International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-3, February 2020

Development of Second-Hand Cars Recommender System Model using the SMOTE and the Random Forest Technique

Sumitra Nuanmeesri, Wongkot Sriurai

Abstract: ***This research aims to develop a model for a second-hand cars recommender system towards the use of the SMOTE approach together with the Random Forest technique. The research team has modified the imbalance of the data by employing the SMOTE method to increase the amount of data to 400% and applying the Random Forest technique to create a decision tree for second-hand car recommendations. After evaluating the model's effectiveness by using the 10-fold validation method, the findings revealed that the second-hand recommender system model applying the SMOTE approach and the Random Forest technique provided an accuracy of 98.84%, a precision of 98.89%, a recall of 98.80% and an F1 score of 98.80%; the overall scores of the model's effectiveness were higher than of the model using only the Random Forest technique. This indicates that the model can be practically used for recommending second-hand cars.***

Keywords: Random Forest, recommender system, second-hand cars, SMOTE.

I. INTRODUCTION

Cars are a kind of facilities in everyday life since today's transportation largely depends on them. Many car dealers, hence, present their cars' information in various ways in order to increase their sales, for example, advertising their cars on television or online platforms. Even though first-hand cars are still desired in the market among those who have middle to high incomes, low-priced used cars are also appreciated as an alternative among low-income consumers. At present, second-hand cars' information is typically provided on different online channels, including websites which recommend cars and give beneficial information to the customers. To illustrate, different car models are suggested based on the website users' personal preferences such as prices, types of functionality, brands, colors, years of production, and so on. The presentation of second-hand cars' information on most websites shares these common features, showing that car dealers' suggestions rather focus on the car specs instead of customers' personal profiles. This research, therefore, proposes the development of a recommender system model for second-hand cars towards the application of the Synthetic Minority Over-sampling Technique (SMOTE) approach and the Random Forest technique; the developed model integrates both car specs and consumer profiles (i.e. careers) in order to suggest cars to the consumers. Subsequently, the model was further refined into a prototype for a second-hand car recommender system, which could be used a tool for making decisions when buying a car with regards to individual preferences.

Revised Manuscript Received on February 24, 2020.
* Correspondence Author
Sumitra Nuanmeesri*, Assistant professor, Department of Information Technology, Faculty of Science and Technology, Suan Sunandha Rajabhat University, Thailand. E-mail: sumitra.nu@ssru.ac.th
Wongkot Sriurai, Assistant Professor, Department of Mathematics Statistics and Computer, Faculty of Science, Ubon Ratchathani University, Thailand. E-mail: wongkot.s@ubu.ac.th

II. RELATED WORKS

In this work, the development of second-hand cars recommender system model using the SMOTE approach and the random forest technique, there are related works as following.

A. Data imbalance resolutions

Data imbalance refers to a disproportionate distribution of classes within a dataset by which the outputs of the data classification are inclined to the majority classes [1]. This research applies the SMOTE in order to resolve the problem mentioned earlier; it is an oversampling technique which increases the number of minority classes, allowing the distribution of data to be more balanced. In practice, this method randomizes a variable and then select its K-nearest neighbor (KNN) before calculating the Euclidean distance between the data points in feature space in order to find the shortest paths between them. After that, dummy data is created between the selected data point and the nearest data point. To illustrate, randomize a variable between five variables, choose one of these neighbors and then place a synthetic point anywhere on the line joining the point under consideration and its chosen neighbor in order to increase the number of minority observations, as shown in (1) [2], [3].

$$X_{new} = x_i + \left(\hat{x}_i - x_i\right)\delta \quad (1)$$

By which

X_{new} refers to a new dataset.

x_i refers to the variable randomized earlier.

$\hat{x}_i$ refers to other random variables (i.e. five more variables are selected).

δ refers to the generated random number between 0-1.

B. Random Forest technique

Random forest technique is a method developing multiple models by randomizing training datasets and features in order to formulate decision trees (DT) [5][6]. Each decision then provides an output. Finally, the outputs from the decision trees are compared in order to find the most suitable answer. Although this method is also decision-tree technique, a variety of training datasets and features produce diverse models or results.

C. The model's effectiveness evaluation

This research evaluated the model's effectiveness by using the 10-fold cross-validation method which split the data into 10 folds; a fold was used as the test set while the rest nine sets were used as training sets. Then, the process was repeated for 10 times by which the test set changed every round until all of the ten sets had been used as the test set. There were four indicators for assessing the model's effectiveness: accuracy, precision, recall, and F-measure, as calculated into the following in (2), (3), (4), and (5) [6] respectively.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (5)$$

By which

TP refers to when the targeted class is "Yes" and the model predicts it "Yes".

TN refers to when the targeted class is "No" and the model predicts it "No".

FP refers to when the targeted class is "No" but the model predicts it "Yes".

FN refers to when the targeted class is "Yes" but the model predicts it "No".

Research studied how to enhance the effectiveness of the decision tree technique when the data was imbalanced by randomizing more minority classes for the data about internet addiction. This research employed the Synthetic Minority Over-sampling Technique (SMOTE) to balance the data sets, and then developed the model by using J48, ID3, Logistic Model Tree (LMT), Classification and Regression Tree (CART), and Random Forest techniques. Subsequently, the model was evaluated by the 10-fold cross validation approach. The findings showed that the Random Forest technique could provide better results than J48, ID3, LMT, and CART techniques [3]. Appling the Random Forest technique to identify and analyze the lung characteristics that could indicate lung cancer. They developed a model and tested 165 cases of Lung Image Database Consortium (LIDC) dataset by comparing three algorithms used for classifying data which were KNN, Support Vector Machine (SVM), and DT. The findings revealed that the Random Forest technique provided better accuracy rate than other techniques by which its accuracy rate was 90.73% [4].

III. RESEARCH METHODOLOGY

This research proposes a second-hand car recommender system model towards the application of the SMOTE and Random Forest techniques. The research process consists of 5 stages as follows: 1) data collection, 2) data preparation, 3) model development, 4) model's effectiveness evaluation, and 5) second-hand car recommender system prototype development. The research process is illustrated in Figure 1.

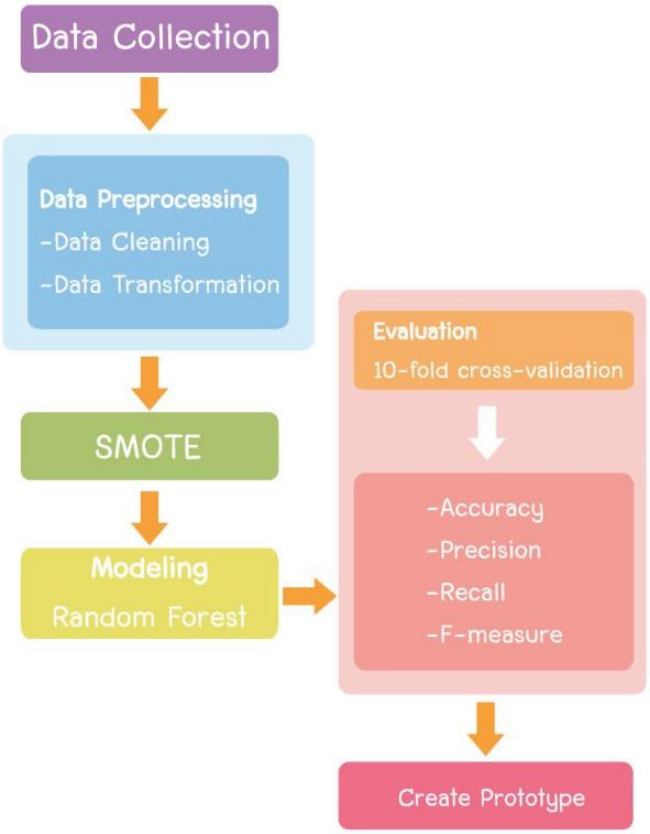

Fig. 1. The research process.

A. Data collection

This research collected data from 600 second-hand car buyers towards the use of questionnaires. The data consisted of the following in Table I.

Table- I: Data collected from questionnaires

No.	Dataset	Dataset information
1	Career Dataset (8 Occupations)	- Doctor
		- Engineer
		- Farmer
		- Merchant
		- Military Officer
		- Nurse
		- Policeman
		- Teacher
2	Car Brand Dataset (5 Brands)	- Ford
		- Honda
		- Mazda
		- Mitsubishi
		- Toyota
3	Car Model Dataset (1997–2012)	- 3000 GTO
		- ACCORD
		- AIRTREK
		- ASPIRE
		- AVANZA

No.	Dataset	Dataset information
		- BIANTE
		- CAMRY
		- CAPRI
		- CHAMP
		- CHARIOT
		- CITY
		- CIVIC
		- COMER
		- COROLLA ALTIS
		- CRV
		- DELICA
		- ESCAPE
		- EVEREST
		- EVOLUTION
		- EXPLORER
		- FIESTA
		- FIGHTER
		- FORTUNER
		- FUSO
		- GALANT
		- GT
		- GTO
		- G-WAGON
		- INNOVA
		- IPSUM
		- JAZZ
		- L200-CYLONE
		- L200-SIRADA
		- LANCER
		- MAZDA 2
		- MAZDA 3
		- MAZDA 6
		- MAZDA 121
		- MAZDA 323
		- MAZDA 323
		- MAZDA 626
		- PRIUS
		- VIOS
		- YARIS
4	Car Price Dataset (4 Ranges of Price)	- Lower than 200,000 THB
		- 200,000–400,000 THB
		- 400,000–600,000 THB
		- Higher than 600,000 THB
5	Car Body Type Dataset (4 Types)	- Compact multi-purpose vehicle (MPV)
		- Pick-up truck
		- Saloon
		- Truck

B. Data preparation

After collecting data, the research team had verified the data and then converted it into .CSV file in order to run it in Weka version 3.9 as illustrated in Figure 2.

	A	B	C	D	E
1	Career	Brands_Ca	V_Car	Price_Car	Type_Car
2	Merchant	Honda	CITY	High	Saloon
3	Military	Mitsubishi	FUSO	High	Truck
4	Military	Mitsubishi	FUSO	High	Truck
5	Merchant	Toyota	COROLLA	VHigh	Saloon
6	Military	Mitsubishi	FUSO	High	Truck
7	Teacher	Toyota	YARIS	Medium	Saloon
8	Military	Toyota	INNOVA	High	Compact MPV
9	Military	Mitsubishi	FUSO	High	Truck
10	Military	Toyota	INNOVA	High	Compact MPV
11	Military	Toyota	INNOVA	High	Compact MPV
12	Policeman	Mitsubishi	CHARIOT	High	Compact MPV
13	Military	Mitsubishi	FUSO	High	Truck
14	Nurse	Toyota	PRIUS	VHigh	Saloon
15	Farmer	Honda	CITY	High	Saloon
16	Doctor	Honda	CIVIC	High	Saloon
17	Military	Toyota	INNOVA	High	Compact MPV
18	Engineer	Honda	ACCORD	High	Saloon
19	Military	Toyota	YARIS	Medium	Saloon
20	Military	Mitsubishi	FUSO	High	Truck

Fig. 2. Data used for modeling.

Table II show the features for classifying and meanings of attributes used for second-hand cars recommendations.

Table- II: Meanings of attributes used for second-hand cars recommendations

Attributes	Symbols	Meanings
Career	Doctor	Doctor
	Engineer	Engineer
	Farmer	Farmer
	Merchant	Merchant
	Military	Military officer
	Nurse	Nurse
	Policeman	Policeman
	Teacher	Teacher
Brands_Car	Ford	Ford car
	Honda	Honda car
	Mazda	Mazda car
	Mitsubishi	Mitsubishi car
	Toyota	Toyota car
V_Car	3000 GTO	3000 GTO model
	ACCORD	ACCORD model
	AIRTREK	AIRTREK model
	ASPIRE	ASPIRE model
	AVANZA	AVANZA model
	BIANTE	BIANTE model
	CAMRY	CAMRY model
	CAPRI	CAPRI model
	CHAMP	CHAMP model
	CHARIOT	CHARIOT model
	CITY	CITY model
	CIVIC	CIVIC model
	COMER	COMER model
	COROLLA ALTIS	COROLLA ALTIS model
	CRV	CRV model
	DELICA	DELICA model
	ESCAPE	ESCAPE model
	EVEREST	EVEREST model
	EVOLUTION	EVOLUTION model
	EXPLORER	EXPLORER model
	FIESTA	FIESTA model
	FIGHTER	FIGHTER model
	FORTUNER	FORTUNER model
	FUSO	FUSO model
	GALANT	GALANT model
	GT	GT model
	GTO	GTO model
	G-WAGON	G-WAGON model
	INNOVA	INNOVA model
	IPSUM	IPSUM model
	JAZZ	JAZZ model
	L200-CYLONE	L200-CYLONE model
	L200-SIRADA	L200-SIRADA model
	LANCER	LANCER model
	MAZDA 2	MAZDA 2 model
	MAZDA 3	MAZDA 3 model
	MAZDA 6	MAZDA 6 model
	MAZDA 121	MAZDA 121 model
	MAZDA 323	MAZDA 323 model
	MAZDA 323	MAZDA 323 model
	MAZDA 626	MAZDA 626 model
	PRIUS	PRIUS model
	VIOS	VIOS model
	YARIS	YARIS model
Price_Car	Low	Low Price (lower than 200,000 THB)
	Medium	Medium Price (200,000-400,000 THB)
	High	High Price (400,000-600,000 THB)
	Vhigh	Very High Price (higher than 600,000 THB)
Type_Car	Saloon	Saloon
	Truck	Truck
	Compact MPV	Compact Multi-Purpose Vehicle
	Pick-up	Pick-up Truck

After collection data towards the use of questionnaires, the data was then verified and analyzed. However, it was found out that there were a few minority classes. Therefore, the research team decided to balance the data sets by using the SMOTE technique to increase the number of minority classes. After modifying the parameters, the test results showed that the best data size that could enhance the model's effectiveness the most was four times of the original size. The number of datasets increased from 260 records to 860 records.

C. Modeling

Following the modification of imbalanced data using the SMOTE technique as explained in data preprocessing, the modified data was then used for developing a second-hand car recommender system model with the application of the Random Forest technique in Weka 3.9.

D. Evaluation

The model's effectiveness was evaluated by the 10-fold cross-validation method, while the decision tree model was developed by the Random Forest technique. There were four indicators for assessing the model's effectiveness: accuracy, precision, recall, and F-measure [7][8].

E. Prototype Development

At this stage, the most effective model was developed into the prototype of a second-hand car recommender system towards the use of the SMOTE and Random Forest techniques, the research team then developed the model into a prototype of a second-hand car recommender system that system could be run by Android smartphones, as illustrated in Figure 3.

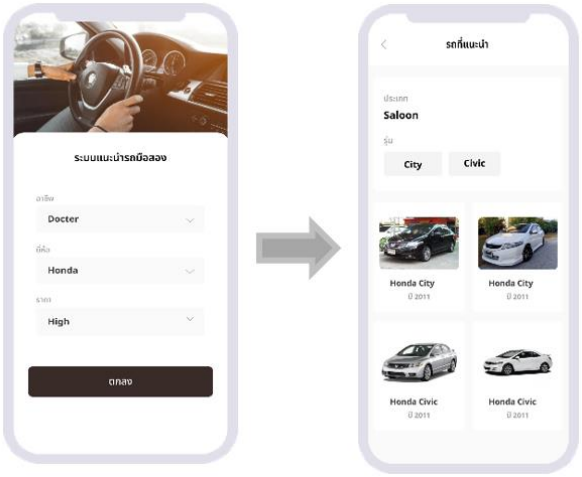

Fig. 3. The display of the second-hand car recommender system.

IV. EXPERIMENTAL RESULTS

The result of presents the development of second-hand cars recommender system model using the SMOTE Approach and the Random Forest technique is following.

A. Data Set Balancing Using the SMOTE Technique

The development of the second-hand car recommender system model applied the SMOTE approach together with the Random Forest technique in which the data was collected from questionnaire respondents who had bought a second-hand car. According to the data collection results, there were a few minority classes of the datasets, so the research team decided to balance the data sets by using the SMOTE method in order to increase the number of minority classes. The test results showed that the best data size that could enhance the model's effectiveness the most was four times of the original size; the number of datasets increased from 260 records to 860 records, as shown in Figure 4.

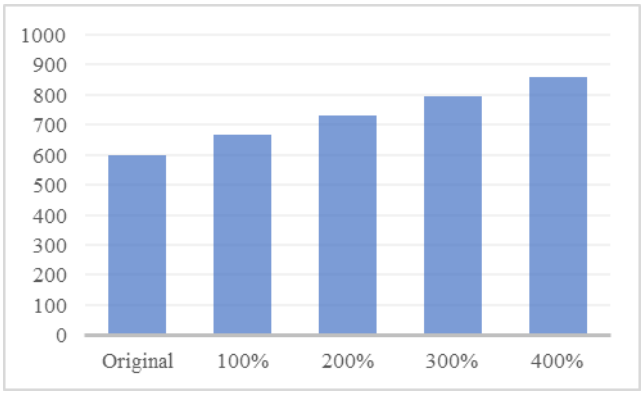

Fig. 4. Results of the evaluation of the model's effectiveness.

B. The Model's Effectiveness Evaluation Results

The imbalanced data was modified towards the application of the SMOTE approach by which the number of minority classes was increased. The test results indicated that the best data size that could enhance the model's effectiveness the most was four times of the original size. After evaluating the dataset with the 10-fold cross-validation method, it was found out that this dataset provided the best accuracy rate (98.84%). Therefore, it was selected to be further developed into a prototype of a second-hand car recommender system, as show Table III.

Table- III: The model's effectiveness evaluation results

Data	Data	Precision (%)	Recall (%)	F-measure (%)	Accuracy (%)
Original	600	98.40	98.30	98.30	98.33
100%	665	98.50	98.50	98.50	98.50
200%	730	98.60	98.60	98.60	98.63
300%	795	98.80	98.87	98.87	98.74
400%	860	98.89	98.80	98.80	98.84

V. CONCLUSION AND DISCUSSION

This article presents the development of a second-hand car recommender system model towards the application of the SMOTE approach alongside the Random Forest technique. After balancing the dataset by using the SMOTE method, it was found out that the most effective data size was four times of the original. Correspondingly, the dataset was used for developing a model by the Random Forest technique.

Then, the model's effectiveness was evaluated by the 10-fold cross-validation method; the evaluation results showed that the modified model, whose data sets were balanced and classified by the SMOTE and the Random Forest techniques,

provided the most accurate results at 98.84% (the accuracy rate of the modified model was higher than of the model using the Random Forest technique solely). The results of this research conform to the study conducted by Palvisut [3] which also applied the SMOTE technique to balancing the data sets and developed a model for data classification of internet addiction by using the Random Forest technique; Palvisut's study also proved that the proposed methodology provided better results than other techniques such as J48, ID3, LMT and CART. The research findings of this study also conform to the study conducted by El-Askary et al. [4] which applied the Random Forest technique in order to identify and analyze the characteristics of lungs which could indicate lung cancer; the aforementioned research presented the data classification technique with more than 90% accuracy.

ACKNOWLEDGMENT

We wish to express our gratitude to the Institute for Research and Development, Suan Sunandha Rajabhat University and Faculty of Science, Ubon Ratchathani University, Ubon Ratchathani Province, who gave us the opportunity to conduct such research.

REFERENCES

1. A. Fernández, S. García, and F. Herrera, "Addressing the classification with imbalanced data: Open problems and new challenges on class distribution," in *Proceedings of the 6th International Conference on Hybrid Artificial Intelligent Systems*, 2011, pp.1–10.
2. N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, 2002, pp. 321–357.
3. P. Palwisut, "Improving decision tree technique in imbalanced data sets using SMOTE for internet addiction disorder data," *Information Technology Journal*, vol. 12, 2016, pp. 54–63.
4. S. N. El-Askary, M. A. M. Salem, and M. I. Roushdy. "Feature extraction and analysis for lung nodule classification using random forest," in *Proceedings of the 2019 8th International Conference on Software and Information Engineering*, 2019, pp. 248–252.
5. L. Breiman, "Random forests," *Machine Learning*, vol. 45, 2001, pp. 5–32.
6. J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random forests and decision trees," *International Journal of Computer Science Issues*, vol. 9, no. 5, 2012, pp. 272–278.
7. D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, 2011, pp. 37–63.
8. S. Nuanmeesri, and W. Sriurai, "Development of the edible and poisonous mushrooms classification model by using the feature selection and the decision tree techniques," *International Journal of Engineering and Advanced Technology*, vol. 9, no. 2, 2019, pp. 3061–3066.

AUTHORS PROFILE

Sumitra Nuanmeesri, received the Ph.D. in Information Technology at King Mongkut's University of Technology North Bangkok, Thailand. She is Assistant Professor in Information Technology Department, Faculty of Science and Technology at Suan Sunandha Rajabhat University, Thailand. Her research interests include speech recognition, data mining, deep learning, image processing, mobile application, supply chain management system, internet of things (IoT), robotics, augmented reality, and virtual reality.

Wongkot Sriurai, received the Ph.D. in Information Technology at King Mongkut's University of Technology North Bangkok. Bangkok, Thailand. She is Assistant Professor in Mathematics Statistics and Computer Department, Faculty of Science, Ubon Ratchathani University, Ubon Ratchathani Province, Thailand. Her research interests include data mining, text mining, web mining, recommender system, information filtering, information retrieval, decision support system, expert system, multimedia technology and computer education.