2020 23rd International Conference on Computer and Information Technology (ICCIT),
19-21 December, 2020

Performance Analysis of Machine Learning Techniques to Predict Hotel booking Cancellations in Hospitality Industry

Md. Shahriare Satu
Dept. of Management Information Systems
NSTU
Noakhali, Bangladesh
shahriarsetu.mis@nstu.edu.bd

Khair Ahammed
Institute of Information Technology
NSTU
Noakhali, Bangladesh
khairahmad6@gmail.com

Mohammad Zoynul Abedin
Dept. of Finance and Banking
HSTU
Dinajpur 5200, Bangladesh
abedinmz@yahoo.com

2020 23rd International Conference on Computer and Information Technology (ICCIT) | 978-1-6654-2244-4/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICCIT51783.2020.9392648

***Abstract*—Hotel booking cancellation is provided a substantial effects on demand management decisions in the hospitality industry. The goal of this work is to investigate the effects of different machine learning methods in hotel booking cancellation process. In this work, we gathered a hotel booking cancellation dataset from Kaggle data repository. Then, different feature transformation techniques were implemented into primary dataset and generate transformed datasets. Further, we reduced insignificant variables using feature selection methods. Therefore, various classifiers were employed into primary and generated subsets. The effects of the machine learning methods were evaluated and explored the best approaches in this step. Among all of these methods, we found that XGBoost is the most frequent method to analyze these datasets. Besides, individual classifiers are generated the highest result for information gain feature selection method. This analysis can be used as the complementary tool to investigate hotel booking cancellation dataset more effectively.**

***Index Terms*—Hotel Booking Cancellation, Machine Learning, Feature Transformation, Feature Selection, Classifier**

I. Introduction

In the hotel management system, booking cancellation act as a major part for decision making regarding associated demands. It affects the services and revenue of the hotel and estimates relevant outputs [6]. The authority of hotel management are concerned about customer's requirement by analyzing these process. There might be many reasons for canceling hotel booking such as lack of facilities, unattractive offers, unexpected occurrences, obligations, illnesses, accidents, etc. Chen et al. [6] represented that the cancellation rates are increasing gradually for the dealing tendency of customers. Sometimes, they evaluate multiple options and finally reject all of them except one. In order to minimize losses, hotel authority employed strict cancellations rules and policies. However, these regulations cannot be applicable in all circumstances because it hinders the quality of services and low refund/non-refund policies decrease revenues as well as number of bookings [5, 16]. Consequently, some hotels are implemented overbooking strategies to get terrible experience where they force to deny service provision [12].

In recent years, machine learning is widely used to extract significant information and predict various aspects in the hotel and tourism management field. Various reasons are happened to cancel hotel booking, but they are not specified more accurately. Morales and Wang [11] stated that it is hard to imagine that one can predict whether a booking will be canceled or not with high accuracy. However, several works were happened where they obtained good result to estimate hotel cancellation. For instance, Huang et al. [9] developed an artificial neural network (ANN) based customer prediction model where backpropagation neural network (BPN) and general regression neural network (GRNN) were investigated 1400 records and detect booking cancellation of Western restaurant chain in Taiwan. Antonio et al. [3] provided a machine learning based prototype that scrutinized instances of hotel management system and employed XGboost (XGB) to predict which booking were likely to cancel and manipulate net demand. Again, they [4] employed mutual information feature selection and several classification methods like boosted decision tree, decision forest, decision jungle, locally deep support vector machine (SVM) and ANN into four hotel databases and predict the cancellation probability. Later, they [1] proposed an automatic machine learning based decision support system which clustered the instances of two hotels and implemented XGboost to capture the changes of cancellation patterns and measured precision of predictions over time. Further, Sanchez et al. [13] implemented random forest, SVM, decision tree, C5.0 and ANN optimized with genetic algorithms in Passenger Name Record (PNR) data and ANN provided 98% accuracy to detect hotel cancellation probability. In this work, we proposed a hotel booking cancellation model that estimates the rejection of customer by investigating customer instances. Then, various feature transformation, selection and classification methods were implemented and explored the best model that can detect hotel cancellation process more appropriately. It estimated the demand of customers and reduced uncertainty in booking cancellation process that is helpful for revenue management, inventory allocation, supplies purchases and pricing decisions as well.

TABLE I: Descriptive Statistics

	Range	Minimum	Maximum	Mean	Std. Deviation	Variance	Skewness	Kurtosis
Canceled	1	0	1	0.37	0.483	0.233	0.537	-1.712
Lead Time	737	0	737	104.01	106.863	11419.722	1.347	1.696
Arrival Year	2	2015	2017	2016.16	0.707	0.501	-0.233	-0.995
Arrival Month	00:00:11	JANUARY	DECEMBER	JUNE	00:00:03.091	9.552	-0.028	-0.995
Arrival Week Number	52	1	53	27.17	13.605	185.100	-0.010	-0.986
Arrival Day of Month	30	1	31	15.80	8.781	77.103	-0.002	-1.187
Stay Weekend Nights	19	0	19	0.93	0.999	0.997	1.380	7.174
Stays Week Nights	50	0	50	2.50	1.908	3.642	2.862	24.285
Adults	55	0	55	1.86	0.579	0.336	18.318	1352.115
Children	10	0	10	0.10	0.399	0.159	4.113	18.674
Babies	10	0	10	0.01	0.097	0.009	24.647	1633.948
Repeated Guest	1	0	1	0.03	0.176	0.031	5.326	26.370
Previous Cancellations	26	0	26	0.09	0.844	0.713	24.458	674.074
Previous Bookings Not Canceled	72	0	72	0.14	1.497	2.242	23.540	767.245
Booking Changes	21	0	21	0.22	0.652	0.426	6.000	79.394
Waiting List	391	0	391	2.32	17.595	309.574	11.944	186.793
adr	5406.38	-6.38	5400.00	101.8311	50.53579	2553.866	10.530	1013.190
Required Car Parking Spaces	8	0	8	0.06	0.245	0.060	4.163	29.998
Total Special Requests	5	0	5	0.57	0.793	0.629	1.349	1.493
Reservation Status Date	1063 00:00:00	10/17/2014	09/14/2017	07/30/2016	229 05:43:17.160		-0.160	-0.903

The organization of this paper is given as follows: section 2 describes the proposed methodology with dataset description, preprocessing, classification and evaluation process briefly. In section 3, the experimental result is shown and described the performance of different classifiers. Then, we summarize this work by indicating some future direction about how to improve this procedure in section 5.

II. Methodology

This work is adopted with several machine learning methods where hotel booking data were processed to obtain good results. Figure 1 illustrates how numerous models were employed into this datset and investigated the effects of individual models. The proposed approach is described more elaborately as follows:

A. Data Description

This data was generated by Antonio, Almeida, and Nunes [2] that contained two hotel booking information like resort (H1) and city hotel (H2). It was gathered from their databases by executing Transact-Structured Query Language (TSQL) query. All instances pertaining hotel or customer identification number were removed. Then, we combined these datasets along with 31 variables where 40,060 observations of H1 and 79,330 observations of H2 were found. However, a city hotel (41.90%) had shown higher cancellation rate than resort hotel (27.69%). It contains three years booking instances from 1^{st} July, 2015 to 31^{st} August, 2017. Table I is shown the descriptive statistics of hotel booking cancellation as follows.

B. Data Preprocessing and Exploration

Data preprocessing is used to clean and manipulate input dataset for further analysis. In addition, different cases such as detecting outliers, missing values and purifying data were scrutinized in order to build well generalized machine learning model. This dataset contains several missing values with 3 columns. As a result, we removed the Company column that leads 94% missing values. Again, the 'reservation_status' column is also removed for being highly correlated (-0.9171) with the target label. Furthermore, several rows with missing values are also considered to be removed for the generalization of the model. Machine learning algorithms can't deal with the categorical variable therefore they are encoded to numerical variables. This process actually converts each unique categorical value corresponding with a numeric number.

C. Feature Transformation and Selection

Feature transformation is created new variables from the existing ones [10]. Three properties were found such as scaling, standardization and normalization [8]. Scaling changes the range of values between 0-1 without affecting the distribution and standardization from mean to 1 [8]. Consequently, normalization converts the distribution into bell-shaped ranges [8]. Likewise, we identified min-max, z-score and square-root methods to change features and evaluate their differences. Min-max method transforms individual features in the range (0-1) and z-score provides the observations of standard deviations below or above the mean. Besides, square root method usually affects the distribution and employed for reducing the right skewness of individual features. Using these techniques, we generated three feature transformed datasets along with primary dataset for further analysis (see Figure 1).

Various feature selection methods were implemented into individual variables and identified them automatically or manually to generate higher outcomes [7]. It detects irrelevant features that reduces time and space complexity, the chances of overfitting and makes machine learning model more simple [7]. There were adopted three algorithms for instance Correlation-Based Feature Selection (CFS), Info Gain Attribute Evaluation (IGAE), Gain Ratio Attribute Evaluation (GRAE) which were also adopted in many previous works

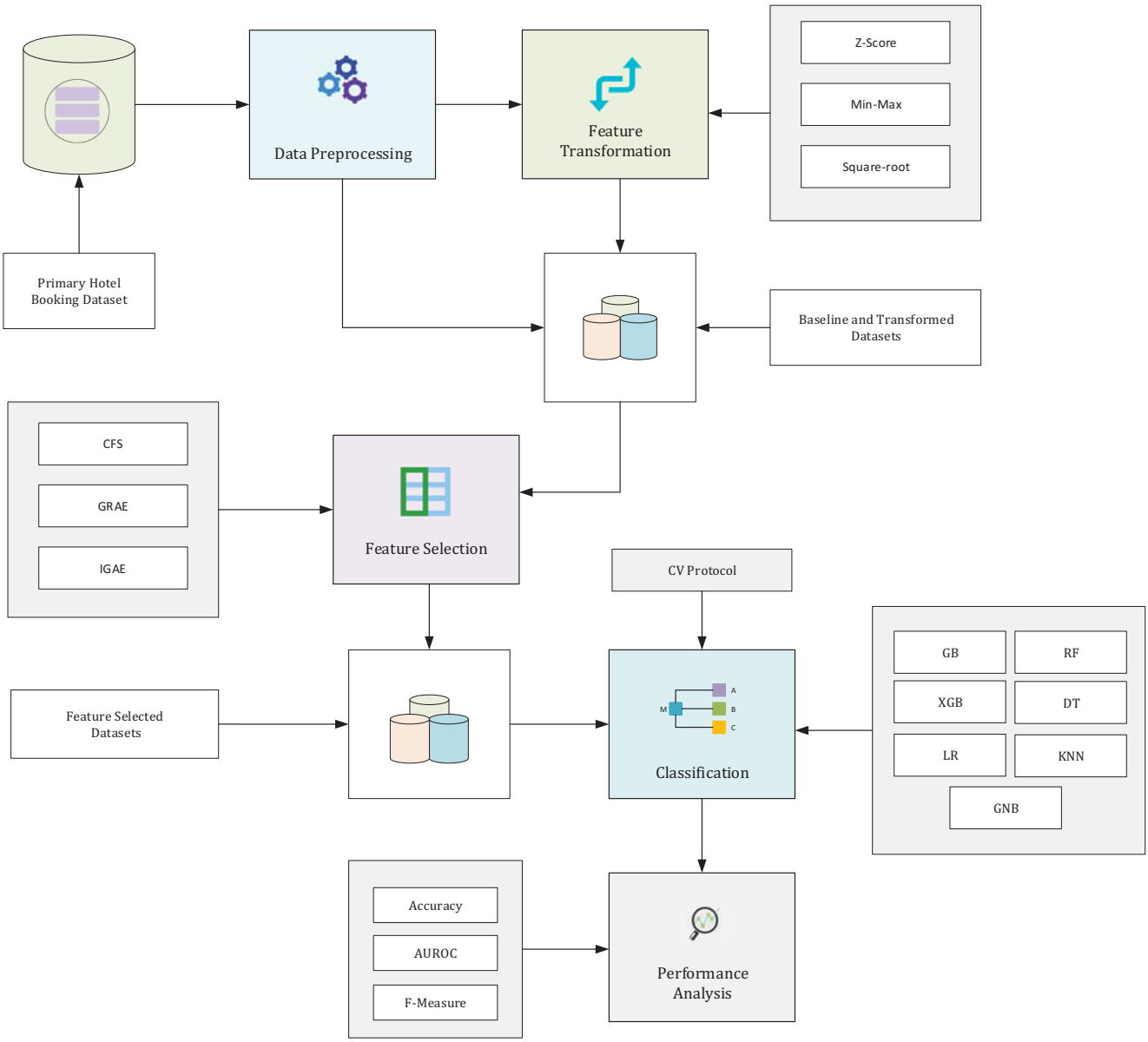

Fig. 1: Proposed Methodology

and generated better results [14, 15]. In CFS, it shows the correlation with the target values and accepts only highly correlated features using best first search method. Therefore, these variables are not maintained Pearson's correlation coefficient or Spearman's. It used different measures like minimum description length (MDL), relief, and symmetrical uncertainty respectively. GRAE takes the size of branches for identifying features and IGAE checks the dependency between the selected and target variable. Both of them were used ranker algorithm to prioritize individual features and manually set a threshold for selecting features. The threshold is defined as 0.0133 and 0.0153 for the IGAE and GRAE respectively. In Figure 1, it illustrates these feature selection methods which were implemented into feature transformed and primary dataset to generate 12 subsets from them.

D. Classification

Classification is a supervising technique that predicts particular classes or labels by analyzing recorded instances. The goal of this work is scrutinized individual related records and estimated booking cancellation of the customer. We employed different classifiers like Gradient Boosting (GB), Random Forest (RF), Xgboost (XGB), Decision Tree (DT), Logistic Regression (LR), K-nearest Neighbour (KNN) and Gaussian Naive Bayes (GNB) into hotel booking dataset and analyze the performance of individual classifiers to estimate how well they

