SMOTE VS. RANDOM UNDERSAMPLING FOR IMBALANCED DATA- CAR OWNERSHIP DEMAND MODEL

Wuttikrai Chaipanha, Patiphan Kaewwichian*

Department of Civil Engineering, Faculty of Engineering, Rajamangala University of Technology Isan, Khon Kaen, Thailand

*E-mail of corresponding author: patiphan.ka@rmuti.ac.th

Resume

Because the numbers of cars reflect each person's travel behaviors for each specific location, the car ownership demand model plays a dominant role in analysis of the travel demand in order to understand each area's individual and household travel behaviors. However, the study project for the master plan of the Khon Kaen expressway represented imbalanced data; namely, the majority class and the minority class were not equal. Before developing a machine learning model, this study suggested a solution to balance the data by using oversampling and under-sampling techniques. The data, which had been improved with SMOTE (Synthetic Minority Oversampling Technique) and kNN (k-nearest neighbors) (k = 5), demonstrated a better effect than the other algorithms that were studied. The TPR (true positive rate) for the rural and suburban areas, which are types of regions with very different imbalance ratios, was calculated before balancing the data at 46.9% and 46.4%. As a result, the TPR values were 63.5% and 54.4%, respectively, following the data balancing.

Article info

Received 28 July 2021
Accepted 31 January 2022
Online 25 March 2022

Keywords:

tour-based model
multiclass classification
k-nearest neighbors
activity-based model

Available online: https://doi.org/10.26552/com.C.2022.3.D105-D115

ISSN 1335-4205 (print version)
ISSN 2585-7878 (online version)

1 Introduction

The most dominant mode of personal travel for a city in extended metropolitan regions is the private car since it is more convenient and flexible. For example, using a vehicle to travel from a suburban or rural area to do business in a CBD (Central Business District) is either a primary destination or an intermediate stop [1]. However, traveling by car influences energy consumption and air quality in an urban area, as well as health problems [2-3].

Regarding the travel demand prediction for transportation planning, household car ownership affects the prognosis of trip frequency choice, destination choice and the mode choice for each trip or tour according to the objectives. It also affects the tour type. Nevertheless, a household's car ownership from each area typically consists of balanced data and imbalanced data [4]. The imbalanced household car ownership was used to predict the Machine Learning Models' travel demand, which represents a state-of-the-art approach [5]. This affects the algorithm training process; specifically, the accuracy and true positive rate would be lower. In the case of the balanced data, the results would be reversed.

Class imbalance is commonly found when the datasets have different members, as most of the minority class contains essential data [6]. Using this imbalanced data to create a model would give an ineffective predicted result due to the decision of the Machine Learning algorithm to rely more on the majority class because it equally focuses on those two classes. In other words, the minority class may be misclassified as the majority class [7-8]. This class disparity is a significant issue in the medical science [9], marketing, banking and manufacturing industries [10], among other fields of study. However, it is still uncommon in transportation planning and is mainly employed when a machine learning model is used to predict car ownership in a household. There has been only one study on Discrete Choice Analysis to analyze the travel demand for business planning, as mentioned in [11]. Therefore, the data imbalance might naturally occur by itself or in limited datasets since a survey require high costs.

A 2-class problem, found in the household's car ownership model, is called a binary classification problem. For example, if there were 100 household car ownership datasets, they would be classified into two groups: 1) households with cars and 2) households

without cars. Meanwhile, if the first group contained 90 datasets and another group had ten datasets, the first group's data with more datasets would be called the majority class. By comparison, the group with fewer datasets would be called the minority class. Using this data to create a model with the classification technique and the machine learning model's basic and standard algorithms (including decision trees (DT), k-nearest neighbors (kNN) and Naive Bayes algorithm), the predicted results might be biased by the majority class.

In contrast, the prediction of another group with fewer datasets seems to be an error and is called data misclassification. It was impossible to correctly classify the datasets in the minority class, or the results might be rarely correct. At the same time, the group with large numbers of sample data has accurate prediction results and an overall high efficiency.

Nevertheless, when the imbalanced data contains more than a 2-class problem; it might be a three-class problem, which is called a multiclass classification. The multiclass classification problem would be transformed into a binary classification problem by assembling every majority class into 1 class. In contrast, the minority class would remain [12].

In order to solve the imbalanced data with an adverse effect on the data classification performance on the minority class before processing the data at a data level, this research offers a helpful technique to improve the data classification for household car ownership with a 3-class problem via two techniques: up-sampling and down-sampling. Weight optimization used these two techniques with feature selection to choose the first ten parameters, which had the optimal weights that could be compared before and after balancing the data. The data classification performance was measured from recall, sensitivity, or from the true positive rate (TPR), F-measure, accuracy and fall-out or false-positive rate (FPR).

Even though the present approach to deal with imbalanced data is well-known, in author's opinion such a strategy has yet to be proven in forecasting a household's automobile ownership with machine learning algorithms. In addition, the Imbalance Ratio (IR), k-fold cross-validation and the variables' main attributes from trip and tour-based models were used in this research. Moreover, it can be used to achieve several vital goals for learning classification by using imbalanced data in travel demand forecasting for transportation planning, such as destination selection and mode selection for each trip or for tour based on the objectives.

This article is composed of the following: an Introduction; Sections 1 and 2, which describe the problem of the imbalanced datasets and the classification performance indicators; Section 3, which presents the solutions to the class imbalance problem at a data level; Section 4, which presents the algorithms selected for the study and the datasets; Section 5, which offers the outcomes; and finally, Section 6, which contains the concluding thoughts, as well as the recommendations for further research.

2 Imbalanced datasets

This section sets forth the imbalanced datasets and later presents the indicators to evaluate the data classification problems in a different manner than a regular classification evaluation.

2.1 The class imbalance problem

In some cases of the data classification problem, the numbers of sample data in every class could differ, especially when the datasets had only 2-class or had more than 2-class transportation planning problems (See [13] for more details). However, it could be stated that the imbalanced data existed when one or more classes had had more datasets than other studies, which would be called the majority class. In contrast, the different types had had fewer datasets, as seen in Figure 1.

This imbalanced dataset had significantly affected the model's prediction performance. Most of the essential and outstanding data had been found in the minority class, such as medical science, marketing, banking, production and in transportation (e.g. choice prediction) [14].

Generally, the data classification algorithm had affected excellent decisions and shown biased results following the majority class. In contrast, the minority class seemed to be misclassified. The machine learning algorithm could not classify the data within the minority class because it had correctly categorized the majority class, showing its high performance. Accordingly, it was challenging to train the algorithm and to accurately predict the data when the datasets were imbalanced. This article focused on the minority class or those households without a car (Class 0). In contrast, the families with one and 2+ vehicles (Class 1 and 2) were determined to be the majority class.

2.2 Evaluation in the imbalanced territory

The performance evaluation of the data classification model from the confusion metrics table, which is one of the broadly used approaches, is typically focused on accuracy, the true positive rate (TPR), the false-positive rate (FPR) and the F-measure.

Namely, accuracy was the accurate value from the model after considering all classes; each class was considered one by one. The TPR was the exact value from the model after considering each category one by one. The F-measure was an evaluation of the precision and the TPR from the model altogether. Each type was

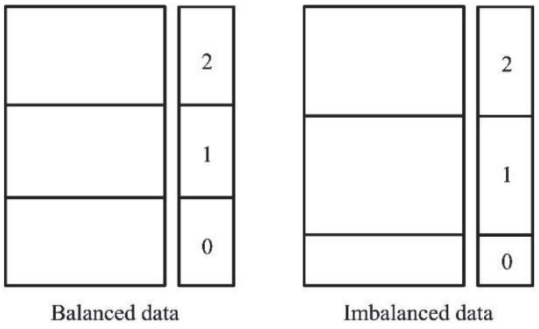


Figure 1 *The illustration of balanced and imbalanced data sets*

Table 1 *The confusion matrix for multiclass classification (for class 0)*

Predicted label classes	True label classes		
	0	1	2
0	TP	FP	FP
1	FN	TN	TN
2	FN	TN	TN

Table 2 *The five-fold cross-validation*

Iteration 1: train on	2	3	4	5	Test on	1
Iteration 2: train on	1	3	4	5	Test on	2
Iteration 3: train on	1	2	4	5	Test on	3
Iteration 4: train on	1	2	3	5	Test on	4
Iteration 5: train on	1	2	3	4	Test on	5

considered one by one. The minority class performance was evaluated mainly from the TPR in the imbalanced datasets since this TPR had described the actual travel distribution [15].

However, this article has mentioned that the TPR had provided an accurate prediction ratio of the minority class (Class 0) and has presented another value for the model's performance evaluation - FPR, which was used to indicate the misclassified majority class ratio.

Table 1 shows the multiclass confusion matrix for Class 0, which was adapted from [16]. True Positive (TP) was the number of the correctly classified data in the target class, Class 0 and it was Class 0. False Positive (FP) was the number of classified data as Class 0, but it was another class. True Negative (TN) was the number of the correctly classified data in any type other than Class 0. False Negative (FN) was the number of the classified data in other categories, but was actually in Class 0. Notably, TN was the opposite of TP and FN was the opposite of FP.

Expression

$$TP+TN/(TP+FN+FP+TN) \quad (1)$$

may be used to calculate accuracy. TPR = TP/(TP+FN) *100; FPR = FP/(FP+TN) *100; and F-measure = (2 * precision * TPR)/(precision + TPR). If these figures were high, precision and TPR would be high as well.

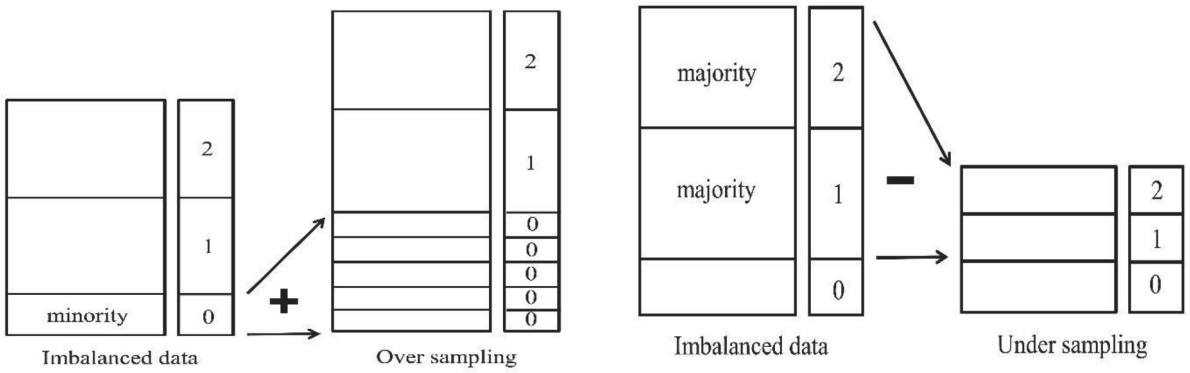
2.3 K-fold cross-validation

The classification model accuracy analysis using the k-fold cross-validation is employed to validate an error in the model's prediction. The K-fold cross-validation was a sampling method, which classified the datasets into several sections (k-fold). Some were tested with the model, while the remainder were used to create the model. After that, the outcomes from the model test were selected for the model's performance evaluation.

The K-fold cross-validation, the datasets were classified equally into several k-folds, e.g., five-folds would find the errors five times. In every calculation round, 1 of 5 folds was singled out to test the model's performance. In contrast, the other four folds were used for algorithm training or model creation. The example is given in Table 2.

3 Solutions in the data level

Before creating the model according to the imbalanced data, the researcher decided to study and balance the data. Therefore, this research suspended the data at a data level by using the Synthetic Minority Oversampling Technique (SMOTE) compared to a random under-sampling technique.



(*a*) (*b*)

Figure 2 *The concept of balancing imbalanced datasets: (a) the oversampling technique and (b) the under-sampling technique*

3.1 Oversampling and under-sampling

Balancing training data is an integral part of data processing. Data imbalances usually exist when the dataset's classes are unequally distributed, which may be a risk while training the model. There are several methods for balancing the data and overcoming the imbalanced training data, which could be performed at either a data level or an algorithm level.

SMOTE, or Synthetic Minority Oversampling Technique, which is a sampling method that is mainly used to fix the imbalanced data with the best results, was developed by [17]. Moreover, it is widely used to solve the imbalanced data sets with statistical knowledge. The SMOTE algorithm is a technique for oversampling the minority classes and minimizing the imbalanced data or equalizing the datasets within the target classes. In contrast, random under-sampling is an approach that is utilized to balance the dataset distribution at each level by randomly deleting the majority class examples. Yet, the major disadvantage of the approach is that some of the necessary and valuable measures might be deleted.

Figure 2 depicts the concepts of the oversampling and under-sampling techniques, which increase the minority class examples (as shown in Figure 2a) and decrease the majority class examples (as shown in Figure 2b). If there were 110 examples in Class 1, there would be 100 examples in Class 2 and 30 examples in Class 0.

4 Experimental framework

In this section, is suggested that the suitable algorithms for study and later describes more imbalanced datasets and related parameters. The final part mentions the statistical test to compare the outcomes from each of the classification algorithms.

4.1 Algorithms selected for the study

When the machine learning model trained the data, it was able to create a model from any of the following three major groups: 1) Geometric models - the models are used for mathematic calculations to define length or weight, including neural network (NN) and k-nearest neighbors (kNN); 2) Probabilistic models - the models are created from the training data probability (e.g. Naive Bayes); and 3) Logical models - these models are used to present data in different logical conditions, (e.g., decision tree (DT)).

4.1.1 k-Nearest Neighbors (kNN)

The kNN algorithm compares the unknown sample to the k training sample, which is the new sample's nearest neighbor. Preliminary theoretical results were published by [18], while a thorough summary was published by [19]. Finding the k closest training examples was the first step in applying the kNN algorithm to a new instance. Then, depending on the number of characteristics in the training example, the "proximity" was calculated from a distance in n-dimensional space.

The distance between the new example and the training examples could be calculated using several metrics, such as the

$$\text{Euclidean distance} = [(x_1-y_1)^2+(x_2-y_2)^2+\ldots+(x_L-y_L)^2]^{1/2}, \quad (2)$$

where x_1 was attributed as 1 of data 1 and y_1 was attributed as 1 of data 2. Thus, the attribute of both data (x and y) was L. However, since the length is frequently based on an absolute value, the data must be normalized before training and using the kNN method.

The kNN algorithm then categorizes the unknown sample by voting on the majority of the neighbors it discovers. In the event of a regression, the predicted

value is equal to the average of the neighbor's found values.

An example of a small class appears scarce in the data space of an imbalanced training dataset. The estimated closest neighbor k is highly likely to identify a sample from a common type given the testing dataset. Small class sizes meant that the test cases would be more likely to be misclassified. This notice is based on research by [20] and [21].

4.1.2 Naive Bayes

Naive Bayes is a method for creating the high-bias, low-variance classifiers and establishing a good model even within a limited dataset. It is a probabilistic classifier that is based on Bayes' theorem. The estimation of a specific component for a given class variable is assumed to be independent by Naive Bayes classifiers;

$$P(C \mid A) = P(A \mid C)*P(C)/P(A), \quad (3)$$

in which P(C | A) is the probability that the data with the attribute A will have Class C. P(A | C) is the chance that attributes A will have Class C in the training data. P(A) is the probability of attribute A, while P(C) is the probability of Class C.

