PREDICTION OF TELECOMMUTING ENGAGEMENT THROUGH MACHINE LEARNING TO ENHANCE TRAVEL SURVEY DATA

Anna Sophie Reiffer, Corresponding Author
Email: anna.reiffer@kit.edu
ORCiD: 0000-0003-1764-0154

Martin Kagerbauer
martin.kagerbauer@kit.edu
ORCiD: 0000-0002-4252-7874

Peter Vortisch
peter.vortisch@kit.edu
ORCiD: 0000-0003-1647-2435

Word Count: 6873 words + 2 table(s) × 250 = 7373 words

Submission Date: August 2, 2023

ABSTRACT

This paper presents a novel approach to enhance household travel survey (HTS) data by predicting telecommuting engagement using machine learning (ML) classification techniques. The study aims to address the debate surrounding the impact of telecommuting on overall travel behavior, considering rebound effects and latent demand. While previous research has primarily relied on questionnaires or HTS data for analysis, few studies have successfully integrated telecommuting behavior into travel demand models. The intricate relationship between telecommuting and travel behavior has been a challenge, limiting the incorporation of telecommuting data into existing models. This study fills this research gap by utilizing ML algorithms to predict telecommuting engagement based on one-day HTS data, employing features such as daily distances traveled and time spent at home.

Three feature selection algorithms, Boruta, VSURF, and Recursive Feature Elimination (RFE), were applied to identify the most relevant features for the ML models. Among the five classification methods tested, the Random Forest (RF) model utilizing features selected by the Boruta algorithm demonstrated superior performance, achieving high accuracy, specificity, F1-Score, and Matthew's Correlation Coefficient (MCC). The Bayesian Network (BN) model, while performing well in sensitivity, underperformed in other metrics due to the unsuitability of continuous data.

To evaluate the proposed approach's generalization, the RF model was applied to a separate HTS dataset from the German Mobility Panel (MOP). The out-of-sample prediction achieved promising results, with a 76% accuracy in predicting telecommuting days. The approach presented in this study has potential applications in enhancing HTS data and can be extended to other data sources to improve activity-based models.

Keywords: household travel survey data, machine learning, data fusion, data enhancement

INTRODUCTION

Telecommuting has been analyzed as a measure to reduce travel and subsequent emissions for almost five decades (*1*) (*1*). Although studies consistently show that working from home decreases the number of commuting trips, it is still debated whether telecommuting decreases overall travel or if rebound effects and latent demand are high enough to offset any benefits. To this day, the majority of studies analyze travel behavior impacts of telecommuting based on questionnaires or household travel surveys (HTS) through descriptive or statistical model analysis (see, e.g., (*2–6*)). Despite the growing interest in incorporating telecommuting behavior into travel demand models, few studies have successfully done so. This is mainly due to the intricate relationship between telecommuting and travel behavior, which even most activity-based models fail to account for. Furthermore, data on telecommuting engagement is limited, particularly when using HTS. Although some surveys ask respondents about the frequency at which they work from home, information on whether they worked from home on the day of the survey is only available in a few surveys, usually when the respective study is focused on in-home activities in addition to travel behavior (*7*). As emphasized by Asgari et al. (*5*), this information is crucial for accurately modeling telecommuting in travel demand models, and in light of the increase in work-from-home activities since the pandemic, it is essential that we examine this issue more closely. However, as appropriate data sources are scarce, new methods have to be explored on how existing data can be leveraged to inform timely research.

This paper aims to predict telecommuting engagement among respondents of a Household Travel Survey (HTS) by employing machine learning (ML) classification techniques. The study will utilize activity-travel behavior and sociodemographic data to train and assess various models using a one-day HTS. The efficacy of the trained models will be assessed by implementing them on a separate seven-day HTS dataset. This study's findings demonstrate that machine learning methods can be used to enhance HTS data using a secondary data source.

The need to account for non-travel activities has been addressed in previous studies on the design of advanced survey methods. Aschauer et al. (*8*) present a "Mobility-Activity-Expenditure-Diary" in which both travel and non-travel activities were reported. Schmid et al. (*9*) conducted a multi-stage survey to account for travel behavior and time allocation. A smartphone-based survey of travel, activities, and time use was presented by Alho et al. (*10*). Similarly, Winkler et al. (*11*) adopted a smartphone-based survey method in which travel is tracked passively, and respondents supplement the data by providing detailed information on activities and expenditures. Although the data collected in these surveys are undeniably valuable, all surveys offered monetary incentives to ensure high-quality responses. However, necessary funds are not always available. In an effort to reduce costs and make the most of existing data, researchers have explored the integration of data from various sources, including mobile phone surveys and web surveys (*12*). The utilization of smart card data can be augmented with HTS data (*13*), while the purpose of trips can be inferred through TNC data via HTS data (*14*). Moreover, HTS data can be supplemented with data from mobile phones (*15*) or social media (*16*). Studies have indicated the potential benefits of combining older, yet information-rich, HTS data with newer HTS data (*17*), as well as how the combination of data from two HTS can minimize biases and underrepresentation (*18*). All of these studies incorporate either weighting techniques (*12, 13, 17*) or econometric approaches (*14–16, 18*) to enhance the respective datasets. Machine learning techniques have primarily been used to improve data that has been collected passively, such as GPS-tracked trips. These techniques have been employed to identify the purpose of trips, the types of destinations, and the mode of transportation

used. (*19*).