perform and build more efficient model. Therefore, most of them were widely applied into different types of datasets and machine learning projects in previous works [3, 4, 1, 15]. Table II shows related parameters of machine learning classifiers in this work. We considered 10 fold cross-validation that is used to evaluate individual model and not biased to solve the overfitting issues.

TABLE II: Associated parameters of machine learning classifiers

Classifier	Parameters
GB	Max_features = 2, max_depth = 2, random_state= 0
XGB	Learning_rate = 0.1, max_depth = 3
RF	Max_depth = none, random_state = 0
DT	Criterion = gini, splitter = best
LR	Liblinear solver, max_iter = 1000
KNN	K = 5, euclidean distance
GNB	Priors = none, var_smothing = 1eˆ-09

E. Evaluation Metrics

Various evaluation metrics is used to execute machine learning models and explored the best model among all of them. Confusion matrix represents individual instances as true positive (TP), false positive (FP), true negative (TN) and false negative (FN) individually. The performance of the model is evaluated by different metrics: Accuracy, AUROC and F-Measure in this work.

- **Accuracy:** represents the efficiency of the classifier that measures the correctness using the following equation.

$$\text{Accuracy} = \frac{TP + TN}{(TP + TN + FP + FN)} \tag{1}$$

- **AUROC:** provides the probabilities how well the classifiers are separated from negative classes. However, it can be determined by following equation where true positive rate (TPR) and true negative rate (TNR) are used:

$$AUROC = \frac{\text{TPR} + \text{TN rate}}{2} \tag{2}$$

- **F-Measure:** Precision defines the number of the similar instances divided by the total number of existing records and recall provides the number of correctly classified similar records divided by the total number of instances. So, F-Measure is manipulated the harmonic mean of precision and recall that represents as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

$$\text{F-Measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

III. Experimental Result and Discussion

In this experiment, we used scikit learn library to implement different feature transformation, selection and classification techniques respectively. Table III shows all of the experimental results. In CFS subset, GB showed the highest accuracy for primary, min-max and z-transformed dataset. Besides, it provided the best AUROC and f-measure for primary dataset. Again, LR showed the highest accuracy for square-rooted dataset. Further, it obtained the upmost AUROC and f-measure for min-max, z-transformed and square-rooted dataset. In GRAE subset, GB showed the highest accuracy for min-max and z-transformed dataset. XGB showed the uppermost accuracy and AUROC for primary dataset. However, LR showed the highest f-measure for primary, the highest AUROC and f-measure for min-max and z-transformation and highest accuracy, AUROC and f-measure for square-rooted dataset. For IGAE subset, XGB showed the maximum accuracy, AUROC and f-measure for all primary and feature transformed datasets.

