iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.52101

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Churn Prediction using Machine Learning Models

Sayee N. Bhoite[1], Vaishnavi D. Gadekar[2], Shashank V. Kapadnis[3], Priynaka R. Ghughe[4]
[1, 2, 3, 4]*BE Information Technology Pune Institute of Computer Technology, Pune*

Abstract: ***The market is expanding quickly across all sectors, giving service providers access to a larger user base. Better offers have led to increased competition, creative new business ideas, and rising costs for acquiring new customers. Service providers understand how crucial it is to keep clients on-site in such a brief setup. Service providers must therefore prevent churn, a condition that occurs when a customer decides not to use a company's services any longer. This study examines the most widely used machine learning algorithms for churn prediction, not just in the banking industry but also in other businesses that place a high value on customer engagement.***
Keywords: Business, churn prediction, service providers, machine learning algorithms.

I. INTRODUCTION

A. *Introduction*

The volume of statistics has been increasing quickly over the past few years due to technological improvements. To handle statistics and uncover valuable information that is hidden in the raw data, numerous cutting-edge techniques and methodologies have been developed. The practice of extracting significant information from data is known as "data mining." In a variety of sectors, numerous data mining techniques had been employed with success.

Because they are viewed as the primary source of income, customers are the most important asset in any business.

These days, businesses have realised that they need to put in a lot of effort not only to win over new customers, but also to keep their existing ones. Churners are people who switch to other businesses for a variety of reasons. The firm should be able to predict customer behaviour effectively, establish links between client attrition, and maintain variables within their control in order to reduce customer turnover. Churn prediction separates churners from non-churners using a binary category venture.

B. *Objectives*

1) To study different types of Machine Learning models.
2) To select best performing models.
3) To build a user-friendly frontend platform.
4) To deploy the application.

II. LITERATURE SURVEY

To address the drawbacks of the general SVM model that creates a black box model, MAH Farquad [3] proposed a hybrid solution (i.e., it does not reveal the knowledge gained during the training in a form understandable to humans). The hybrid strategy is divided into three stages: The feature set is reduced in the first stage using SVM-RFE (SVM Recursive Feature Elimination). The feature-reduced data set is used to create the SVM model, and support vectors are then extracted. The Naive Bayes Tree is used to generate the rules in the final stage (NBTree, which is a combination of the decision tree with a naive Bayes classifier). With 93.24% loyal clients and 6.76% abandoned customers, the ratio is very lopsided. The results of the experiment demonstrated that the model cannot handle very big data sets.

The results that could be compared to SVM, and logistic regression showed that the results are in the highest degree of accuracy. Although it allows for the inclusion of domain knowledge, Ant-Miner results in less sensitive rule-units and intelligible rule-sets that are significantly smaller than those generated by C4.5. Additionally, RIPPER produces manageable rule sets and produces illogical models that disregard spatial information.

In Ning Lu's [6] proposal, clients are divided into clusters based entirely on the weights supplied by the boosting set of rules, which is decorated with boosting algorithms to create a client churn prediction model. A high-risk clientele category has been identified as a result. A churn prediction model is created for each cluster, and logistic regression is utilised as a foundation learner. The experimental results proved that, in comparison to a single logistic regression model, boosting methods offer a substantial separation of churn information.

Benlan He [7] suggested a customer churn prediction method that was entirely based on an SVM model, and he employed a random sampling approach to improve the SVM model by taking into account the client data sets' imbalance attributes. In a high- or infinite-dimensional space, a support vector machine creates a hyper-plane that can be utilised for classification. You can extrade the distribution of data to minimise the dataset's imbalance by using a random sampling strategy.

III. PROPOSED METHODOLOGY

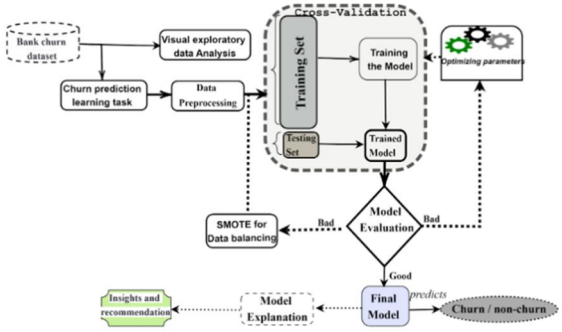
A. Methodology

Customer churn is the term used to describe people switching from one financial organisation to another. Dissatisfaction with the client provider, exorbitant costs, unappealing plans, and poor assistance are the main causes of churn. Since acquiring new customer's costs five to six times more than maintaining existing ones, it is a costly issue in many industries. [1]. A significant additional ability sales source for each organisation is the capacity to predict whether a specific customer is at an elevated risk of leaving. In addition to the immediate loss of sales brought on by a client leaving the business, the original costs of obtaining that client could not have been covered by the client's purchases until this point. Finding consumers with a high propensity to depart a business is the aim of customer churn prediction. In order to keep the current clients, the banking business needs to understand the reasons why consumers depart, which may be seen through the understanding derived from gathered data. Reactive and proactive customer churn management approaches exist, according to Burez and Van den Poel [2]. When an organisation adopts a reactive strategy, it waits until the client asks it to end their provider relationship. In this situation, the company will provide the customer a reason to stay. When a business takes a proactive strategy, it looks for clients who are likely to leave before they actually do. The company then offers extra incentives to keep such customers instead of having them leave.

Machine learning is a data analysis technology that automates the creation of analytical models. Machine learning allows structures to discover hidden patterns without being explicitly told where to look by using algorithms that iteratively evaluate data.

Unsupervised, semi-supervised, and supervised machine learning techniques are the three types available. The goal of supervised learning is to extract hidden patterns from labelled datasets using machine learning. Unsupervised learning is a machine learning endeavor that uses unlabeled statistics to search for hidden patterns. Semi-supervised machine learning is a subset of supervised learning tasks that uses both labelled and unlabeled data for training. A little amount of labelled data is typically mixed with a lot of unlabeled input in semi-supervised machine learning tasks. Semi-supervised learning can be done without supervision.

B. System Architecture



The dataset is been taken from kaggle. Overall proposed system is a Machine Learning model which predicts whether the customer is churned or not from given data. For prediction, data is been preprocessed to remove noisy data (if any), to select the require features. Followed by applying feature importance and feature extraction. Then the splitting of data is performed using k-fold validation. SMOTE library is also used to resample the data. The model is trained and tested over the processed dataset to generate the result. If the desired result is not obtained, then parameter optimization is done and data preprocessing is done again. This is done until desired result is obtained.

The result obtained is displayed through frontend to the user using different visualization techniques. To predict the status of individual customer, a separate form is been created which take all the real-time values.

Popular machine learning algorithm Random Forest is a part of the supervised learning methodology. It can be applied to ML issues involving both classification and regression. It is built on the idea of ensemble learning, which is a method of integrating various classifiers to address difficult issues and enhance model performance. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, it predicts the final output.

As the name implies, ”Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.”

Precision = $\frac{TP}{TP+FP}$ = 85%.

Recall = $\frac{TP}{TP+FN}$ = 90%.

F1 score = $\frac{2*Precision*Recall}{Precision+Recall}$ = 87.4%.

LightGBM is a gradient boosting framework based on decision trees to increases the efficiency of the model and reduces memory usage. Gradient-based One Side Sampling and Exclusive Feature Bundling (EFB), which overcome the restrictions of the histogram-based algorithm that is largely employed in all GBDT (Gradient Boosting Decision Tree) frameworks, are two unique techniques that are used in this framework. The two techniques of GOSS and EFB described
below form the characteristics of LightGBM Algorithm. They comprise together to make the model work efficiently and provide it a cutting edge over other GBDT frameworks.

Precision = $\frac{TP}{TP+FP}$ = 83%.

Recall = $\frac{TP}{TP+FN}$ = 87%.

F1 score = $\frac{2*Precision*Recall}{Precision+Recall}$ = 84.9%.

IV. EXPERIMENTAL RESULTS

	Models	Accuracy_train	Accuracy_test	F1 Score
0	Randomforest	86.01	86.3	0.86
1	Light GBM	84.2	83.04	0.84
2	SVM	72.63	71.21	0.72
3	Logistic Regression	68.14	67.91	0.65

Fig.4.1 Model Accuracy

We have analysed f1_score, training accuracy and testing accuracy, and based on that we have finalised 2 models – LightGBM and RandomForest. Both these models have highest accuracy amongst several other selected models. They have an accuracy of 84% and 86% respectively.

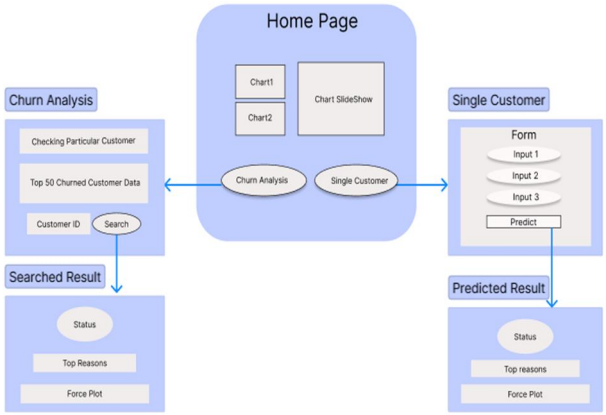


Fig 4.2 Frontend Wireframe

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

To make effective interaction with users, user-friendly frontend is been built, which displays visualized analysis of the data. It also contains two modules, one for searching and predicting the status of already present customer, while the other for predicting the status of new customer that is not present in the dataset.

V. CONCLUSION

In the project, our goal is to foresee which clients will leave. Support Vector Machine and Gradient Boosting techniques are tested in the project, but because the results are not superior to those of Random Forest models, they are not used. The data can be further enriched, and additional variable optimization can be done, with the goal of improving score. Following this, the same models or different methods can be tried again to find more accuracy scores.

We have analysed f1_score, training accuracy and testing accuracy, and based on that we have finalised 2 models – LightGBM and RandomForest. Both these models have highest accuracy amongst several other selected models. They have an accuracy of 84% and 86% respectively.

To make effective interaction with users, user-friendly frontend is been built, which displays visualized analysis of the data. It also contains two modules, one for searching and predicting the status of already present customer, while the other for predicting the status of new customer that is not present in the dataset.

The application is been successfully deployed on onestop.ai, a platform where various applications are been deployed which are open for the organizations to view, choose and make a contract with customization as per requirement.

REFERENCES

[1] Customer churn prediction in telecommunications by T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. Chatzisavvas was published in Simulation Modelling: Practice and Theory 55 (2015) 1-9.

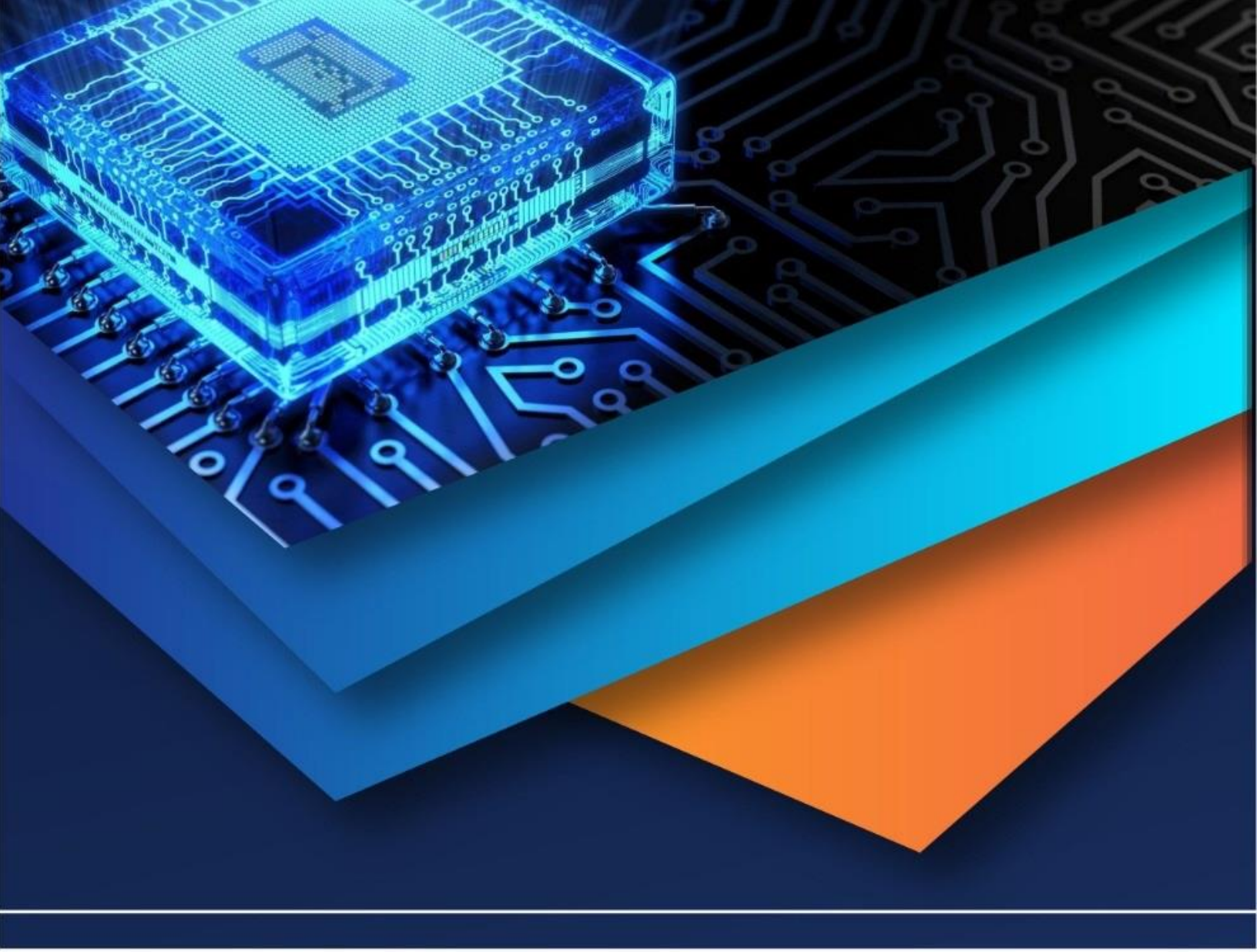
[2] J. Burez, D. Van den Poel, Expert Systems with Applications 32, 277-288, "Crm at a Pay-TV Company: Using Analytical Models to Reduce Customer Attrition by Targeted Marketing for Subscription Services."

[3] Vadlamani Ravi, S. Bapi Raju, and M.A.H. Farquad Applied Soft Computing 19 (2014) 31–40, "Churn prediction with understandable support vector machine: An analytical CRM application."

[4] A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in the Telecom Sector, I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim, IEEE Access, vol. 7, pp. 60134–60149, 2019, doi: 10.1109/ACCESS.2019.2914999

[5] A Customer Churn Prediction Model in the Telecom Industry Using Boosting, IEEE Transactions on Industrial Informatics, vol. 10, no. 2, May 2014. Ning Lu, Hua Lin, Jie Lu, and Guangquan Zhang

[6] "Prediction of customer attrition of commercial banks based on SVM model," by Benlan He, Yong Shi, Qian Wan, and Xi Zhao.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY