Volume: 10 Issue: XI Month of publication: November 2022

DOI: https://doi.org/10.22214/ijraset.2022.47541

Customer Churn Prediction in Banking Sector - A Hybrid Approach

Sreenitya Mandava[1], Anvita Gupta[2], Komanduri Srikar[3]
[1, 2, 3]School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India

Abstract: No institution wants to lose customers and would do anything to prevent it. But banks can't do anything about customer churn unless they predict it first. Predicting customer churn will help them to reach out to such customers and offer the assistance they need. With the support of precise and early churn prediction, customer resource management, and customer experience teams can be more assertive and customer-focused. It has been reported that simply reaching out to customers early enough can prevent 11 percent of attrition. The problem remains about how to predict such behaviour. Surveys are difficult to undertake and not many are interested in answering them. Past data holds the required solution. We can predict the behaviour of present and future customers by using machine learning and data science techniques to learn from this past customer data. In this paper, we aim to conduct a comparative analysis of different churn prediction models for banking institutions. Then we would implement a hybrid approach with the considered advantages from each of the best performing models for a given cluster of the dataset, to get better results.

Keywords: Churn Prediction; Voting Classifier; Random Forest; Logistic Regression; Decision Tree; Support Vector Machine

I. INTRODUCTION

Comparison between all the models provides a clear idea about which model is best suited for the given application, hence helping the banking industry utilize the same when they have to find out their churn prediction. Thus, it will prevent wastage of resources and help them choose the best model for their industry. We have observed through our research and literature review that there are many people in the banking sector who face problems predicting the churn or the model they use does not provide them with the accurate results. Taking into consideration this issue, we decided to compare different classification models.

After comparing the models and studying various papers on similar topics for inspiration, we selected a base paper and implemented a hybrid model with considering advantages from the most impactful models to get better results. The results obtained from this optimized algorithm can thus help the individuals choose the option best suited according to their requirements. This will provide them with accurate results as well as help them reduce the resource wastage in terms of funds, manpower and time.

The models we are planning to compare are: Logistic Regression, K-Nearest Neighbours, Decision Tree Classifier, Naive Bayes, Support Vector Machine, Stochastic Gradient Descent, Random Forest and finally Voting Classifier. These models are built using all the features and then compared. Thus, the comparison provides a bird's eye view to which of the models is the best suitable for predicting the churn in the banking sector.

Comparison on such a large scale using models is the innovation which would help millions of those involved in the banking sector. By learning best practices from each algorithm, we would then implement the hybrid algorithm which leverages the advantages of all models and give better results.

II. LITERATURE REVIEW

The study in paper [1] provides a new model for predicting customer churning in the telecoms industry that is based on a worldwide hybridization of the most prominent machine learning approaches. Each of the models are applied and assessed using cross-validation on a prominent, public domain dataset in the early phase of the studies. The proposed model is described in the second phase to demonstrate the performance increase. Various simulations were run for each method and a wide variety of parameters in order to discover the most effective parameter combinations. The results show that the proposed model is far better than the popular existing ML models. We have considered this paper as our base paper.

Paper [2] proposes novel descriptions and classifications for customer churn identification, as well as challenges in customer churn prediction. Customer churn prediction, customer churn understanding, and customer churn response are the steps used by the authors. The framework also discusses the features and challenges of various customer turnover stages. The results pitch for customers with high churn risk to improve customer service efficiency.

The goal of this research in [3] is to compare the performance of the top two sophisticated machine learning algorithms, the K-Nearest Neighbour and the Decision Tree algorithms, in terms of churn prediction. The results were pretty interesting, revealing a significant difference between the two algorithms in several areas. The decision tree method outperformed the K-NN algorithm in terms of accuracy, precision, recall, F-measure, and the Lift measure, according to statistical results. The AUC for both methods was found to be almost identical, which is due to the low specificity for DT, which causes the AUC values for both algorithms to be virtually identical. However, results shown in the confusion matrix indicate that K-NN had more true positive rates than the DT algorithm, telling us that the K-NN predicted better than the DT regarding real churning customers.

In this paper [4], the Python platform is utilized to develop two supervised machine learning classification models. This comes under the supervised domain of machine learning as the data used here is labelled. The conclusion is that KNN is a better way to forecast customer churn than Logistic Regression, as evidenced by the confusion matrix. When it comes to predicting customer attrition, KNN is 2.0% more accurate than Logistic Regression. The authors conclude that the K-Nearest Neighbors is better to predict the customer churn as compared to the Logistic Regression algorithm.