Therefore, it is really hard to detect the best algorithm for all cases. In CFS subset, GB and LR showed the best performance where GB is the best classifier for primary dataset and LR provided the best result for transformed dataset more frequently. In GRAE subset, XGB showed the highest result for primary dataset, again LR is the best classifier for transformed dataset. In IGAE subset, XGB is the uppermost classifier both for primary and transformed dataset. The result of primary dataset is more improved than transformed dataset. Therefore, feature transformation techniques were not more feasible to generate better result considering primary dataset. But, feature selection provided good outcomes by reducing unwanted features. So, XGB showed the best accuracy for primary GRAE subset. When, we explored the best AUROC and f-measure, XGB showed these outcomes for all IGAE primary and transformed data. Instead, it is found that GB and XGB were more frequent classifier to generate highest result for primary dataset and LR showed the best result when they are transformed except IGAE subset. Besides, different classifiers are represented the topmost result for IGAE subsets. Figure 2 illustrates the average results of individual classifiers where the graph of IGAE subsets too better than another subsets.

IV. Conclusion and Future Works

In this work, we investigated hotel booking instances in order to reduce losses and maximize annual revenue in the hospitality industry. Hence, various feature transformation, selection and classification algorithms were managed to predict the booking cancellation with high accuracy, AUROC and f-measure. This work was usually evaluated the effects different machine learning procedures in hotel booking cancellation dataset. Besides, a few works were happened where different data analytics approaches have been adopted. As a result, the possibility of further analysis and huge research opportunity will remain in this area. Moreover, advanced machine and deep

TABLE III: Experimental Results

CT	Acc	AUROC	F-M	Acc	AUROC	F-M	Acc	AUROC	F-M	Acc	AUROC	F-M
	Baseline			Min-Max			Z-Transformion			Square-root		
	CFS			CFS			CFS			CFS		
GB	**0.7630**	**0.6835**	**0.7288**	**0.7451**	0.6750	0.7183	**0.7451**	0.6750	0.7183	0.7451	0.6750	0.7183
RF	0.7627	0.6828	0.7281	0.5529	0.5236	0.5540	0.5533	0.5239	0.5542	0.5529	0.5235	0.5539
XGB	0.7628	0.6827	0.7281	0.6870	0.6567	0.6841	0.6870	0.6567	0.6841	0.6870	0.6567	0.6841
DT	0.7627	0.6828	0.7281	0.5111	0.4933	0.5172	0.5113	0.4936	0.5174	0.5132	0.4954	0.5192
LR	0.7550	0.6717	0.7163	0.7370	**0.6966**	**0.7291**	0.7362	**0.6986**	**0.7296**	**0.7687**	**0.7305**	**0.7618**
KNN	0.5245	0.5954	0.5021	0.4901	0.4754	0.4974	0.5870	0.5638	0.5892	0.6164	0.5900	0.6167
GNB	0.5629	0.5986	0.5645	0.7273	0.6647	0.7059	0.7438	0.6692	0.7129	0.6950	0.6508	0.6854
	GRAE			GRAE			GRAE			GRAE		
GB	0.7750	0.7086	0.7525	**0.7419**	0.6750	0.7183	**0.7451**	0.6750	0.7183	0.7451	0.6750	0.7183
RF	0.7559	0.7262	0.7522	0.5529	0.5236	0.5540	0.5529	0.5236	0.5540	0.5529	0.5235	0.5539
XGB	**0.7921**	**0.7521**	0.7281	0.6870	0.6567	0.6841	0.6870	0.6567	0.6841	0.6870	0.6567	0.6841
DT	0.7194	0.6996	0.7195	0.5135	0.4956	0.5195	0.5131	0.4951	0.5191	0.5119	0.4937	0.5179
LR	0.7722	0.7250	**0.7612**	0.7370	**0.6966**	**0.7291**	0.7370	**0.6966**	**0.7291**	**0.7687**	**0.7305**	**0.7618**
KNN	0.6697	0.6341	0.6649	0.4901	0.4754	0.4974	0.4901	0.4754	0.4974	0.6164	0.5900	0.6167
GNB	0.6119	0.6572	0.6111	0.7273	0.6647	0.7059	0.7273	0.6647	0.7059	0.6950	0.6508	0.6854
	IGAE			IGAE			IGAE			IGAE		
GB	0.7690	0.7102	0.7514	0.7690	0.7102	0.7514	0.7690	0.7102	0.7514	0.7690	0.7102	0.7514
RF	0.7561	0.7263	0.7524	0.7560	0.7263	0.7523	0.7560	0.7261	0.7523	0.7552	0.7253	0.7515
XGB	**0.7915**	**0.7524**	**0.7843**	**0.7915**	**0.7524**	**0.7843**	**0.7915**	**0.7524**	**0.7843**	**0.7915**	**0.7524**	**0.7843**
DT	0.7189	0.6982	0.7187	0.7180	0.6968	0.7176	0.7190	0.6984	0.7188	0.7204	0.6996	0.7202
LR	0.7688	0.7211	0.7575	0.7691	0.7175	0.7557	0.7691	0.7213	0.7578	0.7805	0.7361	0.7709
KNN	0.6730	0.6373	0.6681	0.7590	0.7324	0.7563	0.7542	0.7270	0.7514	0.7393	0.7144	0.7375
GNB	0.7447	0.6857	0.7263	0.6202	0.6629	0.6206	0.5312	0.6103	0.5017	0.6390	0.6775	0.6410

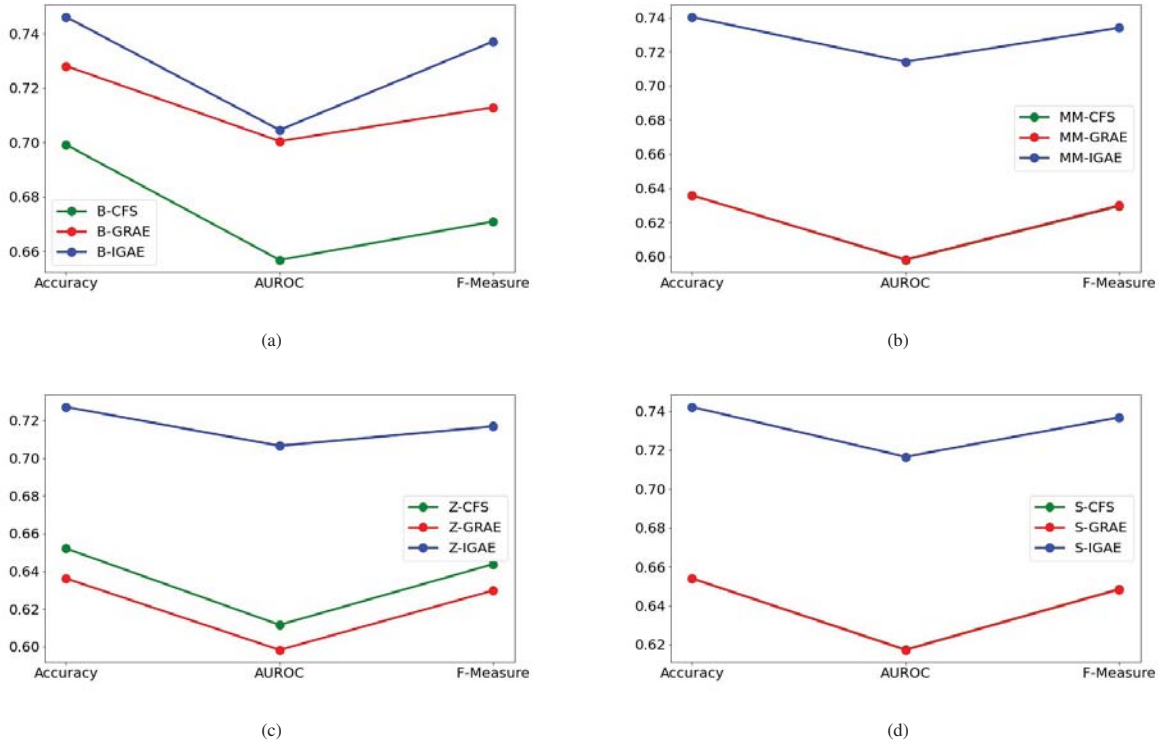

Fig. 2: Average Classification Result of (a) Primary and (b) Min-Max (c) Z-Normalization (d) Square Root Transformed Dataset

