Integrated Churn Prediction and Segmentation

Prof. Nikitha K.S.[1], Mrityunjay Prakash[2], Rahul Kumar Saini[3], Ram Kumar Pandey[4] Rohit Ranjan[5]

[1] *Assistant Professor, Department of CS&E, Bangalore Institute of Technology, Bengaluru, India*
[2,3,4,5] *Department of CS&E, Bangalore Institute of Technology, Bengaluru, India*

ABSTRACT

Customer churn prediction and segmentation are crucial aspects of data-driven marketing. This paper presents a comprehensive approach that encompasses various modules to tackle these challenges effectively. The methodology proposed in this study includes data processing, churn prediction using machine learning algorithms, customer segmentation through K-means clustering, and result analysis. To ensure the quality and consistency of the dataset, the data processing stage performs essential tasks such as data transformation, cleaning, and normalization. Additionally, the Synthetic Minority Over-sampling Technique (SMOTE) is employed to address any issues of data imbalance, enhancing the reliability of the results. The churn prediction module plays a pivotal role in identifying potential churners accurately. By employing bagging tree, extra trees, and random forest algorithms, this module achieves high prediction accuracy. Furthermore, k-fold cross-validation and feature selection techniques are utilized for robust model evaluation and determination of variable importance. The customer segmentation module utilizes Exploratory Data Analysis (EDA) techniques to extract meaningful insights from the data. By employing K-means clustering, customers are grouped based on their similarities and behaviors, enabling businesses to tailor their marketing strategies to different customer segments effectively. Finally, the results obtained from the churn prediction and segmentation models are thoroughly analyzed to assess their effectiveness. This analysis provides valuable insights for businesses seeking to proactively manage customer churn and implement targeted marketing strategies, ultimately leading to improved business performance and enhanced customer satisfaction. This paper offers a comprehensive and practical approach for customer churn prediction and segmentation in data-driven marketing. The proposed methodology and its associated modules provide a structured framework for businesses to effectively manage customer churn and implement targeted marketing strategies, resulting in improved business performance and customer satisfaction.

Keyword *SMOTE, EDA, K-Means, K-Fold Cross Validation, Normalization, Bagging Tree, Extra Tree, Random Forest, Segmentation.*

1. INTRODUCTION

In the dynamic and highly competitive landscape of modern businesses, customer churn has emerged as a critical challenge with far-reaching consequences. Customer churn, also referred to as customer attrition or customer turnover, represents the phenomenon in which customers discontinue their relationship with a company, ceasing to utilize its products or services. Churn poses a significant threat to the sustainability and profitability of businesses operating across diverse industries, spanning telecommunications, e-commerce, financial services, and subscription-based platforms.

The ramifications of customer churn are multifaceted and extend beyond immediate revenue loss. Acquiring new customers typically entails higher costs compared to retaining existing ones, making customer retention a cost-effective strategy for organizations. Moreover, loyal and satisfied customers often serve as brand advocates, promoting positive word-of-mouth and attracting new customers. Conversely, the departure of dissatisfied customers can generate negative reviews and tarnish a company's reputation. Consequently, minimizing customer churn has become a strategic imperative for organizations seeking long-term success and sustainable growth.

In recent years, the convergence of advanced technologies, the proliferation of digital touchpoints, and the availability of vast amounts of customer data have opened new avenues for understanding and predicting customer behavior. Predictive analytics, data mining techniques, and machine learning algorithms have emerged as powerful tools for analyzing large volumes of customer data to forecast and identify customers at risk of churn. By leveraging

historical data, identifying behavioral patterns, and considering relevant customer attributes, organizations can develop accurate churn prediction models that facilitate proactive decision-making and targeted intervention.

However, accurate churn prediction alone is not sufficient to devise effective retention strategies. Organizations must also undertake customer segmentation to gain insights into the unique characteristics, needs, and preferences of different customer groups. Customer segmentation involves dividing a customer base into homogeneous clusters based on demographics, psychographics, purchasing behavior, or other relevant factors. By recognizing distinct customer segments, organizations can tailor their marketing initiatives, personalize communication, and design retention tactics that align with the specific requirements of each segment.

This research paper aims to provide a comprehensive analysis of customer churn prediction and segmentation methodologies, encompassing both theoretical frameworks and practical applications. It will explore various predictive analytics techniques, such as logistic regression, decision trees, random forests, and neural networks, highlighting their strengths, limitations, and applicability to churn prediction. Additionally, the paper will delve into customer segmentation techniques, encompassing traditional methods like demographic and psychographic segmentation, as well as more advanced approaches like clustering and customer lifetime value (CLV) segmentation.

Real-world case studies and industry examples will be incorporated to showcase the successful implementation of churn prediction and segmentation strategies across diverse sectors. Furthermore, the paper will discuss the implications of churn prediction and segmentation on business outcomes, emphasizing the potential benefits of reducing customer churn, enhancing customer satisfaction, and driving customer lifetime value.

In conclusion, understanding and effectively managing customer churn have become critical priorities for organizations striving for sustainable growth in today's fiercely competitive market. By harnessing advanced analytics techniques and segmentation methodologies, businesses can gain valuable insights into customer behavior, anticipate churn, and implement targeted retention strategies. This research paper aims to equip businesses with a comprehensive understanding of customer churn prediction and segmentation, enabling them to make data-driven decisions and optimize their customer retention efforts to foster long-term customer loyalty and business success. By incorporating advanced predictive analytics techniques and effective customer segmentation strategies, organizations can enhance their understanding of customer behavior and preferences, proactively identify churn risk, and implement tailored retention initiatives to mitigate customer attrition.

Moreover, the availability of vast amounts of customer data and advancements in data analytics and machine learning present opportunities to develop accurate churn prediction models. These models leverage historical data, customer attributes, and behavioral patterns to forecast and identify customers who are likely to churn in the future. By identifying potential churners in advance, organizations can allocate resources efficiently, personalize communication efforts, and implement targeted retention strategies.

2. WORK IN THIS AREA

This section offers the detailed depiction of review on various existing techniques employed so far. The author in the paper [2] presents a study on using machine learning algorithms to predict customer churn in the banking sector. The authors argue that predicting customer churn is crucial for banks to retain their customers and increase revenue. The study uses a dataset of 10,000 customers from a Turkish bank, which includes various customer attributes and transactional data. The authors compared the performance of four machine learning algorithms: logistic regression, decision tree, random forest, and gradient boosting machine. They found that the gradient boosting machine algorithm outperformed the other algorithms in terms of accuracy, precision, recall, and F1 score. The authors also used feature importance analysis to identify the most important variables that contribute to customer churn prediction, such as the number of transactions, account balance, and age. The study found that the machine learning algorithm can effectively predict customer churn in the banking sector with a high degree of accuracy. The authors suggested that banks can use the predictive model to identify at-risk customers and take proactive measures to retain them, such as offering personalized promotions and discounts.

The Paper [3] discusses the use of the K-means clustering algorithm to segment customers based on their behaviour and attributes. The authors argue that customer segmentation is an important marketing strategy that allows businesses to tailor their marketing efforts to specific customer groups. The study uses a dataset of customer transactions from a retail company in Indonesia. The dataset includes information such as customer ID, purchase amount, and purchase frequency. The authors used the K-means clustering algorithm to segment customers into four groups based on their purchase behaviour and attributes. The authors found that the K-means algorithm was able to effectively segment customers into distinct groups. The four customer segments identified were "high spenders", "medium spenders", "low spenders", and "infrequent buyers". The authors also used statistical analysis to identify the differences between the customer segments, such as their average purchase amount and frequency.

The paper [4] presents a study on improving the classification performance of imbalanced datasets using a combination of two techniques: the KM++ algorithm for initializing centroids in K-means clustering and the SMOTE algorithm for oversampling the minority class. The authors argue that imbalanced datasets, where one class has significantly fewer samples than the other, can lead to biased classification models. The study uses several imbalanced datasets, including credit card fraud and cancer diagnosis datasets, to evaluate the performance of the proposed algorithm. The authors found that the KM++ SMOTE algorithm outperformed several other oversampling algorithms, such as random oversampling and SMOTE, in terms of classification accuracy, precision, recall, and F1 score. The KM++ SMOTE algorithm was able to improve the performance of classification models on imbalanced datasets by generating synthetic samples of the minority class and initializing centroids in K-means clustering to improve the clustering of the minority class. The study also compared the performance of several classification models, including logistic regression, decision tree, and random forest, on imbalanced datasets with and without oversampling. The authors found that oversampling using the KM++ SMOTE algorithm led to significant improvements in classification performance for all models.

The paper [5] presents a study on predicting customer churn in the telecommunications industry using two machine learning algorithms: decision trees and logistic regression. The authors argue that predicting customer churn is crucial for telecommunications companies to improve customer retention and reduce revenue loss. The study uses a dataset of customer attributes and behaviors from a telecommunications company in India. The authors compared the performance of decision trees and logistic regression algorithms in predicting customer churn. They found that the decision tree algorithm outperformed the logistic regression algorithm in terms of accuracy, precision, and F1 score. The decision tree algorithm was able to identify the most important features for predicting customer churn, such as the number of complaints and customer tenure. The authors also performed feature selection to identify the most relevant features for predicting customer churn. They found that using a subset of features improved the performance of both decision trees and logistic regression models. The study concludes that machine learning algorithms such as decision trees and logistic regression can be effective in predicting customer churn in the telecommunications industry. The authors suggest that telecommunications companies can use the predictive models to identify at-risk customers and take proactive measures to retain them, such as offering personalized promotions or addressing customer complaints.

The paper [6] presents a study on detecting fraud in credit card transactions using exploratory data analysis (EDA) and supervised learning techniques. The authors argue that fraud detection is critical for financial institutions to prevent fraudulent transactions and protect their customers' assets. The study uses a dataset of credit card transactions from a Brazilian bank to demonstrate the effectiveness of the proposed approach. The authors used EDA techniques to explore the dataset and identify features that are most relevant to detecting fraud. They found that features such as transaction amount, time of the day, and location can be useful in identifying fraudulent transactions. The authors then used supervised learning algorithms, including decision trees, random forests, and support vector machines, to build predictive models for fraud detection. They found that the random forest algorithm outperformed the other algorithms in terms of accuracy, precision, recall, and F1 score. The authors also used feature selection techniques to identify the most relevant features for fraud detection. They found that using a subset of features can improve the performance of the predictive models. The study concludes that EDA techniques can be useful in identifying features that are relevant for fraud detection in credit card transactions. The authors suggest that financial institutions can use the predictive models to monitor transactions in real-time and identify potential fraudulent transactions.

The paper [7] presents a comprehensive survey of decision tree algorithms for classification. The authors argue that decision trees are widely used in machine learning for classification tasks due to their simplicity and interpretability. The paper provides an overview of the history and evolution of decision tree algorithms, starting from the earliest algorithm, ID3, to the most recent algorithms such as C4.5, CART, and Random Forests. The authors also discuss the advantages and limitations of each algorithm. The authors provide a detailed description of the decision tree building process, which involves selecting the best attribute for splitting the dataset and recursively building subtrees until a stopping criterion is met. They also discuss the pruning techniques used to avoid overfitting. The paper also compares the performance of decision tree algorithms with other classification algorithms such as logistic regression, support vector machines, and artificial neural networks. The authors found that decision trees perform well on datasets with categorical and continuous variables and can handle missing values and outliers.

The paper [8] presents a study on the use of the K-nearest neighbors (KNN) algorithm for classification tasks. The authors argue that KNN is a popular algorithm for classification due to its simplicity and effectiveness. The study proposes a new KNN model-based approach that uses a discriminant function to improve the performance of the algorithm. The authors compare the proposed approach with the traditional KNN algorithm and other classification algorithms, such as decision trees and support vector machines, on several datasets. The authors found that the proposed KNN model-based approach outperformed the traditional KNN algorithm and other classification algorithms in terms of accuracy, precision, and recall. The authors also compared the performance of the proposed approach with

other model based KNN algorithms, such as probabilistic KNN and weighted KNN, and found that the proposed approach performed better on most datasets. The study also investigated the effect of the number of neighbors on the performance of the KNN algorithm. The authors found that the optimal number of neighbors depends on the characteristics of the dataset, and a small number of neighbors can lead to overfitting, while a large number of neighbors can lead to underfitting. The authors also discuss the limitations of the KNN algorithm, such as its sensitivity to irrelevant and noisy features, and suggest several techniques to address these limitations, such as feature selection and feature weighting.

The paper [9] provides an extensive overview of the research in the field of customer churn prediction, which is the task of identifying customers who are likely to terminate their relationship with a company. The authors argue that customer churn prediction is a critical task for businesses to retain their customers and maintain their profitability. The paper covers various aspects of customer churn prediction, including the definition of churn, the factors influencing churn, the types of data used for prediction, and the machine learning techniques used for prediction. The authors provide an overview of the key research studies in each of these areas, highlighting their contributions and limitations. The paper discusses the importance of feature engineering in customer churn prediction, which involves selecting and transforming the relevant features that can best explain the churn behavior. The authors present various feature engineering techniques, such as clustering, association rule mining, and principal component analysis, that can help in identifying the key features. The paper also provides a comprehensive review of the machine learning techniques used for customer churn prediction, including logistic regression, decision trees, support vector machines, and neural networks. The authors discuss the strengths and weaknesses of each technique and provide recommendations for selecting the appropriate technique based on the characteristics of the dataset and the goals of the prediction.

The paper [10] investigates the application of machine learning techniques, linear models, and Bayesian models for logistic regression in failure detection problems. The authors argue that failure detection is a critical task in many industries, including manufacturing, transportation, and healthcare, and that logistic regression is a popular method for predicting failures. The study presents a comparison of several machine learning techniques, including logistic regression, decision trees, random forests, and support vector machines, for failure detection. The authors evaluate the performance of these techniques on several datasets and compare their accuracy, precision, recall, and F1-score. The results show that logistic regression performs better than the other techniques on most of the datasets. The study also investigates the use of linear models and Bayesian models for logistic regression in failure detection. The authors compare the performance of these models with logistic regression on several datasets and find that the linear models and Bayesian models perform comparably to logistic regression in most cases. The authors discuss the importance of feature selection and regularization in logistic regression for failure detection. They present several techniques for feature selection, such as wrapper-based and filter-based methods and discuss the benefits and limitations of each technique. They also discuss the use of regularization techniques, such as L1 regularization and L2 regularization, to prevent overfitting and improve the generalization performance of the models.

3. PROPOSED METHODOLOFY

This section is a detailed explanation of the proposed system implemented. The proposed system is divided into 3 modules i.e., Data Processing, Churn Prediction, and Customer Segmentation.

Acquiring datasets relevant to Customer Churn is the initial step in the Customer Churn Prediction. As a result, three categories of datasets have been gathered for the investigation of machine learning techniques. The Data Sets used in this paper are the telecom dataset, banking dataset, and hospital dataset.

Data Processing Module

The data processing module plays a crucial role in preparing the raw data for subsequent analysis. It involves several pre-processing steps to ensure the data is in a suitable format and quality for accurate churn prediction and customer segmentation. The first step in data processing is data transformation. This involves reshaping the data into a format that is compatible with the analysis techniques to be used. For example, if the data is originally in a wide format, it may be transformed into a long format or vice versa, depending on the requirements of the subsequent modules. Additionally, variables may be transformed or derived based on domain knowledge or specific feature engineering techniques to capture relevant information or patterns within the data. Following data transformation, the next step is data cleaning. This step aims to identify and address any inconsistencies, errors, duplicates, or missing values present in the dataset. Inconsistencies and errors can arise from data entry mistakes, system errors, or other sources. Duplicates, if present, can lead to biased results and distort the analysis. Missing values, whether occurring due to non-response or other reasons, need to be handled appropriately to ensure the completeness of the dataset. Various techniques, such as

imputation or removal of missing values, can be employed based on the nature and extent of missing data. After data cleaning, data normalization is performed. This step standardizes the variables by bringing them into a consistent range. Variables in the dataset may have different scales, units, or measurement ranges, which can pose challenges in accurately comparing and analyzing them. By applying normalization techniques, such as z-score normalization or min-max scaling, the variables are rescaled to a common range, typically between 0 and 1 or with a mean of 0 and a standard deviation of 1. This normalization process ensures that each variable contributes equally to the subsequent analysis and prevents any dominant variables from unduly influencing the results.