The goal of authors in paper [5] is to create a model that can forecast churn with more accuracy. In order to predict, existing SVM and kernel function models were employed in conjunction with artificial neural networks. When both models were combined, the results were spectacular.

In paper [6], four machine learning models are trained which is Logistic Regression, SVM, Random Forest and Gradient boosted tree, and it can be inferred that Gradient boosting is the best among four models and the Logistic regression and Random forest are average and SVM is underperforming between these models.

A novel complicated user model focusing on user churn intent prediction is proposed in paper [7]. The proposed model's concept is to combine multiple sets of attributes that reflect a user's interaction with a web application. The model's performance is assessed in an indirect way by predicting churn using actual data from online shops. The outcomes demonstrate that predicting churn using the model proposed outperforms predicting churn using baseline models in two domains.

III.METHODOLOGY

A. Design Diagrams

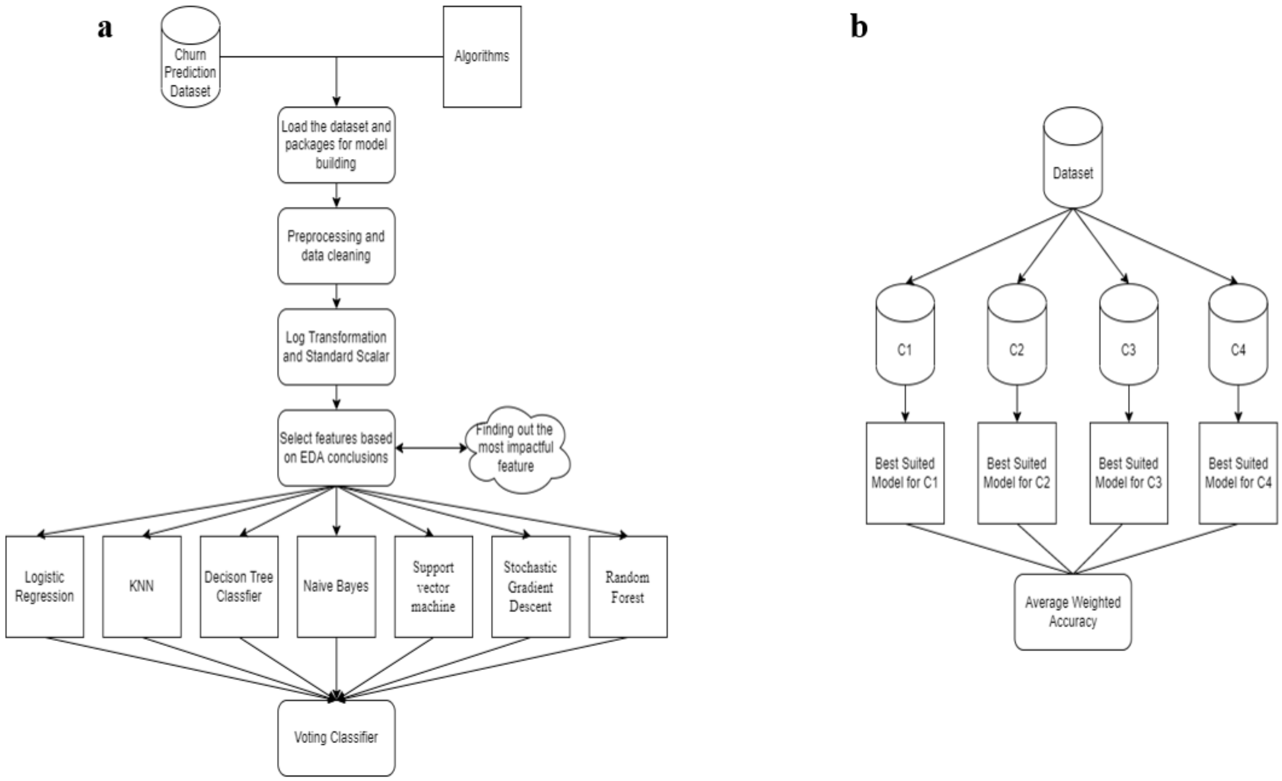


Fig. 1 (a) Initial pre-processing and comparison; (b) The hybrid model approach

B. Model Description

1) *Pre-processing and Algorithms Comparison:* First, we cleaned the data and visualized it to check for relationships between variables. Then we changed all columns to numeric type using dummies function in Python. Further we scaled the data using both log transformation and standard scaler. Next, we would use the popular existing classification models to solve the problem and find out shortcomings in each approach and try to overcome it using a different approach. The final hybrid model, in the next section, will consist of cumulative advantages of all these tested algorithms. The models used are: Logistic Regression, KNN, Decision Tree Classifier, Naive Bayes, Support Vector Machine, Stochastic Gradient Descent and Random Forest. We also constructed a Voting Classifier, which used the numerous models implemented in the previous step, to make better comparisons with the hybrid model. Here, we used all the classifiers and trained them through a voting classifier for the best prediction.
2) *Hybrid Model Implementation:* We implemented the hybrid model in the following steps:

a) We found out the most important feature which impacts the result of prediction based on EDA conclusions, we found it was current_month_debit.
b) Using Qcut, we split the dataset into 4 equal halves and consider them as clusters namely C1, C2, C3, C4.
c) For each cluster, we run all the models and identify the best suited model for the respective cluster.
d) Finally, we implemented the respective models on that particular dataset and calculated the weighted average accuracy.

IV. RESULTS

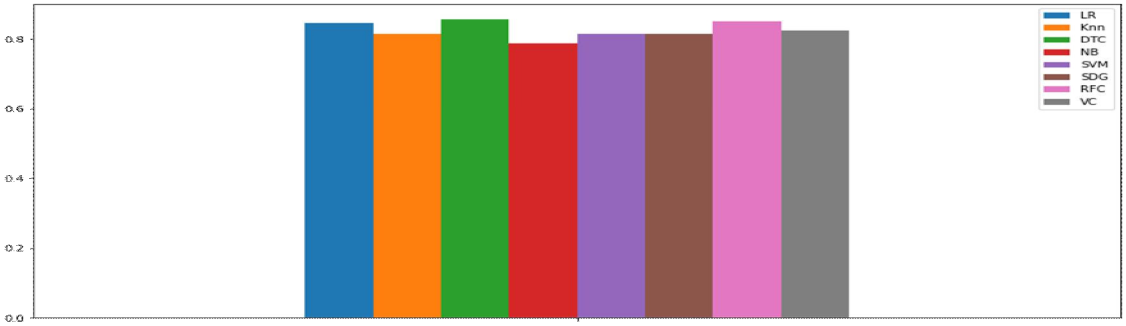


Fig. 2 Comparison of the existing ML models and Voting Classifier

Random Forest Accuracy: **85%**
Voting Classifier Accuracy: **84%**
We can see that Random Forest Classifier still had the leading accuracy over Voting Classifier. So, we then implemented the hybrid model after dividing the dataset into clusters based on the current_month_debit. The best suited models for each cluster were selected on the basis of their testing accuracies as shown below.

TABLE I
ACCURACIES OF CHOSEN MODELS ON EACH CLUSTER

Clusters	Classifiers			
	DTC	SVM	RFC	LR
C1	91%	88%	90%	88%
C2	88%	90%	89%	87%
C3	85%	83%	86%	83%
C4	80%	78%	79%	82%

Overall Accuracy: **87%**

As we can see above, Decision Tree Classifier performed best for C1, Support Vector Machine for C2, Random Forest Classifier for C3 and Logistic Regression for C4. So finally, we first divided the dataset into clusters and then implemented their respective best performing models on each cluster to hybridize the model and make the predictions more accurate. Another observation that we can see is that the cluster C1 has the highest accuracies decreasing to the least accuracies in C4.

V. CONCLUSIONS

After analysing the results, we can conclude that the hybrid approach has the leading accuracy. It performed better than the existing models and voting classifier. Therefore, we can conclude that identifying the most impactful feature and clustering dataset into different classes to apply different models on each cluster provides better prediction compared to traditional classification algorithms. We can extend this idea to form even more clusters and use even more classification algorithms, and try to make each prediction even more accurate. Future work can include performing multiple sub clusters to use different models in the same cluster. This way we can get rid of the outliers' problem.

REFERENCES

[1] Base Paper: F. I. Khamlichi, D. Zaim and K. Khalifa, (2019) “A new model based on global hybridization of machine learning techniques for “customer churn prediction””, 2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS), pp. 1-4.

[2] T. -Y. Tsai, C. -T. Lin and M. Prasad, (2019) “An Intelligent Customer Churn Prediction and Response Framework”, 2019 IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), pp. 928-935.