learning approach will be useful for building this kinds of predictive model. In future, we will gather more data from heterogeneous sources and propose innovative model that can predict hotel booking cancellation more accurately.

REFERENCES

[1] Nuno Antonio, Ana de Almeida, and Luis Nunes. "An automated machine learning based decision support system to predict hotel booking cancellations". In: *An automated machine learning based decision support system to predict hotel booking cancellations* 1 (2019), pp. 1–20.

[2] Nuno Antonio, Ana de Almeida, and Luis Nunes. "Hotel booking demand datasets". In: *Data in brief* 22 (2019), pp. 41–49.

[3] Nuno Antonio, Ana de Almeida, and Luis Nunes. "Predicting hotel bookings cancellation with a machine learning classification model". In: *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE. 2017, pp. 1049–1054.

[4] Nuno Antonio, Ana De Almeida, and Luis Nunes. "Predicting hotel booking cancellations to decrease uncertainty and increase revenue". In: *Tourism & Management Studies* 13.2 (2017), pp. 25–39.

[5] Chih-Chien Chen, Zvi Schwartz, and Patrick Vargas. "The search for the best deal: How hotel cancellation policies affect the search and booking decisions of deal-seeking customers". In: *International Journal of Hospitality Management* 30.1 (2011), pp. 129–135.

[6] Chih-Chien Chen and Karen Lijia Xie. "Differentiation of cancellation policies in the US hotel industry". In: *International Journal of Hospitality Management* 34 (2013), pp. 66–72.

[7] Manoranjan Dash and Huan Liu. "Feature selection for classification". In: *Intelligent data analysis* 1.3 (1997), pp. 131–156.

[8] Jeff Hale. *Scale, Standardize, or Normalize with Scikit-Learn*. Feb. 2020. URL: https://towardsdatascience.com/scale - standardize - or - normalize - with - scikit - learn - 6ccc7d176a02.

[9] Han-Chen Huang, Allen Y Chang, Chih-Chung Ho, et al. "Using artificial neural networks to establish a customer-cancellation prediction model". In: *Przeglad Elektrotechniczny* 89.1b (2013), pp. 178–180.

[10] Huan Liu and Hiroshi Motoda. "Feature transformation and subset selection". In: *IEEE Intell Syst Their Appl* 13.2 (1998), pp. 26–28.

[11] Dolores Romero Morales and Jingbo Wang. "Forecasting cancellation rates for services booking revenue management using data mining". In: *European Journal of Operational Research* 202.2 (2010), pp. 554–562.

[12] Breffni M Noone and Chung Hun Lee. "Hotel overbooking: The effect of overcompensation on customers' reactions to denied service". In: *Journal of Hospitality & Tourism Research* 35.3 (2011), pp. 334–357.

[13] Agustın J Sánchez-Medina, C Eleazar, et al. "Using machine learning and big data for efficient forecasting of hotel booking cancellations". In: *International Journal of Hospitality Management* 89 (2020), p. 102546.

[14] Md Shahriare Satu, Tania Akter, and Md Jamal Uddin. "Performance analysis of classifying localization sites of protein using data mining techniques and artificial neural networks". In: *2017 International Conference on Electrical, Computer and Communication Engineering (ECCE)*. IEEE. 2017, pp. 860–865.

[15] Md Shahriare Satu et al. "Exploring Significant Heart Disease Factors based on Semi Supervised Learning Algorithms". In: *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*. IEEE. 2018, pp. 1–4.

[16] Scott J Smith et al. "Hotel cancelation policies, distributive and procedural fairness, and consumer patronage: A study of the lodging industry". In: *Journal of Travel & Tourism Marketing* 32.7 (2015), pp. 886–906.