iJRASET
International Journal For Research in Applied Science and Engineering Technology



# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 12 **Issue:** VIII **Month of publication:** August 2024

**DOI:** https://doi.org/10.22214/ijraset.2024.64017

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 12 Issue VIII Aug 2024- Available at www.ijraset.com*

# A Novel Methodology for Customer Attrition Prediction

A. Thirunirai Selvi[1], M. Akila[2], I. Muthu Meenatchi[3]
*Assistant Professor, Department of Information Technology, K.L.N College of Engineering, Madurai*

***Abstract:*** ***The project's goal is to use cutting-edge machine learning techniques to create an accurate forecast model for customer attrition. By using important traits and historical data, the Gradient Boosting Classifier algorithm is utilized to predict client attrition. The model detects clients who are at risk of discontinuing by using sequential training combined with ensemble learning. To train and evaluate its prediction skills, the suggested model makes use of past customer data, such as usage trends, engagement metrics, and customer demographics. Model effectiveness is measured by evaluation criteria like accuracy, precision, recall, and F1-score. Forecasts regarding consumer behavior can assist businesses in improving their client relationships and decreasing customer churn, which will ultimately increase the sustainability and profitability of their operations.***

***Keyword – Customer attrition, Gradient Boosting Classifier, Machine learning, Customer retention, Predictive modeling***

## I. INTRODUCTION

Long-term success in today's cutthroat business environment depends on maintaining current customers. Companies frequently struggle with customer churn, which occurs when customers stop doing business with them. Predictive analytics is using sophisticated machine learning techniques more and more to address this. The goal of the research is to employ the Gradient Boosting Classifier algorithm to create an efficient forecasting model for customer attrition. Our goal is to develop a strong model that correctly detects clients who are at risk of leaving by utilizing this technique. The algorithm learns patterns linked to churn behavior by utilizing past customer data, including consumption habits and demographics. It improves its prediction powers through iterative training and validation, assessed by measures like accuracy and precision. These forecasts enable companies to proactively execute retention efforts, which lower

## II. LITERATURE SURVEY

In past years, several studies and projects on Customer Churn Prediction have been presented in the literature covering various aspects of a Client Retention. The insurance sector is a prominent adopter of data forecasting technologies, leveraging vast amounts of data and catering to an ever-expanding customer base. With clients having the ability to switch insurance providers easily, there's a challenge in accurately predicting churn probability. If predictions fall short, it indicates that the insurance company's strategies may not align with the evolving needs of customers.[1] To address the challenges in churn prediction, the discussion revolves around the application of machine learning (ML) and deep learning (DL) methods. These advanced techniques aim to provide more accurate and reliable churn forecasts. The research paper titled "A PCA-AdaBoost model for e-commerce customer churn prediction" authored by Z. Wu, L. Jing, B. Wu, and L. Jin was published in the year 2022.[2] The paper likely introduces a predictive modeling approach for predicting customer churn in the e-commerce domain. The aim is to improve the accuracy of predicting customer churn in e-commerce scenarios. The model mentioned is a combination of Principal Component Analysis (PCA) and AdaBoost, indicating the use of dimensionality reduction and ensemble learning techniques. PCA is often used to reduce the dimensionality of data, while AdaBoost is an ensemble learning method that combines multiple weak learners to create a strong model.The research paper titled "Prediction of employee turnover using random forest classifier with intensive optimized PCA algorithm" authored by A. B. W. Ali and the research paper was published in the year 2021.[3] The authors use a combination of two techniques: a "Random Forest Classifier" and an advanced version of the "Principal Component Analysis (PCA)" algorithm that has been intensively optimized for better performance. The term "intensive optimized" indicates that the researchers have made significant efforts to improve and fine-tune the PCA algorithm for the specific task of predicting employee turnover. In this paper, author Ahmed and Maheswari propose an advanced ensemble classifier for telecom churn prediction, incorporating cost-based uplift modeling. Published in the International Journal of Information Technology, the work addresses telecom industry challenges. The enhanced classifier aims to improve accuracy and efficiency in predicting customer churn.

©IJRASET: All Rights are Reserved 

The study contributes to uplift modeling, emphasizing cost considerations. Published in June 2019, the research showcases a novel approach, providing valuable insights for telecom operators seeking effective strategies for customer retention[4]. These studies showcase the versatility of ML in churn prediction across various domains.

## III. PROPOSED SYSTEM

Customer attrition prediction has become a crucial aspect of customer relationship management in today's business landscape. As companies strive to retain their customer base and maximize profitability, the utilization of advanced machine learning algorithms such as Gradient Boosting Classifier has emerged as a powerful tool for enhancing accuracy in forecasting customer attrition. This essay delves into the application of Gradient Boosting Classifier in customer attrition prediction systems, highlighting its iterative optimization process, robust predictive capabilities, and ability to handle complex data relationships.

1) The customer attrition prediction system utilizes the powerful Gradient Boosting Classifier algorithm to enhance accuracy in forecasting customer attrition.
2) This machine learning approach employs boosting and sequential integration to create a robust predictive model. The standard steps include handling missing values, text-to-number conversion, and normalization.
3) Key client features, including demographics, usage patterns, and transaction history, undergo comprehensive analysis for accurate predictions.
4) Through iterative optimization, the Gradient Boosting Classifier effectively identifies potential churners, enabling proactive customer retention strategies.
5) The success of Gradient Boosting lies in its ability to handle complex relationships within the data, making it a popular choice for accurate and efficient churn prediction in various industries.