Next, we address the issue of imbalanced data, if present, through the data balancing module. The Synthetic Minority Over-Sampling Technique (SMOTE) is employed to generate synthetic samples of the minority class, effectively balancing the dataset. This step ensures that the subsequent churn prediction and segmentation modules are not biased toward the majority class, leading to more accurate and balanced results.

Churn Prediction Module

The churn prediction module is a critical component of the proposed framework, focusing on predicting customer churn using machine learning algorithms. Specifically, bagging trees, extra trees, and random forest algorithms are utilized due to their ability to handle complex relationships and generate reliable predictions. These ensemble-based algorithms combine multiple decision trees to create a powerful predictive model. Bagging tree, also known as bootstrap aggregating, involves creating multiple decision trees trained on different subsets of the data through bootstrapping. Extra trees, or extremely randomized trees, further enhance the ensemble by introducing additional randomness during the tree-building process. Random forest, a widely used algorithm, combines the predictions of multiple decision trees to make a final prediction. To ensure the reliability and generalizability of the churn prediction models, the framework employs k-fold cross-validation. In this evaluation technique, the dataset is divided into k subsets or folds. Each model is trained and evaluated k times, with each fold serving as the validation set once and the remaining folds as the training set. This process allows for a comprehensive assessment of the model's performance across different subsets of the data, mitigating the potential bias of a single train-test split. By averaging the performance metrics obtained from each fold, a more robust estimate of the model's accuracy and predictive power can be obtained. The use of k-fold cross-validation enables the identification of potential overfitting or underfitting issues, as well as the comparison of different models and their variations. It provides insights into the generalizability of the models to unseen data and helps determine the optimal hyperparameters for the algorithms. Additionally, by evaluating the models on different folds, it becomes possible to assess their stability and consistency across different subsets of the data, further enhancing the reliability of the predictions.

Following the churn prediction module, the proposed framework includes a feature selection module to identify the most informative and relevant features for predicting churn. Feature selection plays a crucial role in improving the efficiency, interpretability, and performance of the subsequent analysis steps. Various techniques can be employed within the feature selection module. Features that exhibit significant differences can be considered strong predictors of churn. Another approach is featuring importance analysis, which involves applying machine learning algorithms and analyzing their internal feature importance measures. For example, decision tree-based algorithms can provide feature importance scores based on how much each feature contributes to the predictive performance of the model. Features with higher importance scores are deemed more influential in predicting churn and can be selected for further analysis. Dimensionality reduction techniques, such as principal component analysis (PCA), can also be employed for feature selection. PCA reduces the dimensionality of the feature space by identifying linear combinations of features that capture the most significant variability in the data. By selecting a subset of principal components that explain most of the variance, the framework can retain the most relevant information while reducing the number of features. The feature selection module aims to strike a balance between reducing the dimensionality of the feature space and preserving the most informative features. By eliminating redundant or irrelevant features, the framework reduces computational complexity, improves model performance, and enhances the interpretability of the churn prediction models. The selected subset of features serves as input for the subsequent modules, such as customer segmentation and result analysis. By focusing on the most influential features, these modules can uncover actionable insights and facilitate targeted decision-making. Additionally, the feature selection process contributes to a deeper understanding of the factors driving customer churn, empowering businesses to develop effective retention strategies and allocate resources more efficiently.

Customer Segmentation Module

Moving forward in the proposed framework, the customer segmentation module is a crucial step in understanding customer behavior and preferences. This module aims to group customers into distinct segments based on their characteristics and behavior, allowing businesses to tailor their marketing strategies and retention efforts to specific customer profiles. The customer segmentation module begins with exploratory data analysis (EDA) techniques. EDA involves analyzing the dataset to gain insights into the underlying patterns, relationships, and trends within the data. Various statistical and visualization methods are employed to identify key features, uncover hidden patterns, and understand the distribution and variability of customer attributes. By conducting EDA, the framework can identify relevant variables that significantly impact customer behavior and churn. It helps reveal meaningful relationships between variables and offers insights into customer preferences, demographics, purchase history, or engagement patterns. These insights form the basis for segmentation and subsequent analysis. Subsequently, the framework applies K-means clustering, a widely used unsupervised learning algorithm, to partition customers into homogeneous groups or segments. K-means clustering iteratively assigns customers to clusters based on their proximity to cluster centroids, aiming to minimize the within-cluster sum of squares. The number of clusters, k, is predetermined or determined through techniques like the elbow method or silhouette analysis. The resulting customer segments represent groups of customers who share similar characteristics, behavior, or preferences. Each segment represents a distinct profile within the customer base, allowing businesses to understand the diverse needs, motivations, and preferences of their customers. This understanding enables targeted marketing campaigns, personalized offers, and tailored retention strategies for each segment. The customer segmentation process provides several benefits to businesses. Firstly, it enables effective resource allocation by focusing marketing efforts on the most promising customer segments. By understanding the unique characteristics and preferences of each segment, businesses can tailor their messaging, products, and services to resonate with the specific needs and desires of each group. This personalized approach enhances customer satisfaction, engagement, and loyalty. Furthermore, customer segmentation facilitates market segmentation analysis, enabling businesses to identify lucrative market segments and uncover untapped market opportunities. By identifying segments with high growth potential or unique needs, businesses can develop new products or services to cater to those segments, gaining a competitive edge in the market.

The final step in the proposed framework is result analysis, where the outcomes from the churn prediction and customer segmentation modules are thoroughly analyzed and interpreted. This analysis plays a critical role in extracting valuable insights, understanding the factors influencing churn, identifying high-risk customer segments, and uncovering opportunities for proactive retention strategies. The results from the churn prediction module provide valuable information on customers who are at a high risk of churning. By analyzing the predictions and associated probabilities, businesses can identify the key drivers and indicators of churn. This insight enables businesses to understand the underlying factors that contribute to customer attrition, such as poor customer service, product dissatisfaction, or pricing issues. By addressing these factors, businesses can implement targeted retention strategies to mitigate churn and enhance customer satisfaction. Additionally, the result analysis helps in evaluating the performance and effectiveness of the framework. By assessing the accuracy of the churn prediction models, the quality of the customer segmentation, and the alignment of the results with the business objectives, businesses can validate the efficacy of the framework. This analysis helps identify any shortcomings or areas for improvement, allowing businesses to refine their approach and enhance the overall framework's performance.

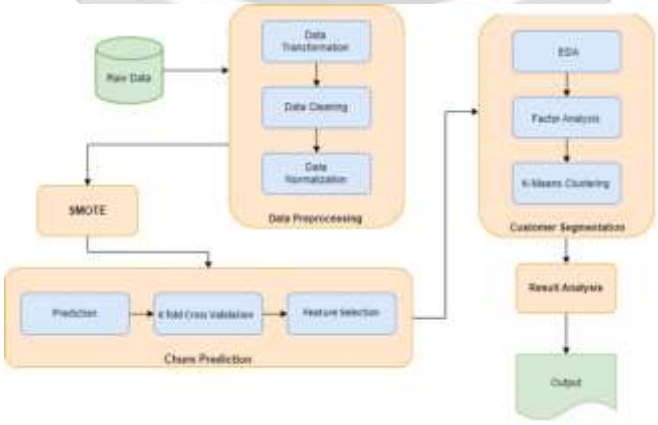

Fig 1. Proposed Methodology

4. EXPERIMENTAL ANALYSIS & RESULTS

The experimental analysis of proposed system is depicted in this section. The outcomes attained are shown in tabulation and graphical representation.

Performance analysis for selecting best model

Machine Learning models are selected for churn prediction based on 4 accuracy scores.

Table 1. Models comparision based on accuracy scores

Performance Metrices	Logistic Regression	Ada Boost Classifier	Gradient Boost Classifier	Bagging Classifier	Extra Trees Classifier	Random Forest Classifier
F1 Score	0.67	0.836	0.847	0.989	1.0	1.0
Precision	0.66	0.831	0.853	0.993	1.0	1.0
Recall	0.68	0.842	0.841	0.985	1.0	1.0
Accuracy	0.67	0.835	0.848	0.989	1.0	1.0

Table 1 is the tabulated values for F1 Score, Precision, Recall and Accuracy for 6+ models. The telecom dataset was used training and testing in the above scenario. It was found that Random Forest and Extra Trees Classifier models performed remarkably well. Three models were selected for further usage.

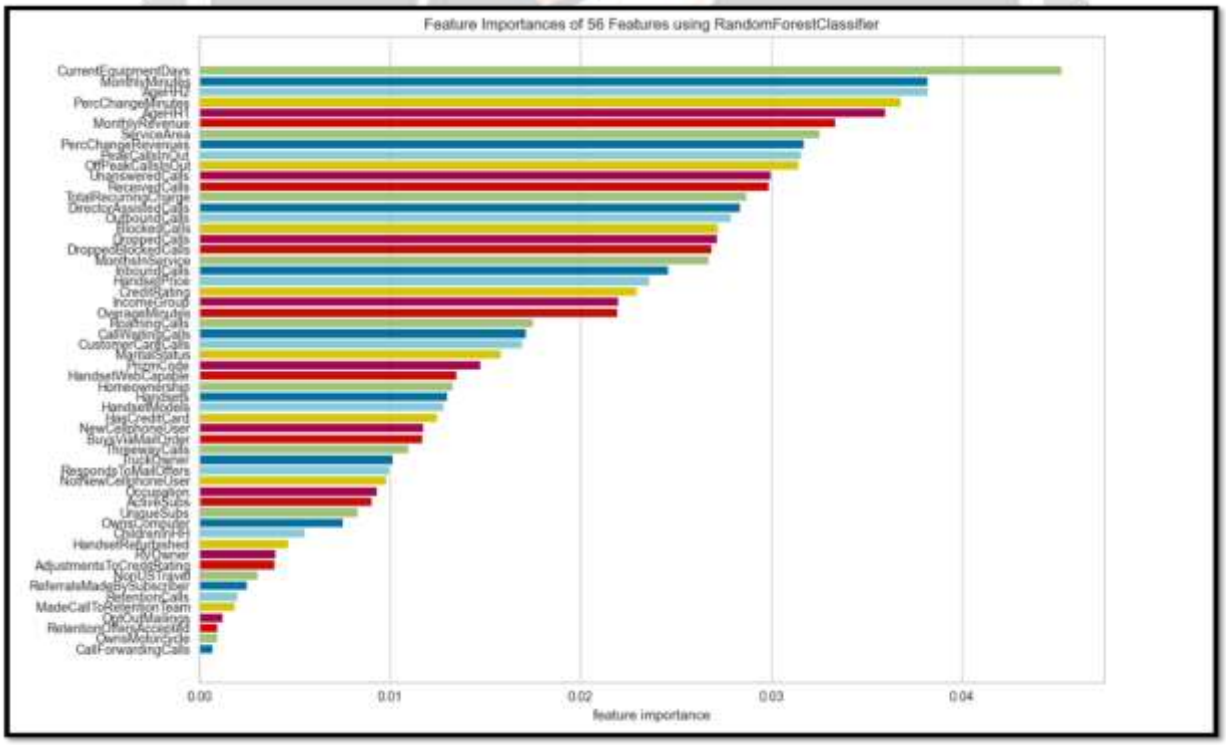

Fig 4.1. Important Features

Fig 4.1 is the representation of important features extracted using Random Forest classifier. Another method Recursive Feature Elimination (RFECV) is also employed for finding the topmost features for each of the telecom and banking datasets.

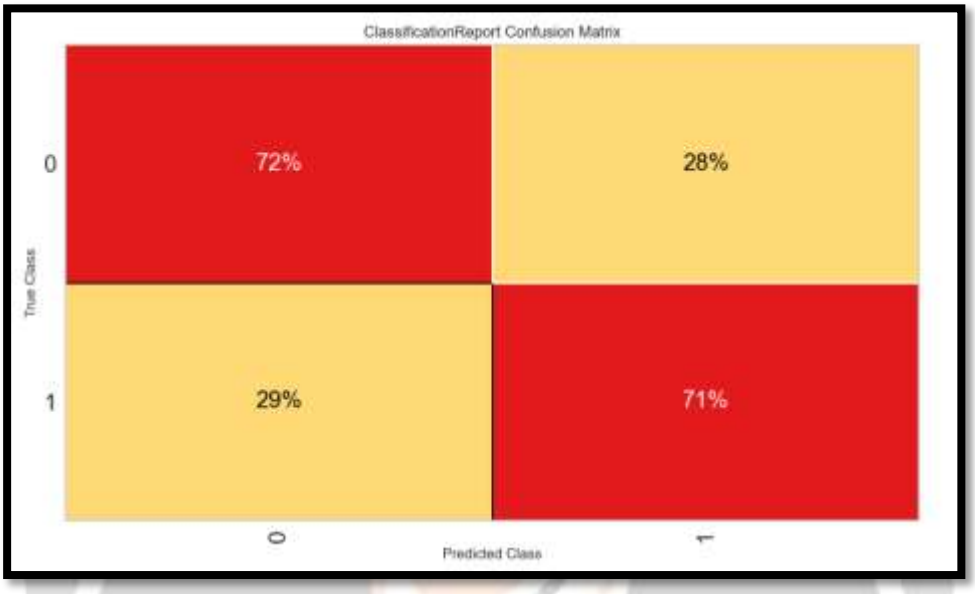

Fig 4.2. Confusion matrix for testing dataset (telecom)

Fig 4.2 shows the confusion matrix for testing dataset. It shows that true positive and false negative rate of people churning (Changing Operator) is 71% and 72% respectively. While false positive and false negative rate of people churning is 28% and 27% respectively.

Handling problem of imbalanced datasets

The imbalanced datasets are handled using Synthetic Minority Over-sampling Technique (SMOTE) which is a data augmentation technique used in machine learning to address class imbalance by generating synthetic samples of the minority class.

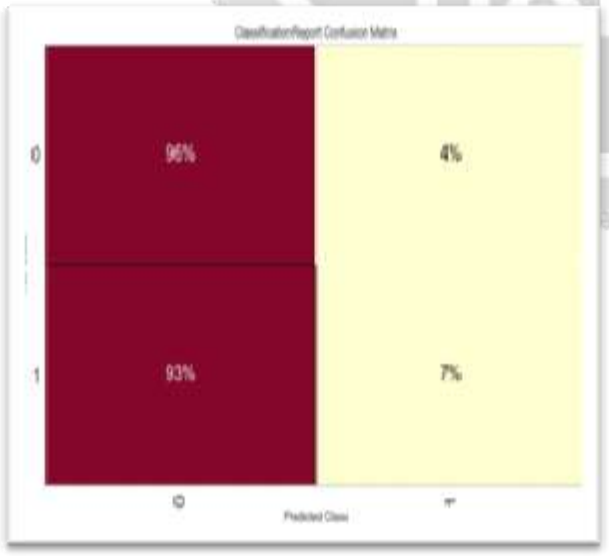

Fig 4.3. Without smote

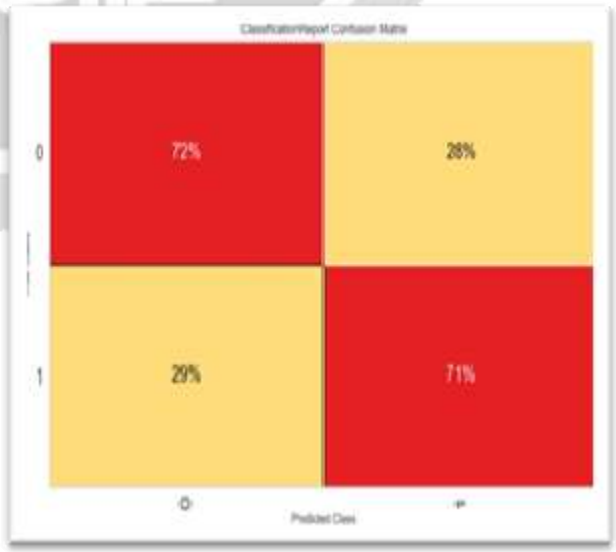

Fig 4.4. With smote

Fig 4.3 represents the classification report confusion matrix and shows remarkably better results for negative test cases (non-churned customers) i.e., 96% and 93% for true negative and false negative respectively. From the classification report it is evident that the dataset is imbalanced where in the positive samples are in minority and negative ones are in majority.

Fig 4.4 represents the classification report confusion matrix with smote being employed on the dataset. The samples are now balanced with 1:1 ratio. It can be inferred from the figure that the performance of the model increased with positive samples being classified correctly. Upon application of smote it can be seen that true positive and false negative samples are classified with 71% and 72% accuracy.

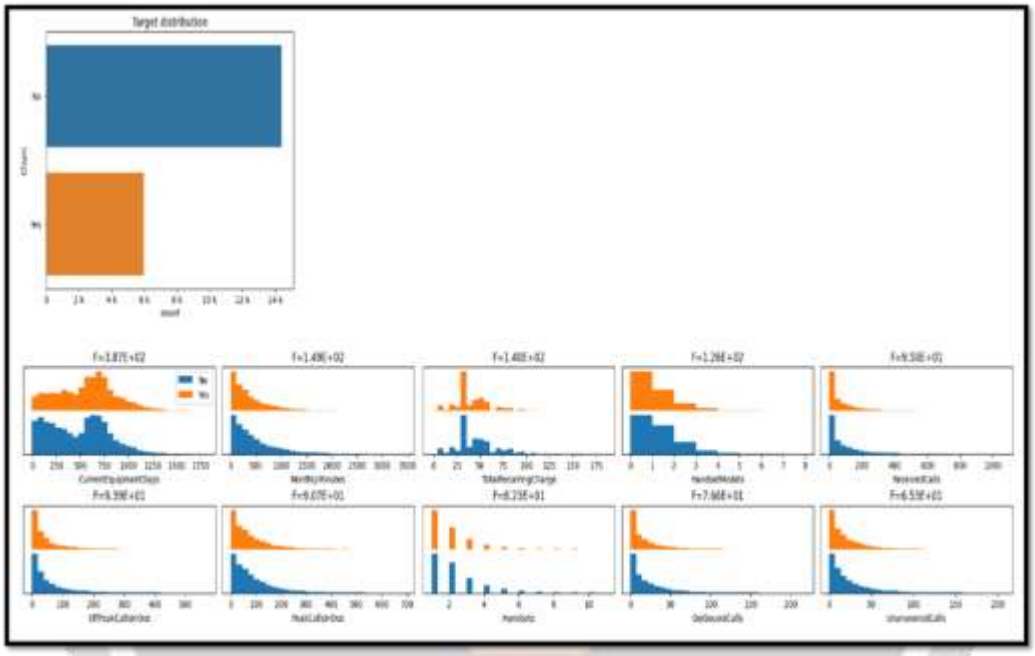

Fig 4.5. Exploratory data analysis with dabl

Fig 4.5 illustrates the exploratory data analysis for the telecom customers. It shows the relationship between different features. It consists of univariate, bivariate and multivariate analysis of the dataset. For example, from the figure it can be inferred that customers have low chances of churning out if the currentEquipmentDays is in the range of 0-200.

K determination for k-means

In order to determine the value of k for k-means clustering. The optimal value of k gives well defined clusters. In this segmentation model Silhouette score and elbow method is used.

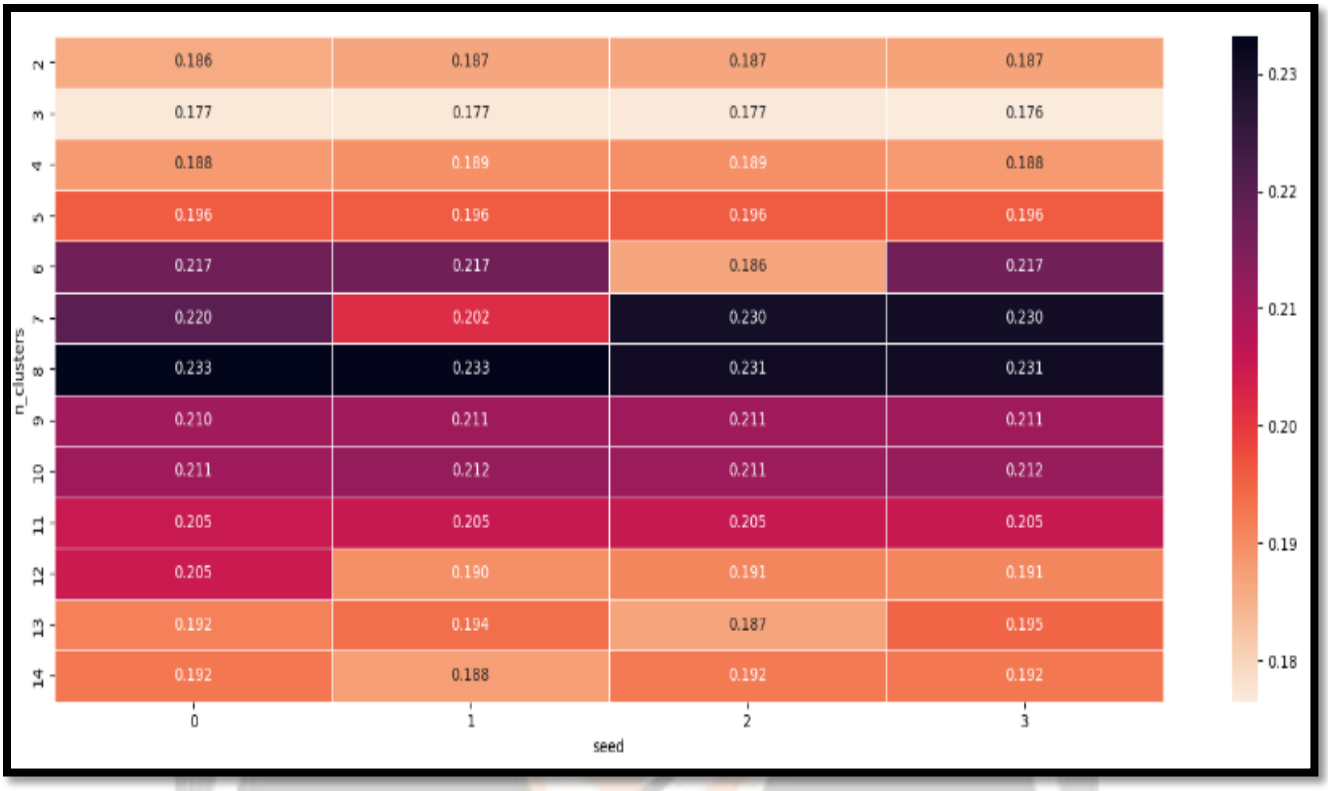

Fig 4.6. Silhouette score matrix.

Fig 4.6 illustrates the silhouette scores for k clusters where k ranges in 2-14. It is found that when k is around 7 or 8, the silhouette score is found to be ~2.33 which is maximum among all k. As a result, k=8 has been chosen for customer segmentation.

K-means clustering for churned customers

The system performs clustering on the customers with churn label = 1, this gives the factors that contribute to churn. It helps to obtain the characteristics of the customers who are likely to churn and thus allow operators, bank or hospitals to have a period of time remediate and implement retention measures for these set of customers.

Table 2 is the tabulated values for each of the important features (chosen from above Fig 4.1) corresponding to each of the 8 clusters ranging 0-7. These clusters shows what are features of customers who have churned and are likely to churn in near future.

Table 2. Clusters formed using k-means

Cluster	Income group	Credit rating	Equipment days	Monthly revenue	Months in service
0	Lower	Lower	High	Low	Medium
1	Mid	Lower	High	Low	Medium
2	Low to mid	Low	Medium	High	High
3	Low to mid	High	Medium	Medium	Medium
4	Mid	Low	Medium	Medium	High
5	Mid	Low	High	Medium	High
6	Mid	Low	Very high	Medium	High
7	High	Low	Very high	Medium	Medium

5. CONCLUSION

The churn prediction and segmentation framework presented in this paper has demonstrated its effectiveness in identifying at-risk customers and predicting churn with high accuracy in major domains such as telecom, banking and hospitals. The proposed framework combines machine learning algorithms with customer segmentation techniques to create a holistic approach that can be easily integrated into existing business processes. The proposed approach was tested on larger and more diverse datasets to evaluate its performance in different scenarios. The potential of the approach was investigated for other types of data analysis tasks, such as classification or clustering. The use of other optimization algorithms or machine learning techniques was explored to improve the efficiency and accuracy of the feature selection and prediction process. An interactive and user-friendly software tool was also developed that implements the proposed approach and can be used by data analysts and domain experts.

6. REFERENCES

[1]. Saumya Saraswat, Akhilesh Tiwari, "A new approach for customer churn prediction in telecom industry", International Journal of Computer Applications (0975 – 8887) Volume 181 – No. 11, August 2018.
[2]. Manas Rahman, V Kumar, "Machine learning based customer churn prediction in banking", Fourth International Conference on Electronics, Communication and Aerospace Technology (ICECA 2020).G. Wolf, J. S. Almeida, M. A. M. Reis, and J. G. Crespo, "Non-mechanistic modelling of complex biofilm reactors and the role of process operation history," J. Biotechnol., vol. 117, no. 4, pp. 367–383, 2005.
[3]. E.Y.L Nandapala, K.P.N Jayasena, "The practical approach in Customers segmentation by using the K-means Algorithm", 15th (IEEE) International Conference on Industrial and Information Systems (ICIIS).
[4]. Li Chen, Ping Dong, Wei Su, &Yan Zhang, "Improving Classification of Imbalanced Datasets Based on KM++ SMOTE Algorithm", 2nd International Conference on Safety Produce Informatization (IICSPI).
[5]. Preeti K. Dalvi, Aditya Bankar, "Analysis of Customer Churn Prediction in Telecom Industry using Decision Trees and Logistic Regression", Symposium on Colossal Data Analysis and Networking (CDAN).
[6]. Vinicius Almendra, Bianca Roman, "Using Exploratory Data Analysis for fraud elicitation through supervised learning", International Symposium on Symbolic and Numeric Algorithms for Scientific Computing.
[7]. Himani Sharma, Sunil Kumar, "A Survey on Decision Tree algorithms of Classification", International Journal of Science and Research (IJSR).
[8]. Gongde Guo, Hui Wang, David Bell, YaxiBi, "KNN Model-based approach in classification", European Commission. Ms. Chinnu P Johny, Mr. Paul P. Mathai, "Customer Churn Prediction: A Survey", International Journal of Advanced Research in Computer Science.
[9]. Ms. Chinnu P Johny, Mr. Paul P. Mathai, "Customer Churn Prediction: A Survey", International Journal of Advanced Research in Computer Science.
[10]. B. Pavlyshenko, "Machine Learning, Linear and Bayesian Models for Logistic Regression in Failure Detection Problems", IEEE International Conference on Big Data.