed that another model feature that is highly interactive with home-work distance is teleworking. Regular teleworking is positively associated with the use of cycling for commuting, especially for people who live more than 5 km distance from their workplace.

At the level of local public policy, our results suggest that some of the measures recently introduced in universities could have a positive impact in terms of increasing cycling and reducing CO2 emissions. For example, the 'Sustainable Mobility Package' (“Forfait de Mobilité Durable”) introduced in 2020 for all public sector (and therefore university) staff, which offers an annual sum of between €100 and €300 to any staff member who regularly uses a bicycle to travel between home and work, should help to achieve this goal, especially as it can be cumulated with partial reimbursement of the cost of public transport season tickets. In particular, this initiative could encourage the use of bicycles by staff living more than 7km away, who contribute significantly to the carbon footprint of academic institutions associated with commuting. Similarly, a more systematic offer of at least 1 or 2 days per week of teleworking to staff in the university community could be expected to have a dual effect: a direct effect in emissions through a reduction in the number of weekly journeys, and an indirect effect through an increase in the use of bicycles on days when staff have to travel to campus. In addition, some of the preliminary results of the MobiCampus-UdL survey have been used as a basis for the new mobility plans for the university campuses in Lyon.

As a future line of research, it would be interesting to carry out new waves of surveys on the same campuses in the near future, in order to assess the changes in the use of bicycles in relation to the improvements made on certain campuses and the general increase in the use of bicycles following the Covid-19 health crisis, as well as changes in the perception of this mode of transport. More generally, the same method of analysis would deserve to be reproduced on a wider population, including categories of people with different behaviours to those of university staff. There are likely to be other determinants, such as convenience or perceptual and economic dimensions of cycling, for populations with lower social status.

Acknowledgement. This work was supported by the Agence Nationale de la Recherche through LABEX IMU [ANR-10-LABX-0088] of the University of Lyon.

References

Aldrich, C. and Auret, L., 2013. *Unsupervised Process Monitoring and Fault Diagnosis with Machine Learning Methods*. London: Springer.

Andersen, LB., Schnohr, P., Schroll, M. and Hein, O., 2000. All-cause mortality associated with physical activity during leisure time, work, sports, and cycling to work. *Archives of Internal Medicine*, 160(11), pp. 1621-1628. https://doi.org/10.1001/archinte.160.11.1621

Azevedo, B.F., Metzger, K. and Pereira, A.I., 2023. A comprehensive data analysis of e-bike mobility and greenhouse gas emissions in a higher education community: IPBike study of case. *SN Applied Sciences* 5, 291. https://doi.org/10.1007/s42452-023-05504-7

Balsas, C., 2003. Sustainable transportation planning on college campuses. *Transport Policy*, 10(1), pp. 35-49. https://doi.org/10.1016/S0967-070X(02)00028-8

Barberan, A., de Abreu e Silva, J. and Monzon, A., 2017. Factors influencing bicycle use: a binary choice model with panel data. *Transportation Research Procedia*, 27, pp. 253-260. https://doi.org/10.1016/j.trpro.2017.12.097

Breiman, L., 2001. Random forests. *Machine Learning*, 45(1), pp. 5-32. http://dx.doi.org/10.1023/A:1010933404324

Breiman, L. and Cutler, A., 2020. Random Forests. Available at: <https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm> [Accessed 25 November 2023]

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J., 1984. *Classification and Regression Trees*. Monterey, CA, Wadsworth and Brooks.

Celbiş, M. G., 2022. Unemployment in rural Europe: A machine learning perspective. *Applied Spatial Analysis and Policy*, 16, pp. 1071-1095. https://doi.org/10.1007/s12061-022-09464-0

Celbiş, M.G., Özgüzel, C., Kourtit, K. and Nijkamp, P. (2023). Industrial Composition, Remote Working and Mobility Changes in Canada and the US During the COVID-19 Pandemic: A

SHAP Value Analysis of XGBoost Predictions. In: *Pandemic and the City Footprints of Regional Science*, pp.189-207. Cham, Springer International Publishing.

Chen, T. and Guestrin, C., 2016. xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 785-794. https://doi.org/10.1145/2939672.2939785

Chen T., He T., Benesty M., Khotilovich V., Tang Y., Cho H., Chen K., Mitchell R., Cano I., Zhou T., Li M., Xie J., Lin M., Geng Y., Li Y., Yuan J. (2023). _xgboost: Extreme Gradient Boosting_. R package version 1.7.6.1. Available at: <https://CRAN.R-project.org/package=xgboost> [Accessed 21 November 2023].

Dill, J. and Voros, K., 2007. Factors affecting bicycling demand: Initial survey findings from the Portland, Oregon, Region. *Transportation Research Record*, 2031(1), pp. 9-17. https://doi.org/10.3141/2031-02

Dimitrova, A., 2021. 10,000 youngsters in Lyon will ride bicycles for free. Available at: <https://www.themayor.eu/en/a/view/10-000-youngsters-in-lyon-will-ride-bicycles-for-free-9174> [Acessed 27 November 2023].

European Cyclists' Federation (ECF), 2016. The benefits of cycling: unlocking their potential for Europe. Available at: <https://ecf.com/sites/ecf.com/files/TheBenefitsOfCycling2018.pdf> [Accessed 27 November 2023].

Friedman, J. H., 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 5, pp. 1189-1232. https://doi.org/10.1214/aos/1013203451

Friedman, J. H., 2002. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4), pp. 367-378. https://doi.org/10.1016/S0167-9473(01)00065-2

Héran, F., 2015. *Le retour de la bicyclette – Une histoire des déplacements urbains en Europe, de 1817 à 2050*. Paris (France): Editis.

Humphreys, D.K., Goodman, A. and Ogilvie, D., 2013. Associations between active commuting and physical and mental wellbeing. *Preventive medicine*, 57(2), pp.135-139. https://doi.org/10.1016/j.ypmed.2013.04.008

James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. *An introduction to statistical learning: with applications in R*. New York: Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

Jensen, P., Rouquier, J-B., Ovtracht, N. and Robardet, C., 2010. Characterizing the speed and paths of shared bicycle use in Lyon. *Transportation Research Part D: Transport and Environment*, 15(8), pp. 522-524. https://doi.org/10.1016/j.trd.2010.07.002

Ji, S., Wang, X., Lyu, T., Liu, X., Wang, Y., Heinen, E. and Sun, Z. (2022). Understanding cycling distance according to the prediction of the XGBoost and the interpretation of SHAP: A non-linear and interaction effect analysis. *Journal of Transport Geography*, 103, 103414. https://doi.org/10.1016/j.jtrangeo.2022.103414

Kelarestaghi, K.B., Ermagun, A. and Heaslip, K.P., 2019. Cycling usage and frequency determinants in college campuses. *Cities*, 90, pp.216-228. https://doi.org/10.1016/j.cities.2019.02.004

Liu, F. T., Ting, K. M. and Zhou, Z. H., 2018. Isolation Forest. 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, pp. 413-422. https://doi.org/10.1109/ICDM.

Lundberg, S. M. and Lee, S.I., 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, pp. 4768-4777.

Lundberg, S. M., Erion, G. and Lee, S-I., 2019. Consistent individualized feature attribution for tree ensembles. *arXiv*, 1802.03888v3. Available at: <https://arxiv.org/pdf/1802.03888.pdf> [Accessed 27 November 2023].

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J., Nair, B., Katz, R., Himmelfarb, J., Bansal, N. and Lee, S-I., 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2, pp. 56-67. https://doi.org/10.1038/s42256-019-0138-9

Ma, L., Ye, R. and Wang, H., 2021. Exploring the causal effects of bicycling for transportation on mental health. *Transportation Research Part D: Transport and Environment*, 93, p.102773. https://doi.org/10.1016/j.trd.2021.102773

Martens, K., 2004. The bicycle as a feedering mode: experiences from three European countries. *Transportation Research Part D: Transport and Environment*, 9(4), pp. 281-294. https://doi.org/10.1016/j.trd.2004.02.005

Martin, A., Goryakin, Y. and Suhrcke, M., 2014. Does active commuting improve psychological wellbeing? Longitudinal evidence from eighteen waves of the British Household Panel Survey. *Preventive medicine*, 69, pp.296-303. https://doi.org/10.1016/j.ypmed.2014.08.023

Mateo-Babiano, I., Tiglao, N.M.C., Mayuga, K.A., Mercado, M.A. and Abis, R.C., 2020. How can universities in emerging economies support a more thriving cycling culture? *Transportation research part D: transport and environment*, 86, 102444. https://doi.org/10.1016/j.trd.2020.102444

Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd ed.). christophm.github.io/interpretable-ml-book/Munoz Lopez, B., 2016. Integrating bicycle option in mode choice models through latent variables. Doctoral Thesis, Universidad Politecnica de Madrid.

Neves, A. and Brand, C., 2019. Assessing the potential for carbon emissions savings from replacing short car trips with walking and cycling using a mixed GPS-travel diary approach. *Transport Research Part A: Policy and Practice*, 123, pp. 130-146. https://doi.org/10.1016/j.tra.2018.08.022

Nieuwenhuijsen, M.J. and Khreis, H., 2019. Urban and Transport planning, environment and health. In Nieuwenhuijsen, Khreis (Eds.), Integrating Human Health into Urban and Transport Planning – A Framework. Cham. Springer Cham.

Nikiforiadis, A., Ayfantopoulou, G. and Stamelou, A., 2020. Assessing the Impact of COVID-19 on Bike-Sharing Usage: The Case of Thessaloniki, Greece, *Sustainability*, 12(9), 8215. https://doi.org/10.3390/su12198215

Parkin, J., Wardman, M. and Page, M., 2007. Estimation of determinants of bicycle mode share for the journey to work using Census Data. *Transportation*, 35(1), pp. 93-109. https://doi.org/10.1007/s11116-007-9137-5

Papon, F., 2003. La ville à pied et à vélo. In :Pumain, D., Mattei, M-F. (Eds), *Données Urbaines*. Insee, Anthropos, CNRS, pp.75-85.

Parsa, A.B., Movahedi, A., Taghipour, H., Derrible, S. and Mohammadian, A., 2020. Towards safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. Accident Analysis and Prevention, 136, 105405. https://doi.org/10.1016/j.aap.2019.105405

Prince, SA., Lancione, S., Lang, JJ., Amankwah, N., de Groh, M., Jaramillo Garcia, A., Merucci, K. and Geneau, R., 2021. Are people who use active modes of transportation more physically active? An overview of reviews across the life course. *Transport Reviews*, 42(5), pp. 645-671. https://doi.org/10.1080/01441647.2021.2004262

Petrunoff, N., Rissel, C. and Wen, L.M., 2016. The effect of active travel interventions conducted in work settings on driving to work: a systematic review. *Journal of Transport & Health*, 3(1), pp.61-76. https://doi.org/10.1016/j.jth.2015.12.001

Raustorp, J. and Koglin, T., 2019. The potential for active commuting by bicycle and its possible effects on public health. *Journal of Transport & Health*, 13, pp.72-77. https://doi.org/10.1016/j.jth.2019.03.012

Raux, C., Zoubir, A. and Geyik, M., 2017. Who are bike sharing schemes members and do they travel differently? The case of Lyon's "Vélo'v" scheme. *Transportation Research Part A: Policy and Practice*, 106, pp. 350-363. https://doi.org/10.1016/j.tra.2017.10.010

Ribeiro, M.T., Singh, S. and Guestrin, C., 2016. Why should I trust you?: Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* August, pp. 1135-1144.

Rondinella, G., 2015. Considering Cycling for Commuting: The role of Mode Familiarity. Doctoral Thesis, Universidad Politecnica de Madrid.

Shapley, L. S., 1953. A value for n-person games. In: H. Kuhn and A. Tucker, eds. *Contributions to the Theory of Games II, Annals of Mathematics Studies Vol. 28.* Princeton, New Jersey. pp. 307-317, Princeton University Press.

Strumbelj, E. and Kononenko, I., 2014. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems,* 41(3), pp. 647–665. https://doi.org/10.1007/s10115-013-0679-x

Synek, S. and Koenigstorfer, J., 2019. Health effects from bicycle commuting to work: Insights from participants of the German company-bicycle leasing program. *Journal of Transport & Health*, 15, p.100619. https://doi.org/10.1016/j.jth.2019.100619

Université de Lyon, 2022. Guide de la décarbonation de l'enseignement supérieur. [online]. Available at: <https://www.universite-lyon.fr/medias/fichier/2022-11-23-guide-decarbonation-enseignement-superieur_1669275258233-pdf?ID_FICHE=104037> [Acessed 25 November 2023].

Van Den Berg, L. and Russo, A., 2017. *The student city: Strategic planning for student communities in EU cities*. London. Routledge. https://doi.org/10.4324/9781315236919

Varian, H.R., 2014. Big Data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2), pp. 3-28. https://doi.org/10.1257/jep.28.2.3

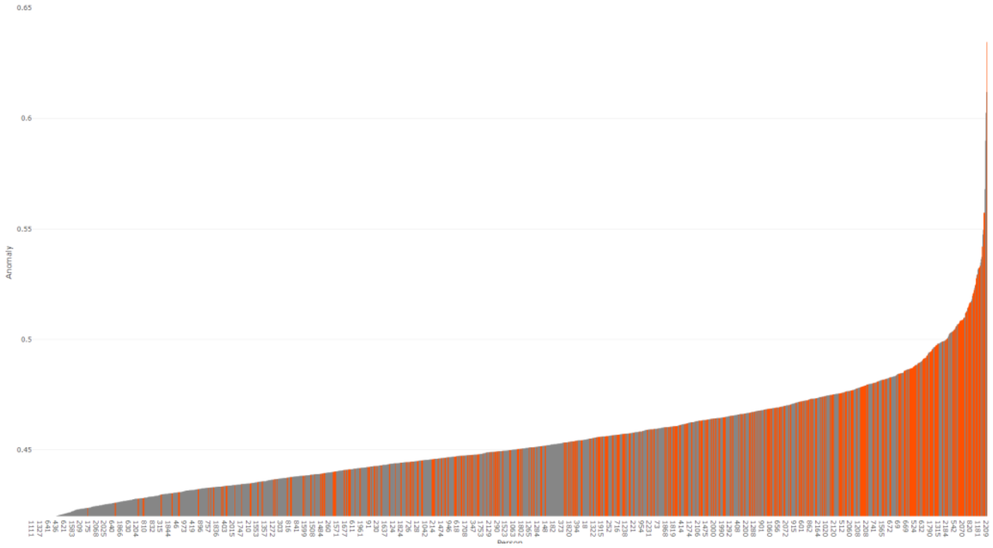
Wild, K. and Woodward, A., 2019. Why are cyclists the happiest commuters? Health, pleasure and the e-bike. *Journal of Transport & Health,* 14, p.100569. https://doi.org/10.1016/j.jth.2019.05.008

Wilson, O., Vairo, N., Bopp, M., Sims, D., Dutt, K. and Pinkos, B., 2018. Best practices for promoting cycling amongst university students and employees. *Journal of Transport & Health*, 9, pp.234-243. https://doi.org/10.1016/j.jth.2018.02.007

Appendix

An analysis of potential anomalies in our data is done from the Figure A, in which the orange color corresponds to bicycle users and the grey color represents the non-bike users. The numbers on the x-axis are respondent identifiers and the values on the y-axis represent their isolation forest anomaly scores. We can state that the vast majority of the data instances fall below the score of 0.5 and the remaining are not too close to 1, and therefore, we do not have highly anomalous observations in our data set.

Figure A. Isolation Forest Anomaly Scores



Source: own processing.