Improving Customer Churn Classification with Ensemble Stacking Method

Mohd Khalid Awang, Mokhairi Makhtar, Norlina Udin, Nur Farraliza Mansor
Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin,
22000 Tembila, Terengganu, Malaysia

***Abstract*—Due to the high cost of acquiring new customers, accurate customer churn classification is critical in any company. The telecommunications industry has employed single classifiers to classify customer churn; however, the classification accuracy remains low. Nevertheless, combining several classifiers' decisions improves classification accuracy. This article attempts to enhance ensemble integration via stack generalisation. This paper proposed a stacking ensemble based on six different learning algorithms as the base-classifiers and tested on five different meta-model classifiers. We compared the performance of the proposed stacking ensemble model with single classifiers, bagging and boosting ensemble. The performances of the models were evaluated with accuracy, precision, recall and ROC criteria. The findings of the experiments demonstrated that the proposed stacking ensemble model resulted in the improvement of the customer churn classification. Based on the results of the experiments, it indicates that the prediction accuracy, precision, recall and ROC of the proposed stacking ensemble with MLP meta-model outperformed other single classifiers and ensemble methods for the customer churn dataset.**

Keywords*—Stacking ensemble; customer churn prediction; bagging; boosting

I. Introduction

The rapid development of wireless telecommunications has altered the course of Malaysia's telecommunications industry [1]. Customers may choose and switch between the packages of various service providers. Churn is a term used to describe the behaviour of customers who switch service providers, and it has become a significant issue for Malaysian network providers.

Numerous researchers have attempted to develop various classifiers to predict customer churn, including the decision tree [2], genetic algorithm [3], and regression analysis [4]. However, the conventional approach of using single classifiers for churn prediction is ineffective. It should be improved, as various uncertainty factors such as customer service, network coverage, product quality, packaging prices, and reception quality can all contribute to customer churn [5].

Furthermore, a set of classifier methods referred to as the ensemble method may be used to improve prediction accuracy. The ensemble approach performs better than individual classifiers because of their divergence or independent character. The ensemble technique combines the choices of many classifiers to enhance classification performance [6].

Multi-classifier ensemble techniques, also known as many classifiers, are machine learning algorithms that include training many base classifiers and then aggregating their output to get the highest possible prediction accuracy [7]. Combining the predictions of several classifiers, such as bagging [8], boosting [9], stacking [10] and ensemble selection [11], maybe a practical approach for improving classification performance.

The rest of this article is structured as follows: Section 2 discusses the review of related literature, including ensemble methods such as bagging, boosting, and stacking. Section 3 covers the research methodology, including the data set and the proposed ensemble stacking. Section 4 presents the experimental setup and results from the discussion. The conclusion of this research is discussed in Section 5.

II. Literature Review

A. Predictive Analytics

Predictive analytics is the most often used technique of predicting customer turnover in the business world. When it comes to predictive modelling, it is a model that can be used to forecast or estimate the target values of future instances [12]. In the context of this research, it is described as the process of forecasting or identifying consumers who are likely to abandon their current purchases in the near future [13].

Predictive analytics is made up of a variety of techniques such as statistical prediction modelling, machine learning modelling, and data mining that analyse previous information and make predictions about future events or something completely new and unknown [14]. Predictive modelling is a technique in which a classifier is usually built based on certain information in order to anticipate the result of a given situation. In accordance with [15], predictive modelling may be divided into four subcategories, as follows:

1) Classification is used when the predicted result is categorical in nature.

2) A regression analysis is used when the prediction results in a numerical value as the result of the analysis.

3) Clustering is the term used to describe the process of grouping a certain collection of items based on their characteristics as a result of the analysis.

4) When the result is the discovery of intriguing connections between data, this is referred to as association rules.

Predictive models are frequently employed in business because they may detect threats and opportunities by identifying trends in historical and transactional data that are inherent in the database. When used correctly, predictive models may discover connections between numerous variables, allowing for risk assessment or possibly linked with a set of particular circumstances, and therefore assist in the decision-making process for a transaction, among other things [16]. Predictive models are capable of overcoming some of the challenges associated with conventional data analysis, such as dealing with large amounts of data and characteristics with a high degree of dimensionality. Making an effective prediction model requires a number of steps that must be completed in order for it to be successful. These steps include data preparation, data quality checking, feature selection, modelling, prediction, and data analysis. It is sometimes called data mining or knowledge discovery to refer to the whole process [12].

B. Data Mining

According to [16], data mining is a logical process used to mine a vast quantity of information to discover a significant piece of information. In order to get information that is usable, quicker, and more productive [17], data mining methods must be used due to the availability of vast quantities of data and the difficulty of the information retrieval process being prohibitively complex. Apart from that, when compared to statistical techniques, this strategy has emerged as one of the most effective options for forecasting future trends [18]. This data mining method has been successfully used in a variety of important sectors. For example, the need for physicians to enhance their prediction models for specific patients necessitates the use of data mining methods to build and improve risk models [18]. There are a variety of data mining methods accessible, each with a different level of appropriateness based on the domain application. Business data mining applications, such as customer churn forecasts, have great promise and are already in widespread usage and application [19]. A potential client who wishes to terminate the service is identified and detected automatically using this tool. Classification, regression, grouping, and association are just a few of the tasks that are involved in data mining [16].

Classification is one of the most important tasks in the field of data mining. Because the output of the predictive model falls into one of two categories (churn or non-churn), the categorisation activity is regarded in this research as customer churn classification. The goal of customer churn classification is to explain the relationships between a variety of variables, such as the customer profile, call history, and payment information. Essentially, there are twenty (20) characteristics that identify the most significant variables that lead to client turnover [20]. When predicting the behaviour of a new unknown consumer, the relationships between characteristics are taken into consideration.

C. Classification in Data Mining

Classification is described as a component of functional learning that assigns a new object to one of many predefined classes. Classification is a two-step process that begins with the creation and training of a classifier model using any classification method. Then, in the second phase, the model is evaluated using a set of test data to determine the classifier's performance and accuracy. Classification is a general term that refers to the process of defining class labels for a data set whose class labels are unknown. Classification techniques are employed in knowledge discovery applications for a variety of purposes, including categorising financial market movements and automatically identifying interesting items in big picture collections [16].

D. Classification Algorithms in Data Mining

When doing data analysis or data mining, classification is a fundamental activity that involves the development of a classifier [12]. It is possible to create a classifier by using a collection of characteristics to describe instances and then assigning them a class label. Classifier induction from data sets including previously classified cases is a fundamental issue in machine learning. Various functional representations, such as decision trees, decision lists, neural networks, decision graphs and rules, are used in a variety of methods to solve this issue.

E. Ensemble Methods

A key concept of the ensemble technique is that it seeks to combine ideas from many individual classifiers in order to get superior results that complement one another [21]. The majority of prior research agrees that accuracy increases when employing an ensemble approach rather than a single classifier, with the condition that the mixers in the combinations must be accurate and varied in order for the accuracy to improve [17], [22]. The idea of this ensemble technique is comparable to the concept of the decision-making process, in which individuals are urged to have a conversation with their colleagues before making any decisions about anything. Before making any major choices, it is common for people to seek second or third views. In general, before a decision is made, individual opinions that may be slightly different from each other will be considered, and then their opinions will be combined to reach the final decision [23]–[25].

The results of ensemble techniques are a set of complementary hypotheses whose predictions are consistent with the evidence that has been seen. When multiple classifiers are fitted to the training data, or when a single classifier is fitted under different training circumstances, these hypotheses are generated. For example, the ensemble approach may be implemented by including randomisation methods into the learning algorithm or by using a variety of heuristics for the estimate of the classifier parameters. In the next step, the ensemble prediction is calculated using averaging or voting procedures to combine the choices of the various components in the ensemble to produce a single prediction [26], [27]. In a discrete variable environment, voting rules are nothing more than simple averages.

F. The Fundamental of Ensemble Methods Data

The ensemble approach for classification problems is shown in Fig. 1, which shows a typical structure. Each phase of the framework is split into four sections, which are as follows:

1) Training set.
2) Base inducer.
3) Diversity generator.
4) Combiner or composer.

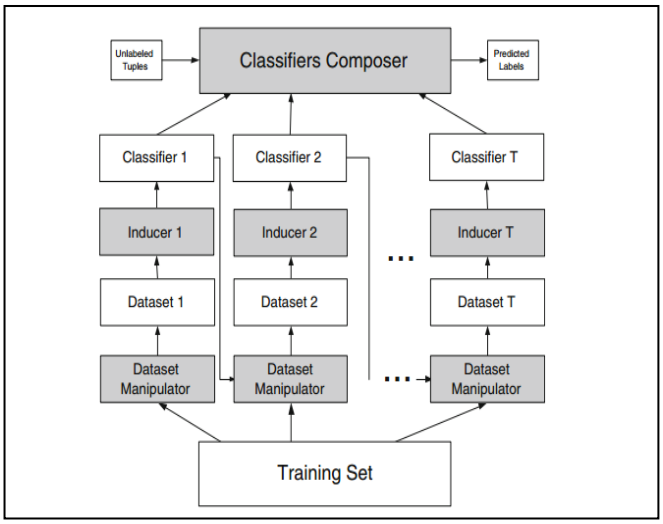

Fig. 1. Ensemble Framework.

The selection of the data set for the training set was the first step in the ensemble's development. Following the selection of the training data set, the subsequent phase involves the generation of the base inducer or ensemble creation, during which the classification algorithms are chosen and trained using the training data set. The diversity generator will guarantee that the basic classifiers have a diverse set of characteristics. At the end of the process, the several classifiers are merged to create the final ensemble.

A study by [28] identifies three kinds of motivations for why ensemble techniques may be better than a single classifier in certain situations. Fig. 2 depicts the problems that need to be addressed.

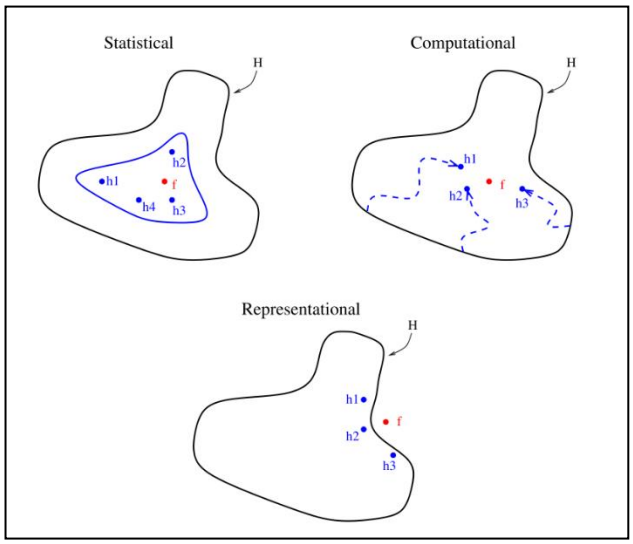

Fig. 2. Three Fundamental Reasons why an Ensemble may Work Better than a Single Classifier.

Statistical issue: When the hypothesis space is too vast to investigate and the available training data is restricted, statistical problems emerge, and there may be many hypotheses that provide the same accuracy on the training data. The issue arises when the learning algorithm selects one of these hypotheses, and there is a chance that the selected hypothesis is incorrect, and therefore the system will be unable to accurately predict future data. Ensemble techniques, on the other hand, suggested combining various ideas, as illustrated in Fig. 2. Combining the hypotheses may minimise or eliminate the statistical issue, as well as the danger of selecting the incorrect hypothesis [29].

Computational issue: A machine learning algorithm, such as a neural network or decision tree, may become trapped in local optima because of the way the search progresses. Finding the optimal hypothesis is always challenging, even if there are ample training data. Instead of searching sequentially from a single location, we use an ensemble approach where we begin at several remote sources. The resulting approximation is thus likely to be closer to the true unknown hypothesis. According to the findings presented in Fig. 2, selecting the incorrect local minimum's risk may be reduced [30].

Representational issue: Even for the vast majority of machine learning problems, no hypothesis can accurately represent the unknown hypothesis in the hypothesis space. When using the ensemble technique, the results presented in Fig. 2 may be feasible to represent even more functions. Since the learning algorithm may be able to formulate a more accurate approximation to the unknown hypothesis, it may be able to get a more accurate solution [28].

Generally, conventional learning methods fail to address difficulties pertaining to the three issues of statistical, constitutional, and representational in nature [29]. In the statistical domain, "high variance" issues are defined as situations in which traditional learning methods fail to address statistical problems. In contrast, the failure of conventional learning methods in computing problems is referred to as a "high variance calculation." A further distinction may be made between classifiers and learning algorithms that suffer from representational problems and those that suffer from a very "high bias." Because of this, ensemble techniques have the potential to mitigate or eliminate the three major shortcomings of conventional learning algorithms.

It is possible to divide the ensemble methods into two main phases: the construction phase and the merging phase. There are at least two main phases in each of the ensemble methods, according to [5]. The creation of ensemble categories should be the first step in the ensemble's growth. It is associated with the combination of the predictions of each classification in an ensemble that the second phase, known as ensemble integration or combination, is performed. However, some researchers recommend ensemble methods that are divided into three phases [5]. Ensemble construction, ensemble trimming, and ensemble combination are the three phases.

1) When the ensemble building phases are completed, they create a collection of heterogeneous base learner classifiers that are used to predict the final output using a given learning technique.

2) As part of the ensemble pruning phase, some fundamental classifiers are eliminated using a variety of mathematical techniques in order to improve the overall accuracy of the ensemble.

3) The third step is the selection and combining of ensembles. During the ensemble selection and combination phase, the filtered learner models are combined to form a single or subset of classifiers, which may provide results that are more accurate than the average of all the individuals' basic classifiers. The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities.

G. Homogeneous Ensemble

The term homogeneous refers to the employment of the same learning technique. Various variables are utilised in the same learning process to create different homogeneous models that are generated from different executions [5], and boosting are two popular methods for creating homogeneous models. The approaches for constructing a homogenous ensemble are as follows:

1) Manipulation of the learning algorithm's parameters.

2) Injection of randomness into the learning process; or

3) Manipulation of the training cases; or

4) Manipulation of the input characteristics and classifier outputs.

Bagging

"Bootstrap" is implied by the term "bagging" [31]. Bagging is based on two basic principles: bootstrap and aggregate. Because the use of several independent base classifiers generally results in a substantial decrease in error, the basis classifiers must be as self-contained as possible. Bagging encourages diversity and variety of classifications by randomly using a part of training data sets to train each classifier in the ensemble. There must be no overlap between the data sets used. The random forest approach, for example, combines this methodology with random decision-making trees to get very high classification accuracy.

Boosting

The boosting algorithm's strength rests in its ability to turn weak categories into strong classifiers. The weak classifier is somewhat better than random predictions, whereas the strong classifier is intuitively close to the optimum performance. The method's origins may be traced back to a basic question: can weak and strong classifiers be combined to achieve a perfect result? Because the number of poor classifiers usually exceeds the high criterion, this idea is very important. According to the boost, every bad classification may be upgraded to strong classification. Getting a bad learner is very easy, while getting a strong learner is more challenging [32].

H. Stacking Ensemble

On two aspects, stacking differs from bagging and boosting. First, stacking often takes into account heterogeneous weak learners, while bagging and boosting mostly take into account, homogeneous weak learners. Second, stacking uses a meta-model to combine the basic models, while bagging and boosting use stochastic methods to combine weak learners.

The heterogeneous ensemble model is created when the classifier uses multiple learning methods on the same data set [33]. Because of the many learning methods, the classifier has a variety of perspectives and predictions. This technique is one way to create many ensembles while guaranteeing excellent ensemble merging outcomes. Each algorithm has its own set of benefits and drawbacks. For example, as compared to the nearest k-neighbor method, neural networks are stronger for noise. The use of a combination of categories may improve categorisation performance. The boosting algorithm's strength rests in its ability to turn weak categories into strong classifiers. The weak classifier is somewhat better than random predictions, whereas the strong classifier is intuitively close to the optimum performance.

I. Literature Review on Customer Churn

Future customer behaviour prediction is one of the most important tasks in company operations, as it serves as the foundation for all strategic choices and planning. According to [34], customer retention leads to higher revenues while simultaneously lowering marketing expenses when compared to selling to new clients. Rather than seeking additional clients who would raise expenses, the long-term profitability is determined by maintaining the appropriate customer base. With growing rivalry from strong rivals in the telecommunications sector, client retention and loyalty management problems are becoming essential. Predicting client behaviour is very difficult due to the fact that they are human and that their happiness is dependent on the quality of customer service and goods provided to them. In order to forecast customer turnover, several academics have attempted to develop different classifiers. These include the decision tree, support vector machine (SVM), neural network, and logistic regression, among others. In the present state of prediction models, most methods are based on single classifiers, which have poor accuracy. The use of numerous classifiers is introduced in some recent studies; however, the methods used are based on various combinations that make use of all the basic classifiers in order to create the final outcome. The common algorithms of the single classifier and the multiple classifiers method in customer churn models are discussed in detail in the following sub-section of this document.

J. Single Classifiers Approach in Customer Churn Prediction

The models of customer attrition prediction that are most often used by researchers are presented in this subsection. Single classifiers such as logistic regression models, decision tree models, support vector machine models, Bayesian models, and artificial neural network models are among the most often used.

Churners were predicted using decision trees and logistic regression, according to a study conducted by [35]. The researchers concluded that logistics regression is an appropriate choice of classifier for incorporating domain knowledge into the model because, in the analysis of the two sets of data, model performance remains relatively stable even after the introduction of domain restrictions when the AUC measure is taken into consideration.

Based on a dataset collected from the 2009 KDD Cup, [36] presented a J48 decision tree and logistic regression. Customers of a French telecoms firm are studied to determine their marketing preferences. They discovered that the accuracy achieved with the decision tree method was much greater than that obtained with the logistic regression technique, indicating that the decision tree technique is superior.

According to [37] study, linear models, such as logistic regression, are a good choice for modelling customer churn prediction, while decision trees are unstable and should not be used. A linear model, according to the study, has a higher level of stability than decision trees, which tend to age quickly and see their performance deteriorate because of this.

Comparison between logistic regression with other algorithms was performed by [38], who sought to discover the most accurate predictors of churn and to assess the accuracy of various data mining methods; their findings were also confirmed. When compared to decision trees and neural networks models, logistic regression demonstrated better performance in their research.

Unlike the decision tree, logistic regression, and other classification algorithms, the neural network, which replicates our human thinking, is a new kind of classification method. It is possible to forecast customer turnover using the neural network learning algorithm in several different ways. Using a Feed Forward Back Propagation (FFBP) Neural Network, [39] developed a classification model for classification problems. The highest level of precision was reached with a 92.35 percent rate of success. There are three neurons in the hidden layer of the prediction model and two neurons in the output layer for churners, and no neurons for non-churners. In order to achieve a balance between churn and non-churn customers, no data pre-processing or sampling technique was used in the proposed model, which includes all characteristics.

The study by [40] asserted that a neural network could achieve maximum output accuracy and demonstrate that it is superior to decision trees and logistic regression. They also claimed that a neural network could achieve maximum output accuracy and demonstrate that it is superior to decision trees and logistic regression. The efficiency of the algorithm, on the other hand, is not only determined by the accuracy of the output but also by other variables such as the time required to make a prediction and the amount of memory resources required to accomplish the job. Although the neural network algorithm was successful in generating high accuracy in this research, the time required and the amount of memory used by the neural network method were both excessive.

The authors of a research [41] developed the particle classification optimisation-based Back Propagation neural network for telecoms customer churn prediction (PBCCP) method, which was published in Nature Communications. They conducted extensive tests with large amounts of telecommunications data and concluded that the PBCCP algorithm provides a significant increase in accuracy when predicting customer turnover when compared to existing classification methods. The author in [42] conducted research in which they used decision trees, artificial neural networks, and support vector machines (SVM) to reduce customer turnover for an Iranian mobile business. Specifically, they discovered that the neural network model outperformed alternative categorisation methods. However, according to [38], logistic regression outperforms the neural network method in terms of accuracy. However, a study conducted by [41] found that decision trees outperformed neural network models on a churn data set for a Taiwanese telecom firm and that this was the case even after controlling for other factors.

The support vector machine (SVM), which has full theoretical underpinnings, is extensively utilised in a broad range of applications. The author in [42] developed a hierarchical reference model for SVM-based classification in customer churn prediction, which is based on a hierarchical reference model. Their experimental design comprised a variety of different classifiers, including logistic regression, classification, and regression trees, among other things. SVM outperformed all other classifiers, according to the researchers, in terms of predictive performance. This result on SVM has also been supported by other research, such as the one conducted by which examined the performance of neural networks, support vector machines, and Bayesian networks. The data set includes all 21 characteristics, and no data pre-processing or sampling methods were employed in the collection of the data. The results of the tests indicate that SVM outperforms all other algorithms used in the experiments. Customer churn prediction accuracy is also influenced by feature factors. One of the model's drawbacks is that it did not make use of any feature selection methods, and it is probable that the accuracy of predictions will be improved if the variable selection is carried out.

K. Ensemble Method Approach in Customer Churn

Customers churn prediction models have been improved by using ensemble methods, which have been suggested by academics to enhance their predictive ability. [43] published one of the first ensemble methods used in a customer churn prediction model, which was one of the first to be used. Back-propagation artificial neural networks and self-organising maps were suggested by the authors as hybrid artificial neural network models, which are a combination of both. It was discovered via the experiments that ensemble models beat the basic model of a single neural network when it came to the accuracy of predictions, the total number of predictions, the total number of errors, and the total number of predictions per second. In particular, the artificial neural network hybrid models perform to their highest potential.

Enhancing is an ensemble technique that tries to create a strong classifier from a collection of weak classifiers in a given situation. Based on these findings, [44] investigated the effects of boosting customer churn prediction models by

utilising logistic regression as a base learner and building separate customer churn prediction models for each cluster of customers. It is compared against a single logistic regression model to see how well it works. Following the results of the experimental assessment, it was discovered that boosting outperformed any single logistics regression model.

A hybrid model based on clustering and ensemble classifiers has been proposed, which was used in several studies [45]. In particular, the self-organising map clustering method, as well as four additional classifier techniques, such as the support vector machine, the decision tree, artificial neural networks, and K-nearest neighbours, were utilised in this study. The authors created 14 models, and the ensemble classifier incorporates all of the basic classifiers. They then examined the accuracy, sensitivity, and specification performance of the various models they created. Compared to other single classification models, the findings indicated that combining the self-organising map with heterogeneous boosting produced the highest performance.

Customers churn prediction was made possible by [46], who developed an intelligent hybrid model based on Particle Swarm Optimization and a Feedforward neural network. If the suggested ensemble model is used in conjunction with other states of the art classification methods, the assessment outcomes of churn consumers are shown to be substantially improved. Another significant result from the proposed model is that the weights of the input characteristics were automatically allocated and optimised by the algorithm. Despite this, it gave weight to each of the input characteristics, and no feature selection was performed prior to the ensemble building process. The second disadvantage of the model is that it makes use of all the basic classifiers in the ensemble combination. An ensemble may be composed of models that are both homogeneous and heterogeneous in nature. This section will go into more depth on each of these major categories, which are homogeneous and heterogeneous, respectively.

III. Methodology

This study is based on a customer dataset obtained from one of the local telecoms providers. There are a total of 272 entries in the datasets, which were subsequently split into two groups: training and testing. Table I includes the specifics of the dataset's input characteristics as well as the label for the dataset's output. The output indicates if the client is a churner or not.

A. Proposed Stacking Ensemble

As shown in Fig. 3, stacking utilises the meta-classifier idea (level-2 classifier) to aggregate the output of the basic classifiers (level-1 classifiers).

Cross-validation is used to prevent overfitting. The following is a broad explanation of the suggested stacking model:

1) Split the customer dataset into training and testing datasets.

2) For the training dataset and split them into k-folds. (test for k=5, k=10, and k=20)

3) For each of the 1st level models (Base classifiers model, test for model 1 to model 6)

a) Train a base model on the k-1 parts

b) Prediction is made on the k^{th} part.

4) Training data set predictions are employed as features for the 2nd level model (meta-model).

5) Then the predictions are made on the test dataset.

TABLE I. The Customer Churn Dataset

Input Features
input X1= The State Code
input X2= The Account length
input X3= The Area code
input X4= The Customer Phone number
input X5= Choice of International Plan
input X6= Choice of Voice Mail Plan
input X7= The Number of voice mail messages
input X8= The Total day minutes
input X9= The Number of day calls
input X10= The Total day charge
input X11=The Total evening minutes
input X12= The Number of evening calls
input X13=The Total evening charge
input X14= The Total night minutes
input X15= The Number of night calls
input X16= The Total night charge
input X17= The Total international minutes
input X18=The number of international calls
input X19= The Total international charge
input X20=The number of calls to customer service
Output Feature Y1=actual result

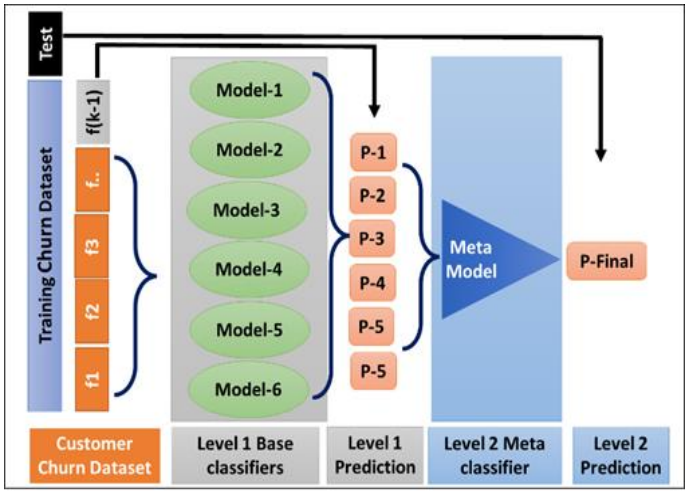

Fig. 3. The Proposed Ensemble Stacking Model for Customer Churn.

B. Level-1 Base Classifiers Construction

The study began with the creation of a level1 model, which is the base classifiers. The first step in creating a successful ensemble technique is to create a varied set of base classifiers in the repository. Different learning algorithms are often used to generate base models, which in turn form the basis of ensembles. Therefore, these ensembles included several kinds of models, all of which provide a desirable degree of variety when it comes to predictions. The pool of classifiers in this study is made up of heterogeneous classifiers created using six different classification learning methods. The selected learning algorithms are as follows:

1) Model-1 = KNeighborsClassifier()
2) Model-2 = DecisionTreeClassifier()
3) Model-3 = SVC()
4) Model-4 = GaussianNB()
5) Model-5 = AdaBoostClassifier()
6) Model-6 = BaggingClassifier

The outputs of the level-1 base classifiers are then used to train a level-2 meta-classifier.

C. Level-2 Meta Classifier Construction

Normally, the meta-model is constructed based on a basic linear model, such as linear regression or logistic regression for regression issues or classification problems. However, any machine learning model or algorithm may act as the meta learner. In this research, various learning algorithms have been employed to evaluate their classification performance and aims to find the best meta-learner model. The selected meta-learners are listed as follows:

1) Meta-Model-1 = KNeighborsClassifier()
2) Meta-Model-2 = MLPClassifier ()
3) Meta-Model-3 = SVC()
4) Meta-Model-4 = GaussianNB()
5) Meta-Model-5 = LogisticRegression()

D. Performance Measurements

The performance of classifiers is an essential part of data mining activities. Generally, the most common performance measure in classification tasks is the percentage of accuracy, which describes the ratio of a total number of correct classifications over the total number of cases. Accuracy is considered an excellent statistic, but only when we have symmetrical datasets with near-identical values for false positives and false negatives. Therefore, we should consider additional factors while evaluating our model's performance. In this experiment, we will consider four types of performance measurements which are as follows:

1) Accuracy
2) Precision
3) Recall
4) ROC

Accuracy equals TP+TN/TP+FP+FN+TN

Precision - Precision is defined as the ratio of properly predicted positive observations to anticipated positive observations in total. This statistic answers the query, "Of all customers classified as churned, how many really churned?" Precision refers to the low incidence of false positives.

Precision is equal to TP/TP+FP.

Recall (Sensitivity) - Recall is defined as the ratio of properly predicted positive observations to all observed positive observations in the actual class - yes. The recall question is: How many customers who really churned did we label?

Recall equals to TP/TP+FN.

ROC - The receiver operating characteristic curve (ROC curve) is a performance metric for classifying issues at different threshold levels. The receiver operating characteristic (ROC) curve indicates the degree or measure of separability, whereas the area under the curve (AUC) represents the degree or measure of separability. It indicates the degree to which the model can discriminate between classes. The larger the AUC, the more accurately the model predicts 0 classes as 0 and 1 classes as 1. For example, the higher the AUC, the more accurate the model is at differentiating churners from non-churners. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

IV. Result and Discussion

To examine the performance of the classification algorithms, we utilised a variety of performance measures. Accuracy, precision, recall, sensitivity, and the ROC curve are the measures in question. The accuracy metric indicates the proportion of properly classified instances; however, it is insufficient for evaluating the classifier's performance. Table II and Fig. 4 show the overall performance of the base model, while Table III presents the performance of different meta-models.

Based on Table II and Fig. 4, we could notice that the best base model is DecisionTreeClassifier with an accuracy of 93.5 percent, precision of 95.1 percent, recall of 93.6 percent and ROC of 94.0 percent. The BaggingClassifier, on the other hand, has an amazing ROC performance of 95.1 percent, but its accuracy of 90.4 percent is somewhat lower than that of the DecisionTreeClassifier.

TABLE II. The Overall Performance of Base-Model Classifiers

Level 1 – Base-model Classifier	**Performance Measurement**			
	Accuracy	Precision	Recall	ROC
KNeighborsClassifier	66.3	69.8	73.3	81.6
DecisionTreeClassifier	93.6	95.1	93.6	94.0
SupportVectorMachine	57.4	57.4	1.0	0.5
GaussianNB	65.7	67.2	79.9	72.9
AdaBoostClassifier	86.7	88.5	88.7	92.8
BaggingClassifier	90.4	92.3	89.5	95.1

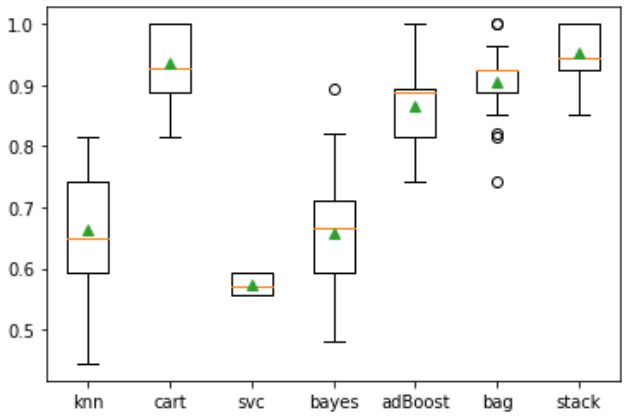

Fig. 4. The Performance of Stacking Ensemble.

Another finding in Table II is that there are few single classifiers with low accuracies, such as the SupportVectorMachine (with a 57.4 percent accuracy), perform poorly compared to other classifiers. The same low performance is also revealed by KNeighborsClassifier and GaussianNB classifiers. Therefore, these ensembles included several kinds of models, all of which provide a desirable degree of variety when it comes to predictions. Consequently, we may suppose that our base-model pool is made up of both excellent and bad classifiers and that the rule of meta-model at the next level is to merge them in order to create a superior model.

TABLE III. THE OVERALL PERFORMANCE OF LEVEL2, META-MODEL CLASSIFIERS

Level 2 – Meta-model Classifier	**Performance Measurement**			
	Accuracy	Precision	Recall	ROC
Stacking Ensemble (KNeighborsClassifier)	94.3	94.7	94.3	95.5
Stacking Ensemble (MultiLayerPerceptron)	95.4	95.9	94.9	97.8
Stacking Ensemble (SupportVectorMachine)	93.9	94.9	95.7	97.2
Stacking Ensemble (GaussianNB)	94.5	95.5	95.1	96.4
Stacking Ensemble (LogisticRegression)	95.3	95.9	95.1	96.8

Based on Table III, we have developed, tested and compared with the six base-model classifiers, KNeighborsClassifier, DecisionTreeClassifier, SupportVectorMachine, GaussianNB, bagging, and boosting, our proposed stacking ensemble classifier has achieved excellent classification results. All meta-models of stacking ensemble classifiers gained significantly better performance than individual classifiers, bagging and boosting.

The Stacking Ensemble (SupportVectorMachine) is the weakest meta-model, with an accuracy of 93.9 percent, although it performs better than the best model in the base-model (DecisionTreeClassifier). According to Table III, the Stacking Ensemble (MultiLayerPerceptron) meta-model classifier surpassed all other models with a classification accuracy of 95.4 percent. In addition, it had the highest ROC of 97.8 percent. Stacking Ensemble (LogisticRegression) performance might also be considered since it attained almost the same accuracy (95.3 percent) as the top meta-model.

The proposed stacking ensemble method to classify customer churn has a high performance since the base classifiers are stacked, combining their predictive power. Different classifiers in this model compensate for the shortcomings of other classifiers, resulting in an overall improvement in performance. The suggested stacking ensemble is a one-of-a-kind combination of heterogeneous base classifiers and meta classifiers that perform best at classification.

V. CONCLUSION

In this study, we employed six different learning algorithms as the base classifiers, and we tested several different meta-model classifiers. It was discovered that the MultiLayerPerceptron meta-model classifier performed the best among the other classifiers. A large number of research studies are being conducted in the area of ensembles of classifiers, and many of them are proposing various kinds of classifiers at the base level and at the meta-level, depending on the type of application being investigated. This study contributes to the area of data mining research by suggesting an effective combination of base and meta-level classifiers for a customer churn classification. Thus, this study strongly indicates that the proposed ensemble stacking model outperformed any single classifiers, bagging and boosting ensemble, which is also in accordance with the previous research findings in other application areas.

When compared to single and ensemble techniques for predicting customer churn, our proposed ensemble stacking model has proven to be superior. However, we have only tested our proposed model on the selected customer churn dataset, and we intend to validate it on additional datasets in the future, both in terms of customer churn datasets and other types of datasets, in order to determine whether our approach can be applied to different kinds of problems.

REFERENCES

[1] M. A. Hajar, D. N. Ibrahim, and M. A. Al-shara, "Value Innovation in the Malaysian Telecommunications Service Industry: Case Study," in International Conference of Reliable Information and Communication Technology, 2018, pp. 892–901.

[2] A. De Caigny, K. Coussement, and K. W. De Bock, "A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees," Eur. J. Oper. Res., vol. 269, no. 2, pp. 760–772, 2018.

[3] E. Stripling, S. vanden Broucke, K. Antonio, B. Baesens, and M. Snoeck, "Profit maximising logistic model for customer churn prediction using genetic algorithms," Swarm Evol. Comput., vol. 40, pp. 116–130, 2018.

[4] H. Jain, A. Khunteta, and S. Srivastava, "Churn Prediction in Telecommunication using Logistic Regression and Logit Boost," Procedia Comput. Sci., vol. 167, no. 2019, pp. 101–112, 2020.

[5] N. N. Y. Vo, S. Liu, X. Li, and G. Xu, "Leveraging unstructured call log data for customer churn prediction," Knowledge-Based Syst., vol. 212, p. 106586, 2021.

[6] L. Rokach, Pattern Classification Using Ensemble Methods. World Scientific, 2010.

[7] L. Rokach, "Ensemble-based classifiers," Artif. Intell. Rev., vol. 33, no. 1–2, pp. 1–39, 2010.

[8] Q. L. Zhao and Y. H. Jiang, "Incremental learning based on ensemble pruning," in 2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2011, pp. 377–381.

[9] D. Opitz and R. Maclin, "Popular Ensemble Methods: An Empirical Study," J. Artif. Intell. Res., vol. 11, pp. 169–198, 1999.

[10] U. Sultan et al., "Master Machine Learning Algorithms: discover how they work and implement them from scratch.," Appl. Soft Comput. J., vol. 77, pp. 188–204, 2016.

[11] Y. Yang, G. Wang, Z. Zhang, and K. Tian, "A novel emotion recognition approach based on ensemble learning and rough set theory," 9th IEEE Int. Conf. Cogn. Informatics, pp. 46–52, 2010.

[12] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The KDD process for extracting useful knowledge from volumes of data," Commun. ACM, vol. 39, no. 11, pp. 27–34, 1996.

[13] V. Lazarov and M. Capota, "Churn Prediction," TUM Comput. Sci., 2007.

[14] X. Wu, T. Ma, J. Cao, Y. Tian, and A. Alabdulkarim, "A comparative study of clustering ensemble algorithms," Comput. Electr. Eng., vol. 68, no. August 2017, pp. 603–615, 2018.

[15] C. Elkan, Predictive analytics and data mining. 2010.

[16] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," AI Mag., vol. 17, no. 3, p. 37, 1996.

[17] I. H. Witten and E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques," Data Min. Pract. Mach. Learn. Tools Tech., 2016.

[18] T. G. Dietterich, "Ensemble methods in machine learning," Lect. Notes Comput. Sci., vol. 1857, pp. 1–15, 2000.

[19] A. A. Ahmed, "Methods For Customer Retention In Telecom Industries," 2017.

[20] M. K. Awang, M. R. Ismail, M. Makhtar, and M. A. Nordin, "Performance Comparison of Neural Network Training Algorithms for Modeling Customer Churn Prediction," p. 94.

[21] M. Mohamad, M. Y. M. Saman, and N. A. Hamid, "Complexity Approximation of Classification Task for Large Dataset Ensemble Artificial Neural Networks," Lect. Notes Electr. Eng., vol. 520, no. April, pp. 195–202, 2019.

[22] M. Mohamad, M. Y. M. Saman, and M. S. Hitam, "The use of output combiners in enhancing the performance of large data for ANNs," IAENG Int. J. Comput. Sci., vol. 41, no. 1, pp. 38–47, 2014.

[23] R. Polikar, "Ensemble based systems in decision making," Circuits Syst. Mag. IEEE, vol. 6, no. 3, pp. 21–45, 2006.

[24] C. F. Tsai and M. Y. Chen, "Variable selection by association rules for customer churn prediction of multimedia on demand," Expert Syst. Appl., vol. 37, no. 3, pp. 2006–2015, 2010.

[25] R. Rosly, M. Makhtar, M. K. Awang, and M. A. Nordin, "The Study on the Accuracy of Classifiers for Water Quality Application," Int. J. u- e-Serv. Sci. Technol., vol. 8, no. 3, pp. 145–154, 2015.

[26] M. Wozniak and M. Zmyslony, "Chosen problems of designing effective multiple classifier systems," in 2010 International Conference on Computer Information Systems and Industrial Management Applications, CISIM 2010, 2010, pp. 42–47.

[27] M. Makhtar, D. C. Neagu, and M. J. Ridley, "Comparing multi-class classifiers: On the similarity of confusion matrices for predictive toxicology applications," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 6936 LNCS, pp. 252–261, 2011.

[28] T. G. Dietterich, "Machine-learning research," AI Mag., vol. 18, no. 4, pp. 97–136, 1997.

[29] Zhi-Hua Zhou, Ensemble Methods Foundations and Algorithms. Cambridge, UK AIMS: Chapman & Hall/CRC, 2014.

[30] Z. Zhou and W. Tang, "Selective ensemble of decision trees," Lect. Notes Comput. Sci., vol. 2639, pp. 476–483, 2003.

[31] L. Breiman, "Bagging Predictors," Mach. Learn., vol. 24, no. 421, pp. 123–140, 1996.

[32] Y. Freund, "Boosting a weak learning algorithm by majority," Inf. Comput., vol. 121, no. 2, pp. 256–285, 1995.

[33] G. Tsoumakas, L. Angelis, and I. Vlahavas, "Selective fusion of heterogeneous classifiers," Intell. Data Anal., vol. 9, no. 6, pp. 511–525, 2005.

[34] E. Ascarza, "Retention Futility: Targeting High-Risk Customers Might Be Ineffective," J. Mark. Res., vol. 55, no. 1, pp. 80–95, 2016.

[35] E. Lima, C. Mues, and B. Baesens, "Domain knowledge integration in data mining using decision tables : case studies in churn prediction," J. Oper. Res. Soc., vol. 60, pp. 1096–1106, 2009.

[36] K. Dahiya, "Customer Churn Analysis in Telecom Industry," 4th Int. Conf. Reliab. Infocom Technol. Optim. (ICRITO)(Trends Futur. Dir., pp. 1–6, 2015.

[37] M. Owczarczuk, "Churn models for prepaid customers in the cellular telecommunication industry using large data marts," Expert Syst. Appl., vol. 37, no. 6, pp. 4710–4712, 2010.

[38] A. A. Khan, S. Jamwal, and M. M. Sepehri, "Applying Data Mining to Customer Churn Prediction in an Internet Service Provider," Int. J. Comput. Appl., vol. 9, no. 7, pp. 8–14, 2010.

[39] S. Babu, N. R. Ananthanarayanan, and V. Ramesh, "A Study on Efficiency of Decision Tree and Multi Layer Perceptron to Predict the Customer Churn in Telecommunication using WEKA," Int. J. Comput. Appl., vol. 140, no. 4, pp. 26–30, 2016.

[40] S. Khodabandehlou and M. Zivari Rahman, "Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior," J. Syst. Inf. Technol., vol. 19, no. 1–2, pp. 65–93, 2017.

[41] C. F. Tsai and M. Y. Chen, "Variable selection by association rules for customer churn prediction of multimedia on demand," Expert Syst. Appl., vol. 37, no. 3, pp. 2006–2015, 2010.

[42] S. Lessmann and S. Voß, "Computational Intelligence and Information Management A reference model for customer-centric data mining with support vector machines," Eur. J. Oper. Res., vol. 199, no. 2, pp. 520–530, 2009.

[43] F. T. Chih and H. L. Yu, "Data mining techniques in customer churn prediction," Recent Patents Comput. Sci., vol. 3, no. 1, pp. 28–32, 2010.

[44] N. Lu, H. Lin, J. Lu, and G. Zhang, "A customer churn prediction model in telecom industry using boosting," IEEE Trans. Ind. Informatics, vol. 10, no. 2, pp. 1659–1665, 2014.

[45] M. Fathian, Y. Hoseinpoor, and B. Minaei, "Offering a hybrid approach of data mining to predict the customer churn based on bagging and boosting methods," Kybernetes, vol. 45, no. 5, pp. 732–743, 2016.

[46] H. Faris, "A hybrid swarm intelligent neural network model for customer churn prediction and identifying the influencing factors," Inf., vol. 9, no. 11, pp. 1–18, 2018.