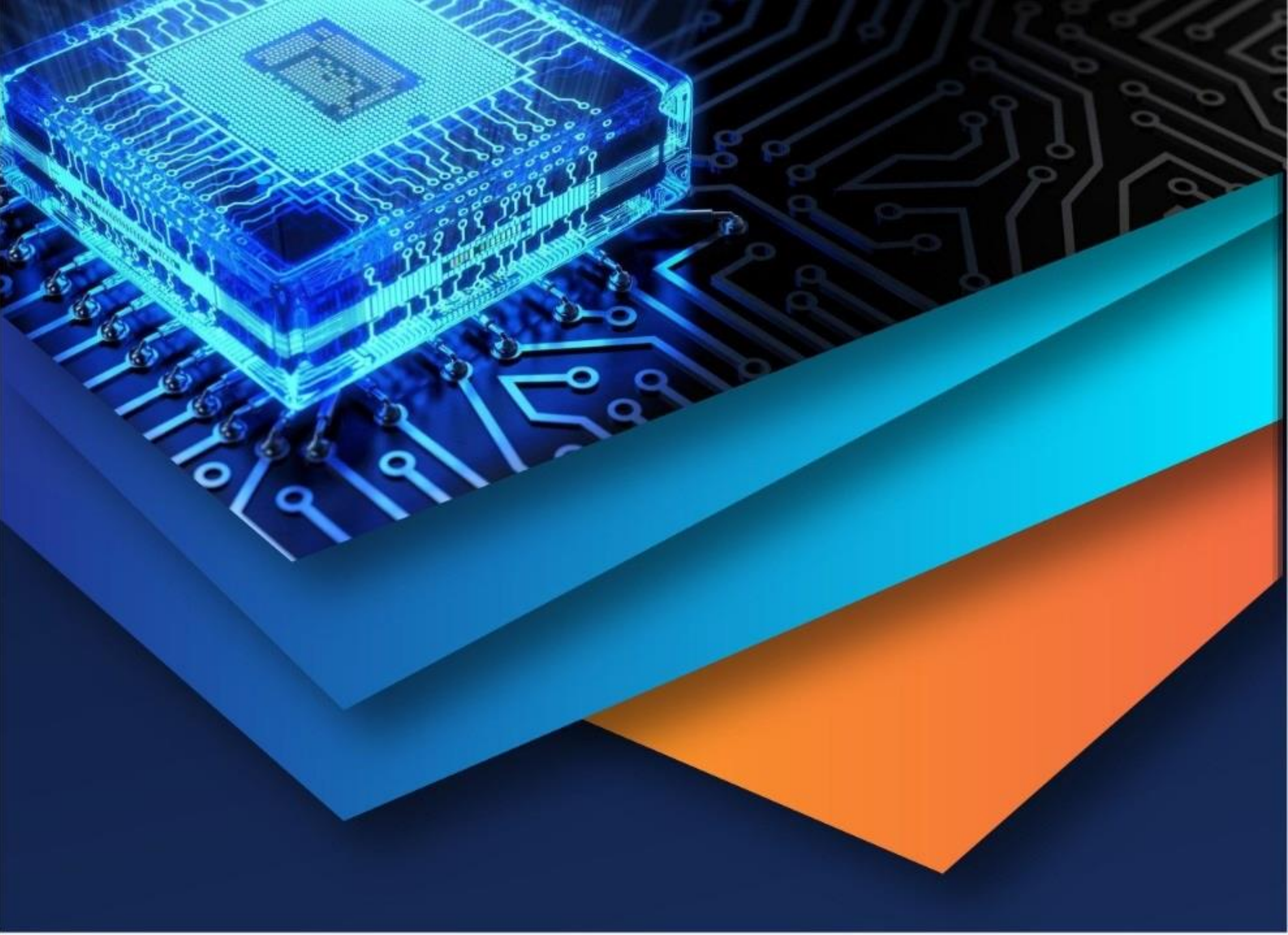
[3] M. A. Hassonah, A. Rodan, A. Al-Tamimi and J. Alsakran, (2019) “Churn Prediction: A Comparative Study Using KNN and Decision Trees”, 2019 Sixth HCT Information Technology Trends (ITT), pp. 182-186.

[4] FA A. Bhatnagar and S. Srivastava, (2019) “A Robust Model for Churn Prediction using Supervised Machine Learning”, 2019 IEEE 9th International Conference on Advanced Computing (IACC), pp. 45-49.

[5] P. Hemalatha and G. M. Amalanathan, (2019) “A Hybrid Classification Approach for Customer Churn Prediction using Supervised Learning Methods: Banking Sector”, 2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN), pp. 1-6.

[6] A. Gaur and R. Dubey, (2018) “Predicting Customer Churn Prediction In Telecom Sector Using Various Machine Learning Techniques”, 2018 International Conference on Advanced Computation and Telecommunication (ICACAT), pp. 1-5.

[7] P. Berger and M. Kompan, (2019) “User Modeling for Churn Prediction in E-Commerce”, in IEEE Intelligent Systems, **34** (**2**): 44-52.



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY