A Comparative Analysis of The Performance of Homogenous Ensembles on Customer Churn Prediction

Ramoni Tirimisiyu Amosa, Fabiyi Aderanti Alifat, Olorunlomerue Adam Biodun, Oluwatosin Adefunke Oluwatobi & Ugwu Jennifer Ifeoma

Computer Science Department, Federal Polytechnic Ede, Osun State, Nigeria

Abstract: **Customer retention is a challenging and critical issue in telecommunication and service-based sectors. Various researchers have established the need for a service-based company to retain their existing customers much cheaper than acquiring new ones. However, the predictive models for observing customers' behavior is one of the great instruments in the customer retention process and inferring the future behavior of the customers. Selecting the right and best model is another herculean task because the performances of predictive models are greatly affected when the real-world dataset is highly imbalanced. The study analyses the performance of homogeneous ensembles; bagging, boosting, rotation forest, cascade, and dagging. These ensembles were applied to both raw and balanced datasets to compare the performance of the models. The data sampling method (oversampling) was adopted to balance the raw dataset. The primary metric used for the evaluation of the performance of the models was Accuracy and ROC/AUC (Receiver Operating Characteristics/Area Under Curve). Weka 3.8.5 machine learning tool used to analyze and develop the models. The study reveals that Bagging had the best performance having an AUC of 0.987, followed by boosting and Rotation Forest both with an AUC of 0.985.**

Keywords: **Customer, Dataset, Ensemble, Homogeneous, Model.**

I. INTRODUCTION

Due to dominance of mobile communication in telecommunication sector, new ideas, technologies and players are emerging daily and this has made it necessary and important to predict the customers and client who may have to shift from one service provider to a new one. Churn occur when a customer leave a service provider and move to the new service provider, in certain situation customer churn is also refers to as customer attrition. If a customer switches a service provider's company then face loss occurs in the company's revenue. Prediction can be performed to identify the potential churners and retention solutions may be provided to them (Bilal et al., 2022). A large number of mining algorithms are available which classify the behavior of customers into churner and non-churners.

The telecommunication sector is one of the major source of revenue and very crucial to the economy development in developed countries for almost two decades (Ullah et al., 2019; Adnan, 2021). Data mining plays a vital role for prediction and analysis in the telecom industry due to availability of huge data. The basic application area is to perform prediction of churner in order to save customer retention and to make a high-profit rate. Data mining techniques are used in the telecom sector to observe the churn behaviour of the customers.

The rate of telecom user keep on increasing daily, hence telecom companies, they now offer variety of services for the retention of customers as various researcher have established that retaining the existing customers is more economical than attracting new customers (Ahmad, 2019; Almuqren, 2021; Amin et al., 2016). Retaining existing customer will require the telecommunication companies to reduce the number of customer that may leave the services to the next service providers. However, it becomes imperatives to predict and prevent prospective churners in order to prevent loss of customer and revenue. Through retention solution. Telecom customers churn for reasons which include; Price hike (Al-Rifaie & Alhakbani, 2016), poor service (Ahmed & Maheswari, 2017), relocation by customer, privacy issue Entry of new competitor into the market. A large number of mining algorithms ranging from tree based, function, lazy, rule based and so on are available which classify the attitude of clients/customers into churner and non-churners. A more reliable and accurate prediction model is highly important for right identification of customer's churn and therefore performs a crucial role in making decisions about their retention (Diala, 2019). Various prediction model in existence were built with imbalance dataset which definitely affect the performance of the classifiers and algorithm and there result in biased classification. Class imbalance occurs when the classes in a dataset are skewed or not equal.

This research proposed churn prediction models using five different ensemble and five tree-based classifiers. Ensemble include; bagging, boosting, cascade, rotation forest and dagging while the classifiers are CS-Forest, Random Forest, REPTree (Reduced Error Pruning Tree), BFTree (Best-First Tree) and Sysfor (Systematically Developed Forest of Multiple Decision Tree). The result of the model were evaluated and compared in order to fix out the best ensemble here that can reliably and rightly predict the potential churners. This research has contributed in the following ways;

i. Identify the best ensemble method using tree-based classifiers
ii. Established that class imbalance is a serious challenge that need to be addressed before building a prediction model in order to have a reliable and accurate prediction model.