This study contributes to the current body of research on leveraging existing data sources to improve the value of the data by inferring non-travel activities. By utilizing ML-based data fusion of multiple sources, we can determine if in-home activities are linked to telecommuting. This will improve the data required for integrating telecommuting into travel demand models.

The rest of this paper is structured as follows. We first present the two HTS used in this study. We further provide an overview of the machine learning classifiers analyzed in this research, including feature selection algorithms. Subsequently, we describe the performance metrics used to assess the suitability of each classifier for the proposed task. We go on to describe and discuss our findings. We conclude this paper by providing a summary and general implications of our study.

MATERIALS AND METHODS

In this section, we first provide an overview of the research design applied in this study. Further, we describe the travel diary data used in our models, including a descriptive analysis of the key variables. Subsequently, we elaborate on the classification methods we applied to predict telecommuting participation on the survey day, including the utilized feature selection algorithms.

Considering that ML methods have not yet been utilized for data fusion of HTS, there is no apriori way of knowing which ML classifier will perform best. Similarly, it is unclear which features should be included or left out. We have, therefore, opted to train and test multiple machine-learning classifiers based on different feature configurations. The overall research framework employed in this study is illustrated in figure 1. The data in this study stems from an HTS, which was split into a train and a test data set with a ratio of 70% (train) to 30% (test). Based on the test data, we applied three feature selection methods. For each of the three selected feature configurations, we trained five machine-learning classifiers. Subsequently, we predicted the telecommuting engagement in the test data. We determined the confusion matrix for each model and feature configuration pair and calculated six performance metrics. We chose the overall best scoring model, which was then used to predict telecommuting engagement in another HTS.

In the next subsections, we will provide more detail on the different parts of the research framework.

Household Travel Survey Data

This paper is based on data from a household travel survey conducted in the metropolitan area of Stuttgart, Germany. The survey was conducted in the fall of 2021 among 9,543 respondents in 4,567 (households. The scope of the survey was twofold: the data is used to update the travel demand model of the region, which is supplemented to account for telecommuting. Thus, the second scope of the survey was to gather detailed information on working from home behavior of respondents. The data used in this study stems from the travel diaries, which respondents kept for one day, and household- and person-level information. Based on the travel diaries, we determined variables of respondents' activity-travel patterns, namely duration of the activities *home, work, work-related, shopping/errands, education, escorting someone, roundtrips, and other*. The number of trips was determined by each category, as well as the overall travel time and distances traveled. Telecommuting engagement on the survey day was measured by one question, which asked where respondents worked on the given day. Additionally, we tested if the sociodemographic information of respondents influenced the models.

We further used data from the German Mobility Panel (MOP). The MOP is a longitudi-

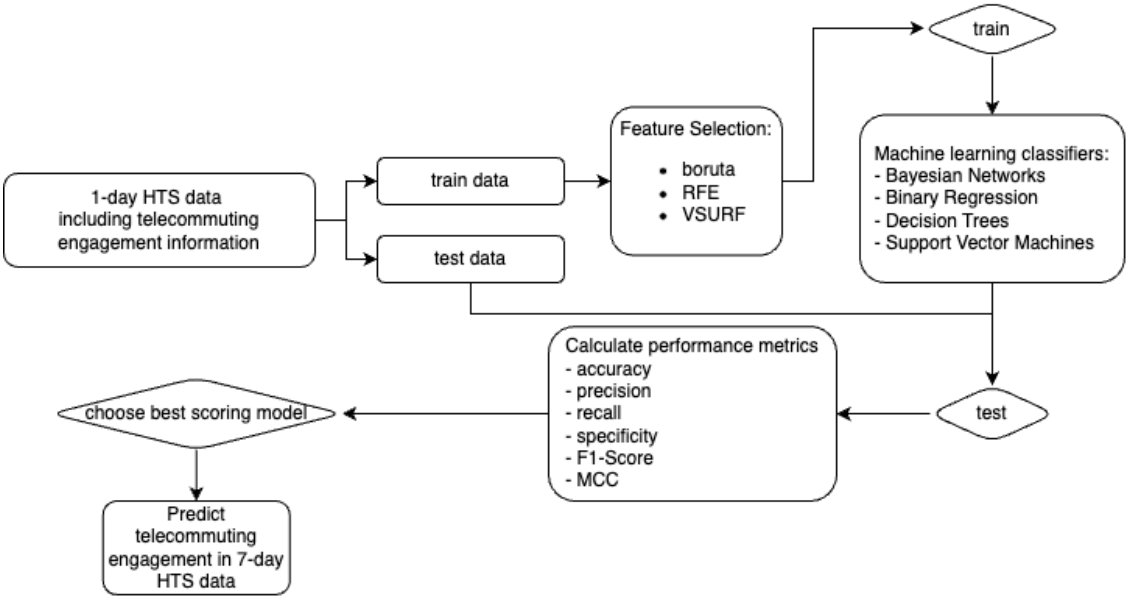


FIGURE 1: Overview of the research framework employed in this study

nal household travel survey that has been conducted annually since 1994. Because of its panel design, the questionnaires are relatively rigid to ensure the comparability and compatibility of survey waves. Furthermore, respondents are asked to keep a travel diary for seven consecutive days, which puts a high response burden on them, and person- and household-level questionnaires have to be kept as concise as possible. Thus, only a few questions concerning working from home are included. In the 2022 wave, one additional question was included asking how many days respondents worked from home in the survey week; however, not on which specific days they did. We use this data to evaluate how the models perform on out-of-sample data. The data was prepared similar to the 1-day HTS data. However, in this data, respondents were not asked where they worked on a given survey day. Instead, they provided the number of days worked from home in the survey week. Thus, the models' ability to predict telecommuting engagement can only be evaluated at an aggregate level.

Data preparation

Because work from home is limited to - axiomatically - to employed respondents and furthermore, by the jobs they hold or their employer. Thus the number of respondents dropped considerably in both surveys after applying the respective filters. However, we were still able to leverage a large enough sample for our study. The prevalence of working from home is relatively low in both surveys. In the training data set, 473 out of the total used sample of 1380 attested they had worked from home on the survey date. This amounts to about 37%. A similar proportion is provided in the MOP data. Overall, 3896 telecommuting days were reported out of the 10515 reported survey days, which equals a proportion of 34%.

In machine learning classification tasks, dealing with unbalanced classes can present significant challenges. Unbalanced classes refer to situations where the distribution of class labels is highly imbalanced, with one class having a much larger number of instances than the others. When this occurs, standard classifiers tend to be biased toward the majority class, leading to poor per-

formance on the minority class and reduced overall predictive accuracy. Consequently, the model might struggle to generalize well to real-world scenarios where the minority class is of primary interest. One common issue is that the classifier might achieve high accuracy simply by always predicting the majority class while completely overlooking the minority class. One strategy to mitigate this issue is to apply the synthetic minority over-sampling technique (SMOTE) (*20*). SMOTE works by creating synthetic samples of the minority class by interpolating between existing instances. It selects a random instance from the minority class, identifies its k-nearest neighbors, and then generates new samples by combining the features of the selected instance with those of its neighbors.

Classification Algorithms

In this section, we provide a concise overview of each classification algorithm investigated in this study.

Bayesian Networks

Bayesian networks are probabilistic graphical models that represent the dependencies between random variables using a directed acyclic graph (DAG) (*21, 22*). Each node in the graph corresponds to a random variable, and the edges between nodes encode conditional dependencies. The conditional probability distribution for each variable given its parents is modeled using Bayes' rule. Let X_i denote the i-th random variable, and $\text{Pa}(X_i)$ be the set of parent nodes of X_i in the graph. The joint probability distribution of all variables can be written as:

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \,|\, \text{Pa}(X_i)) \tag{1}$$

Inference in Bayesian networks involves computing probabilities or making predictions based on observed evidence. Popular algorithms for inference include variable elimination and Markov chain Monte Carlo (MCMC) methods.

Binary Regression

Binary regression, also known as logistic regression, is a popular supervised learning algorithm for binary classification tasks. Given a dataset with input-output pairs $(\mathbf{x}_i, y_i)$, where $\mathbf{x}_i$ is a feature vector and $y_i \in \{0, 1\}$ is the binary class label, the goal is to learn a model that estimates the probability of the positive class, i.e., $P(y_i = 1 \,|\, \mathbf{x}_i)$. The logistic regression model assumes a linear relationship between the features and the log-odds of the positive class:

$$\log\left(\frac{P(y_i = 1 \,|\, \mathbf{x}_i)}{1 - P(y_i = 1 \,|\, \mathbf{x}_i)}\right) = \mathbf{w}^T \mathbf{x}_i + b \tag{2}$$

where $\mathbf{w}$ is the weight vector and b is the bias term. To obtain probabilistic predictions, the logistic function is applied to the output of the linear model:

$$P(y_i = 1 \,|\, \mathbf{x}_i) = \frac{1}{1 + e^{-(\mathbf{w}^T \mathbf{x}_i + b)}} \tag{3}$$

Decision Trees

Decision trees are non-linear, hierarchical models used for both classification and regression tasks (*23*). They recursively split the data into subsets based on the values of individual features, aiming

to maximize the information gain or Gini impurity at each split. Each internal node in the tree represents a decision based on a feature, and each leaf node corresponds to a predicted class or regression value. The decision tree can be represented as a set of rules, and the final prediction for a given input is determined by following the path from the root to the appropriate leaf node.

Random Forest

Random forests are ensemble learning methods that combine multiple decision trees to improve predictive performance and reduce overfitting (*24, 25*). The key idea is to build a collection of decision trees by training each tree on a random subset of the training data (bootstrap sampling) and selecting a random subset of features at each split. The final prediction is made by aggregating the predictions of all individual trees, often using majority voting for classification problems or averaging for regression problems. Random forests tend to be more robust and accurate than individual decision trees, and they can handle high-dimensional data and capture complex relationships between variables.

Support Vector Machines

Support Vector Machines (SVMs) are powerful supervised learning algorithms used for both classification and regression tasks. SVMs aim to find the optimal hyperplane that best separates the data points of different classes while maximizing the margin between the classes (*26*). In the case of binary classification, given a training dataset $(\mathbf{x}_i, y_i)$, where $\mathbf{x}_i$ is the feature vector and $y_i \in \{-1, 1\}$ is the class label, SVMs find the weight vector $\mathbf{w}$ and bias term b that define the decision boundary:

$$\mathbf{w}^T \mathbf{x}_i + b = 0 \tag{4}$$

The margin is computed as the distance between the hyperplane and the closest data points (support vectors) from each class. SVM aims to maximize this margin while penalizing misclassifications. For non-linearly separable data, SVM can use kernel tricks to map the data into a higher-dimensional space, where linear separation becomes possible. Common kernel functions include polynomial, radial basis function (RBF), and sigmoid kernels.

Feature Selection

We further present the feature selection algorithms applied. We have trained and tested all algorithms on the dataset described above using the features determined by the respective feature algorithm.

Boruta

Boruta is a method for determining the importance of variables in a system using random forests. The system involves replicating each descriptive variable and randomly permuting the values of replicated variables across objects (*27*). The randomization is different for each run of the random forest algorithm. The importance of each variable is computed for each run, and a statistical test is performed to determine if the variable is significant or not. An attribute is considered important for a single run if its level of importance is greater than the highest level of importance among all randomized attributes. If a variable is deemed unimportant, it is removed from the system along with its replicated mirror pair. The procedure is repeated for a predefined number of iterations or until all attributes are either rejected or deemed important. The algorithm was applied using the R

package *boruta* (*28*).

Variable Selection Using Random Forests - VSURF

Another method based on RF is the VSURF algorithm, which is short for "Variable Selection Using Random Forests" (*29*). The procedure consists of two steps. In the first step, the variables are ranked based on their importance, estimating a threshold value for variable importance (VI) using the standard deviation of VI for less important variables and retaining only the variables with an averaged VI value above the threshold. In the second step, a sequence of ascending RF models is used to make predictions. Variables are added to each model only if they significantly decrease the error rate, using a threshold based on the out-of-bag (OOB) error decrease. The final set of variables comes from the last model. In this study, we applied the VSURF method using the R package with the same name (*29*).

Recursive Feature Elimination - RFE

Recursive Feature Elimination (RFE) was first introduced by Guyon et al. (*30*). It is a method for feature selection similar to backward feature elimination (*31*) but allows for the elimination of multiple variables simultaneously instead of having to eliminate one feature at a time through exhaustive enumeration. In the RFE procedure, a model is first built on all features. In the second step, a ranked feature list is created by ranking the combination of each feature. Lastly, features are eliminated if they do not meaningfully contribute to the model. We applied the RFE method using the R package *caret* (*32*).

Performance Metrics

We utilize several quantitative metrics to assess the performance of the ML classifiers. They all rely on the true positive (TP), true negative (TN), false positive (FP), or false negative (FN) values in one way or another. In the context of this study, these values are defined as:

- true positive (TP): model correctly classifies a telecommuting day as a telecommuting day
- true negative (TN): model correctly classifies a non-telecommuting day as a non-telecommuting day
- false positive (FP): model incorrectly classifies a non-telecommuting day as a telecommuting day
- false negative (FN): model incorrectly classifies a telecommuting day as a non-telecommuting day

Accuracy

The accuracy of a classification model is the percentage of sample objects that are correctly classified and labeled. This is done by calculating the ratio of the total number of true predictions to the sum of all observations.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

Precision

Precision, also known as the positive predictive value (PPV), is defined as the ratio of correctly classified positive cases over all classified positive cases (*33*).

$$Precision = PPV = \frac{TP}{TP+FP} \tag{6}$$

Recall

Recall, also referred to as sensitivity or true positive rate (TPR), is defined as the ratio of correctly classified positive cases over all actually positive cases (*33*). Recall and precision are often trade-offs of each other.

$$Recall = Sensitivity = TPR = \frac{TP}{TP+FN} \tag{7}$$

Specificity

Specificity, also known as the true negative rate (TNR), is determined like the TPR, except that negative cases are now relevant. The TNR is defined as the ratio of correctly classified negative cases over all actually negative cases (*33*).

$$Specificity = TNR = \frac{TN}{TN+FP} \tag{8}$$

F1 Score

The F1-score is determined by calculating the harmonic mean of the precision (eq. 6) and recall (eq. 7) (*33*). The F1-score can take values between 0 and 1, where 1 constitutes a perfect classification. As can be seen in the formula, this is achieved if the sum of false positives and false negatives is zero.

$$F1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} = \frac{2 \cdot Precision \cdot TPR}{Precision + TPR} \tag{9}$$

Matthew's Correlation Coefficient

Although many studies use accuracy as the gold standard of model evaluation, it is very sensitive to unbalanced data, which can lead to a false sense of model performance (*33*). A way to counteract the issue of class imbalance when evaluating model performance is to calculate the Matthews Correlation Coefficient (MCC) (*34*). The MCC is calculated similarly to the Pearson product-moment correlation coefficient based on the confusion matrix of the model. The MCC can take on values between -1 and 1, where -1 is the worst possible value ($TP = TN = 0$) and 1 is the best possible value (i.e., $FP = FN = 0$).

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP) \cdot (TP+FN) \cdot (TN+FP) \cdot (TN+FN)}} \tag{10}$$

RESULTS AND DISCUSSION

We first assess the results of the feature selection algorithms, which are presented in 1. Most strikingly is the similarity between boruta and rfe. The only difference in the two feature set is car access (included in rfe but not boruta) and the number of work-related trips (included in boruta bot not in rfe). The vsurf feature set ist the smallest, consisting of eight features. Features included in all three feature sets are *daily distance traveled, travel time, the hour of the last trip of the day,*

the number of work trips the time spent at home, spent for leisure activities, spent shopping, and on work-related activities.. While it is expected that time spent at home is an important feature, we initially assumed that this would also be the case for time spent at work, which was deemed unimportant by the vsurf algorithm.

TABLE 1: Selected features by selection method. Values indicate if feature was selected by the method (1) or not (0)

value	boruta	rfe	vsurf
age	1	1	0
car access	0	1	0
telecommuting Frequency	1	1	0
distance traveled	1	1	1
travel time	1	1	1
escorting someone	1	1	0
first trip of the day	1	1	0
last trip of the day	1	1	1
home	1	1	0
leisure	1	1	0
other	1	1	0
round trip	1	1	0
shopping	1	1	0
work	1	1	1
work-related	1	0	0
time use escorting someone	1	1	0
time use home	1	1	1
time use leisure	1	1	1
time use round trip	1	1	0
time use shopping	1	1	1
time use work	1	1	0
time use work-related	1	1	1

After selecting the features, we trained each model with the three feature sets on 70% of the 1-day HTS data. Subsequently, we tested the models on the remaining 30% by predicting whether a respondent in the test data worked from home on the respective survey day. In order to measure the performance of these models, we calculated the confusion matrix of predicted and real values, which provided us with values for true positive, true negative, false positive, and false negative. Figure 2 presents the confusion matrices for each classifier and each feature selection method.

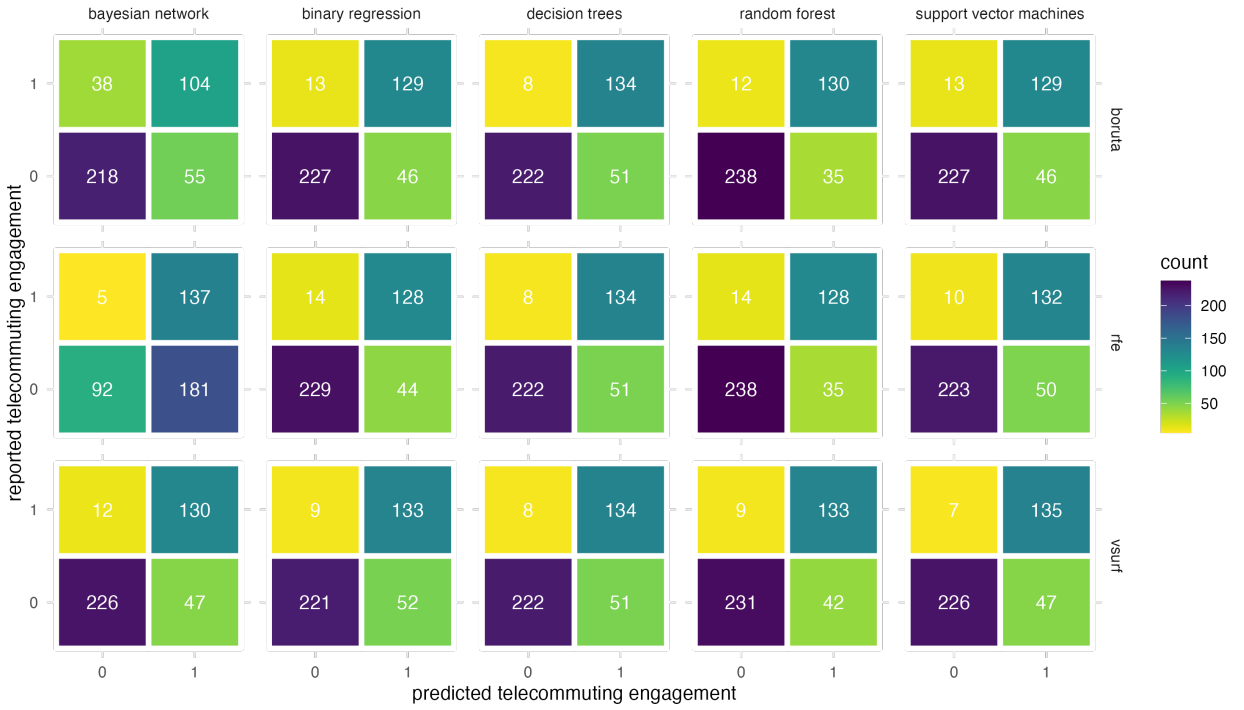


FIGURE 2: Confusion matrix by classifier and feature selection algorithm

The highest true positive values are predicted by the random forest model with 238 correctly classified true values based on the boruta and rfe feature sets, and 231 based on the vsurf feature set, respectively. All models have only few false negative vales, with bayesian networks based on the rfe feature set. This seems to come at the price of a very high value for false positive predictions. These are comparatively low in all other models, with random forest, again, performing best. Finally, true positive predictions are highest in the Bayesian network model based on the rfe feature set and lowest also for the Bayesian network when considering the boruta feature set. This is an interesting finding, as the boruta and rfe feature sets are almost identical (see Table 1), highlighting the need for pre-processing of data as Bayesian networks are often unsuited for continuous data or outliers (*35*). All other models show almost identical confusion matrices for these two feature sets.

To further assess the performance of the model, we put the values of each confusion matrix into context with each other. The performance metrics for each model based on the values in the confusion matrices are presented in table 2.