It performs pretty well on large datasets, in which this condition is assumed and holds, even though it demands an unrealistic need that the attribute values are restrictively free [22].

4.1.3 Decision trees (DT)

The DT is an explanation strategy that builds the rule of the DT by summarizing truths or related materials. This method has been used the most frequently because it aids the model in interpreting and making the data more understandable [23]. In this case, repetitive attribute partitioning was used to build the model.

The approach would find each attribute or feature's information gain ratio (IG) at each tree level (starting at the root node) and would then compare it to the class to identify the attribute with the highest IG. It would then be assigned as the decision tree's root (the chosen feature might categorize the data samples for model building and assign them to the same class if it were feasible (maximizing class homogeneity)). The ultimate objective of the decision trees algorithm is to divide all the data into subgroups with similar responses or classes (i.e. the sequence of slicing data to create appropriate if-then rules). The resultant rules can be used to explain an example from root to leaf. All of the information, which is provided, is accurate. In other words, this procedure is repeated until the last node (leaf node). Each node can categorize the sample into separate subgroups with a homogenous class. This procedure then comes to a halt and a decision tree model is generated.

Trees are usually trimmed to increase the predictability of decision structures and to minimize overfitting (See [24] for more details.).

4.2 Datasets

The trip data from the Khon Kaen Expressway Master Plan 2015 (Thailand) Engineering, Economical, Financial and Environmental Feasibility Study was used in this study. Since the population of the research region shares comparable features, systematic random sampling was utilized to obtain a total of 2,015 households consisting of 2 percent of the total households in the target area (4,757 people provided travel information, 616 people provided no travel information). In addition, face-to-face interviews were conducted to obtain information. The participants were

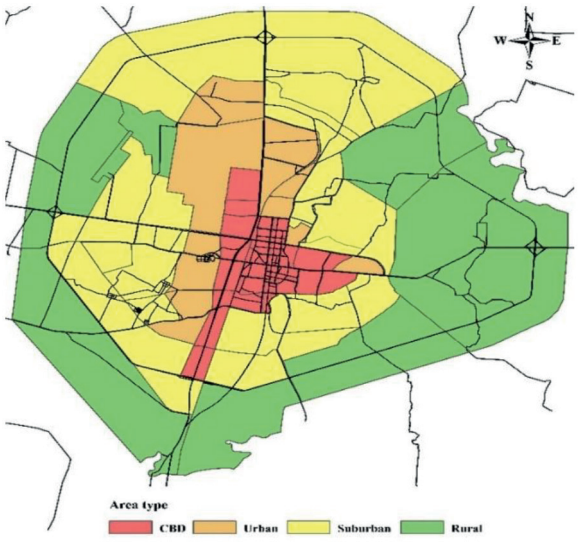


Figure 3 *The Area types in Khon Kaen, Thailand*

Table 3: *The explanatory variables*

Variables[1]	Definitions	Values
Socio-demographic attributes		
Gender	Gender of traveler	Male; female
TDwelling	Type of dwelling	Detached house; commercial buildings; townhouse; condominium/flat, etc.
HHT#	Household type	Two categories of worker vs. non- worker variable and two groups of dwelling type
Income	Household income (US$.)	No income; 0.029-74.5; 74.53-149.0; 149.03-223.5; 223.53-299.0; 299.03-448.5; 448.53-598.0; 598.03-747.5; 747.53-897.0; 897.03-1196.0.; 1196.03-1495.0; 1495.03-2242.5; 2242.53-2990.0; 2990.03-4485.0; >4485
Empstatus	Employment status	A number of full-time workers in the HH; Non-workers; Self-employed; Students; etc.
Apptype	Type of appointment	Number of CEOs; white-collar, blue-collar, red-collar, pink-collar, student, etc.
Kids	Number of children in HH	Numbers
HHS	Household size	Numbers
AGE	Number of people within an age category	<6; 7-19; 20-39; 40-59; 60-79; >80
Zone attributes		
Areatype	Origin or destination area type	CBD (central business district); urban; suburban; rural
PoDwell	Percentage of detached houses	Percentage
HHSlow[2]	Percentage of low-income households	PHHS1to2; PHHS3to4; PHHS5to6
Tour attributes		
Tour type	Number of stops on a tour	1: 1 (more) stop; 0: no stop
Main mode	Mode choice in a tour	Car; motorcycle; motor tricycle; minibus; train; bicycle; walking; other
Numbtrip	Number of trip segments within each tour	Numbers
Accessibility attributes		
Accessibility[3]	Accessibility Measurement	Acci; aij time

Note: [1]The factors were divided into socio-demographic characteristics, tour characteristics and accessibility characteristics. These variables were used to build and assess the models' performance. [2] The percentage of HHS had 2, 4, or 6 family members and the average income was less than 448.5, 897.0, or 1495.0, respectively. [3] A commuter's attempt to overcome the physical and time barrier between zones was measured by accessibility. The aij was the free-flow travel time between traffic zones i and j; acci was the average journey time between location i and a random place within the region [26].

recruited from 73 zones. The GIS information was utilized to categorize the research region. In addition, 10 more zones from suburban and urban areas were included, which increased the number of zones to 83. As illustrated in Figure 3, the residential density divided these zones into four area types: the central business district (CBD), the urban region, the suburban area and the rural area. These area types indicated the source region type and primary destination location for each trip under one tour. Thus, they were variables that indicated the source region type for each trip and the primary destination location under one tour.