II. LITERATURE REVIEW

Bilal et al. (2022) in their research titled an ensemble based approach using a combination of clustering and classification algorithms to enhance customer churn prediction in telecom industry. The researchers built hybrid churn prediction model that is based on a combination of clustering and classification algorithms using an ensemble. The hybrid models were introduced by combining the clusters with seven different classification algorithms individually and then evaluations were performed using ensembles, clustering algorithms used include K-means, K-medoids, X-means and random clustering. The accuracy achieved for the two datasets used are 94.7% and 92.43%.

Ullah et al. (2019) proposed the JIT-CCP model, which predicts churn. Data preprocessing comes first in this paradigm, followed by binary classification and performance evaluation using the values of the confusion matrix that correspond to true positives, true negatives, false positives, and false negatives. Probability of detection is computed using these terms. The accuracy of various classifiers is determined using the probability of detection (PD). The results of the classifier are significantly better if the PD value is close to 1, and vice versa. The suggested model, however, cannot handle a lot of data. The modal goes through various stages. Data is cleaned and various deviations from data are removed in the first stage. The following stage involves gathering testing and training data sets from several clusters. Algorithms for prediction are then used. Accuracy, specificity, and sensitivity are tested in the last stage to gauge the effectiveness of the suggested model.

Vijaya & Sivasankar (2018) provided proof that a business' success depended on its capacity to retain clients. It increases profitability while also maintaining the company's standing in the telecom industry. It is less expensive to keep current customers than to win new ones. The two criteria that determine whether a business succeeds are keeping customers and managing customer associations. In this study, a hybrid model combining supervised and unstructured techniques is used to predict churn and the accuracy of the model was 87.61%.

Höppner et al. (2017) mentioned that several prediction algorithms are used in client retention policies. The most recent development (EMPC), which chooses the most lucrative churn model, is anticipated to maximize the profit. In this study, a novel classification technique is presented that incorporates the (EMPC) matrix directly into the churn model. ProfTree is the name of this method. This model's key benefit is that a telecom firm can make the most money possible. In comparison to other models, the suggested model performs and is more accurate, the accuracy of this model was 89% which a bit higher that the previous model. In the future, this model might be incorporated into various algorithms to improve forecast accuracy even more.

Ali et al. (2018) employed a variety of mining techniques and algorithms to forecast churners. Different classifiers are applied using WEKA software. This model starts with data preprocessing, where missing values are eliminated. Following preprocessing, feature reduction is carried out using the FSS (Feature Subset Selection) stages. The cost of data security was also decreased. The target dataset is then ranked using information gain ratio. The benefit of this research is the discovery of intriguing patterns for churner behavior prediction. The drawback of this research is that it will become slow and take more time to predict if the dataset is extended.

Bharat (2019) developed a model based on client behavior patterns. By calculating the typical idleness duration and frequency, it is possible to measure the customer's activities specifically. Churn prediction can be done in additional areas using the suggested method.

In Gajowniczek, Orłowski & Ząbkowski (2019) For the consumer prediction model, an artificial neural network with entropy cost functions was used. The simplest way to apply the new q-error functions to resolve the issue is through the use of numerical methods like classification trees or SVM, which offer improved classification accuracy.

Customer churn is beneficial for telecom businesses to maintain significant users, according to Zhang et al. (2018). For decisions about client retention, a CCP(Customer Churn Prediction) model with more accuracy is crucial. SVM approach is also employed in this paper because it has a much higher level of precision. It resolves problems with samples in a low-dimensional space that is linearly inspirable in two dimensions. A drawback of the proposed strategy is that it is very challenging to quantify lost clients. Therefore, a further in-depth research is necessary. In order to construct hybrid ensembles models for prediction, Ahmed et al. (2020) proposed a model based on the integration of various classifiers. Learners are suggested in this stack of paper bags. Two datasets pertaining to telecom businesses are used for experimentation. High precision is attained. This model has the advantage of not functioning with generalized data sets.