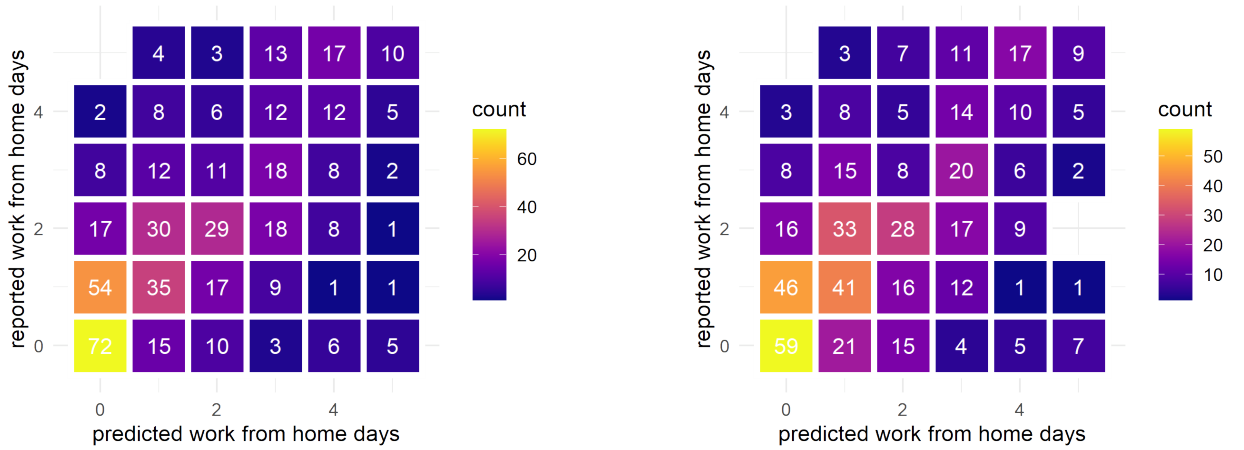
TABLE 2: Performance metrics by classification model and feature selection algorithm

classifier	feature selection	accuracy	precision	sensitivity	specificity	F1-Score	mcc
Bayesian Network	rfe	0.55	0.43	0.96	0.33	0.59	0.34
Random Forest	rfe	0.88	0.79	0.91	0.87	0.84	0.76
Support Vector Machnie	rfe	0.86	0.73	0.92	0.82	0.81	0.71
Binary Regression	rfe	0.86	0.75	0.91	0.84	0.82	0.72
Decision Trees	rfe	0.86	0.72	0.94	0.81	0.82	0.72
Bayesian Network	vsurf	0.86	0.73	0.92	0.83	0.82	0.71
Random Forest	vsurf	0.87	0.75	0.94	0.84	0.83	0.74
Support Vector Machnie	vsurf	0.87	0.74	0.95	0.83	0.83	0.74
Binary Regression	vsurf	0.85	0.72	0.94	0.81	0.81	0.71
Decision Trees	vsurf	0.86	0.72	0.94	0.81	0.82	0.72
Bayesian Network	boruta	0.77	0.65	0.70	0.80	0.68	0.50
Random Forest	boruta	0.89	0.79	0.92	0.88	0.85	0.77
Support Vector Machnie	boruta	0.87	0.74	0.93	0.83	0.82	0.73
Binary Regression	boruta	0.86	0.74	0.91	0.84	0.82	0.71
Decision Trees	boruta	0.86	0.72	0.94	0.81	0.82	0.72

Overall, we can see that almost all models achieved high rates of accuracy with only two models achieving an accuracy below 0.85. In both cases, bayesian network classification performed much worse compared to the other models. The recall/sensitivity metric shows even higher values, with, again, only the bayesian network model based on the features selected through the boruta algorithm reaching values below 0.91. Specificity/the true negative rate is not as high as the previous two metrics, but overall, almost consistently values of over .80 are reached. Regarding the F1-Score and the Matthew's Correlation Coefficient (MCC), a similar trend concerning Bayesian networks is detectable: overall, the metric values are relatively high and over 0.80 and 0.70, respectively. However, both metrics are much lower for the Bayesian network model based on the features selected through rfe and boruta. Our analysis indicates that the random forest model utilizing features from the boruta selection exhibited the most superior performance overall. This model achieved the highest accuracy, specificity, F1-Score, and mcc values in comparison to the other models. However, it did present a comparatively low sensitivity value. On the other hand, the bayesian network model based on rfe feature selection demonstrated the highest sensitivity value, but underperformed in all other metrics.

To further evaluate the performance of our proposed approach to data enhancement, we predicted the telecommuting engagement in a separate HTS. For this purpose, we leveraged data from the German Mobility Panel in which respondents keep a travel diary for seven days. As the random forest models performed best, we used those for prediction. To evaluate how the comparatively large number of features from the best model (boruta feature set) compares to the smaller feature set from the vsurf algorithm, we conducted two out-of-sample predictions. Because information on telecommuting engagement is provided at the week-level and not the day-l, we cannot calculate the aforementioned performance metrics. Instead, we predicted the telecommuting engagement for each day and added them over the week for each respondents to get a comparative measure. At the aggregate level over the entire dataset, the model based on the boruta feature set performs slightly better. This model detects 803 telecommuting days out of the 915 (87.7%) reported days. Whereas the model based on the smaller vsurf feature set predicts 792 work from home days (85.6%) The confusion matrices by the number of telecommuting days are presented

in Figure 3. On the left (Figure 3a), the results based on the boruta feature set are depicted, and on the right (Figure 3b) those on the vsurf feature set. Values above the diagonal are likely false negatives while values below the diagonal are most likely false positives.



(a) based on the boruta feature set

(b) based on the vsurf feature set

FIGURE 3: Predicted and actual number of work-from-home days in German Mobility Panel dataset.

The two models perform very similar. Both have a high rate of predicted non-telecommuting days over reported non-telecommuting days but also a relatively large rate of predicted non-telecommuting days over the reported one day of telework per week. In the latter case, vsurf performs slightly better than boruta. This performance difference is negated for a larger number of telework days per week.

Overall, the application of the models on the MOP dataset shows promising results and shows that the approach is viable to be applied to other HTS data. Random forest models are best suited for this approach and perform well even on a relatively small feature set. To the best of our knowledge, this is the first study testing different machine learning models to enhance in-home activity information in HTS data. Thus, we cannot compare our results to other studies. However, other studies comparing the performance of classification methods also find random forests to be one of the best performing methods (*36, 37*).

CONCLUSION

In this study, we evaluated how machine learning methods can be leveraged to classify telecommuting engagement as an in-home activity to enhance data from a household travel survey. We evaluated five classification methods, namely Bayesian Networks (BN), Binary Regression (BR), Decision Trees (DT), Random Forest (RF), and Support Vector Machines (SVM). Overall, the NB reached the highest specificity; however, it was outperformed by the other models on all other metrics. Overall, the RF model based on features selected using the *boruta* method performed best. The out-of-sample prediction based on data from the German Mobility Panel shows promising

outcomes with an aggregate ratio of predicted telecommuting days over reported telecommuting days of 76%.

The research framework proposed is versatile and can be applied to various data sources and research questions. The models mainly utilized features that focused on the travel patterns of the respondents, making them suitable for GPS-based data. Variables that were particularly important in the models included daily distances traveled and time spent at home, which can easily be derived from GPS-based data. However, because additional features were still important in model prediction, we anticipate that the performance of ML models might be slightly lower than our application. Moreover, the method presented in this paper can be utilized to improve HTS data. The approach can be applied to time-use data, which will provide more detailed information on activities and enhance HTS data.

There are also some limitations worth noting. In our study, we applied the same feature selection algorithms for all machine learning methods without any feature engineering. This approach was adopted to evaluate the performance of unprocessed data. However, this technique proved to be especially limiting for Bayesian network models. Therefore, future work will need to explore more advanced feature engineering techniques to optimize the performance of these models. Further, the surveys used were not solely conducted for the purpose of this study, meaning this was not a controlled experiment. It would be interesting to repeat this effort with, e.g., GPS-based data that was enriched with information on telecommuting engagement as a controlled study to further evaluate the proposed approach.

Future work will, thus, include evaluating how feature engineering can improve the models. Further, we intend to apply the approach to other similar survey data to enhance existing data sources. Namely, we will evaluate how time-use data can be merged with household travel survey data to make the best of both worlds.

ACKNOWLEDGEMENTS

The data on which the models are trained was collected within the project "Traffic reduction through new forms of work and mobility technologies", funded by the German Federal Ministry of Education and Research under grant number 01UV2091C.

AUTHOR CONTRIBUTIONS

The authors confirm the contribution of the paper as follows: study conception and design: A. Reiffer; data collection: A. Reiffer, M. Kagerbauer; analysis and interpretation of results: A. Reiffer; draft manuscript preparation: A. Reiffer; supervision: M. Kagerbauer, P. Vortisch. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

1. Nilles, J., F. Carlson, P. Gray, and G. Hanneman, Telecommuting - An Alternative to Urban Transportation Congestion. *IEEE Transactions on systems, man, and cybernetics*, Vol. SMC-6, No. 2, 1976, pp. 77–84.
2. Elldér, E., Telework and Daily Travel: New Evidence from Sweden. *Journal of Transport Geography*, Vol. 86, 2020, p. 102777.
3. He, S. Y. and L. Hu, Telecommuting, Income, and out-of-Home Activities. *Travel Behaviour and Society*, Vol. 2, No. 3, 2015, pp. 131–147.
4. Shi, Y., S. R. Sorrell, and T. J. Foxon, Do Teleworkers Have Lower Transport Emissions? What Are the Most Important Factors? *SSRN Electronic Journal*, 2022.
5. Asgari, H., X. Jin, and A. Mohseni, Choice, Frequency, and Engagement: Framework for Telecommuting Behavior Analysis and Modeling. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2413, No. 1, 2014, pp. 101–109.
6. Reiffer, A., M. Magdolen, L. Ecke, and P. Vortisch, Effects of COVID-19 on Telework and Commuting Behavior: Evidence from 3 Years of Panel Data. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2677, No. 4, 2023, pp. 478–493.
7. Asgari, H., X. Jin, and Y. Du, Examination of the Impacts of Telecommuting on the Time Use of Nonmandatory Activities. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2566, No. 1, 2016, pp. 83–92.
8. Aschauer, F., R. Hössinger, K. W. Axhausen, B. Schmid, and R. Gerike, Implications of Survey Methods on Travel and Non-Travel Activities: A Comparison of the Austrian National Travel Survey and an Innovative Mobility-Activity-Expenditure Diary (MAED). *European Journal of Transport and Infrastructure Research*, Vol. 18, No. 1, 2018, pp. 4–35.
9. Schmid, B., M. Balac, and K. W. Axhausen, Post-Car World: Data Collection Methods and Response Behavior in a Multi-Stage Travel Survey. *Transportation*, Vol. 46, No. 2, 2019, pp. 425–492.
10. Alho, A., C. Cheng, D. T. Hieu, T. Sakai, F. Zhao, M. Ben-Akiva, and L. Cheah, Online and In-Person Activity Logging Using a Smartphone-Based Travel, Activity, and Time-Use Survey. *Transportation Research Interdisciplinary Perspectives*, Vol. 13, 2022, p. 100524.
11. Winkler, C., A. Meister, B. Schmid, and K. W. Axhausen, TimeUse+: Testing a Novel Survey for Understanding Travel, Time Use, and Expenditure Behavior, 2022, p. 17 p.
12. Verreault, H. and C. Morency, Integration of a Phone-Based Household Travel Survey and a Web-Based Student Travel Survey. *Transportation*, Vol. 45, No. 1, 2018, pp. 89–103.
13. Grapperon, A., B. Farooq, and M. Trépanier, *Information Fusion of Smart Card Data with Travel Survey*. CIRRELT - Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation, 2016.
14. Hossain, S. and K. N. Habib, Inferring the Purposes of Using Ride-Hailing Services through Data Fusion of Trip Trajectories, Secondary Travel Surveys, and Land Use Data. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2675, No. 9, 2021, pp. 558–573.
15. Bwambale, A., C. F. Choudhury, S. Hess, and M. S. Iqbal, Getting the Best of Both Worlds: A Framework for Combining Disaggregate Travel Survey Data and Aggregate

Mobile Phone Data for Trip Generation Modelling. *Transportation*, Vol. 48, No. 5, 2021, pp. 2287–2314.

16. Maghrebi, M., A. Abbasi, T. H. Rashidi, and S. T. Waller, Complementing Travel Diary Surveys with Twitter Data: Application of Text Mining Techniques on Activity Location, Type and Time. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, IEEE, Gran Canaria, Spain, 2015, pp. 208–213.
17. Verreault, H. and C. Morency, Methodology to Add Value to Ageing Travel Survey Data. *Transportation Research Record: Journal of the Transportation Research Board*, 2023, p. 036119812311591.
18. Wang, K., S. Hossain, and K. N. Habib, A Hybrid Data Fusion Methodology for Household Travel Surveys to Reduce Proxy Biases and Under-Representation of Specific Sub-Group of Population. *Transportation*, Vol. 49, No. 6, 2022, pp. 1801–1836.
19. Koushik, A. N., M. Manoj, and N. Nezamuddin, Machine Learning Applications in Activity-Travel Behaviour Research: A Review. *Transport Reviews*, Vol. 40, No. 3, 2020, pp. 288–311.
20. Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, Vol. 16, 2002, pp. 321–357.
21. Pearl, J., *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. The Morgan Kaufmann Series in Representation and Reasoning, Morgan Kaufmann, San Francisco, Calif, rev. 2. ed., transferred to digital printing ed., 1997.
22. Neapolitan, R. E., *Probabilistic Reasoning in Expert Systems: Theory and Algorithms*. publisher not identified, Place of publication not identified, 2012.
23. Dattatreya, G. R. and L. Kanal, Decision Trees in Pattern Recognition. University of Maryland, 1984.
24. Tin Kam Ho, Random Decision Forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, IEEE Comput. Soc. Press, Montreal, Que., Canada, 1995, Vol. 1, pp. 278–282.
25. Tin Kam Ho, The Random Subspace Method for Constructing Decision Forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 8, Aug./1998, pp. 832–844.
26. Cortes, C. and V. Vapnik, Support-Vector Networks. *Machine Learning*, Vol. 20, No. 3, 1995, pp. 273–297.
27. Kursa, M. B., A. Jankowski, and W. R. Rudnicki, Boruta – A System for Feature Selection. *Fundamenta Informaticae*, Vol. 101, No. 4, 2010, pp. 271–285.
28. Kursa, M. B. and W. R. Rudnicki, Feature Selection with the Boruta Package. *Journal of Statistical Software*, Vol. 36, 2010, pp. 1–13.
29. Genuer, R., J.-M. Poggi, and C. Tuleau-Malot, VSURF: An R Package for Variable Selection Using Random Forests. *The R Journal*, Vol. 7, No. 2, 2015, pp. 19–33.
30. Guyon, I., J. Weston, S. Barnhill, and V. Vapnik, Gene Selection for Cancer Classification Using Support Vector Machines. *Machine Learning*, Vol. 46, No. 1/3, 2002, pp. 389–422.
31. Kohavi, R. and G. H. John, Wrappers for Feature Subset Selection. *Artificial Intelligence*, Vol. 97, No. 1-2, 1997, pp. 273–324.
32. Kuhn, M., Building Predictive Models in *R* Using the **Caret** Package. *Journal of Statistical Software*, Vol. 28, No. 5, 2008.

33. Chicco, D. and G. Jurman, The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation. *BMC Genomics*, Vol. 21, No. 1, 2020, p. 6.
34. Baldi, P., S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen, Assessing the Accuracy of Prediction Algorithms for Classification: An Overview. *Bioinformatics*, Vol. 16, No. 5, 2000, pp. 412–424.
35. Cheng, J. and R. Greiner, Comparing Bayesian Network Classifiers, 2013.
36. Zhang, C., C. Liu, X. Zhang, and G. Almpanidis, An Up-to-Date Comparison of State-of-the-Art Classification Algorithms. *Expert Systems with Applications*, Vol. 82, 2017, pp. 128–150.
37. Chen, R.-C., C. Dewi, S.-W. Huang, and R. E. Caraka, Selecting Critical Features for Data Classification Based on Machine Learning Methods. *Journal of Big Data*, Vol. 7, No. 1, 2020, p. 52.