The variables' main attributes from the tour and trip-based models were used in this research to create the model shown in Table 3.

4.2.1 Parameters

Each model used the kNN, Naive Bayes and decision trees algorithms to classify the data and the RapidMiner Studio Educational 9.7 Software Tool's default parameters. That is, the default parameters for the decision tree algorithm consisted of "criterion, gain ratio, 20 maximal depth of a tree, 0.25 the confidence level, 0.1 the minimal gain of a node and 2.0 minimal leaf size". The default parameters for the kNN algorithm were "k = 5 [25], measure types: mixed measures, mixed measure: Euclidean Distance"; and the default parameters for Naive Bayes algorithm were "5.0 number of neighbors, 1000 up-sampling size and a 0.5 nominal change rate".

Table 4 *The details of the imbalanced data sets for each area type*

Area types	Instances	Class0	Class1	Class2	Imbalance ratio, IR
Rural	809	211	373	225	2.83
Total	4852	874	2321	1657	4.55
Urban	1054	177	472	405	4.95
CBD.	1358	223	667	468	5.09
Suburban	1631	263	809	559	5.20

Note: As the consideration values for each region type, the imbalance ratio (IR) is defined as the negative class example or majority class divided by the number of positive class examples or minority class. If the IR was more than 9 [27], the dataset would have been significantly skewed. If the IR was less than 9, the dataset imbalance would have been considered as mild or low.

The car ownership ratio for the households in the study area is presented in Table 4, in which Class 0 was the minority class and Classes 1 and 2 were in the majority classes. The data was primarily used to construct and test the performance of the car ownership demand model via each machine learning algorithm before and after the data balancing. In addition, all of the data was compared using the statistical significance tests (T-Test) (alpha = 0.05).

5 Results

This research used a Synthetic Minority Oversampling Technique, SMOTE, to make the minority class datasets equal to the majority class and employed random under-sampling to balance the datasets after the data normalization.

Simultaneously, the weight optimization was selected to choose the first ten variables with the optimal weight to create the model. Meanwhile, the k-fold cross-validation was used to create and validate the data classification models with kNN, Naive Bayes and DT algorithms. Finally, the commission mainly considered the model's performance from balancing and from the imbalanced data in order to predict Class 0 as the minority class.

5.1 The results before and after the data balancing

The results from the model's performance validation, which were obtained by using different algorithms before the data was balanced, is presented in Table 5. The kNN algorithm showed high performance in predicting the minority class (Class 0) when considering the optimal accuracy and the TPR of all the area types and the FPR was low. However, the TPR percentage of each area-type was still low (the lowest = 44.3% for the total area type), in which the data analysis indicated that this problem had occurred due to the imbalanced data as the Imbalanced Ratio (IR) [28]. In each area, the type was in the range between 2.83 - 5.20.

After determining the best performing algorithm, the researcher solved the imbalanced dataset with a 2-part validation. In Part 1, the problem was solved at a data level using SMOTE and random under-sampling. In Part 2, the optimal k-parameter of the kNN algorithm was defined in order to improve the model performance so that the data in Part 1 could be classified.

By using the kNN algorithm to create a model for household car ownership, by performing the data balancing with SMOTE and by employing an under-sampling technique in each area type, the model's performance to predict Class 0 was illustrated as shown in Table 6. From the table, it can be seen that SMOTE demonstrated a higher TPR when compared to the time prior to balancing, but it was lower than the under-sampling technique. The TPR values for each area type were 63.5%, 64.0%, 67.2%, 66.4%, 54.4% and 78.7%, 83.2%, 85.9%, 85.2% and 74.5% for SMOTE and the under-sampling technique, respectively. Nevertheless, the under-sampling techniques might have deleted some important data, resulting in a high FPR value. Therefore, even though the TPR from the under-sampling technique was higher than the results from SMOTE, several misclassified cases were often found (FPR value was high.) [29].

In conclusion, after balancing the dataset with SMOTE and using the kNN algorithm to create the model for household car ownership demand, the model performed highly when predicting the minority class (Class 0) with the essential data compared to the imbalanced data and the balanced data with the under-sampling technique. Furthermore, an oversampling approach was found to be superior to an undersampling strategy for the inadequate data [30]. After balancing the data with SMOTE, another critical problem was selecting the appropriate k value for the kNN method because the neighbors from the k nearest neighbors had been chosen randomly depending upon the amount of oversampling desired.

This step began with a fixed size k = 5 (k-neighborhood) as the default parameter that was used to classify the training data by calculating the distance between the examples and adjusting the other measures for the following calculations (adding different k parameters). Any k parameter lower than 5 (e.g. 1 or 3) would not be counted due to low discrimination power [31]. Finally,

Table 5 *The performance before balancing the training data for the minority class (Class 0)*

Algorithms	Area types	Accuracy (%)	TPR (%)	FPR (%)	F-measure (%)
K-nearest neighbors (kNN)	Rural	63.4	46.9	9.2	54.2
	Total	67.7	44.3	6.7	50.6
	Urban	70.0	53.6	6.1	58.3
	CBD.	69.0	53.8	6.6	57.4
	Suburban	67.2	46.4	7.2	50.5
Naive Bayes	Rural	50.3	20.8	5.8	30.3
	Total	50.5	20.1	5.5	27.7
	Urban	52.8	28.2	7.7	33.9
	CBD.	50.4	36.7	10.5	38.7
	Suburban	51.2	25.1	8.3	29.9
Decision tree	Rural	52.3	15.1	0.3	26.1
	Total	47.8	0.8	0.2	1.6
	Urban	48.3	13.5	0.2	23.6
	CBD.	50.6	8.0	0.6	14.5
	Suburban	50.3	0.0	0.0	0.0

Table 6 *The results of TPR and FPR, when balancing the training data with the kNN algorithm*

Algorithms	Area types	TPR_n (%)	FPR_n (%)	TPR_o (%)	FPR_o (%)	TPR_u (%)	FPR_u (%)
K-nearest neighbors (kNN)	Rural	*46.9*	9.2	*63.5*	16.0	*78.7*	45.0
	Total	*44.3*	6.7	*64.0*	11.0	*83.2*	43.9
	Urban	*53.6*	6.1	*67.2*	7.7	*85.9*	34.7
	CBD.	*53.8*	6.6	*66.4*	9.4	*85.2*	41.7
	Suburban	*46.4*	7.2	*54.4*	12.3	*74.5*	50.1

Note: Subscript "n" refers to the pre-balancing data set, "o" refers to the balancing method by the SMOTE method and "u" refers to the balancing method using a random under-sampling technique

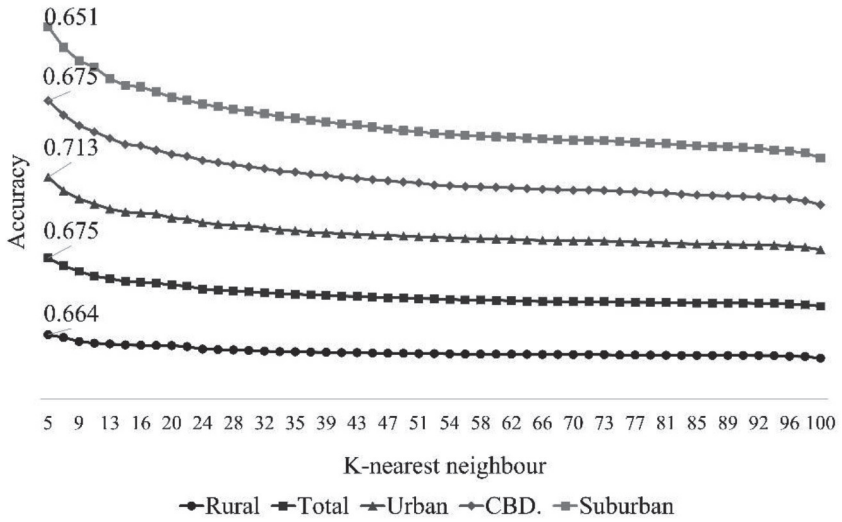


Figure 4 *The results of k-neighborhood parameterization as appropriate for each area type*

the outcome affirmed that the k neighborhood algorithm was suitable with 5 points for the training datasets in all the area types (IR= 2.83 - 5.20) within the study area, as illustrated in Figure 4, representing accuracy from all the area types with different k parameters.

Additionally, when k = 5 in all the area types, it was observed to give the best accuracy for the classification of household car ownership; the accuracy values for each area type were 65.1%, 67.5%, 71.3%, 67.5% and 66.4% for a rural area, the total area, an urban area, CBD and the suburban area, respectively.

It is likely that the target class label had not been uniformly distributed among the categorization jobs. A dataset like this is described as "imbalanced data." Data imbalances might make training a data science model difficult. For example, if the model is trained primarily on the majority class in imbalanced class problems, it will bias the model's prediction towards the majority class.

As a result, it was found that dealing with the imbalanced class is necessary before moving onto the modeling process. Many class balancing approaches tackle class imbalance by either sampling the minority class once again or by eliminating some samples from the majority class. The method for handling class balance strategies is divided into two categories: over-sampling and under-sampling.

One consequence of utilizing under-sampling approaches is that many majority class data points are lost in balancing the class. Over-sampling strategies do compensate for this flaw. However, producing several samples within the minority class may lead to model overfitting.

SMOTE is a common and well-known oversampling technique, which is used by data scientists to generate false minority data points within a cluster of minority class samples. The research outcomes also affirmed that balanced data facilitates the machine learning model, which leads to a more accurate minority class classification.

6 Conclusions and future work

This research created a helpful model for the classification of household car ownership in 5 area types with the imbalanced datasets. However, more than 2-class problems had affected the model's classification of the minority class (Class 0). The researcher recognized the significance of the preparatory step. Therefore, the parameters from the trip-based and tour-based models with weight optimization were selected to find the first ten parameters with optimal model creation. Later, cross-validation was used to create and test the performance of the data classification model before using the over and under-sampling technique to balance the datasets.

Using a geometric model, the k-Nearest Neighbors (kNN) algorithm created a model with balanced data by using SMOTE or the oversampling technique. As a result, the model was able to better classify the datasets because the data balancing had prevented biased results from occurring when the data was imbalanced. Consequently, the efficiency of the classification of the minority class was higher.

Future work should focus on developing the model to have better performance in order to solve the class imbalance with a hybrid sampling method at the data level, at the ensemble classifier, at the semi-supervised classifier at an algorithm level and in the feature selection at the feature level. In addition, the findings should be tested in order to find opportunities to improve the data prediction and to define policies for better urban transportation planning via the machine learning model and households with favorable characteristics.

References

[1] KAEWWICHIAN, P., TANWANICHKUL, L., PITAKSRINGKARN, J. Car ownership demand modeling using machine learning: decision trees and neural networks. *International Journal of GEOMATE* [online]. 2019, **17**(62), p. 219-230. ISSN 2186-2982, eISSN 2186-2990. Available from: https://doi.org/10.21660/2019.62.94618

[2] FEI, C., LIU, R., LI, Z., WANG, T., BAIG, F. N. (2021). Machine and deep learning algorithms for wearable health monitoring. In: *Computational intelligence in healthcare* [online]. MANOCHA, A. K., JAIN, S., SINGH, M., PAUL, S. (eds.). Cham: Springer international publishing, 2021. ISBN 978-3-030-68722-9, eISBN 978-3-030-68723-6, p. 105-160. Available from: https://doi.org/10.1007/978-3-030-68723-6_6

[3] JANDACKA, D., DURCANSKA, D., KOVALOVA, D. Concentrations of traffic-related pollutants in the vicinity of different types of urban crossroads. *Communications - Scientific Letters of the University of Zilina* [online]. 2019, **21**(1), p. 49-58. ISSN 1335-4205, eISSN 2585-7878. Available from: https://doi.org/10.26552/com.C.2019.1.49-58

[4] DOUZAS, G., BACAO, F. Effective data generation for imbalanced learning using conditional generative adversarial networks. *Expert Systems with Applications* [online]. 2018, **91,** p. 464-471. ISSN 0957-4174. Available from: https://doi.org/10.1016/j.eswa.2017.09.030

[5] BASU, R., FERREIRA, J. Understanding household vehicle ownership in Singapore through a comparison of econometric and machine learning models. *Transportation Research Procedia* [online]. 2020, **48**, p. 1674-1693. ISSN 2352-1465. Available from: https://doi.org/10.1016/j.trpro.2020.08.207

[6] JOHNSON, J. M., KHOSHGOFTAAR, T. M. Survey on deep learning with class imbalance. *Journal of Big Data* [online]. 2019, **6**(1), p. 1-54. ISSN 2196-1115. Available from: https://doi.org/10.1186/s40537-019-0192-5

[7] SMITH, M. R., MARTINEZ, T. The robustness of majority voting compared to filtering misclassified instances in supervised classification tasks. *Artificial Intelligence Review* [online]. 2018, **49**(1), p. 105-130. ISSN 0269-2821, eISSN 1573-7462. Available from: https://doi.org/10.1007/s10462-016-9518-2

[8] BRANCO, P., TORGO, L., RIBEIRO, R. P. A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys (CSUR)* [online]. 2016, **49**(2), p. 1-50. ISSN 0360-0300. Available from: https://doi.org/10.1145/2907070

[9] MAZUROWSKI, M. A., HABAS, P. A., ZURADA, J. M., LO, J. Y., BAKER, J. A., TOURASSI, G. D. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural Networks* [online]. 2018, **21**(2-3), p. 427-436. ISSN 0893-6080. Available from: https://doi.org/10.1016/j.neunet.2007.12.031

[10] VUTTIPITTAYAMONGKOL, P., ELYAN, E., PETROVSKI, A. On the class overlap problem in imbalanced data classification. *Knowledge-Based Systems* [online]. 2021, **212**, 106631. ISSN 0950-7051. Available from: https://doi.org/10.1016/j.knosys.2020.106631

[11] DENG, J., LORENZINI, K., KRAUS, E., PALETI, R., CASTRO, M., BHAT, C. Business process and logical model to support a tour-based travel demand. 2014.

[12] HOSENIE, Z., LYON, R. J., STAPPERS, B. W., MOOTOOVALOO, A. Comparing multiclass, binary and hierarchical machine learning classification schemes for variable stars. *Monthly Notices of the Royal Astronomical Society* [online]. 2019, **488**(4), p. 4858-4872. ISSN 0035-8711, eISSN 1365-2966. Available from: https://doi.org/10.1093/mnras/stz1999

[13] KAEWWICHIAN, P. Multiclass classification with imbalanced datasets for car ownership demand model - cost-sensitive learning. *Promet – Traffic and Transportation* [online]. 2021, **33**(3), p. 361-371. ISSN 1848-4069. Available from: https://doi.org/10.7307/ptt.v33i3.3728

[14] WANG, S., WANG, Q., ZHAO, J. Deep neural networks for choice analysis: extracting complete economic information for interpretation. *Transportation Research Part C: Emerging Technologies* [online]. 2020, **118**, 102701. ISSN 0968-090X. Available from: https://doi.org/10.1016/j.trc.2020.102701

[15] BIAGIONI, J. P., SZCZUREK, P., NELSON, P., MOHAMMADIAN, A. Tour-based mode choice modeling: using an ensemble of (un-) conditional data-mining classifiers. In: 88th Annual Meeting of the Transportation Research Board: proceedings. 2008.

[16] RIVAS-PEREA, P., COTA-RUIZ, J., PEREZ VENZOR, J. A., CHAPARRO, D. G., ROSILES, J.-G. Lp-SVR model selection using an inexact globalized quasi-newton strategy. *Journal of Intelligent Learning Systems and Applications* [online]. 2013, **5**(1), p. 19-28. ISSN 2150-8402, eISSN 2150-8410. Available from: https://doi.org/10.4236/jilsa.2013.51003

[17] CHAWLA, N. V., BOWYER, K. W., HALL, L. O., KEGELMEYER, W. P. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* [online]. 2002, **16**, p. 321-357. ISSN 1076-9757. Available from: https://doi.org/10.1613/jair.953

[18] COVER, T., HART, P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* [online]. 2006, **13**(1), p. 21-27. ISSN 0018-9448. Available from: https://doi.org/10.1109/TIT.1967.1053964

[19] AGARWAL, Y., POORNALATHA, G. Analysis of the nearest neighbor classifiers: a review. In: Advances in Artificial Intelligence and Data Engineering: proceedings [online]. 2021. ISSN 2194-5357, eISSN 2194-5365. Available from: https://doi.org/10.1007/978-981-15-3514-7_43

[20] MANI, I., ZHANG, I. kNN approach to unbalanced data distributions: a case study involving information extraction. In: Workshop on Learning from Imbalanced Datasets: proceedings. Vol. 126. 2003.

[21] HRIC, M., CHMULIK, M., JARINA, R. Comparison of selected classification methods in automatic speaker identification. *Communications - Scientific Letters of the University of Zilina* [online]. 2011, **13**(4), p. 20-24. ISSN 1335-4205, eISSN 2585-7878. Available from: http://komunikacie.uniza.sk/index.php/communications/article/view/873

[22] AGRAWAL, R. Predictive analysis of breast cancer using machine learning techniques. *Ingenieria Solidaria* [online]. 2019, **15**(3), p. 1-23. ISSN 2357-6014. Available from: https://doi.org/10.16925/2357-6014.2019.03.01

[23] WOSYKA, J., PRIBYL, P. Decision trees as a tool for real-time travel time estimation on highways. *Communications - Scientific Letters of the University of Zilina* [online]. 2013, **15**(2A), p. 11-16. ISSN 1335-4205, eISSN 2585-7878. Available from: http://komunikacie.uniza.sk/index.php/communications/article/view/648

[24] WETS, G., VANHOOF, K., ARENTZE, T., TIMMERMANS, H. Identifying decision structures underlying activity patterns: an exploration of data mining algorithms. *Transportation Research Record* [online]. 2000, **1718**(1), p. 1-9. ISSN 0361-1981. Available from: https://doi.org/10.3141/1718-01

[25] ZHANG, S., LI, X., ZONG, M., ZHU, X., WANG, R. Efficient kNN classification with different numbers of nearest neighbors. *IEEE Transactions on Neural Networks and Learning Systems* [online]. 2018, **29**(5), p. 1774-1785. ISSN 2162-237X. Available from: DOI: 10.1109/TNNLS.2017.2673241

[26] ALLEN, W. B., LIU, D., SINGER, S. Accessibility measures of U.S. metropolitan areas. *Transportation Research Part B: Methodological*. 1993, **27**(6), p.439-449. ISSN 0191-2615.

[27] GARCIA, S., HERRERA, F. Evolutionary undersampling for classification with imbalanced datasets: proposals and taxonomy. *Evolutionary Computation* [online]. 2009, **17**(3), p. 275-306. ISSN 1063-6560. Available from: https://doi.org/10.1162/evco.2009.17.3.275

[28] BUDA, M., MAKI, A., MAZUROWSKI, M. A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks* [online]. 2018, **106**, p. 249-259. ISSN 0893-6080. Available from: https://doi.org/10.1016/j.neunet.2018.07.011

[29] FERNANDEZ, A., GARCIA, S., HERRERA, F., CHAWLA, N. V. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research* [online]. 2018, **61**, p. 863-905. ISSN 1076-9757. Available from: https://doi.org/10.1613/jair.1.11192

[30] KRISHNAVENI, C. SOBHA RANI, T. On the classification of imbalanced datasets. *International Journal of Computer Science and Technology IJCST* [online]. 2011, **2**(SP1), p. 145-148. ISSN 0976-8491, eISSN 2229-4333. Available from: https://doi.org/10.13140/RG.2.2.14964.24961

[31] NAPIERALA, K., STEFANOWSKI, J. Types of minority class examples and their influence on learning classifiers from imbalanced data. *Journal of Intelligent Information Systems* [online]. 2016, **46**(3), p. 563-597. ISSN 0925-9902, eISSN 1573-7675. Available from: https://doi.org/10.1007/s10844-015-0368-1