In conclusion, the application of Gradient Boosting Classifier in customer attrition prediction systems represents a significant advancement in customer relationship management. By leveraging boosting and sequential integration, organizations can develop highly accurate predictive models that identify potential churners and inform proactive retention strategies. With its ability to handle complex data relationships and interpret results transparently, Gradient Boosting offers a powerful solution for enhancing accuracy and efficiency in churn prediction across various industries. As businesses continue to prioritize customer retention and loyalty, the adoption of Gradient Boosting Classifier stands poised to revolutionize the way organizations approach customer attrition prediction and customer relationship management.

## IV. WORKING METHODOLOGY

Customer attrition prediction, a critical aspect of customer relationship management, employs advanced machine learning algorithms like Gradient Boosting Classifier (GBC) and LightGBM to enhance accuracy and efficiency in forecasting customer attrition. The methodology involves several key steps aimed at data preprocessing, model training, evaluation, and optimization.

1) The process begins with data preprocessing, where the dataset is cleaned and prepared for analysis. This involves handling missing values, converting categorical variables into numerical representations, and scaling features to ensure uniformity across the dataset. Once the data is preprocessed, it is split into training and testing sets to facilitate model training and evaluation.
2) For both Gradient Boosting Classifier and LightGBM algorithms, the training phase involves sequential learning to build an ensemble of weak learners. In GBC, each weak learner corrects the errors of its predecessors by optimizing a predefined loss function. LightGBM, on the other hand, utilizes a gradient-based approach with a novel tree-boosting algorithm that focuses on leaf-wise growth, enabling faster training and improved accuracy.
3) After training the models on the training data, their performance is evaluated using various metrics such as accuracy, precision, recall, and F1-score. Additionally, a confusion matrix is generated to visualize the model's predictions and assess its ability to classify instances accurately.
4) To optimize the performance of the models further, hyperparameters are tuned using techniques like grid search cross-validation. This involves systematically exploring different combinations of hyperparameters to identify the configuration that maximizes model performance. By fine-tuning parameters such as learning rate, maximum tree depth, and regularization strength, the models can achieve higher accuracy and better generalization to unseen data.

Once the models are trained, evaluated, and optimized, they can be deployed in a production environment to predict customer attrition. Real-time data can be fed into the models, allowing businesses to identify potential churners and implement proactive retention strategies effectively.

## V. MODULES

### A. *Project Planning And Deployment*

Select Python as the programming language for developing the customer attrition predictive model. Analyze and choose appropriate machine learning models for churn prediction. Common choices include logistic regression, decision trees, random forests, gradient boosting machines, and neural networks. Use 'pip' to install libraries and frameworks. For example, to install popular libraries like NumPy, Pandas, Scikit-learn and Matplotlib: "pip install numpy pandas". To install and use an integrated development environment (IDE) for a more convenient development experience. Popular Python IDEs include PyCharm, Visual Studio Code, Sublime Text and Jupyter Notebook.

### B. *Data Collection And Preprocessing*

Data collection for customer attrition prediction involves gathering relevant customer data such as demographics, purchase history, customer ID, gender, age, balance, and interaction logs. Data preprocessing for missing values involves strategies like imputation, where missing data is estimated or filled in with statistical measures such as mean, median, mode, and standard deviation. Data preprocessing for outlier detection in customer attrition prediction involves using statistical method like interquartile range (IQR) detection to identify and handle abnormal data points. Scaling ensures numerical features are on similar scales using z-score normalization method, helping machine learning model performance.

### C. *Data Analysis*

Exploratory Data Analysis, is the process of analyzing data to understand its patterns, relationships, and distributions to gain insights and inform decision-making. Univariate analysis examines variables individually, using histograms for numerical data and bar plots for categorical data. Bivariate analysis explores relationships between variables, comparing them with the churn label through scatter plots or correlation matrices. Feature engineering creates new metrics like average transaction amount, aiding in predicting churn by providing insights into customer behaviour.

### D. *Implementing The Gradient Boosting Classifier Algorithm*

Gradient Boosting Classifier sequentially builds an ensemble of weak learners, correcting errors of predecessors by optimizing a loss function, enhancing prediction accuracy through iterative learning. Split the data into training and testing sets, then train the Gradient Boosting Classifier model on the training data. Evaluate the model's performance using metrics such as accuracy, precision, recall and F1-score. Additionally, generate a confusion matrix to understand the model's predictions. Tune the model hyperparameters using technique like Grid search cross-validation to find the best combination of parameters that optimize model performance.