Calzada-Infante, Skarsdóttir, and Baesens (2020) compared two methods. To assess the prediction effectiveness of the similarity forest classifier with each centrality metric, Time-Order-Graph and Aggregated-Static-Graph with forest classifier use three threshold measures. Focal Loss and Weighted Loss are two cost-sensitive learning algorithms that Nguyen and Duong (2021) compared against two prediction systems, SMOTE (Synthetic Minority Oversampling TEchnique) and Deep Belief Network (DBN). Results indicate that Focal Loss and Weighted Loss perform more effectively than SMOTE and DBN.

Vural, Okay & Yildiz (2020) proposed a new method based on ANN for churn prediction. In this method two layers of ANN are used to predict churn and the model was able to achieve the accuracy of 89%. Jain, Khunteta & Srivastava (2020) discussed overview of different classification algorithms. These algorithms are Multi-Layer-Perception, KNN measure, Fuzzy Cluster and Deep Learning CNN. After comparison of results Deep Learning CNN shows better results as compared to other classification algorithms.

III. METHODOLOGY

3.1 Dataset Source and Description

In building the churn prediction model, the first and critical stage is data acquisition, this is so because the privacy of the customer information must be respected and must not be misused. Hence Telecom operators don't release the database of their customer to the public. The dataset used in this research was acquired from public repository telecommunications business, which included demographic data including information on the services that customers got. The dataset was organized in relational form using Microsoft excel sheet with CSV (comma-separated values) file format and 20 Attributes as shown in figure 1.0.

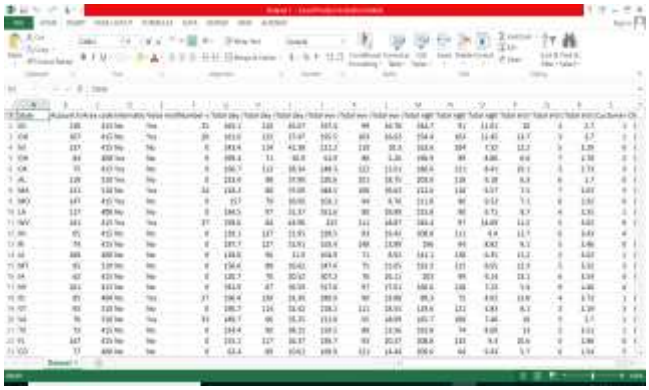

Figure 1.0: Model Dataset

Because the telecom dataset is quite large and active, it is usually updated on a regular basis, it is accessible at https://www.kaggle.com. The description of the dataset is shown in figure 2.0.

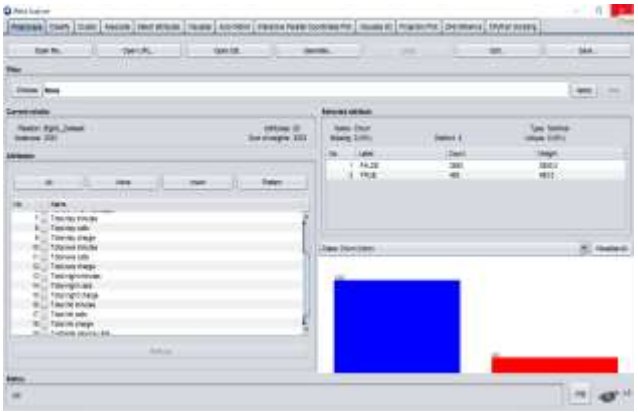

Figure 2.0: Description of the dataset

Weka 3.8.5 data mining was used to perform the experiment. Weka is a graphical user interface that comprises a set of data analysis and predictive modeling visualization tools and algorithms, it is a collection of machine learning techniques for data mining. Data preparation, categorization, regression, clustering, association rule mining, and visualization tools are all included. It is observed from figure 1.0 that the original dataset is highly imbalanced and that class distribution of churners to non churners is 483 to 2850 this is graphically represented in figure 2.0. The class distribution of the dataset represent the quantity of data in each category (Churn & Non-Churn). The red bar in the figure represent churner(Yes) while blue bar represent non-churner(No) and the churn percentage is 14.9%.

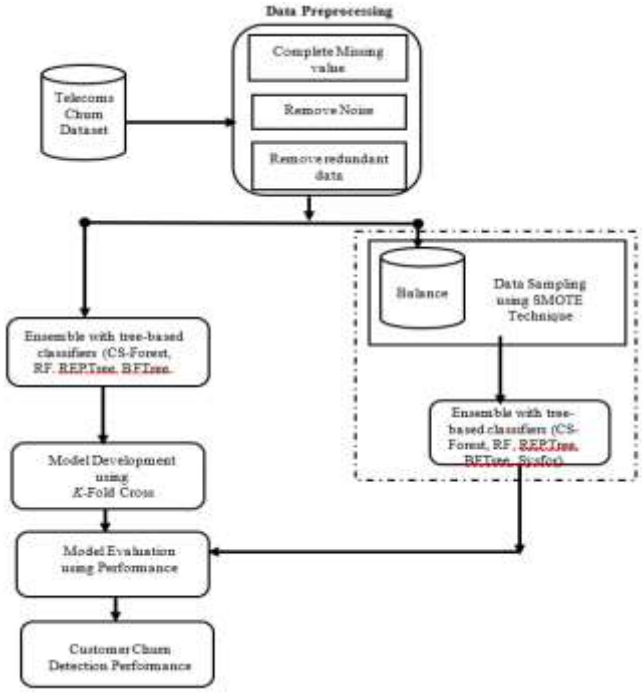

Figure 3.0: Research Framework

3.2 Research Design

This section explained the practical steps and framework used to build the model. Five homogenous ensembles Bagging, Boosting, Cascade, Rotation Forest and Dagging were applied to the dataset individually with five tree-based (CS-forest, RF, REPTree, BFTree and Sysfor) were combined with the ensemble. The result of the model such as accuracy, AUC, F-Measure and MCC were recorded and evaluated. After the acquisition of the dataset, it was discovered that the dataset is highly imbalanced as shown in figure 2.0. Then the dataset was balanced using SMOTE (Synthetic Minority Ovesampling Technique) so as to improve the efficiency and accuracy of the model in order to ascertain the submission of some of the past researcher that class imbalance high affect the performance and accuracy of the prediction model (Ullah et al., 2019; Vural et al., 2020; Zhang et al., 2018; Almuqren et al., 2021; Bharat, 2019).

3.3 Data Preparation

In order to increase the model's accuracy and effectiveness in predicting customer attrition, data preparation is vital, necessary, and time consuming. This entails deleting all ambiguity, mistakes, and irrelevant data from the dataset. Customer churn data preparation may involve different operations like removing attributes that contains text data such as state, Date of birth etc., removing attributes with null values, removing attributes that have same values and convert the data to usable format e.g binary format so the dataset was properly prepared as appropriate data preparation reduce the data size to a reasonable format as larger data size leads to high computational time.

3.4 Balancing the Dataset

Prior to building a prediction model, it is important to make a choice on the problem of class imbalance, however, it is important as well to state that minority class is of interest in this study. As shown in table 1.0, the churn percentage of dataset1 is 14.49% which shows high level of imbalance. In order to eliminate data imbalance at data pre-processing stage, different methods to data imbalance are available but in this research we have adopted Random Oversampling (ROS) precisely SMOTE, SMOTE creates synthetic minority class samples. It was used to synthetically balance the training dataset before training the classifier. Instead of making duplicates of data from the minority class, it developed synthetic samples from the minor class. In order to balance the data, it chose similar records from the minority class and changed them one column after another by a random cost. Records were simply added to the minority class records.

Figure 4.0: Balanced Dataset

IV. RESULTS & DISCUSSION

As stated in section 3. Weka 3.8.5 is the data mining tool used for the building of the models. The first process is to load the dataset into the data mining application as shown in figure 1. The application read all the attributes of the dataset, therefore the classification can ten start by clicking on the classification tab. the ensemble and classification algorithm to be used is selected. Figure 5 shows the classification model result for bagging ensemble using CS-Forest algorithm for raw/imbalanced dataset. The Accuracy, AUC/AOC, F-Measure and MCC of the result is then extracted from the output and the result is then insert into table 1.0.

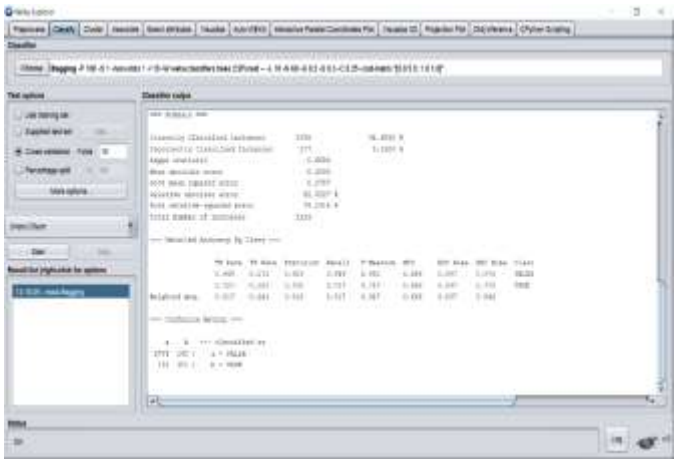

Figure 5.0: Bagging Ensemble with CS-Forest Classifier for imbalanced dataset

After the extraction of the result into table 1.0, the balanced dataset is then loaded into the Weka and the same Bagging ensemble with CS-Forest classifier is also used and the result is shown in figure 6.0. The two datasets (Balanced & imbalanced) were used to build the model so as to compare the result of the two models in order to investigate the effects of imbalanced dataset on customer churn classification models.

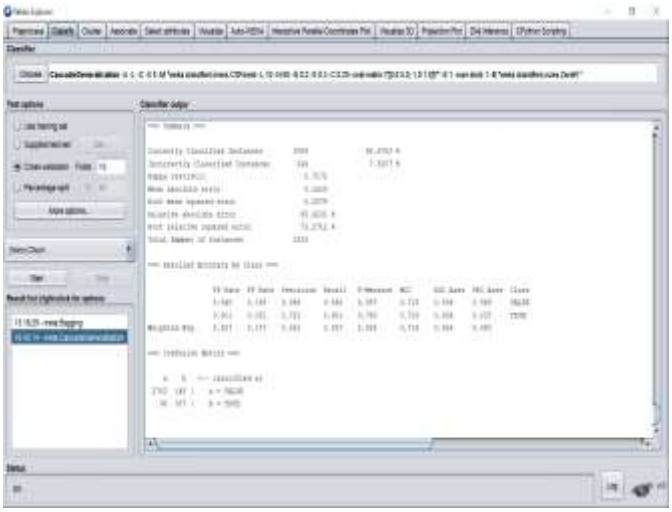

Figure 6.0: Bagging Ensemble with CS-Forest Classifier for balanced dataset

This process is repeated for other ensembles (Boosting, Cascade, R-Forest & Dagging) with the classification algorithms listed in section 3.2. the results are extracted and inserted into table 1.0 accordingly.

