Global Social Sciences Review (GSSR)
URL: http://dx.doi.org/10.31703/gssr.2021(VI-III).12

Modeling and Predicting Commuters' Travel Mode Choice in Lahore, Pakistan

Fariha Tariq [*] Nabeel Shakeel [†]

Abstract *The travel mode preference exists in both culture and the environment. The wide scale of people's mobility makes our cities more polluted and congested, eventually affecting urban assets. Understanding people's mode choice is important to develop urban transportation planning policies effectively. This study aims to model and predict the commuter's mode choice behaviour in Lahore, Pakistan. A survey was conducted, and the data was used for model validation. The comparative study was further done among multinomial logit model (MNL), Random Forest (RF), and K-Nearest Neighbor (KNN) classification approaches. It's common in existing studies that vehicle ownership is ranked as the most important among all features impacting commuters' travel mode choice. Since many commuters in Lahore own no vehicle, it's unclear what the rank of factors impacting non-vehicle owners is. Other than the comparison of predicting the performance of the methods, our contribution is to do more analysis of the rank of factors impacting the different types of commuters. It was observed that occupation is ranked as the most important among all features for non-vehicle owners.*

- DOI: 10.31703/gssr.2021(VI-III).12
- Vol. VI, No. III (Summer 2021)
- Pages: 106 – 118
- p- ISSN: 2520-0348
- e-ISSN: 2616-793X
- ISSN-L: 2520-0348

Key Words: Travel Behavior, Machine Learning, Multinomial Logit Model, Random Forest, K-nearest Neighbor, the Travel Mode Choice

Introduction

Recent theoretical advancements have revealed that dealing with the present bottleneck and creating a sustainable transportation system, is the greatest challenge for urban transportation planners. Studying travel mode choice plays an integral part, which gives the understanding to travel mode choice preferences of commuters and helps to validate the introduction of the new transport system to existing ones. Moreover, interest in understanding people's transportation behavior has risen dramatically. The human scale of Pakistani metropolitan cities is being overwhelmed by traffic congestion and urban sprawl. A growing number of urban policy analytics and planners are advocating city reorganization and renovation as a means of reducing the problems associated with the auto-dominated transport modes. Plans and forms of community planning and development that empathize with pedestrian enforcement and comfort, these advocates say, would facilitate increased use of walking and public transportation, thus reducing vehicle use and congestion on the freeway.

Machine Learning (ML) is the better alternative to statistical methods for predicting travel mode choice behavior because these methods do not make rigid assumptions. Instead, these techniques learn to represent complex associations in a data-driven manner. On the other hand, statistical methods are good for inference about the relationships between features, while the ML methods are good to make the utmost accurate predictions. Now, many experts from the fields of planning, transportation, economics, and geography are shifting toward more advanced and precise methods to study people's behavior to form effective urban policies. The ML techniques have been demonstrated in many existing studies to solve different transportation problems. Existing studies show that the RF is an outstanding

[*] Department of City and Regional Planning, University of Management and Technology, Lahore, Punjab, Pakistan. Email: fariha.tariq@umt.edu.pk
[†] Department of City and Regional Planning, University of Management and Technology, Lahore, Punjab, Pakistan.

Citation: Tariq, F., & Shakeel, N. (2021). Modeling and Predicting Commuters' Travel Mode Choice in Lahore, Pakistan. *Global Social Sciences Review, VI*(III), 106-118. https://doi.org/10.31703/gssr.2021(VI-III).12

technique, which has the capability of better classify the choices, although only fewer studies found for travel mode choice analysis. To increase the predictive performance, we tune the model parameters on different stages to get more robust and accurate results of model validation. After analyzing the model parameters, we train our model on the survey data. The results indicate that the RF is showing better predictive performance than the MNL and the KNN classification methods.

The researchers from developed and high-income countries applied several techniques to study transportation problems using different means of data. Unfortunately, such data sources and transportation data management systems are not well maintained in developing countries like Pakistan. That is why the researchers always need to look for alternative ways to gather data, e.g., stated preference (SP) survey (Belgiawan et al., 2019; Sperry et al., 2017). Therefore, we suggest transport organizations of Pakistan maintain data management systems, and further studies need to be carried out for the cities of Pakistan where there is the need for the hour to focus on the transportation sector. The commuter's mode choice is both tour and trip based decision, but this study is limited to trip based mode choice decision assuming each trip's origin as home and destination as a work-related place. This trip based modelling of mode choice behavior in Lahore, Pakistan, is essential to develop a framework for policymakers to weigh the travel demand before introducing any new transport systems. In a developing country like Pakistan, cities are growing at a faster rate as compared to past years, and it requires the well-established association of travel demand and its influencing factors. Unfortunately, very little attention has been given to studying this association in developing countries, especially in the cities of Pakistan.

This study demonstrates the modelling and predicting of trip based mode choice behavior of daily commuters in Lahore, Pakistan. As this is the first study to be carried out for any city of Pakistan, so first the MNL was used to analyze the mode choice preferences, and then the survey data was split into training and testing sub-datasets that were analyzed using the MNL, the RF, and the KNN classifiers. The survey data used includes socio-demographic and travel attributes. More details on the features selection have been explained in the "Data" section of this study.

This study is not only based on modelling and predicting mode choice but distinguished from the previous studies in several ways. First, it is common in the previous studies that vehicle ownership is ranked most valuable in modelling mode choice, but the ranks of other factors in the absence of vehicle have not been fully addressed. Hence, this study investigated the ranks of factors in detail under three different scenarios. Second, the study of Lanzini et al. (2017) investigated 58 studies of commuters' behavior, and psychological determinants of mode choice, and all of those 58 studies were either based on the USA or European countries but not include any of developing countries. The mode choice outcome has surely been different in developing countries compared to most of the existing studies carried out. Unfortunately, no attention has been given before to study mode choice in any city of Pakistan. Hence, there is clear studies gaps in investigating the mode choice in developing and emerging countries. Third, the accuracy has not been given much concern in the studies of mode choice, but the abilities of ML techniques to take into account the out-of-bag observations make this technique more robust in choice behavior studies.

Literature Review

The mode choice study is a fundamental task for transport policy formulation, which is based on many socio-demographic and travel factors, e.g., income, gender (Giuliano et al., 2006), age (Zahabi et al., 2012) and trip origin-destination distance (Pucher et al., 2006). The effectiveness of using road spaces is different for different travel mode choices, which plays a crucial role in policy formulation. Thus, the choice of transport mode has become an essential indicator of transport policy. In addition, the efficiency of daily travel is influenced by choice of travel mode, making it one of the most significant elements feeding into transport policies. On the other hand, travel mode choice behavior helps to develop transport policies that define the optimal location of urban elements (Ewing et al., 2010), e.g., parking spaces, to ensure a healthy urban environment for people.

The modelling of travel mode choice has been a topic of interest among many researchers for a long period (Buehler et al., 2011; Assi et al.,

2018). The large quantity of existing studies on travel mode choice contains many travel-related factors that have been investigated, e.g., the commuters' attitude and habits (Lo et al., 2016) and household factors (Gao et al., 2017) in selecting travel mode. The investigating of the effect of mode choices under different weather conditions in the Netherlands, i.e., shift from bike to car and public transport in high temperature while to walk or cycling in low temperature (Weinberger et al., 2019), and mode shifts between two choices (Cumming et al., 2019). Most of the existing studies applied the statistical model to study the discrete choice behavior of travel mode. These statistical methods are better for casual studies and have no ability to consider out-of-bag observations, which eventually decreases the performance of the model. In recent years, artificial intelligence and ML algorithms have been the better alternatives for predicting individual mode choice behavior. Because of this, the interest has increased, particularly among transportation researchers, in expanding the practicability of applying ML algorithms to address transportation problems.

The RF and the KNN, the most used powerful supervised ensemble ML methods, are popular because they have better capabilities of making predictions and solving classification problems for small datasets (Sharifi et al., 2019; Shakeel et al., 2019). In place of making a single decision tree (that over-fits when the tree size becomes large and demonstrate the poor performance of the model), the ensemble ML technique and the RF form several decision trees and combine them to get the best prediction performance. The RF has been considered being one of the most precise and accurate ML techniques available in data mining techniques (Genuer et al., 2010). In the study of individual travel mode choices, the RF makes multiple decision trees, and every single decision tree may have a different variance in data and the final decision made by voting. By this technique, the RF helps to enhance the accuracy of the model. The process abilities of RF allows us to better distinguish results, and it is significant to analyze the relations between travel mode choice and its causal factors. The capabilities of RF are thus explored in the mode choice behavior study.

The studies that applied the RF for solving different transportation problems have been categorized into four types: mode choice behavior, traffic instance predictions, traffic flow predictions, and pattern recognition, but only a few studies are available for travel mode choice behavior. This technique has been applied to study the traffic sign recognition (Zaklouta et al., 2012), for traffic postures recognition (Zhao et al., 2012), for the vehicle type recognition (Zhang et al., 2012), and for the drivers stop or run behavior at yellow indication on traffic signals (Elhenawy et al., 2014). Some existing studies used the GPS tracker data to identify the trip purpose of travelers (Montini et al., 2014) and for the predictions of traveller's behavior, driving performance at the start of a yellow signal at signalized connections and travel mode choice (Rasouli et al., 2014; Ermagun and Samimi, 2015). Some existing studies were able to treat mixed types of data and able to make predictions using multi-category classification problems and applied this technique to predict air traffic delays and proposed a method that is suitable for complex nonlinear relationships while requiring small data preprocessing (Rebollo et al., 2014). Some existing studies found multiple models to make forecasting of long and short-term traffic flows (Hou et al., 2014). Some existing studies applied this technique to predict travel mode recognition using cell phones sensor data (Jahangiri et al., 2015) and to study pattern recognition, traffic signals recognition and travel mode recognition (Shafique et al., 2015).

Data

The questionnaire-based survey was designed and conducted in three traffic assessment zones (TAZs) of Lahore city, based on their contributions to daily transport as shown in Fig. 1. Lahore is the cultural capital of the province Punjab and the second most populous city of Pakistan, with a population of 11,126,285 inhabitants (Pakistan Population Census, 2017). The survey was spread to individuals of the selected areas, comprised of close-ended questions to get the data of commuters socio-demographic and travel attributes. Initially, 470 individuals from different age groups and occupations voluntarily contributed to fill the survey during the period of January 9th to January 23rd, 2021. The survey was conducted under controlled conditions directly from the targeted population, and data diversity was ensured. This survey method of data collection commonly contains conflicted observations but is the only potential method of data collection

when no other (open) data sources are available for studying transportation problems in developing countries like Pakistan. We further set several criteria to remove the choice conflicts. The features, choices with description is shown in Table 1.

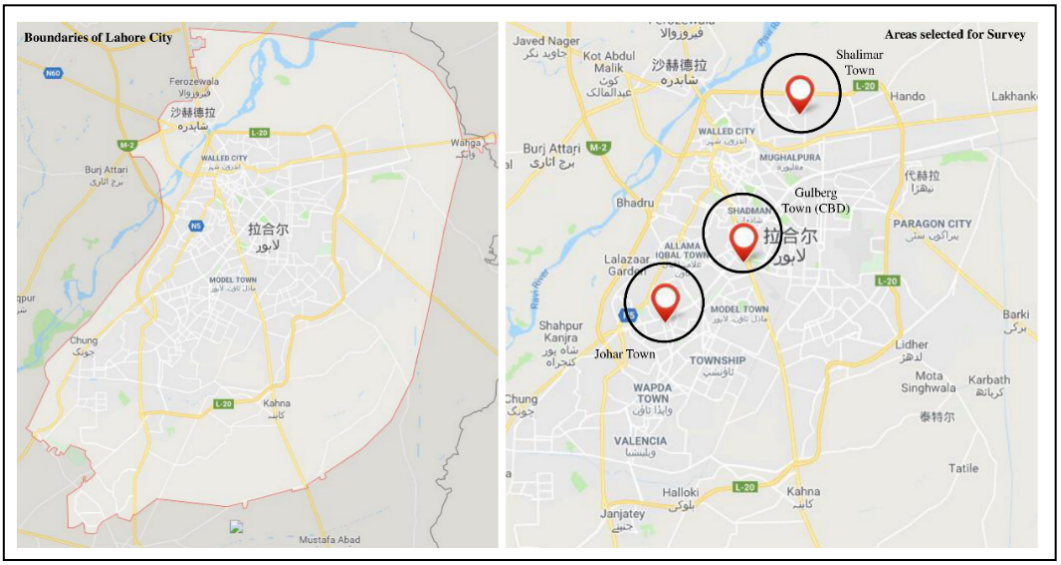

Figure 1: Boundary of Lahore city and Selected Areas for Data Collection

Table 1. Data Exploration and Description

Features	Choices		Description
	Name	Sample	
	Four discrete choices:		The age of commuter. The major contribution to daily trips is made by the age group of 21-30 years and 31-40 years because Pakistan has the youngest generation with an average life span of 50-60 years (United Nations Development Program, 2017).
	11-20 years	5.5%	
Age	21-30 years	53.3%	
	31-40 years	40.8%	
	41 and more years	0.5%	
	Two dummy choices:		The gender of commuter.
Gender	Male	71.5%	
	Female	28.5%	
	Five discrete choices:		The occupation/profession of commuter. In some cases, it is possible that commuters are engaged in more than one profession at the same time, e.g., a student can also be a private employee or do some part-time job. However, in this study, we assumed that each commuter is engaged in only one dominant profession.
	Student	24.8%	
	Government employee	15%	
Occupation	Home-based work	2.8%	
	Own business	10.3%	
	Private employee	47.3%	
	Three discrete choices:		The type of vehicle owned by commuters for daily commute from origin to destination. In some cases, it is also possible that the commuters own more than one vehicle, e.g., the commuter can own a bike and car both at the same time. But in this study, our vehicle ownership means is the vehicle owned and used by commuters for daily commute.
	No vehicle	30.2%	
Vehicle ownership	Bike	28.8%	
	Car	41%	
	Two dummy choices:		The driving license holding status by a commuter.
Driving license	No	50.2%	
	Yes	49.8%	
	Six discrete choices:		Monthly income of commuter. Here, 'Rs.' is the representation of the Pakistani Rupee (currency). Individuals with no income and 1-25,000 Rs. is considered as lower class, 25,001 – 50,000 Rs. as lower middle class,
	No income	6%	
	1-25,000 Rs.	30.3%	
	25,001-50,000 Rs.	14.5%	

Features	Choices		Description
	Name	Sample	
Income	50,001-75,000 Rs.	22.5%	50,001 – 75,000 Rs. as middle class, 75,001-100,000 Rs. as upper-middle-class and more than 100,000 Rs. as high-class income commuters.
	75,001-100,000 Rs.	15.3%	
	More than 100,000 Rs.	11.5%	
Trip purpose	Five discrete choices:		The purpose of the trip for which commuter commute daily from origin to destination.
	Study	25%	
	Work	72%	
	Recreational	2%	
	Shopping	1%	
Trip length	Five discrete choices:		The estimated distance between origin and destination or how long a commuter needs to travel to reach the destination.
	0-5 km	25.5%	
	6-10 km	29%	
	11-15 km	17.3%	
	16-20 km	14.5%	
	More than 20 km	13.8%	
Mode choice	Five discrete choices:		Travel mode choice by commuter for daily trips.
	Walk	3.5%	
	Pick up and drop off by others	6.8%	
	Public transport	22.8%	
	Bike	27.3%	
	Car	39.8%	

Removing Choice Conflicts

In data-driven modeling approaches the data quality control is a fundamental task to increase the applicability of the model. In the online web-based survey, the observation selectivity faults are common. To overcome this issue, we critically checked each observation and built the criteria against possible selectivity fault and discarded from the data as follows:

- If individuals with an age range of 11-20 years selected the mode choice 'car', then discard the observation because, in the local context, 'car' is not or less affordable among individuals' ranges in this age group.
- The individuals with vehicle ownership status as 'no vehicle' can either choose 'walk', 'public transport' or 'pick up and drop off by others' for a trip. But, if chosen vehicle ownership is 'no vehicle' and chosen mode choice is either 'bike' or 'car', then discarded the observation. In the same ways, if an individual's vehicle ownership status is 'bike' and chosen mode choice is 'car' or vice versa, then discard those particular observations from the data.
- If the individuals with occupation status as 'student' and income status as 'no income', choose mode choice as 'car' then discard the observation because of the wrong interpretation.
- The individuals with occupation status as 'government employee', 'home based' and 'private employee' can only choose trip purpose as 'work', 'shopping' or 'recreational' but not the 'study'. So we discarded such observations because of their wrong interpretations.
- The individuals with occupation status as 'student' most probably have income status as 'no income' or '1-25,000 Rs.' in the local context but not more than this. So, if individuals with occupation status as 'student' choose other than this income status, we discarded those particular observations from data.
- Furthermore, for individuals with mode choice as 'walk' and trip length as '11-15 km', '16-20 km' or 'more than 20 km' is very rare (Althoff et al., 2017). So we discarded such observations from data.

Sampling

After removing the misinterpreted and conflicted observations from survey data, we left with 403 observations. To fit the model and to specify our results to the entire city's population, we applied Kohran's formula to calculate the exact sample size as written in Equation 1 (Altares et al., 2003; Aziz et al., 2018).

$$n = \frac{N}{1+Ne^2} \quad (1)$$

Where n is the sample size, N is the total population, and e is the marginal error of sample size. By using e as 5% to ensure 95% accuracy in sampling, we calculated the sample size of $399.98 \approx 400$ that was further used for model validation and prediction of the trip based mode choice behavior of daily commuters in Lahore, Pakistan.

Resolving the Issue of a Small Dataset

ML methods are mostly used to analyze the data big in observations. But in many cases, researchers only have small experimental data to deal with and applied ML techniques to analyze and make predictions. One issue with a small dataset might be the overfitting, but data scientists have proposed several ways to address the issue of small data with less overfitting and outliers. From ML perceptions, small data needs algorithms that have low complexity to avoid overfitting the data (Zhang et al., 2018). The suggestion to that is to choose the right and powerful ensemble ML algorithms with fewer parameters to tune to decrease the bias and variance (Shaikhina et al., 2015). The MNL, the RF and the KNN are the most powerful supervised classifiers with fewer parameters to tune among predictive classifiers, which has been used in this study in comparison to predict travel mode choice. For the analysis, data were randomly split into training and testing. 80% of the total dataset was used in training, and the remaining 20% of the dataset was used for testing to estimate the predictive accuracy.

The ML Parameters Tuning

The optimum parameters of the ML classifiers need to be set to achieve higher predictive accuracy of the model, which varies with different choices of parameters. We set parameters of the RF by following the work of Breiman et al. (2001). The splitting variables value sets as 3 using $log_2 P$Where P is the number of variables. For *n_estimators,* as shown in Fig. 2, the lower the number of trees, the higher the error rate. As noticed, the number of trees from 30 to 50 is most appropriate to use for training of the RF classifier, so we randomly selected 40 as *n_estimators* value to train the RF, although, above this, there is no change in the predictive performance. The other important parameter is *max_depth*, which decides the depth of the forest. The default value *none* was used because of the small dataset. This allows the growth of forest until the maximum limits reach.

One of the useful features of the RF is that this technique estimates the relative importance of different features based on the Gini Impurity (GI) Index. The GI measures how frequently a randomly selected variable from the data has been wrongly labeled if it was randomly labeled, conferring to the circulation of labels in the dataset. We used GI to find the relative importance of variables. For splitting variable X_i, with the number of categories $L_1, \ldots, L_j$, which was calculated using Equation 2.

$$G(X_i) = \sum_{j=1}^{j} P(X_i = L_j)\left(1 - P\left(X_i = L_j\right)\right) \quad (2)$$

Where $G(X_i)$ is the GI index, X_i are the variable and $P\left(X_i = L_j\right)$ represents the estimated probabilities of category $X_i = L_j$.

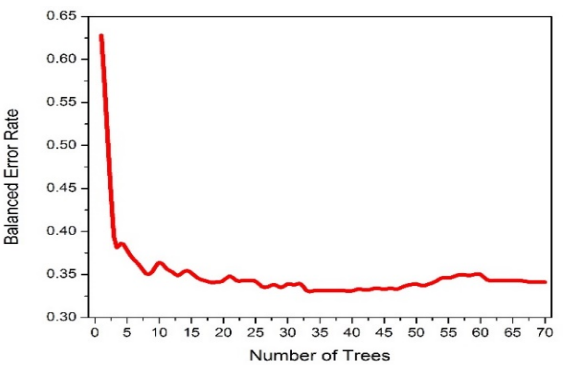

Figure 2: Number of Trees vs Error Rate

The selection of the k value in the KNN is a very important step of parameter tuning (Hand et al., 2001). The classifier shows more blind behavior towards classes for small k value and more outlier for large k value, so finding an optimal value is an important task. One of the most used methods to find the k value is k-fold cross-validation. In general, the best for the selection of k-folds is 5 or 10 folds (James et al., 2013). The dataset was examined to calculate the error for k value from a random range of 1-30 as shown in Fig. 3. The lowest mean error was observed at k values of 5, 8, and 9. As the even k value always confused the classifier to decide class assigning to the data points, so we ignored the selection of k value as 8. To choose one optimum k value from these values, we applied the rule of thumb, that is, take the square root of the testing data's size and choose the closest odd k value. So, we selected 9 as the optimum k value to train our classifier.

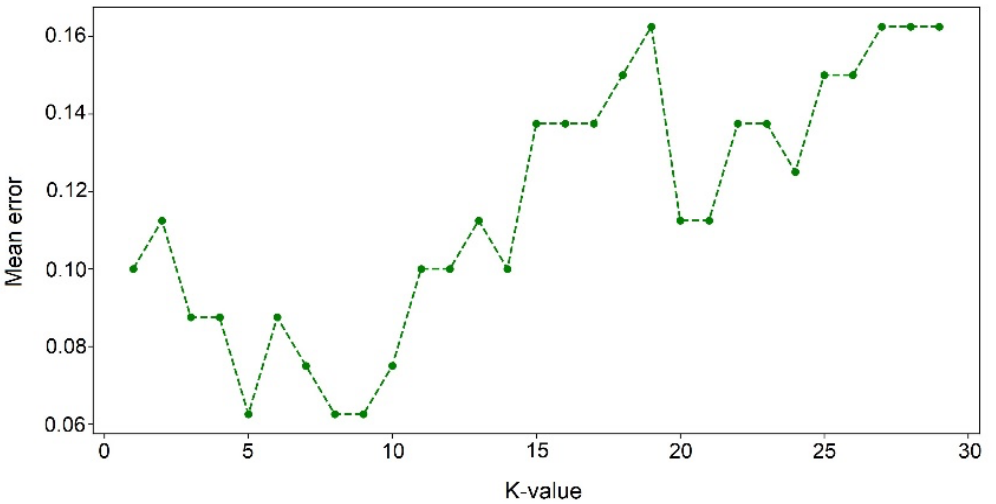

Figure 3: Mean Error for k Values

Findings and Discussion

The MNL Results

The MNL was used to analyze the mode choice behavior in this study and to estimate the best mode choice preferences. The 'public transport' was considered as a baseline or reference mode choice against the others. We assumed the independent effects of all the choices, which is one of the assumptions of the MNL. The MNL results delivered coefficients for each feature used in each mode choice preference. Following the work of Train et al. (2009) and Field et al. (2009), the coefficient value was used to deliver information about the likelihood of falling into a particular group compared to the baseline mode choice and p-value to check the statistical significance level of features. Both the significant and insignificant features contribute to understanding the mode choice preference. On the other hand, the odds ratio was used to estimate the ratio of the probability of commuters' mode choice to the probability of baseline mode choice. In several cases, the estimated odds ratio is smaller enough to become zero when rounded. It is because one of the odds being compared is close to zero and indicates a strong negative association (Abelson, 1995).

The results of the MNL model are shown in Table 2. It is observed that commuters in the age group of 11-20 years are more likely to walk and pick up and drop off by others and less to choose the car, while commuters in other age groups are less likely to walk and pick up and drop off by others and prefer to choose car compared to commuters in the reference age group. The p-value is higher than the significance level showing no significant difference between commuters of all ages group and commuters in

reference age group in choosing walk and pick up and drop off by others. Males are less likely to pick up and drop off by others and more likely to drive compared to commuters in the reference gender group. This shows that females are more dependent on the pick and drop off by other family members while males are more likely to drive. Students are more likely to pick up and drop off by others and ride a bike and less prefer to walk, while government employees are more likely to drive a bike compared to the reference occupation group. Commuters having small home-based work are more likely to walk and pick up and drop off by others while commuters who own businesses are less likely to walk but showed an insignificant difference from commuters in the reference occupation group. Commuters are more likely to drive owned vehicles compared to non-vehicle owners. Commuters having no driving license are less likely to pick up and drop off by others and drive a vehicle compared to the reference driving license group. Commuters having no income are more likely to walk, while the income group of 1-25,000 Rs. are more likely to ride a bike compared to the reference income group. The chances of choosing a walk, pick up and drop off by others and driving a vehicle is higher for the income group of 25,001 – 50,000 Rs. While the chances of choosing to walk and drive a vehicle are higher for the income group of 75,001 – 100,000 Rs. Compared to the reference income group. Commuters in the income group of more than 100,000 Rs. are more likely to drive a car. Commuters with the trip purpose of study are less likely to walk, pick up and drop off by others and drive a vehicle. Commuters are more likely to ride a bike and drive a car for recreational and shopping, respectively, compared to the reference trip purpose group. For shorter trips, i.e., 0-5 km commuters are more likely to walk and pick up and drop off by others while less likely to walk and more to choose other choices for longer trips, i.e., 11-15 km and above compared to the reference trip length group.

Features Importance

We used the RF to find the relative importance of features. The higher value of relative importance shows the stronger influence of feature on the target variable. In this study, three different scenarios were set to check the relative importance of features for mode choice behavior. The three different scenarios are as follows:

Scenario A: In this scenario, we considered all the features and choices, which were analyzed and ranked based on their relative importance for mode choice prediction.

Scenario B: In this scenario, we considered only those observations from the sample dataset with no vehicle ownership status to check the rank of factors impacting non-vehicle commuters. 30.2% sample observations of the total dataset with no vehicle ownership status were analyzed in this scenario.

Scenario C: In this scenario, we considered only those observations from the sample dataset with no driving license status to observe the features ranks in the absence of a driving license. 50.2% observations of the total dataset with no driving license status were analyzed in this scenario.

The results of the relevant importance of these scenarios are shown in Table 3. *Scenario A* results indicate that vehicle ownership is playing a significant role in predicting commuters' mode choice. As the trend in Pakistani cities, people prefer to use their private vehicles for their daily commute. *Scenario B* results indicate that, in the absence of vehicle ownership, occupation plays a significant role in predicting commuters' mode choice behavior as, in this scenario, we excluded the records of those who own vehicles. It is estimated that the majority of the commuters own a vehicle, and every commuter prefers to use the owned vehicle as a daily travel mode. For the remaining 30.2%, who do not own a vehicle, occupation plays a significant role in predicting commuter's mode choice behavior. *Scenario C* results indicate that vehicle ownership is playing a significant role in predicting commuters' mode choice; similar to the results of *Scenario A,* as in this scenario, we only considered the records for those who do not have a driving license. Further, it is estimated that 41.79% of sample analyzed in this scenario drive a vehicle for their daily commute even without holding driving license. It is because, in the cities of Pakistan, people can own a vehicle even without holding driving license and many commuters drive a vehicle for daily trips

Table 2. The MNL Analysis Results

Feature	Choices	Walk			Pick up and drop off by other			Bike			Car		
		B	p-value	OR	B	p-value	OR	B	p-value	OR	B	p-value	OR
Age	11-20 years	0.049	0.95	1.05	0.19	0.75	1.2	-1.31	0.02*	0.26	-15.59	0.98	0
	21-30 years (RC*)												
	31-40 years	-0.47	0.49	0.61	-15.35	0.98	0	-1.08	0.002**	0.33	1.65	0.00**	5.25
	41 and above years	-0.11	.***	0.88	-0.39	0.99	0.67	-0.34	0.99	0.7	17.49	0.99	-
Gender	Male	0.87	0.14	2.4	-0.87	0.84	0.91	3.13	0.00**	22.8	1.55	0.00**	4.7
	Female (RC)												
Occupation	Student	-0.98	0.29	0.37	0.61	0.001**	0.15	0.3	0.003**	0.4	37.55	0.74	0
	Government employee	1.21	0.11	3.37	110.97	0.91	0	1.05	0.0006**	0.02	0.37	0.48	0.77
	Home-based	22.51	0.00**	.***	0	.***	0	0	.***	0.51	0	.***	0.58
	Own business	0.32	0.99	1.38	190.91	0.99	0.53	123.7	0.99	0.51	91.32	0.89	0
	Private employee (RC)												
Vehicle ownership	No vehicle (RC)												
	Bike	0.34	0.75	1.41	-19.55	.***	0	25.11	0.99	0	2.57	0.99	13.1
	Car	1.95	0.17	7.08	2.69	0.01*	14.78	5.15	0.99	173.01	26.72	.***	0
Driving license	No	-0.27	0.8	0.75	-1.36	0.04*	0.25	-2.29	0.00**	0.1	-5.44	0.00**	0.004
	Yes (RC)												
Income	No income	1.18	0.26	3.27	0.96	0.2	2.61	0.93	0.12	2.54	-19.68	0.99	0
	1-25,000 Rs.	0.25	0.78	1.28	0.87	0.12	2.4	1.77	0.00**	5.9	-19.29	0.99	0
	25,001-50,000 Rs.	3.58	0.01*	36	3.36	0.005**	28.79	4.52	0.00**	91	3.27	0.00**	26.52
	50,001-75,000 Rs. (RC)												
	75,001-100,000 Rs.	4.49	0.0006**	90	-15.99	0.99	0	3.46	0.002**	32	3.79	0.00**	44.52
	More than 100,000 Rs.	1.01	.***	2.76	0.75	.***	2.13	1.56	0.99	4.79	21.2	0.99	0
Trip purpose	Study	-0.56	0.34	0.57	-0.19	0.65	0.81	-0.45	0.12	0.63	-17.69	0.97	2.06
	Work (RC)												
	Recreational	-18.79	.***	0	-18.65	0.99	0	-1.52	0.06	0.21	-19.21	0.99	0
	Shopping	-0.3	.***	0.73	-0.16	0.99	0.85	-0.24	0.99	0.78	16.97	0.99	0
Trip length	0-5 km	1.54	0.02*	4.66	2.03	0.01*	7.63	0.85	0.01*	2.35	-1.45	0.00**	0.23
	6-10 km (RC)												
	11-15 km	-17.19	0.99	0	3.51	0.00**	33.6	2.12	0.00**	8.4	1.81	0.00**	6.17
	16-20 km	-17.78	0.99	0	2.57	0.00**	13.12	1.43	0.00**	4.2	1.13	0.01*	3.1
	More than 20 km	-17.01	.***	0	-15.94	0.99	0	2.41	0.00*	11.2	2.33	0.00**	10.28

** Reference choice*
*** p-value < 0.05, ** p-value < 0.05*
**** Floating point overflow occurred while computing this statistic. Its value is therefore set to system missing.*

Table 3. Features Importance in Three Different Scenarios

Features	Scenario A		Scenario B		Scenario C	
	Rank	Importance	Rank	Importance	Rank	Importance
Age	7	4.72%	6	5.67%	6	5.87%
Driving license	2	12.62%	7	1.72%	8	0.00%
Gender	6	4.87%	4	11.05%	4	9.59%
Income	3	10.94%	3	20.12%	2	12.37%
Occupation	4	8.90%	1	30.60%	3	10.69%
Trip length	5	5.43%	2	24%	5	8.99%
Trip purpose	8	3.41%	5	6.85%	7	4.41%
Vehicle ownership	1	49.11%	8	0.00%	1	48.08%

Evaluation

The predictive performances of the MNL, the RF, and the KNN classifiers were compared with the help of mean average percentage error (MAPE), training and testing accuracy. Fig. 4 shows the MAPE, training and testing accuracy of the classifiers. It is observed that the RF has relatively better performance with high training and testing accuracy and low MAPE. This is because the RF has the ability to interrelate the complicated relationship among features than the KNN and the MNL. The RF performs better as the multiple trees are formed, and the results are selected as majority voting, which helps to reduce bias and variance.

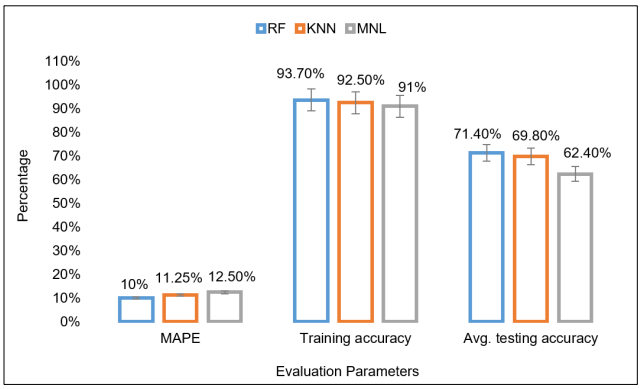

Figure 4: MAPE, Training and Average Testing Accuracy of Classifiers

Conclusion

This study focused on commuters' mode choice behavior in the city of Lahore, Pakistan. It was carried out to analyze mode choice preference and then comparison with ML techniques, which has powerful predictive performance. The RF classifier used in this study is more accurate in classifying the travel mode choice as compared to other ML techniques. To our knowledge, very little attention has been paid before to applying these ML techniques to study mode choice behavior in any city of Pakistan. The data was analyzed using both statistical and ML techniques, and based on that; this study proposed a robust RF model to predict commuters' travel mode choice. The comparative study was also done to compare the predictive performance of the MNL, the RF and the KNN classifiers. It was observed that the RF performs better than the KNN and MNL.

The relative importance of variables provides a framework for studying the significance of variables in impacting the travel mode choice. The results in this study indicate that vehicle ownership is the most important feature in studying commuter mode choice behavior from socio-demographic and travel attributes. It indicates that these features are essential to study mode choice and can be considered as the keys to estimate present and future travel demand. As the results of RF, vehicle ownership is ranked as the most important among all features impacting vehicle owners'

travel mode choice whilst occupation is that for non-vehicle owners in Lahore, Pakistan. Meanwhile, special attention should be given to these when evaluating the transport planning and policy formulation for the cities of Pakistan. Prior to introducing any new transport system, travel mode choice behavior studies are very significant, especially in the case of underdeveloped countries like Pakistan.

In the future, researchers can bring recently advanced technology datasets, observations big in nature, household survey data, built environment data, smartphone data, and GPS location data to study mode choice in large cities of Pakistan using ML techniques. Further, land-use and built environment characteristics can be merged with socio-demographic and travel attributes to study people's mode choice behavior in metropolitan cities of Pakistan.

References

Abelson, R. P. (1995). *Statistics as Principled Argument*. Psychology Press, *New York*, USA.

Altares, P.S. et al. (2003). *Elementary Statistics: A modern Approach*. Rex Book Store Manila, Philippines, p. 13.

Althoff, T., Hicks, J. L., King, A. C., Delp, S. L., & Leskovec, J. (2017). Large-scale physical activity data reveal worldwide activity inequality. *Nature*, *547*(7663), 336-339.

Assi, K. J., Nahiduzzaman, K. M., Ratrout, N. T., & Aldosary, A. S. (2018). Mode choice behavior of high school goers: Evaluating logistic regression and MLP neural networks. *Case Studies on Transport Policy, 6*(2), 225-230.

Aziz, A., Nawaz, M. S., Nadeem, M., & Afzal, L. (2018). Examining suitability of the integrated public transport system: A case study of Lahore. *Transportation Research Part A: Policy and Practice*, *117*, 13-25.

Belgiawan, P. F., Ilahi, A., & Axhausen, K. W. (2019). Influence of pricing on mode choice decision in Jakarta: A random regret minimization model. *Case Studies on Transport Policy*, *7*(1), 87-95.

Breiman, L. (2001). *Random Forests*. Machine Learning *45*, 5–32.

Buehler, R. (2011). Determinants of transport mode choice: a comparison of Germany and the USA. *Journal of Transport Geography*, *19*(4), 644-657.

Cumming, I., Weal, Z., Afzali, R., Rezaei, S., & Idris, A. O. (2019). The impacts of office relocation on commuting mode shift behaviour in the context of Transportation Demand Management (TDM). *Case Studies on Transport Policy*, *7*(2), 346-356.

Elhenawy, M., Rakha, H. A., & El-Shawarby, I. (2014). Enhanced modeling of driver stop-or-run actions at a yellow indication: Use of historical behavior and machine learning methods. *Transportation Research Record, 2423*(1), 24-34.

Ermagun, A., & Samimi, A. (2015). Promoting active transportation modes in school trips. *Transport Policy*, *37*, 203-211.

Ewing, R., & Cervero, R. (2010). Travel and the built environment: A meta-analysis. *Journal of the American Planning Association*, *76*(3), 265-294.

Field, A. (2009). *Discovering statistics using SPSS*, Sage Publications Ltd.

Gao, Y., Chen, X., Li, T., & Chen, F. (2017). Differences in pupils' school commute characteristics and mode choice based on the household registration system in China. *Case Studies on Transport Policy*, *5*(4), 656-661.

Genuer, R., Poggi, J. M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, *31*(14), 2225-2236.

Giuliano, G., & Dargay, J. (2006). Car ownership, travel and land use: a comparison of the US and Great Britain. *Transportation Research Part A: Policy and Practice*, *40*(2), 106-124.

Hand, D., Mannila, M., & Smyth, P. (2001). *Principles of Data Mining*. United States of America: The MIT Press.

Hou, Y., Edara, P., & Sun, C. (2014). Traffic flow forecasting for urban work zones. *IEEE Transactions on Intelligent Transportation Systems*, *16*(4), 1761-1770.

Hu, H., Xu, J., Shen, Q., Shi, F., & Chen, Y. (2018). Travel mode choices in small cities of China: A case study of Changting. *Transportation Research Part D: Transport and Environment*, *59*, 361-374.

Jahangiri, A., & Rakha, H. A. (2015). Applying Machine Learning Techniques to Transportation Mode Recognition Using Mobile Phone Sensor Data. *IEEE Transactions on Intelligent Transportation Systems*, 1-12.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning with Application in R. *New York: Springer*.

Lanzini, P., & Khan, S. A. (2017). Shedding light on the psychological and behavioral determinants of travel mode choice: A meta-analysis. *Transportation Research Part F: Traffic Psychology and Behaviour*, *48*, 13-27.

Lo, S. H., van Breukelen, G. J., Peters, G. J. Y., & Kok, G. (2016). Commuting travel mode choice among office workers: Comparing an Extended Theory of Planned Behavior model between regions and organizational sectors. *Travel Behaviour and Society*, *4*, 1-10.

Montini, L., Rieser-Schüssler, N., Horni, A., & Axhausen, K. W. (2014). Trip purpose identification from GPS tracks. *Transportation Research Record*, *2405*(1), 16-23.

Pakistan Population Census. (2017). Bureau of Statistic Govt. of Punjab, Pakistan.

Pucher, J., & Buehler, R. (2006). Why Canadians cycle more than Americans: A comparative analysis of bicycling trends and policies. *Transport Policy*, *13*(3), 265–279.

Rasouli, S., & Timmermans, H. J. (2014). Using ensembles of decision trees to predict transport mode choice decisions: Effects on predictive success and uncertainty estimates. *European Journal of Transport and Infrastructure Research*, *14*(4), 412-424.

Rebollo, J. J., & Balakrishnan, H. (2014). Characterization and prediction of air traffic delays. *Transportation Research Part C: Emerging Technologies*, *44*, 231-241.

Shafique, M. A., & Hato, E. (2015). Use of acceleration data for transportation mode prediction. *Transportation*, *42*(1), 163-188.

Shaikhina, T., Lowe, D., Daga, S., Briggs, D., Higgins, R., & Khovanova, N. (2015). Machine learning for predictive modelling based on small data in biomedical engineering. *IFAC-PapersOnLine*, *48*(20), 469-474.

Shakeel, N., Baig, F., & Saddiq, M. A. (2019). Modeling Commuter's Socio-demographic Characteristics to Predict Public Transport Usage Frequency by Applying Supervised Machine Learning Method. *Transport Technic and Technology*, *15*(2), 1-7.

Sharifi, F., & Burris, M. W. (2019). Application of machine learning to characterize uneconomical managed lane choice behaviour. *Case Studies on Transport Policy*, *7*(4), 781-789.

Sperry, B. R., Burris, M., & Woosnam, K. M. (2017). Investigating the impact of high-speed rail equipment visualization on mode choice models: Case study in central Texas. *Case Studies on Transport Policy*, *5*(4), 560-572.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge university press.

United Nations Development Program. (2017). Pakistan National Human Development Report. Pakistan: UNDP.

Weinberger, R., & Goetzke, F. (2019). Automobile ownership and mode choice: Learned or instrumentally rational?. *Travel Behaviour and Society*, *16*, 153-160.

Zahabi, S. A. H., Miranda-Moreno, L. F., Patterson, Z., & Barla, P. (2012). Evaluating the effects of land use and strategies for parking and transit supply on mode choice of downtown commuters. *Journal of Transport and Land Use*, *5*(2), 103-119.

Zaklouta, F., & Stanciulescu, B. (2012). Real-time traffic-sign recognition using tree classifiers. *IEEE Transactions on Intelligent Transportation Systems*, *13*(4), 1507-1514.

Zhang, B. (2012). Reliable classification of vehicle types based on cascade classifier ensembles. *IEEE Transactions on Intelligent Transportation Systems*, *14*(1), 322-332.

Zhang, R., Yao, E., & Liu, Z. (2017). School travel mode choice in Beijing, China. *Journal of Transport Geography*, *62*, 98-110.

Zhang, Y., & Ling, C. (2018). A strategy to apply machine learning to small datasets in materials science. *Npj Computational Materials*, *4*(1), 1-8.

Zhao, C. H., Zhang, B. L., He, J., & Lian, J. (2012). Recognition of driving postures by contourlet transform and random forests. *IET Intelligent Transport Systems*, *6*(2), 161-168.