## VI. DATA FLOW DIAGRAM

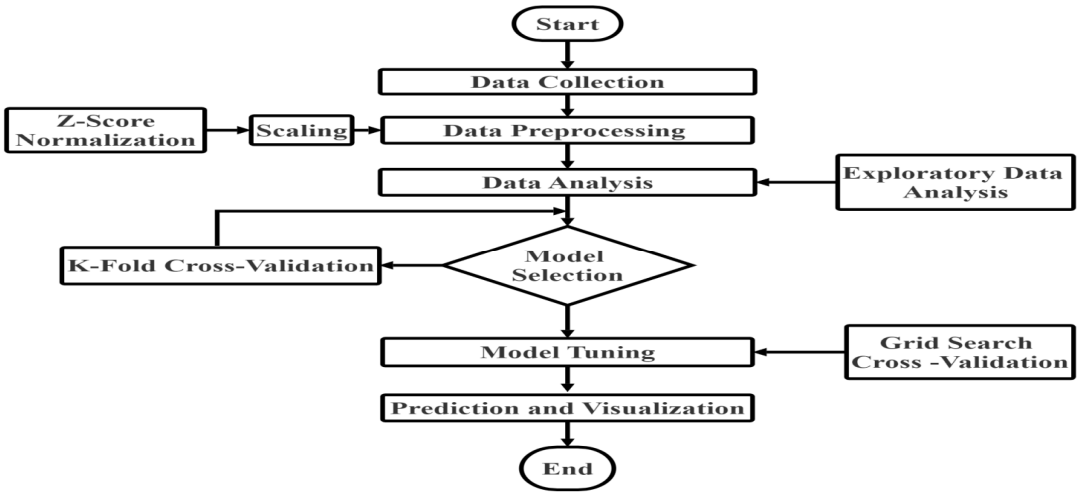



Figure:1.1 Flow Diagram

## VII. SCREENSHOTS

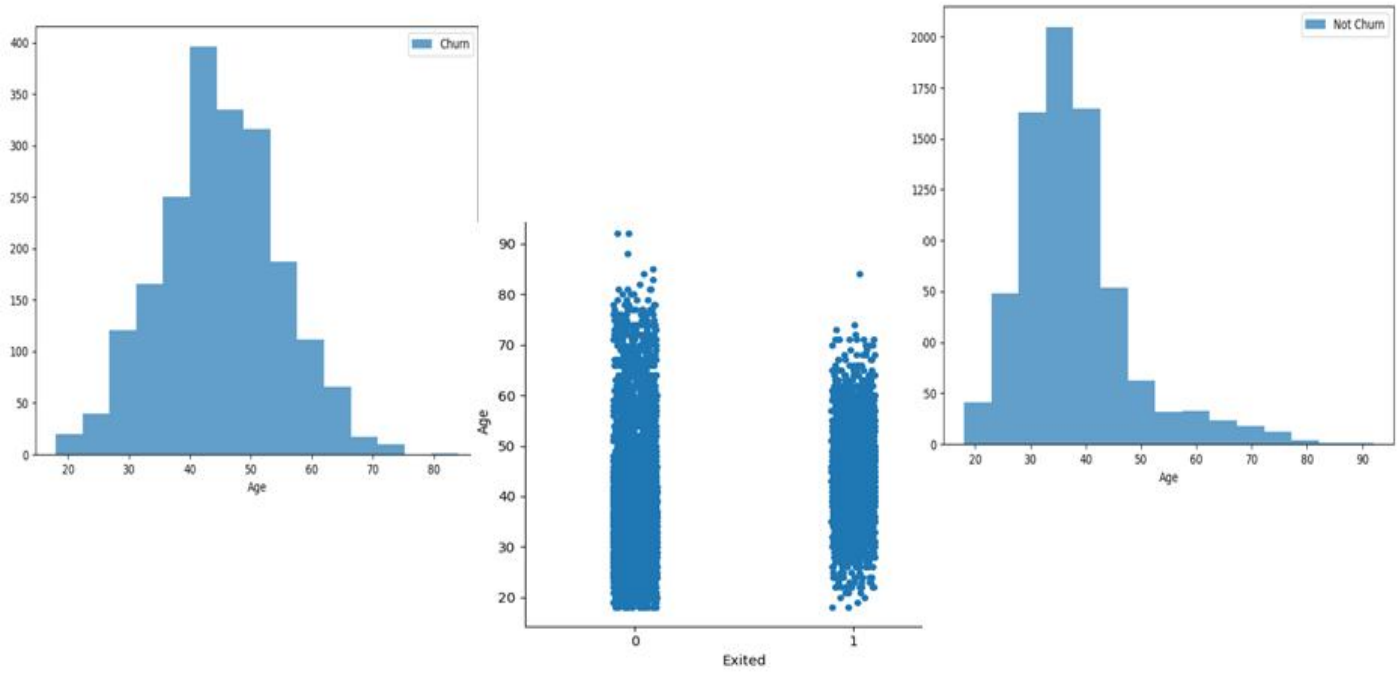



**Figure:2.1 Screenshot for Age Analysis**

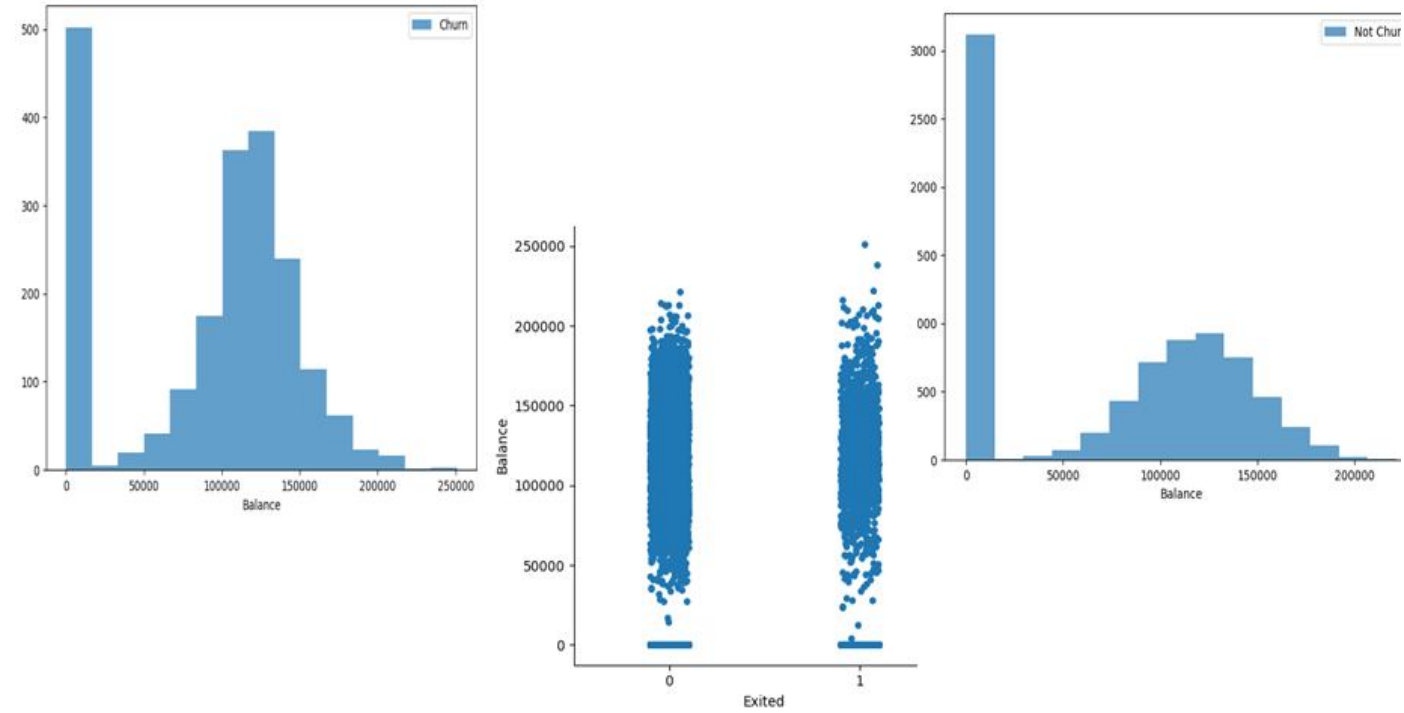



**Figure:2.2 Screenshot for Balance Analysis**

©IJRASET: All Rights are Reserved 

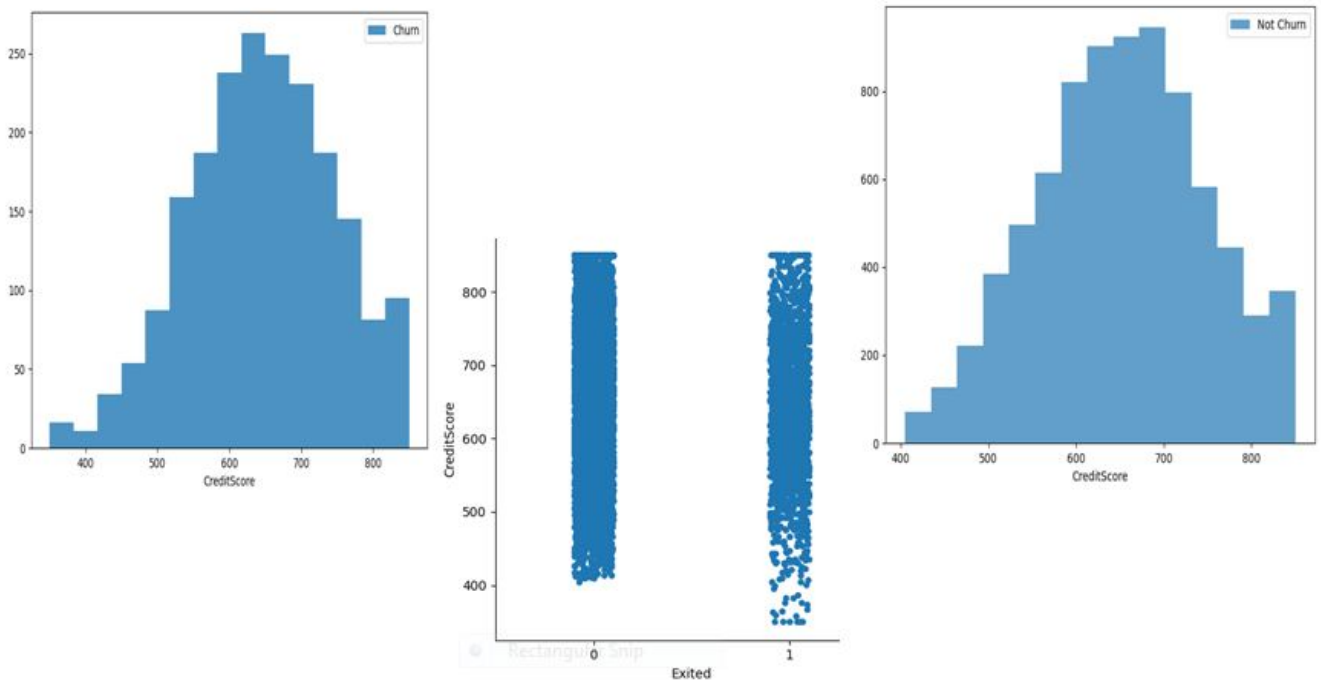



**Figure:2.3 Screenshot for Credit Score Analysis**

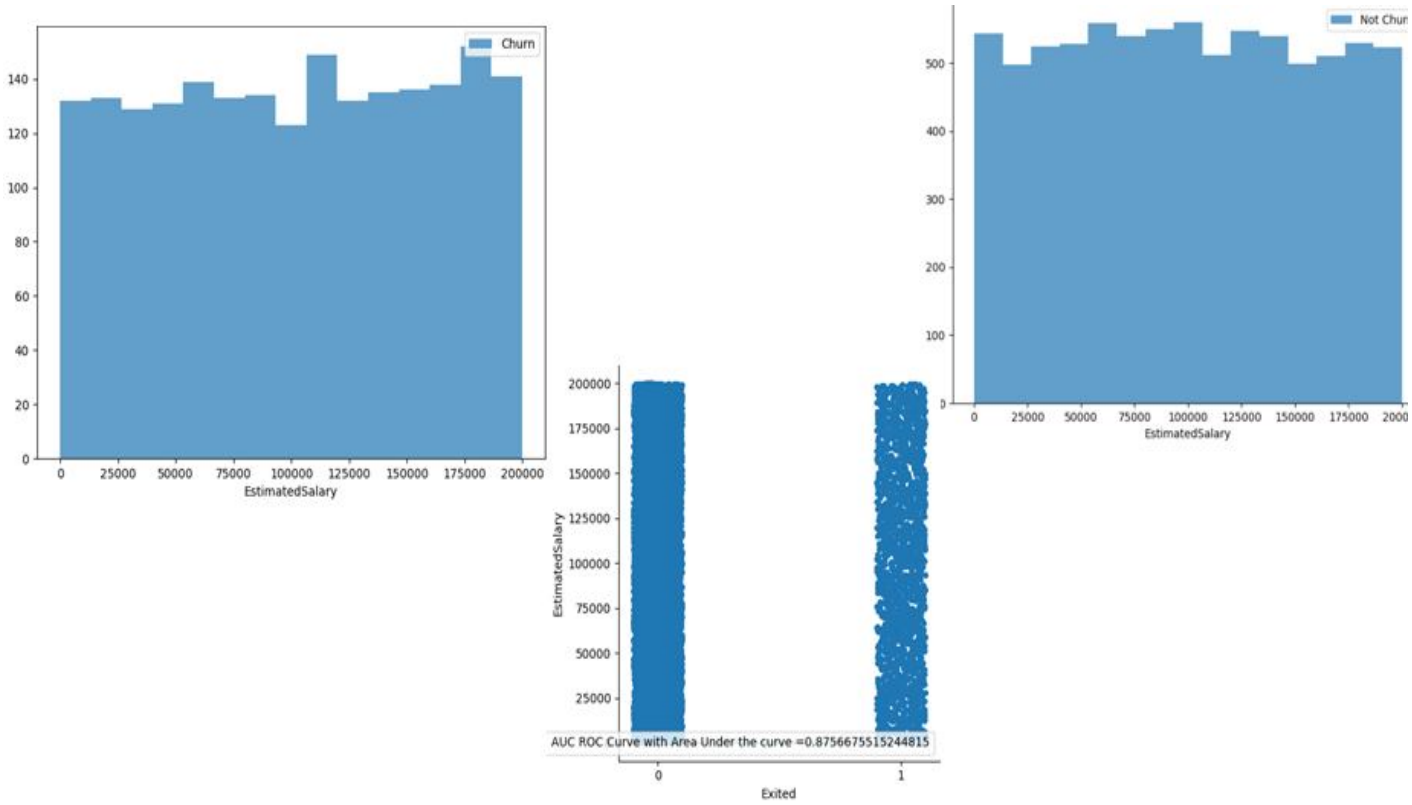



**Figure:2.4 Screenshot for Estimated Salary Analysis**

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 12 Issue VIII Aug 2024- Available at www.ijraset.com*

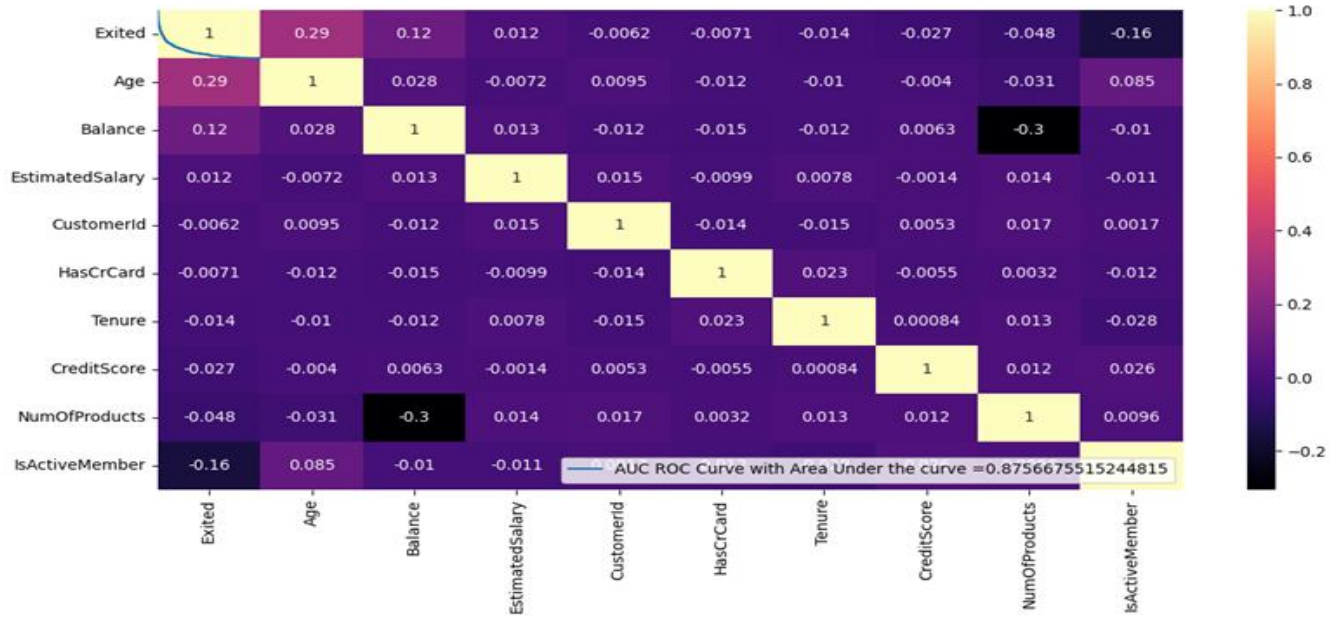



**Figure:2.5 Screenshot for Correlation Matrix Heatmap**

```
Index: 10000 entries, 1 to 10000
Data columns (total 13 columns):

 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   CustomerId       10000 non-null  int64
 1   Surname          10000 non-null  object
 2   CreditScore      10000 non-null  int64
 3   Geography        10000 non-null  object
 4   Gender           10000 non-null  object
 5   Age              10000 non-null  int64
 6   Tenure           10000 non-null  int64
 7   Balance          10000 non-null  float64
 8   NumOfProducts    10000 non-null  int64
 9   HasCrCard        10000 non-null  int64
 10  IsActiveMember   10000 non-null  int64
 11  EstimatedSalary  10000 non-null  float64
 12  Exited           10000 non-null  int64
dtypes: float64(2), int64(8), object(3)

CreditScore has  None Outliers
Age has  None Outliers
Balance has  None Outliers
EstimatedSalary has  None Outliers
```

```
LR: 0.825100 (0.009710)
KNN: 0.821000 (0.007071)
CART: 0.794500 (0.010595)
RF: 0.860700 (0.006724)
SVR: 0.847000 (0.009230)
GB: 0.864300 (0.009371)
LightGBM: 0.863600 (0.009952)

confusion matrix for GB:
[[1520  230]
 [  53  197]]
True Positive :  197
True Negative :  1520
False Positive:  230
False Negative:  53

              precision    recall  f1-score   support

           0       0.97      0.87      0.91      1750
           1       0.46      0.79      0.58       250

    accuracy                           0.86      2000
   macro avg       0.71      0.83      0.75      2000
weighted avg       0.90      0.86      0.87      2000

GB: 0.866600 (0.008273)
LightGBM: 0.866900 (0.009741)
Accuracy for LightGBM:  97.88%
```

**Figure:2.6 Screenshot for Model Accuracy**

©IJRASET: All Rights are Reserved 

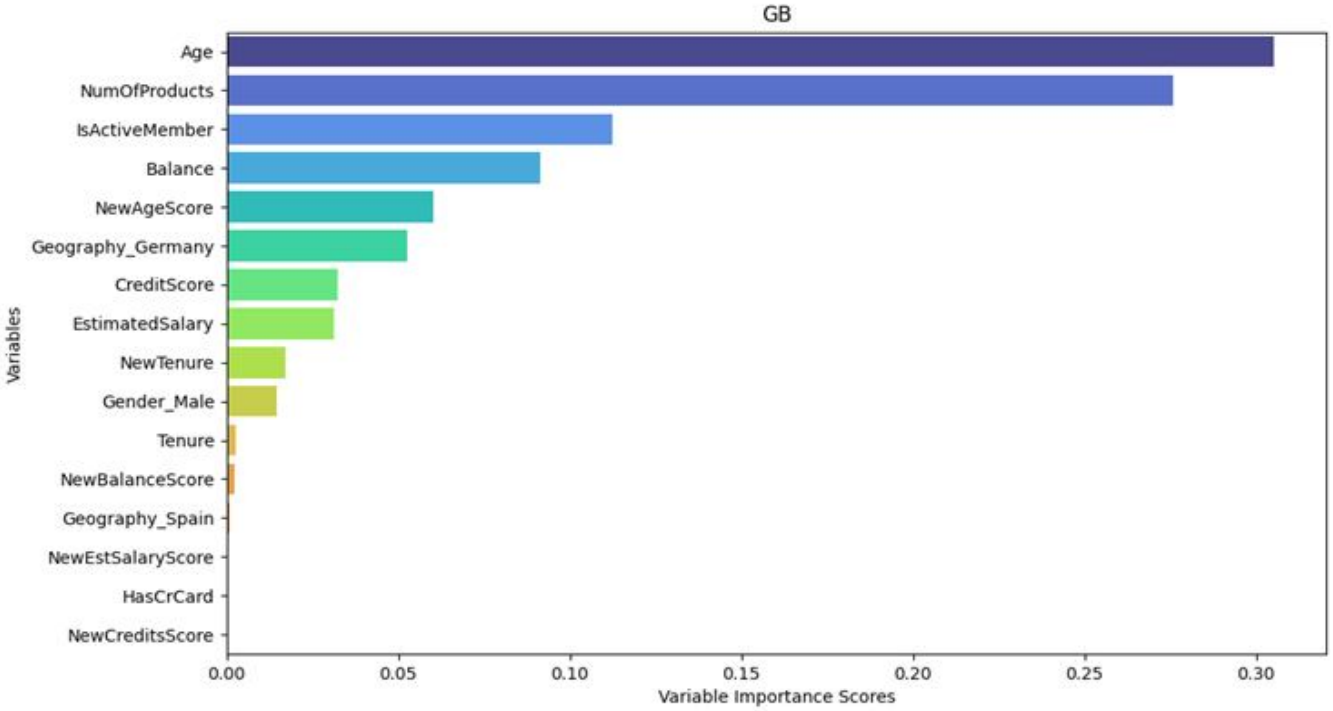



**Figure:2.7 Screenshot for Churn Prediction using Gradient Boosting Classifier**

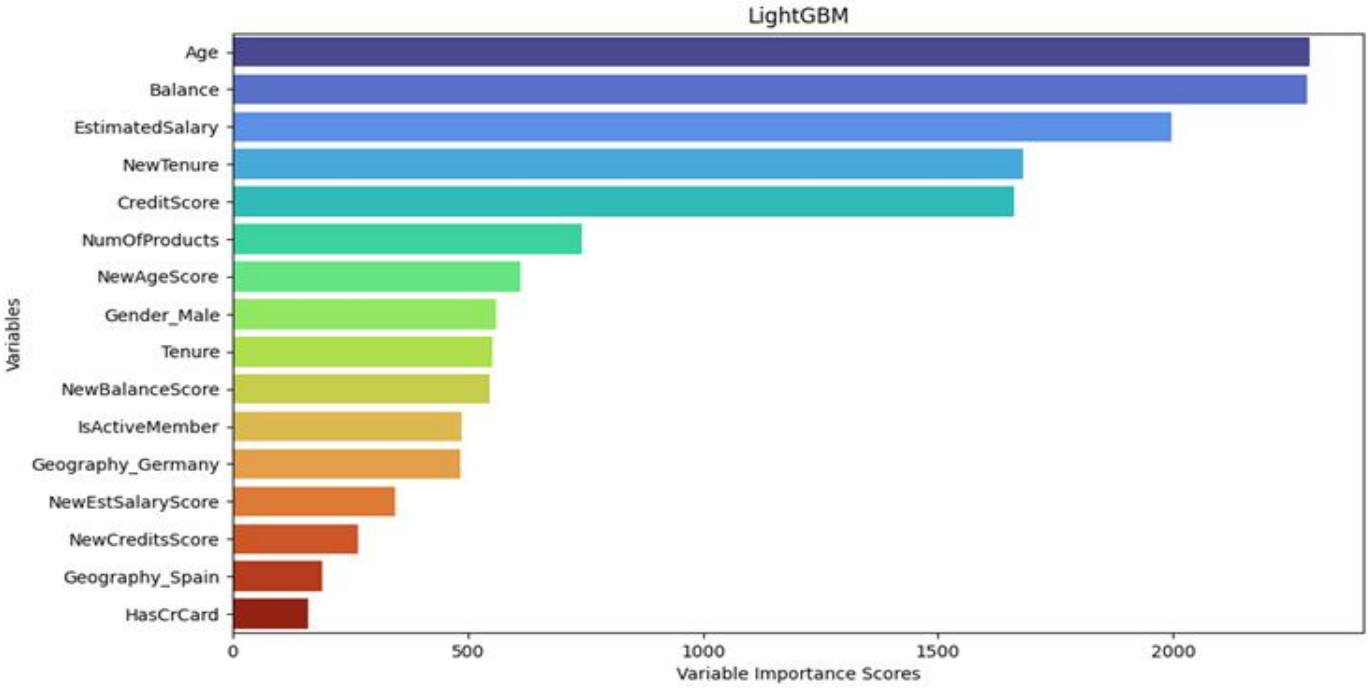



**Figure:2.8 Screenshot for Churn Prediction using LightGBM**

©IJRASET: All Rights are Reserved 

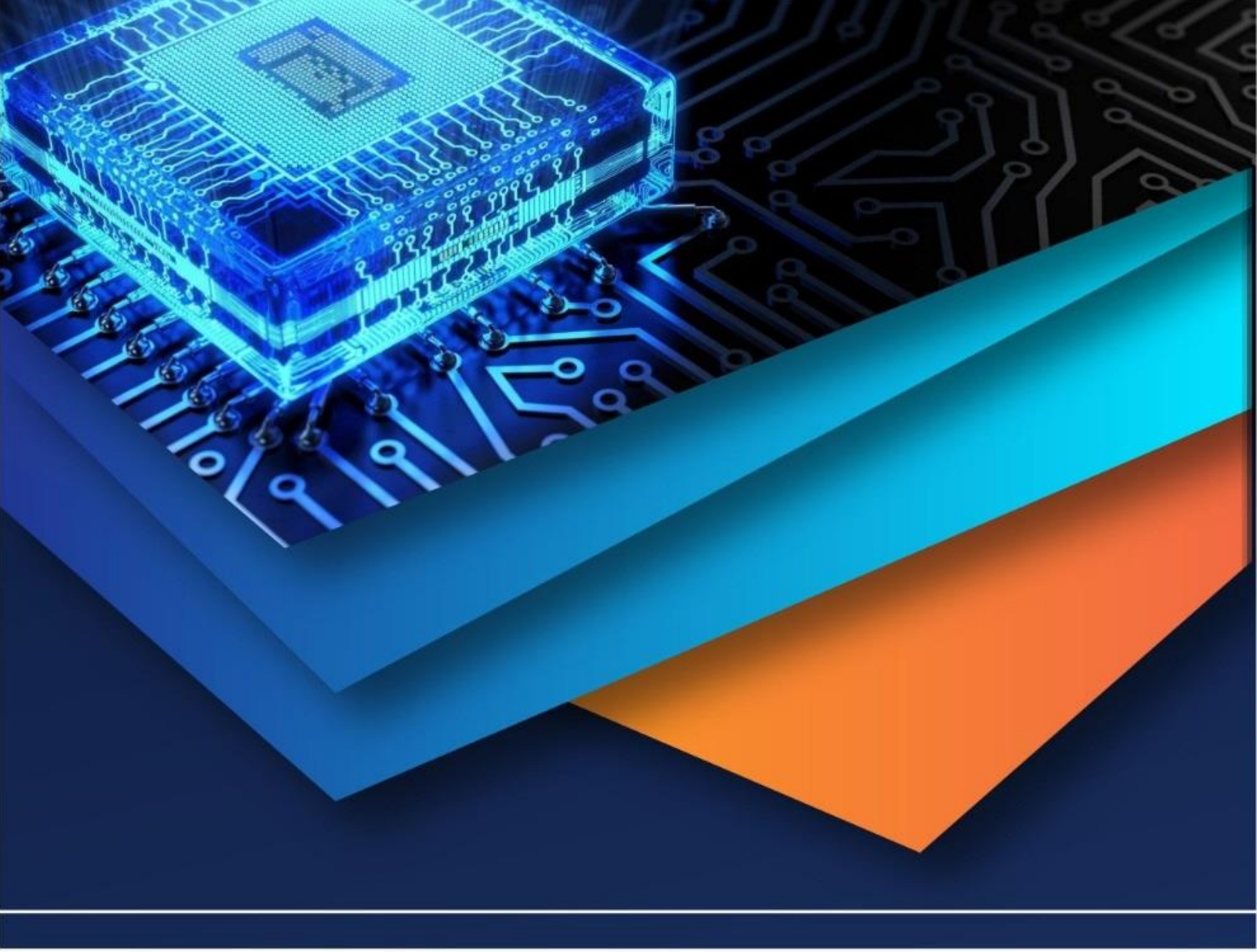
## VIII. CONCLUSION

In order for businesses to keep valuable customers, they must be able to predict customer attrition. Machine learning analysis helps identify factors that influence customer attrition and provides a reliable framework for identifying patterns and predictors. By knowing these factors, businesses can implement targeted retention strategies that minimize attrition and maximize customer lifetime value. This analysis also paves the way for ongoing efforts to optimize customer retention and boost business performance.

## REFERENCES

[1] N. Jajam and N. P. Challa, ‘‘Customer churn detection for insurance data using blended logistic regression decision tree algorithm (BLRDT),’’ Int. J. Intell. Syst. Appl. Eng., vol. 11, no. 1, pp. 72–83, 2023.
[2] R. A. de Lima Lemos, T. C. Silva, and B. M. Tabak, ‘‘Propension to customer churn in a financial institution: A machine learning approach,’’ Neural Comput. Appl., vol. 34, no. 14, pp. 11751–11768, Jul. 2022.
[3] A. B. W. Ali, ‘‘Prediction of employee turn over using random forest classifier with intensive optimized PCA algorithm,’’ Wireless Pers. Commun., vol. 119, no. 4, pp. 3365–3382, Aug. 2021.
[4] E. Zdravevski, P. Lameski, C. Apanowicz, and D. Ślęzak, ‘‘From big data to business analytics: The case study of churn prediction,’’ Appl. Soft Comput., vol. 90, May 2020, Art. no. 106164.
[5] A. De Caigny, K. Coussement, K. W. De Bock, and S. Lessmann, ‘‘Incorporating textual information in customer churn prediction models based on a convolutional neural network,’’ Int. J. Forecasting, vol. 36, no. 4, pp. 1563–1578, Oct. 2020.
[6] A. A. Q. Ahmed and D. Maheswari, ‘‘An enhanced ensemble classifier for telecom churn prediction using cost based uplift modelling,’’ Int. J. Inf. Technol., vol. 11, no. 2, pp. 381–391, Jun. 2019, doi: 10.1007/s41870-018- 0248-3.
[7] 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, Sep. 2018.
[8] V. Umayaparvathi and K. Iyakutti, ‘‘Automated feature selection and churn prediction using deep learning models,’’ Int. Res. J. Eng. Technol., vol. 4, no. 3, pp. 1846–1854, 2017.
[9] V. Umayaparvathi and K. Iyakutti, ‘‘A survey on customer churn prediction in telecom industry: Datasets, methods and metrics,’’ Int. Res. J. Eng. Technol., vol. 3, no. 4, pp. 1065–1070, 2016.
[10] A. Keramati, R. Jafari-Marandi, M. Aliannejadi, I. Ahmadian, M. Mozaffari, and U. Abbasi, ‘‘Improved churn prediction in telecommunication industry using data mining techniques,’’ Appl. Soft Comput., vol. 24, pp. 994–1012, Nov. 2014.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)