When building a predictive model, it is important to select a metric that is appropriate for the classification problem to evaluate how well our models are performing. Commonly, confusion matrix is used to evaluate the performance of a model. The accuracy and the AUC-ROC are the parameter used to measure the performance of our models.

$$\text{Accuracy} \quad \frac{TP + TN}{TP + TN + FP + FN}$$

Note: TP represents true positives, TN is true negatives, FP shows false positives and FN shows false negatives

The result in table 1.0 shows that boosting outperform other ensemble having the accuracy of 98.06% followed by rotation forest with the accuracy of 95.36% for balanced dataset. However for the raw dataset the highest accuracy is 95.14% produced by bagging ensemble. This also affirm the submission of some past researcher that data imbalance highly affects the accuracy and performance of the prediction model.

Table 1.0: Model Results

		Original Dataset					Balanced Dataset				
Ensemble	**Metric**	**CS-Forest**	**RF**	**RepTree**	**BFTree**	**SysFor**	**CS-Forest**	**RF**	**RepTree**	**BFTree**	**SysFor**
Bagging	**Accuracy**	91.69	90.16	88.3	93.85	95.14	80.14	92.30	86.42	86.42	94.88
Boosting		86.62	91.36	87.82	92.17	94.03	88.08	92.30	87.90	93.3	98.06
Cascade		92.68	91.09	89.74	92.11	94.66	77.38	92.70	83.70	90.93	93.80
R-Forest		93.58	88.90	92.65	92.68	94.09	91.68	92.70	94.67	94.79	95.36
Dagging		88.39	85.90	85.51	86.47	86.95	67.38	87.9	75.24	85.53	80.64
Bagging	**AUC**	0.897	0.898	0.851	0.900	0.916	0.926	0.971	0.924	0.924	0.982
Boosting		0.788	0.899	0.789	0.893	0.902	0.948	0.973	0.941	0.977	0.985
Cascade		0.906	0.904	0.824	0.794	0.914	0.924	0.974	0.870	0.910	0.979
R-Forest		0.910	0.898	0.908	0.917	0.917	0.98	0.974	0.982	0.982	0.985
Dagging		0.877	0.880	0.558	0.826	0.892	0.858	0.949	0.817	0.928	0.932
		Original Dataset					**Balanced Dataset**				
Bagging	**F_Measure**	0.917	0.881	0.874	0.936	0.949	0.798	0.923	0.864	0.864	0.949
Boosting		0.861	0.900	0.865	0.918	0.937	0.881	0.923	0.879	0.933	0.961
Cascade		0.928	0.897	0.889	0.917	0.944	0.768	0.927	0.837	0.909	0.938
R-Forest		0.935	0.86	0.918	0.918	0.936	0.917	0.927	0.947	0.948	0.954
Dagging		0.884	0.798	0.922	0.822	0.826	0.643	0.878	0.752	0.855	0.804
Bagging	**MCC**	0.668	0.533	0.470	0.736	0.793	0.625	0.846	0.729	0.729	0.898
Boosting		0.420	0.602	0.426	0.659	0.743	0.763	0.847	0.758	0.866	0.921
Cascade		0.719	0.587	0.534	0.657	0.772	0.581	0.854	0.675	0.819	0.876
R-Forest		0.735	0.453	0.671	0.673	0.743	0.837	0.855	0.894	0.896	0.908
Dagging		0.533	0.152	?	0.251	0.291	0.431	0.757	0.805	0.711	0.625

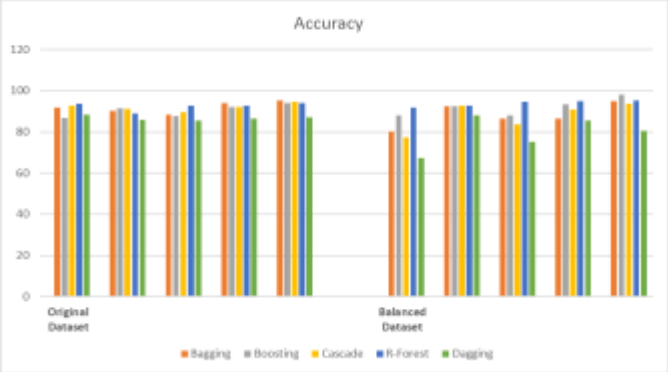

Figure 7.0: Graphical Representation of the accuracy of the models

Figure 7.0 shows that bagging outperform other ensemble in building model for original/raw dataset with the accuracy of 95.14% followed by cascade ensemble.

Accuracy of the model cannot only be used to measure performance and reliability of a prediction model due to its inefficiency so AUC is also used here as a major parameter/metric to measure the performance of the model. AUC is the likelihood that a classifier would score a randomly selected positive instance higher than a randomly selected negative instance. AUC-ROC curve is a visual depiction of our machine learning classifier's performance. This evaluation technique is employed for this research due to its effectiveness and reliability for binary classification, In addition, a successful classifier should have an AUC score of closer to 1. Table 1.0 show that cascade outperform other ensemble with the AUC of 0.979 while boosting and rotation forest have the same AUC of 0.9895 for balanced dataset. However, for original/raw dataset rotation forest outperform other ensembles having the AUC of 0.917.

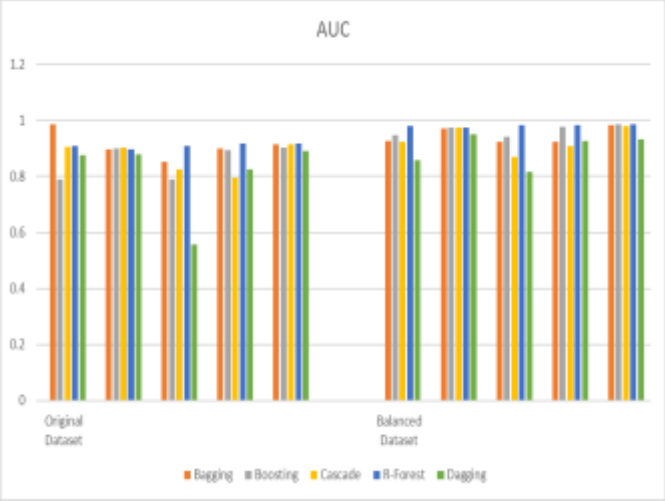

Figure 8.0: Graphical Representation of the AUC of the models

As stated earlier bagging ensemble outperform other with raw/original dataset, the AUC of bagging ensemble is 0.789 indicating that the imbalance dataset has little or no effects on the performance of this ensemble. However boosting and rotation forest has the same AUC of 0.985 for balanced dataset but this two ensemble has lower performance on original dataset/raw dataset. Cascade also perform excellently on balanced dataset with the AUC of 0.979.

V. CONCLUSION AND RECOMMENDATION

Churn prediction is a challenging and critical problem need to be addressed for valuable customers retention in telecommunication sector, this is as a result of competition and daily entry of new players into the sector. The researchers have critically and logically analysed the accuracy and performance of five different ensemble (Bagging, Boosting, Cascade, Rotation Forest and Dagging) with tree-based classifiers for in predicting customer churn. The major metrics used in this research is the accuracy and AUC for both original and balanced dataset. It is concluded that bagging perform excellently on both raw and balanced dataset especially when it is applied on CS-forest and Sysfor classifiers with the AUC of (0.987 & 0.982) and this indicate that class imbalance has very little effects on Bagging ensemble. However, other ensembles other ensemble perform excellently with good accuracy and AUC when applied on balanced dataset but are poorly performed on raw dataset. This outcome affirm the finding of previous researchers that imbalance dataset is an impediment that should be eliminated so as to improve the performance of prediction model. Finally it is observed that this ensemble give better accuracy and AUC when combine with the Sysfor classifier indicating that any of these ensemble with sysfor classifier will greatly help us in predicting prospective churners in telecommunication sector.

In future, we will implement these ensemble with other categories of classifier such as lazy, function, rules classifier to find out the classifiers that will give us better performance. Also, further research may involve using larger dataset so as to establish whether the size of the dataset will have effect on the performance of prediction model.

REFERENCES

[1] Adnan A., Feras A., Babar S., May A., Changez K., Hamood Ur Rehman D., Sajid A.(2021). Just-in-time customer churn prediction in the telecommunication sector

[2] Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data, 6(1), 1-24.

[3] Ahmed, A. A., & Maheswari, D. (2017). Churn prediction on huge telecom data using hybrid firefly based classification. Egyptian Informatics Journal, 18(3), 215-220.

[4] Ahmed, M., Afzal, H., Siddiqi, I., Amjad, M. F., & Khurshid, K. (2020). Exploring nested ensemble learners using overproduction and choose approach for churn prediction in telecom industry. Neural Computing and Applications, 32(8), 3237-3251.

[5] Bharat, A. (2019). Consumer engagement pattern analysis leading to improved churn analytics: an approach for telecom industry. In Data Management, Analytics and Innovation (pp. 203-211). Springer, Singapore.

[6] Ali, M., Rehman, A. U., Hafeez, S., & Ashraf, M. U. (2018, August). Prediction of churning behavior of customers in telecom sector using supervised learning techniques. In 2018 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE) (pp. 1-6). IEEE.

[7] Almuqren, L., Alrayes, F. S., & Cristea, A. I. (2021). An Empirical Study on Customer Churn Behaviours Prediction Using Arabic Twitter Mining Approach. Future Internet, 13(7), 175.

[8] Al-Rifaie, M. M., & Alhakbani, H. A. (2016, July). Handling class imbalance in direct marketing dataset using a hybrid data and algorithmic level solutions. In 2016 SAI Computing Conference (SAI) (pp. 446-451). IEEE.

[9] Amin, A., Anwar, S., Adnan, A., Nawaz, M., Howard, N., Qadir, J., ... & Hussain, A. (2016). Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study. IEEE Access, 4, 7940-7957.

[10] Bilal, S. F., Almazroi, A. A., Bashir, S., Khan, F. H., & Almazroi, A. A. (2022). An ensemble based approach using a combination of clustering and classification algorithms to enhance customer churn prediction in telecom industry. PeerJ Computer Science, 8, e854.

[11] Calzada-Infante, L., Óskarsdóttir, M., & Baesens, B. (2020). Evaluation of customer behavior with temporal centrality metrics for churn prediction of prepaid contracts. Expert Systems with Applications, 160, 113553.

[12] Diala, P. (2019). Prediction of Customer Churn in Telecommunication using Machine Learning Algorithms (Doctoral dissertation, University of the Witwatersrand, Johannesburg).

[13] Gajowniczek, K., Orłowski, A., & Ząbkowski, T. (2019). Insolvency modeling with generalized entropy cost function in neural networks. Physica A: Statistical Mechanics and Its Applications, 526, 120730.

[14] Höppner, S., Stripling, E., Baesens, B., & Verdonck, T. (2017). Profit driven decision trees for churn prediction. arXiv preprint arXiv:1712.08101.

[15] Jain, H., Khunteta, A., & Srivastava, S. (2020). Churn prediction in telecommunication using logistic regression and logit boost. Procedia Computer Science, 167, 101-112.

[16] Nguyen, N. N., & Duong, A. T. (2021). Comparison of Two Main Approaches for Handling Imbalanced Data in Churn Prediction Problem [J]. Journal of advances in information technology, 12(1).

[17] Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., & Kim, S. W. (2019). A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector. IEEE Access, 7, 60134-60149.

[18] Vijaya, J., & Sivasankar, E. (2018). Improved churn prediction based on supervised and unsupervised hybrid data mining system.

In Information and Communication Technology for Sustainable Development (pp. 485-499). Springer, Singapore.

[19] Vural, U., Okay, M. E., & Yildiz, E. M. (2020). Churn prediction for telecommunication industry using artificial neural networks. International Journal of Computer and Information Engineering, 14(11), 396-399.

[20] Zhang, X., Zhang, Z., Liang, D., & Jin, H. (2018, August). A novel decision tree based on profit variance maximization criterion for customer churn problem. In 2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC) (Vol. 1, pp. 20-23). IEEE.