Final Year Project Mid Defense

Studying Customers' Attrition in Telecom using Modern Learning Algorithms

Group members:

Fiza Maqsood | 355790

Hira Absar khan | 348206

Hadeesa Muskan | 349285

Project supervisor:

Dr. Adnan Idris

Department of Computer Science,
NUST Baluchistan Campus

Table of Contents

Studying Customer churn of telecom customers using Modern Learning Algorithms

I. Introduction:

Customer attrition is a major problem for the telecom industry. it has attracted much attention since it affects revenue and stability in the market. When customers discontinue services provided by a company it is called churning. A churn poses an important challenge for telecom operators worldwide. This problem has arisen due to the intensifying rivalry in the sector and the growing expenses incurred in obtaining new clients as opposed to retaining the existing ones.

Telecommunication companies are currently undergoing significant changes. According to the 2022 State of Customer Churn in Telecom survey, customer loyalty to telecom providers has decreased by 22% since the pandemic. The customer experience now plays a more crucial role than ever in retaining customers, and there is a notable increase in price sensitivity, with 58% of customers considering telecom offerings as expensive. The average turnover rate in the telecom sector is very high. In Western European markets, the typical turnover rate for the more significant utilities industry is approximately 12–15%. However, the average churn rate for telecoms is substantially higher, ranging from 30% to 35% [1]. Several causes, such as poor customer service, flawed products, and high prices, influence this trend. Customer attrition has significant consequences, including higher costs for acquiring new customers and products, fewer recommendations, and a decline in customer lifetime value (CLV). Reducing churn becomes more critical in our changing economic environment, marked by rising interest rates, skyrocketing inflation that affects the price of goods, and a deteriorating labor market. Telecom companies that ignore this factor run the risk of continuously raising user acquisition expenses, which creates a harmful cycle of financial hardship. In the end, unchecked attrition means slow death for telecom companies.

In the telecom sector, where technological advancements and service innovations are common, customer loyalty is the foundation of sustainable growth and profitability. As a result, telecom firms are devoting more resources to detecting, comprehending, and reducing client attrition. This has led to the evolution of churn prediction as a vital strategic imperative. Churn prediction involves forecasting the likelihood of customers discontinuing services based on behavioral, usage, and contextual data.

Understanding the dynamics of churn in the telecom sector involves coping with numerous challenges. The sector's expansive customer base and the wide range of services offered present a complex environment for analysis. Large amounts of data, including call logs, subscription information, usage trends for services, and demographic data, provide a complex yet rich environment for predictive analytics.

Effective churn prediction strategies must navigate this complex data setting, leveraging advanced analytical methodologies and machine learning techniques. The objective is to create proactive solutions that predict, reduce, and preferably prevent churn events rather than predict them. Telecom companies

aim to use predictive models that can identify consumers who are likely to leave and provide retention strategies specific to each customer group.

The significance of churn prediction in the telecom industry cannot be overstated. It affects both sales and customer satisfaction, brand reputation, and market placement. Thus, pursuing precise and effective churn prediction models continues to be the primary focus in telecom analytics, driving advancements in data science, machine learning, and predictive analytics techniques.

II. Literature survey

Churn prediction, a critical facet in industries like telecommunications, finance, and subscription-based services, focuses on forecasting customer attrition or the likelihood of clients discontinuing their relationship with a business. It is vital as retaining existing customers is often more cost-effective than acquiring new ones.

Numerous studies and publications have delved into churn prediction, employing various methodologies, including machine learning, reinforcement learning, deep learning, and ensemble techniques. These approaches aim to leverage historical data to develop models capable of identifying patterns, behaviors, or indicators that precede customer churn. The diversity of techniques reflects the complex nature of customer behavior and the quest for accurate predictive models to assist businesses in proactive churn management strategies.

One study highlights the telecom industry's struggle with customer retention amid fierce competition, emphasizing the significance of retaining existing customers due to higher acquisition costs. This research proposes a novel churn prediction framework utilizing WEKA Data Mining software. It compares the efficacy of Decision Trees and Logistic Regression techniques, aiming to address the industry's high churn rates and resulting losses. The study concludes that decision trees outperform logistic regression, emphasizing their efficiency in churn prediction and aiming to manage and reduce churn rates to stabilize the industry's value [2]

Another study explores the potential of deep learning algorithms to overcome traditional churn prediction limitations. It aims to streamline the process by eliminating manual feature engineering and comparing three deep learning models on real-world telecom datasets (CrowdAnalytix and Cell2Cell). Findings reveal that these models perform comparably to traditional methods like SVM and random forest, showcasing deep learning's effectiveness without manual feature selection [3].

Addressing limitations in existing churn-prediction systems, a separate study introduces a novel technique utilizing subscriber contractual information and call pattern changes for churn prediction. This method identifies potential churners at the contract level and employs a multi-classifier class-combiner approach to handle dataset imbalances. Empirical evaluations demonstrate satisfactory predictive power, particularly when recent call details are used, presenting comparable performance to previous demographics-based systems [4]. However, a comprehensive comparative analysis against diverse churn prediction models could enhance the validation of this approach.

The paper explores customer churn prediction (CCP) within the telecom sector, leveraging advancements in machine learning across six distinct phases. It meticulously covers data preprocessing, feature analysis,

and feature selection via the gravitational search algorithm. It then applies various models such as logistic regression, naive Bayes, SVM, and ensemble techniques on training sets. Emphasizing Adaboost and XGBoost classifiers with an 84% AUC score, the study underscores their superior accuracy, precision, recall, and F-measure performance for churn prediction [5]. Despite this comprehensive exploration and methodical approach, it overlooks providing specific insights into the industry's contextual factors, which could enhance the proposed methodology's accuracy and relevance by considering telecom-specific nuances.

Another study focuses on customer churn prediction within the telecom sector using a Deep-BP-ANN model, showcasing superior accuracy compared to various machine learning techniques on IBM Telco and Cell2Cell datasets. The study employs feature selection methods and parameter tuning to identify critical attributes impacting churn prediction. While demonstrating impressive results and a systematic approach, the paper could benefit from more profound insights into industry-specific nuances influencing churn and discussing practical implementation challenges [6]. Moreover, although the model exhibits high accuracy, emphasizing the interpretability of identified features could enhance its practical business application.

A separate study addresses customer churn prediction in the telecom industry as a critical concern impacting business earnings. The research introduces a churn prediction model employing SVM, MLP, RF, and NB, along with a novel feature selection method. It evaluates the model's accuracy, precision, and F-measure using 10-fold cross-validation, showcasing competitive performance compared to existing methods, particularly in precision and F-measure [7]. However, the study must evaluate projected churn customers' characteristics crucial for business decisions, potentially limiting the model's real-world applicability. Future research on understanding churn customers' attributes and lifetime value could significantly enhance its impact.

Additionally, another paper highlights the importance of analyzing customer behaviors to predict subscription cancellations in the telecom sector, emphasizing the role of data mining techniques. It explores various machine learning models for churn prediction, identifying Gradient boosting as the most effective among them. However, the paper could provide deeper insights into the identified model performances and elucidate churn prediction factors within the telecom industry for improved applicability and relevance [8]

The investigation into churn prediction within the telecom industry employed machine learning models like Logistic Regression, SVM, Random Forest, and Gradient Boosting. The study unveiled a significant improvement of up to 26.2% in AUC and a 17% increase in F-measure. However, the paper needs more detailed insights into each method's individual impact and practical implementation guidance for industry application [9]. On the other hand, exploring customer churn prediction in the telecommunications sector, this paper utilized data mining techniques, including Decision Trees, Artificial Neural Networks (ANN), K-nearest neighbors (KNN), and Support Vector machines (SVM). The study on an Iranian mobile company's dataset highlighted ANN's superior performance, showcasing over 95% accuracy in Precision and Recall. The proposed hybrid methodology outshined individual classifiers, offering flexibility between Precision and Recall.

Additionally, a novel dimensionality reduction technique identified influential features like frequency of use and total complaints, providing valuable insights for telecom companies. The study's adaptable approach extends beyond telecommunications, showcasing its potential applications in various customer-centric business domains. Despite its promising hybrid methodology, the paper could benefit from more detailed explanations regarding the feature selection and dimensionality reduction methods employed [10].

The paper [11] introduces a machine learning model employing a 2-D convolutional neural network (CNN) to predict churned users in e-businesses. The model exhibits high accuracy (96.3%) and efficient performance, with true-positive and true-negative rates at 95% and 94%, respectively. The study emphasizes comprehensive data preprocessing for CNN success and meticulously addresses the Telco Customer Churn dataset to enhance customer churn prediction.

The telecom industry's competitiveness necessitates accurate churn prediction to retain loyal customers. Another paper advocates an Artificial Neural Network (ANN) model, using diverse customer data like demographics, billing, and usage patterns, achieving a 79% accuracy in predicting telecom churn in Pakistan [12]. ANN results effectively highlight churn factors, facilitating proactive steps to address attrition. The study underscores the importance of timely churn prediction in a rapidly evolving telecom market, emphasizing ANN's superiority over other classification techniques and leveraging backpropagation for model training. Additionally, it assesses variables' impact on telecom churn for a comprehensive understanding.

The paper [13] introduces a hybrid churn prediction model, combining clustering and classification algorithms, exhibiting high accuracy rates of 94.7% on the GitHub dataset and 92.43% on the Bigml dataset, surpassing existing state-of-the-art models. This comprehensive approach addresses the critical issue of identifying dissatisfied customers, enabling companies to mitigate churn factors effectively. However, while the proposed model demonstrates significant advancements, the study could benefit from addressing scalability concerns, considering the potential challenges in applying these models to larger datasets or other industry sectors. For deeper customer satisfaction insights, big data analytics and social network analysis could further enhance the model's efficacy and real-world applicability.

Another study thoroughly analyzes machine learning methods for telecom customer churn prediction [14]. Initially, the comparison involved popular classifiers without boosting, showcasing the top performers as the Backpropagation Network and the Decision Tree, achieving approximately 94% accuracy and 77% F-measure. However, Naïve Bayes and Logistic Regression fell short with around 86% accuracy and considerably lower F-measures. Upon implementing boosting, the classifiers experienced performance enhancement, notably improving accuracy by 1% to 4% and F-measure by 4.5% to 15%. The boosted SVM-POLY with AdaBoost emerged as the best classifier, achieving nearly 97% accuracy and over 84% F-measure. The study's findings underscore the significance of boosting techniques in improving classification performance. Future research aims to delve deeper into parameter simulations for AdaBoost, explore additional boosting algorithms, and utilize more extensive, more comprehensive telecom datasets to bolster the statistical significance of the results. This work offers valuable insights into the efficacy of various machine-learning approaches for churn prediction and sets the stage for further advancements in this domain.

Furthermore, a comprehensive study delves into churn prediction by extensively evaluating supervised machine learning techniques and data sampling methods across diverse datasets[15]. With a focus on the Area Under the Curve (AUC) metric, the research explores the interplay between models like Random Forest, Logistic Regression, Gradient Boosting, and others, alongside various sampling techniques. Nemenyi tests and Correspondence Analysis shed light on associations between algorithms, sampling methods, and dataset characteristics. The study provides an in-depth overview of churn analysis research, encompassing dataset descriptions, oversampling, undersampling, hybrid approaches, and a range of classifiers employed, including Random Forest, Logistic Regression, and Gradient Boosting. It culminates in a practical recommendation: an ensemble-based churn prediction pipeline applicable to different churn-like datasets. Visualizations effectively illustrate relationships between classifiers, sampling methods, and dataset behavior, offering a valuable reference for those navigating machine learning options in churn prediction.

This paper addresses churn prediction in the telecom sector, acknowledging the high cost of acquiring new customers[16]. The proposed model combines classification and clustering techniques to identify churn customers and understand the reasons behind their churn. Utilizing Random Forest for classification achieved an 88.63% correct classification rate, enabling effective retention policy formulation. Beyond classification, the model segments churn customers using cosine similarity for group-based retention offers. It pinpoints churn factors essential for understanding customer behavior, enabling tailored retention strategies. The evaluation demonstrates the model's effectiveness in churn classification and customer profiling based on accuracy, precision, recall, F-measure, and ROC area. The study underscores the significance of churn prediction for telecom companies, offering insights for CRM and decision-makers and proposing future extensions involving advanced learning approaches and behavior pattern analysis for churn customers using Artificial Intelligence techniques.

Customer churn is a pressing issue in the telecommunications sector; the study highlights the need for existing models to swiftly adapt to evolving customer behavior and decisions[17]. An adaptive learning approach employing the Naïve Bayes classifier and Genetic Algorithm-based feature weighting is proposed to address this challenge. Evaluation of public datasets (BigML Telco churn, IBM Telco, and Cell2Cell) showcases significant enhancement in prediction performance compared to various baseline classifiers. Precision, recall rates, F1-score, Matthews Correlation Coefficient (MCC), and accuracy metrics demonstrate notable improvements, indicating the effectiveness of the proposed approach. Given the telecom industry's saturation, predicting and retaining potential churners becomes pivotal for effective relationship management and sustained growth [Adnan Amin].

This study delves into advanced churn prediction using the kernel Support Vector Machines (SVM) algorithm within the telecom domain[18]. Baseline SVM models were initially built to ascertain suitable kernel types, followed by a comparison with other methods. Techniques like Sequential Forward Selection (SFS), Sequential Backward Selection (SBS), and resampling strategies (e.g., SMOTE Tomek, SMOTE ENN) were employed for dimension reduction and handling imbalanced data. Results surpassed previous works, achieving impressive F1-score (99%) and accuracy (98.9%). The objective is to minimize misclassifications in churn prediction, refining kernel SVM models using varied techniques like feature selection, resampling, and hyperparameter tuning. The RBF kernel SVM, enhanced with parameter tuning, SFS/SBS, and SMOTE ENN, showcased superior performance, reaching an accuracy of 99.01% and an

F1 score of 98.88%. Despite this high accuracy, the model still needs to beimal due to reliance on hyperparameters. Future research aims to expand the search space for hyperparameter values and explore alternative algorithms for feature selection and resampling to refine the model further.

This study delves into addressing customer churn, particularly in the telecom sector[19]. It focuses on crafting a churn prediction model crucial for telecom operators to forecast potential churners. Leveraging machine learning on a big data platform, the model achieved an impressive 93.3% AUC, further enhanced to 93.3% with Social Network Analysis (SNA) features. They are tested on a comprehensive nine-month dataset from SyriaTel, the model employed Decision Tree, Random Forest, GBM, and XGBOOST algorithms, with XGBOOST yielding the optimal 93.301% AUC. Overcoming data imbalance, practical feature engineering, and proactive churn prediction underscore its value for telecom companies in revenue maximization. Periodic retraining for non-stationary data remains a consideration for future refinement.

The study explores the intricate landscape of customer churn prediction in the telecom industry, shedding light on its substantial impact on revenue and customer retention[20]. Examining research spanning 2005 to 2020, it covers churn effects, causes, and the techniques used, emphasizing machine learning's role and CNN's potential in feature extraction for larger datasets. While it adeptly discusses various performance measures beyond accuracy, such as confusion matrix and precision, the study could further elaborate on the practical application of these measures. Moreover, it details telecom dataset attributes, underlining the importance of efficiently handling extensive data and the necessity for robust feature extraction methods. Overall, while providing a comprehensive overview, the study could benefit from a more practical demonstration of how these findings can be implemented in real-world scenarios.

The study focuses on developing a robust financial crisis prediction (FCP) model using ant colony optimization (ACO) to select relevant variables effectively and improve classification accuracy[21]. This model involves two key phases: ACO-based feature selection (ACO-FS) and ACO-based data classification (ACO-DC). Notably, the proposed ACO-FCP model outperforms existing methods in benchmark datasets, showcasing its superiority in feature selection and classification. However, while the study highlights the significant improvements in accuracy, sensitivity, specificity, and F-score achieved by the ACO-FCP model, it could elaborate more on these findings' practical implications and real-world applications. Additionally, the study could benefit from discussing potential limitations or challenges in deploying the ACO-FCP model in diverse financial institutions, offering a more comprehensive view of future implementations.

The study uses Fisher discriminant equations and logistic regression to predict customer churn in the telecom sector[22]. It achieves a high prediction accuracy of 93.94%, aiding telecom companies in foreseeing and addressing customer churn. However, it could diversify its modeling techniques for a more comprehensive approach and offer deeper insights into influential variables for improved practical application. Additionally, exploring practical constraints in implementing recommendations and discussing the complexities of gauging customer opinions could enhance the study's real-world applicability.

This study introduces a novel Customer Churn Prediction (CCP) approach leveraging a distance factor to estimate classifier certainty[23]. It partitions data into high and low certainty zones, revealing a strong correlation between the distance factor and classifier accuracy. Notably, it showcases higher accuracy in

zones with more significant distance factors across various publicly available datasets from the Telecommunication Industry (TCI). The study pioneers this CCP model based on the distance factor, shedding light on its impact on classifier certainty and presenting valuable insights for churn prediction in TCI.

However, while demonstrating promising results, the study could benefit from more comprehensive evaluations across balanced datasets and exploring diverse models. Furthermore, a more detailed statistical comparison among various model results might provide a deeper understanding of the performance of this novel approach. Moreover, incorporating adaptations for critical node identification in social media or other domains could extend the application scope of this technique, enhancing its versatility.

The paper introduces a novel set of features for land-line customer churn prediction in telecommunication services, encompassing various data types like Henley segmentation, call details, demographic profiles, and service-related information [24]. Seven prediction techniques were employed to assess these features, revealing their superiority over existing sets in predicting customer churn. The experiments offered insights into the effectiveness of different modeling techniques and the comparative advantages of the new feature set.

Despite its contributions, the study acknowledges several limitations. It suggests future improvements by incorporating additional information like complaint and contract details to enhance the feature set. Furthermore, the study recognizes the need to address dimensionality issues by exploring feature selection and extraction methods for more refined inputs. Additionally, the paper acknowledges the imbalance classification problem and suggests exploring alternative methods beyond sampling techniques to tackle this challenge in future research. This self-awareness of limitations sets a clear path for further advancements and strengthens the study's credibility.

The paper introduces a groundbreaking approach to predicting daily customer churn in mobile telecom operators, considering the dynamic shifts in customer behavior. It proposes distinct models for daily churn prediction[25]. One model leverages Recency, Frequency, and Monetary value (RFM) functions applied to the multivariate time series to extract meaningful features and utilizes a Random Forest classifier for predictions. Another model extracts statistical features from the multivariate time series and feeds them into a traditional machine learning model, such as Random Forest, for churn predictions. Additionally, the study introduces deep learning models, including Long Short-Term Memory (LSTM) networks that learn representative features from the multivariate time series and Convolutional Neural Networks (CNN) that automatically extract features from the data to predict churn. The study highlights the superiority of daily models over traditional monthly predictions in detecting potential churners earlier and more accurately. However, it might benefit from providing comparative statistics or benchmarks to quantify the significance of this improvement. Furthermore, the paper acknowledges the need for future research, suggesting incorporating diverse features and causality-driven models to enhance predictive power and refine targeting strategies based on customer responses, opening promising avenues in churn prediction research.

The study introduces a novel approach, CCPBI-TAMO, blending text analytics with metaheuristic optimization for Customer Churn Prediction (CCP) in business intelligence[26]. Leveraging the Chaotic

Pigeon Inspired Optimization for feature selection, the CPIO-FS technique efficiently reduces computational complexity. Furthermore, it integrates a Long Short-Term Memory (LSTM) model with a Stacked Auto Encoder (SAE) to classify the reduced features, combining SAE's compact feature detection with LSTM's classification abilities. To refine performance, the model undergoes Sunflower Optimization (SFO) for hyperparameter tuning. The comprehensive analysis of benchmark datasets underscores the superior performance of this model, achieving remarkable accuracy rates of 95.56%, 93.44%, and 92.74% on datasets 1-3, respectively. However, despite showcasing impressive accuracy, providing further insights or comparative metrics could enhance the study's depth and value in understanding the model's advantages in different scenarios or against varying datasets.

This paper introduces an integrated framework that merges churn prediction and customer segmentation in the telecom industry[27]. The framework offers a comprehensive churn analysis using six components - data pre-processing, exploratory data analysis (EDA), churn prediction, factor analysis, customer segmentation, and behavior analytics. Experiments conducted on three datasets using six machine learning classifiers revealed that AdaBoost performed best in Dataset 1, achieving an accuracy of 77.19% and an F1-score of 63.11%. Random Forest led in Dataset 2 with an accuracy of 93.6% and an F1-score of 77.20%. In Dataset 3, Multi-layer Perceptron showed the highest F1-score at 42.84%. Bayesian Logistic Regression was subsequently employed for factor analysis, while K-means clustering segmented churn customers into distinct groups, aiding in targeted retention strategies. The research's contributions lie in seamlessly connecting churn prediction and segmentation, leveraging Bayesian Analysis, providing operators with cluster-wise churning probabilities, and fostering a deeper understanding of customer groups.

Exploring various oversampling and undersampling methods beyond SMOTE could enhance dataset balance. Additionally, delving into ROC analysis to determine optimal churn prediction thresholds and employing Bayesian optimization for hyperparameter tuning could refine the accuracy of models using SMOTE. These future directions can fortify the framework's predictive capabilities, enabling more nuanced customer churn predictions and strategic decision-making in the telecom sector.

Finally, here is the detailed literature survey for the Telecom Churn Prediction of all these years from 2023 to 2012. The features included are Year, Title, Author, Methods, Dataset Used, Performance measure and Key points.

	A	B	C	D
1	year	title	author	methods
2	2023	A review on customer segmentation methods for personalized customer targeting in e-commerce use cases	Miguel A	commonly used methods for customer representation (feature selection, RFM) and segmentation (k-means).
3	2022	Anovel customer churnprediction model for the telecommunication industry using data transformation methods and feature selection	JoydebK	Various machine learning models: Not explicitly mentioned in the excerpt.
4	2022	Customer Churn Prediction in Telecommunication Industry Using Deep Learning	Samah W	Model: Deep-BP-ANN with specific configurations. Feature selection: Variance Thresholding and Lasso Regression compared
5	2022	Asurveyonmachinelearningmethodsforchurnprediction	Louis Ge	11 supervised and semi-supervised algorithms
6	2022	Churnprediction in telecommunication industry using kernel Support Vector Machines	Nguyenl	Kernel SVM with different kernel types evaluated
7	2022	An ensemble based approach using a combination of clustering and classification	Syed Fal	K-means, K-medoids, X-means, and random clustering to group customers
8	2022	AData-Driven Approach to Improve Customer Churn Prediction Based on Telecom Customer Segmentation	Tianyua	Fisher discriminant equations: A statistical technique for finding linear combinations of features that best separate different groups (churners and non-churners). Logistic regression analysis: A statistical technique for modeling the probability of an event (churn) based on independent variables (customer features).
9	2021	Dynamic customer churn prediction strategy for business intelligence using text analytics with evolutionary optimization algorithms	Irina V. Pustok hina C. Jeyalak	CCPBI-TAMO: A novel metaheuristic optimization algorithm, CPIO-FS: Feature selection technique inspired by chaotic pigeon optimization. LSTM-SAE: Combines LSTM's sequence learning capabilities with SAE's feature extraction strengths. SFO: Sunflower optimization
10	2021	Customerchurnprediction system: a machinelearning approach	Praveen	Data Preprocessing, ISMOTE: balance the dataset OWELM: Classification algorithm for CP

	E	F	G
1	data set	performance	key points
2	The paper reviews studies with various datasets	It focuses on identifying commonly used methods and their trends rather than comparing their effectiveness.	Review of customer segmentation methods: The paper provides a comprehensive overvi segmentation methods
3	Publicly available TCI datasets	Up to 26.2% increase in AUC (Area Under ROC Curve). Up to 17% increase in F-measure.	This paper highlights the potential of combining data transformation, feature selection, hyperparameter
4	IBM telco abd cell2cell	high accuracy for the Deep-BP-ANN model	This paper presents a Deep-BP-ANN approach for customer churn prediction with feature overfitting prevention, and data balancing strategies.
5	16 publicly available churn-like datasets with dive	Performance heavily depends on specific data characteristics and no single method/sampling combination works best across all datasets.	Extensive evaluation on a variety of datasets and methods.
6	Information about the telecom company dataset	F1-score and accuracy of 99%and 98.9%respectively.	Kernel SVM: Powerful machine learning algorithm for classification. Feature selection: SFS and SBS for identifying important features. Data balancing: SMOTETomek and SMOTEENN to address imbalanced dataset.
7	Two publicly available telecom datasets from Gith	achieve 94.7% accuracy on GitHub dataset and 92.43% on Bigml dataset	Ensemble classification: Combining clusters with seven different classification algorithr
8	Collected from three major Chinese telecom comp	Logistic regression achieved higher prediction accuracy (93.94%) compared to Fisher discriminant equations.	Uses real-world data from major Chinese telecom companies. Compares two different modeling techniques and finds the most effective one. Highlights the potential for increased profits through customer retention.
9	benchmark customer churn pre diction dataset	maximum accuracy of 95.56%, 93.44%, and 92.74% on the applied dataset 1-3 respectively	The paper emphasizes the importance of innovative customer churn prediction (CCP) for
10	Three unspecified telecom datasets are used, representing different customer base and network characteristics.	Accuracy of 0.94, 0.92, and 0.909 on the three datasets,	a new CCP model "ISMOTE-OWELM" for telecom data that combines: Improved SMOTE (ISMOTE): A data balancing technique for handling imbalanced dataset inactive than active churners). Optimal Weighted Extreme Machine Learning (OWELM): A classification algorithm for chu

	A	B	C	D
11	2020	Predictive analytics using big data for increased customer loyalty: Syriatel Telecom Company case study	Wissam	Time-frequency-monetary (TFM)
12	2020	Optimum profit-driven churn decision making: innovative artificial neural networks in telecom industry	Ruholla	ChP-SOEDNN combines: Artificial Neural Networks (ANNs) for self-organizing and error-driven learning. Clustering to differentiate individual customer behavior. Misclassification cost analysis for profit-
13	2019	Customer churn prediction in telecom using machine learning in big data platform	Abdelra	Decision Tree, Random Forest, Gradient Boosted Machine Tree "GBM" and Extreme Gradient Boosting
14	2019	Customers Churn Prediction using Artificial Neural Networks (ANN) in Telecom Ind	Yasser K	Artificial Neural Network (ANN) model used
15	2019	A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector	IRFAN U	Random Forest (RF) algorithm, K-means clustering
16	2017	Intelligent churn prediction for telecom using GP-AdaBoost learning and PSO undersampling	Adnan I	ChP-GPAB system: Combines GP, AdaBoost, and PSO
17	2015	A comparison of machine learning techniques for customer churn prediction	T. Vafei	decision trees, random forests, logistic regression, etc
18	2014	Improved churn prediction in telecommunication industry using data mining techniques	A. Keran	Decision Tree, Artificial Neural Networks (ANN), K-Nearest Neighbors (KNN), Support Vector Machine (
19	2012	Genetic Programming and Adaboosting based churn prediction for Telecom	Adnan I	Genetic programming (GP),Adaboost style boosting.
20	2020	Dynamic behavior based churn prediction in mobile telecom	Nadia A	Four prediction models compared: RFM-based: Uses Recency, Frequency, and Monetary Value features extracted from time series. Statistics-based: Analyzes statistical properties of time series data. LSTM-based: Deep learning model using Long Short-Term Memory networks for automatic feature extraction. CNN-based: Deep learning model using Convolutional Neural Networks for automatic feature

	E	F	G
11	SyriaTel	The paper does not explicitly mention specific performance metrics like accuracy, precision, recall, etc., for the classification models.	Focus on different-value customers: The paper proposes a methodology for telecom com segment customers based on the "Time-frequency-monetary (TFM)" approach and target them with appropriate offers and TFM metric: This new metric combines customer behavior across time, frequency of servi and monetary value generated to segment customers into meaningful groups.
12	not mentioned	Capable of devising cost-efficient retention strategies for individual customer clusters	Individuality of customers: Captures unique characteristics beyond hidden patterns.
13	SyriaTel	AUC 93.3%	The use of SNA enhanced the performance of the model from 84 to 93.3% against AUC sta
14	Telecom company dataset including demographic data,	accuracy of 79% for predicting customer churn in Pakistan	Using ANN as a modeling approach is not unique, but its application to Pakistan's telec
15	not mentioned	88.63% accuracy	Identifies churn customers and provides group-based retention offers. Discovers churn factors for targeted retention strategies and improved marketing campa
16	cell2cell & orange	AUC values (0.91 and 0.86) are quite high, indicating strong performance on both datasets	Develops a new customer churn prediction system (ChP-GPAB) by combining: Genetic Programming (GP): To search for effective churn prediction features. AdaBoost: To enhance classification accuracy. Particle Swarm Optimization (PSO): To add
17	Public domain dataset	highest accuracy (nearly 97%) and F-measure (over 84%).	Boosted models demonstrated clear superiority in prediction accuracy and F-measure cc
18	Customer data from an Iranian mobile company	exceeding 95% for both recall and precision.	Hybrid methodology: Based on insights from the comparisons, the paper proposes a hyb methodology
19	Orange Telecom,cell2cell.	Used 10 fold cross validation,0.89 score of AUC	Adaboost style boosting is used to evolve a number of programs per class.
20	150-day customer data from MTN operator in a spe	All three models (LSTM, CNN, RFM) outperform the statistics-based model.	Monthly churn prediction misses valuable timing information: Existing churn models often predict monthly, ignoring daily changes in customer behavic up to churn.

III. Contributions

Members	Implementation	Conference paper	Report	Contributions
Fiza Maqsood	✓		✓	Decision Tree, Random Forest, Shallowed Random Forest, Write up of Report
Hira Absar Khan	✓	✓		Random Forest, Conference paper write up
Hadeesa Muskan			✓	Literature review, Write up

IV. Methodology

Prediction of churners in the telecom industry is a binary classification problem as there are churners and non-churners class distribution. It is crucial for the companies to retain customers and maintaining business stability. This methodology aims to forecast churn by employing the Decision Trees, Random Forest, Balanced Random Forest and Shallowed Random Forest on two distinct datasets: Cell2Cell and Orange. For enhancing telecom service providers' operational efficiency and customer retention strategies its crucial to predict churners at right time. This methodology provides a comprehensive approach to predict customer churn in the telecom industry. This methodology involves employing tree-based classifiers machine learning techniques to anticipate customer churn, a critical telecom industry concern.

In this methodology we are relaying on Tree-based classifiers that are broadly used in classification problems due to their ability to handle complex relationships within dataset therefore they can provide insights into feature importance. These classifiers include Decision Trees and their ensemble methods like Random Forests, construct hierarchical structures to segment the feature space. Decision Trees are the fundamental unit, partition the dataset based on feature thresholds, creating a tree-like structure that facilitates intuitive decision-making. However, their susceptibility to overfitting is mitigated by strategies like pruning or limiting tree depth. On the other hand, Random Forests are derived from Decision trees so they serve as an ensemble of Decision Trees, offer superior performance by creating random subsets of the dataset and features and then constructing multiple trees on subsets. This technique aggregates predictions from individual trees, providing more accurate and stable results. Tree-based models excel in handling high-dimensional data, assessing feature importance, making them one of ideal choice for telecom churn prediction tasks where interpretability and predictive accuracy are paramount.

Decision Trees is one of foundational classification models used extensively in prediction of telecom customer churners because of their interpretability and simple implementation. The model of Decision tree basically performs segmentation of telecom dataset into a tree-like structure, after that it can

intuitively make decisions based on various features. However, their susceptibility to overfitting, creating excessively complex trees can be resolve by using strategies like tree pruning or restraining the tree's depth to enhance reliability in churn prediction tasks.

In the context of telecom churn prediction, Random Forests emerge as robust ensemble models that somehow uses the concept of Decision trees and is built on multiple Decision Trees. The effectiveness of Random Forest lies in dealing and handling complex relationships between the features of telecom datasets. They address complexity of dataset by constructing multiple trees where each tree is trained on random subsets of the dataset and features. This randomization enables to generate a diverse set of trees, reducing overfitting and enhancing the model's ability to generalize to unseen data.

Balanced Random Forests is further an extension of the Random Forest technique, specifically customized to address the challenges of telecom datasets. These models preserve a balanced representation of churners and non-churners within each tree during creation of different trees on randomized data. By addressing this Balanced Random Forests significantly improve predictive performance for accurate prediction. It enables to capture essential patterns from both classes more effectually.

Moreover, Shallowed Random Forests is also an optimized variant of Random Forests which restricts the depth of each generating tree to mitigate overfitting issue. This adjustment of depth aims to strike an equilibrium between extracting essential patterns within the dataset and preventing excessive complexity. As telecom datasets hold huge feature space so Shallowed Random Forests are particularly beneficial in scenarios involving large feature spaces or datasets, offering improved computational efficiency without compromising predictive accuracy.

These Models allows telecom companies for customer churn prediction effectively. Leveraging the strengths of Decision Trees, Random Forests, Balanced Random Forests, and Shallowed Random Forests empower companies to gain insights into customer's behavior, optimize retention strategies, and make significant business decisions.

The block diagram given in *Figure 1* shows the proposed churn prediction system.

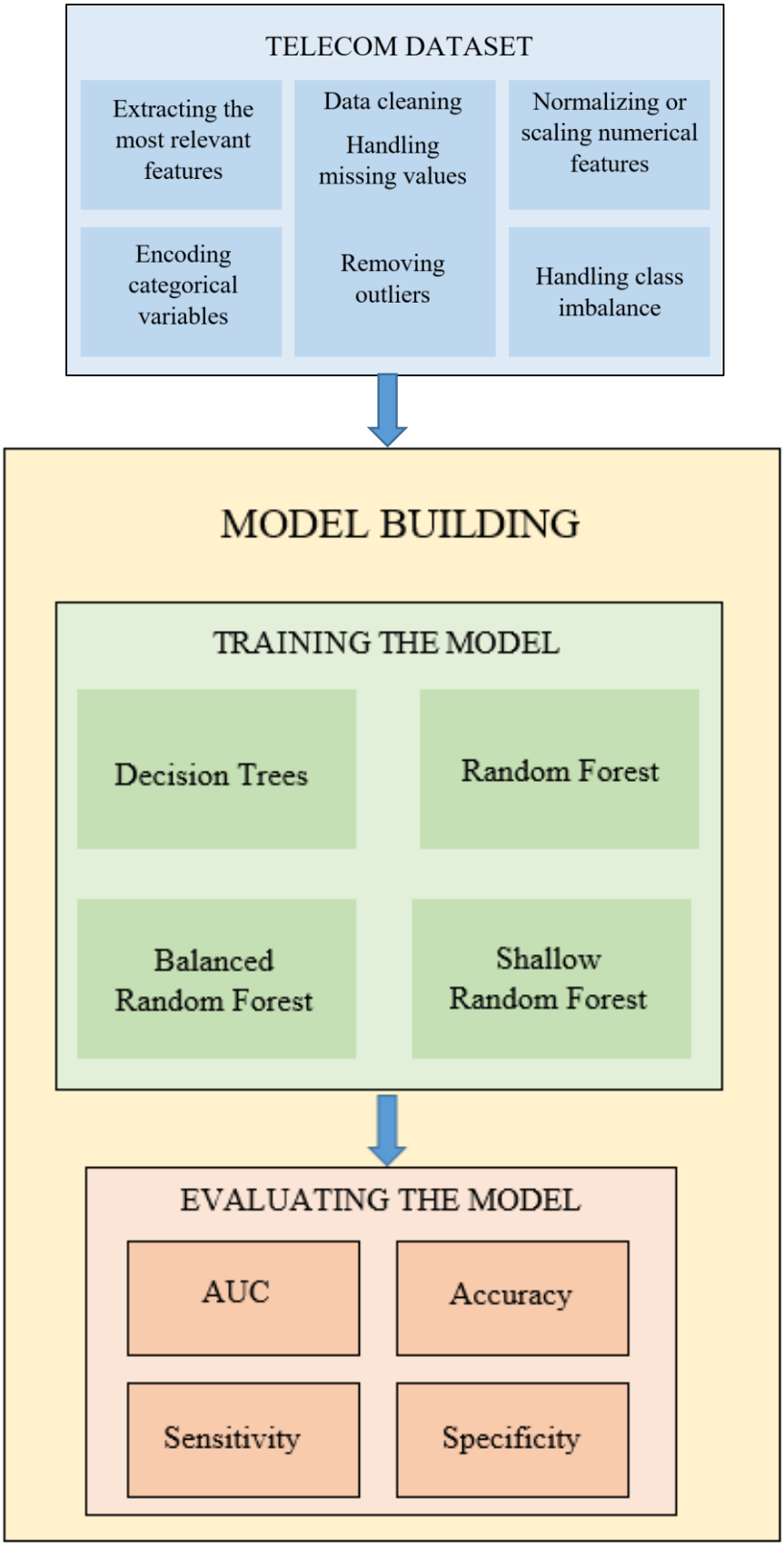


Figure 1: methodology

The methodology for classification of churners in telecom datasets encompassed several sequential steps given below:

1. Data cleaning**:** Telecom dataset contained large number of missing values that were identify and address within the dataset by using strategies like imputation where missing values are filled with mean, median, or mode or deletion of records/attributes might be applied if there is not large percentage of data loss by deleting those instances. Furthermore, both datasets contained Outliers. Outliers are data points that usually differ from the majority of the data in a dataset. They are extreme values that fall far outside the from the range or distribution of the remaining of the data. Identifying outliers and handling them is crucial in prediction process of churners is because they can skew statistical analyses and machine learning models, leading to inaccurate classification of churners and non-churners. To handle the Outliers techniques such as statistical methods or clustering to detect outliers and decide on appropriate actions (e.g., removing outliers or transforming their values) were used.

2. Categorical Data Encoding**:** The models we used uses numerical values so it was necessary to convert the categorical features into numerical. This conversion could be attain using many techniques one of the most used techniques is one-hot encoding that is used in this methodology. Label encoding, or target encoding techniques can also use for data encoding purpose. This transformation ensures compatibility with machine learning algorithms.

3. Handling Class Imbalance**:** Due to the nature of Telecom datasets, where class distribution is imbalanced, techniques like oversampling the minority class (churners) or under sampling the majority class (non-churners) may be applied to achieve a balanced representation of classes because the models can suffer while training. So, in this methodology we experiment with sampling methods like SMOTE (Synthetic Minority Over-sampling Technique) class to balance the dataset while ensuring no loss of critical information.

4. Train-Test Split: Before model training, split the datasets into training and testing sets with 70% for training purpose and 30% for testing purpose. This step ensures the model's performance is evaluated on unseen data, preventing overfitting and allowing for accurate assessment of its predictive abilities.

5. Model Training and Testing**:** This stage involves the actual training of the models on the training dataset and subsequently testing their performance on the testing dataset. All four tree-based classifiers were applied to build and assess the models' predictive capabilities using 10-fold cross validation.

6. Performance Evaluation**:** Assessing the models' performance metrics like accuracy, AUC (Area Under the ROC Curve), specificity and sensitivity were evaluated to determines model's effectiveness in predicting telecom churn. Evaluation metrics help gauge how well the models differentiate between churners and non-churners, guiding the selection of the best-performing model for deployment.

a. Decision Trees

Decision Trees are hierarchical structures used in machine learning for classification and regression tasks. These trees recursively partition the dataset into smaller subsets based on the values of input features. At each node of the tree, a decision is made regarding which feature and value will best split the data, aiming to maximize the homogeneity of resulting subsets concerning the target variable (classification label or regression value).

The tree-building process starts at the root node, representing the entire dataset. The feature that best separates the data is chosen to create child nodes. This process continues iteratively, branching out and creating new nodes until certain stopping criteria are met, such as reaching a specified maximum depth, achieving a minimum number of samples in a leaf node, or when no further improvement in homogeneity is possible.

Decision Trees offer transparency and interpretability, allowing users to visualize the decision-making process. Each split in the tree represents a decision point based on a specific feature, making it straightforward to understand the logic behind the model's predictions. However, they are susceptible to overfitting, especially when allowed to grow deep or when dealing with noisy data. Deep trees tend to capture noise and intricacies present in the training data, impacting their ability to generalize to unseen examples.

Strategies like pruning, limiting tree depth, or setting minimum sample sizes in leaf nodes help mitigate overfitting. Pruning involves removing branches that do not significantly improve the tree's predictive accuracy on validation data, leading to simpler and more generalized trees. Balancing between complexity and predictive performance is crucial in building decision trees that generalize well to unseen instances while effectively capturing essential patterns in the data.

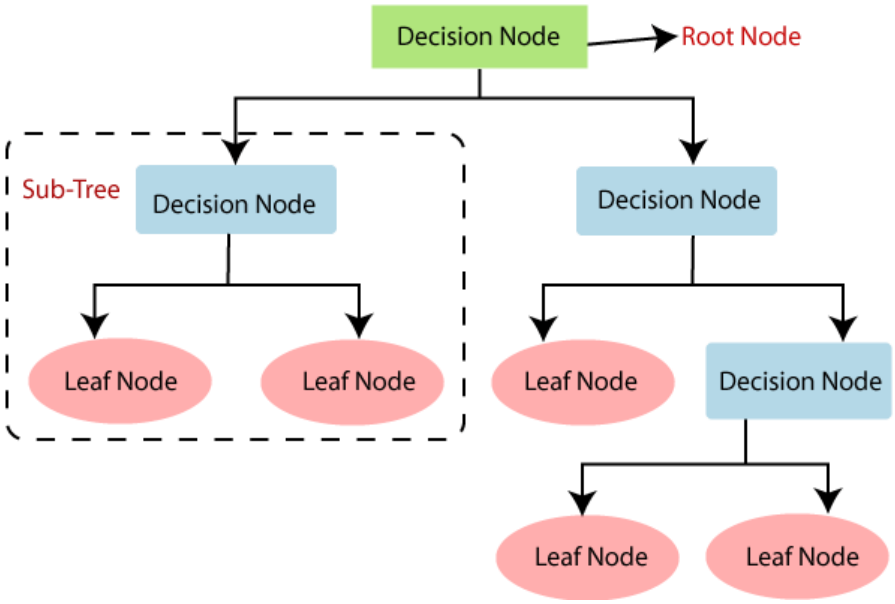
Sure, the ID3 algorithm is one of the fundamental algorithms used for decision tree learning. When applied to predict telecom customer churn, it follows these basic steps:

1. Data Collection and Preparation: Gather relevant data on telecom customers, including demographics, usage patterns, service subscriptions, billing information, customer service interactions, etc. Cleanse and preprocess the data to handle missing values, outliers, and ensure it's ready for analysis.
2. Attribute Selection**:** Identify the attributes/features that could potentially influence customer churn, such as call duration, plan type, contract length, customer tenure, satisfaction scores, etc.

These attributes should be chosen based on their relevance and potential impact on predicting churn.

3. Entropy Calculation: Calculate the entropy of the target variable (churn) and its corresponding attribute values. Entropy measures the impurity or randomness in the data. The ID3 algorithm uses entropy to determine the best attribute at each node to split the data.
4. Attribute Selection using Information Gain: Calculate the information gain for each attribute. Information gain measures the effectiveness of an attribute in classifying the data. Attributes with higher information gain are more valuable for predicting churn. The attribute with the highest information gain becomes the root node of the decision tree.
5. Splitting Data: Split the dataset based on the selected attribute. Each branch represents a different value of the selected attribute.
6. Recursive Splitting: Repeat the process recursively for each branch by considering subsets of data. At each node, select the best attribute to split the data until a stopping criterion is met. This could be a predefined tree depth, a minimum number of samples in a node, or other criteria to prevent overfitting.
7. Tree Pruning (Optional): After building the tree, prune unnecessary branches to avoid overfitting. Pruning involves removing branches that provide little predictive power or could potentially cause overfitting on the training data.
8. Prediction: Once the decision tree is constructed, use it to predict customer churn by inputting new data through the tree. The path taken through the tree's nodes will lead to a predicted outcome (churn or non-churn) for a given customer.
9. Model Evaluation: Assess the performance of the decision tree model using evaluation metrics such as accuracy, precision, recall, F1-score, or ROC curve to determine how well it predicts churn on unseen data.

The ID3 algorithm iteratively creates a tree by selecting the best attribute at each node based on information gain, aiming to classify telecom customers into churn or non-churn categories accurately.



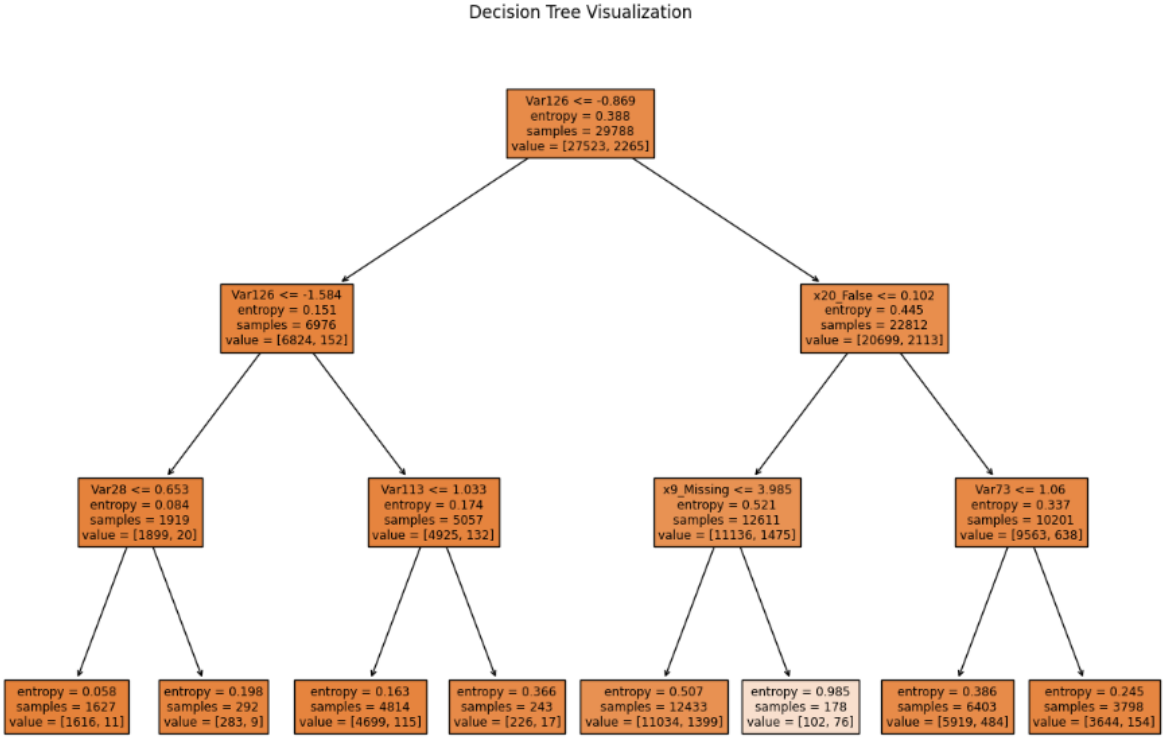


Figure 2: Decision Tree visualization for orange dataset with depth 3

b. Random Forest:

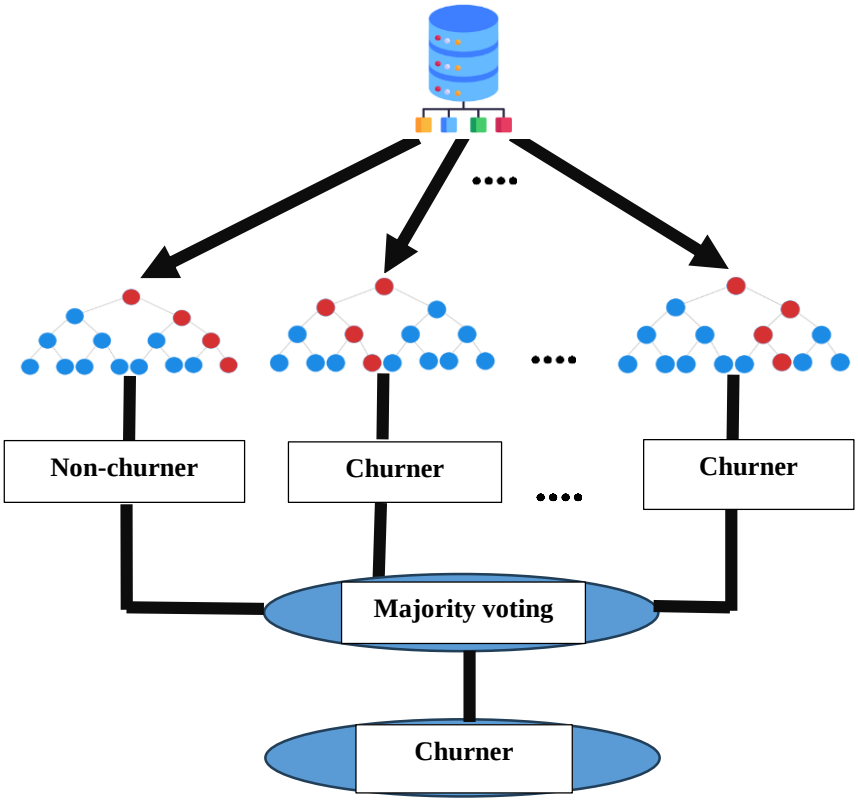
Random Forest is an ensemble learning method that leverages the power of multiple decision trees to improve predictive accuracy and reduce overfitting. It operates by creating an ensemble or a collection of decision trees, each trained on a different subset of the dataset and using a different set of features.

The key idea behind Random Forest is twofold: randomness and aggregation. The randomness aspect involves creating subsets of both instances (bootstrap sampling) and features (feature subsampling) from the original dataset. This randomness injects diversity into each tree's training process, ensuring that individual trees capture different aspects and patterns within the data.

During training, each decision tree in the forest is constructed independently but based on these randomly selected subsets. These trees grow by recursively splitting the data according to the best features available within the subsets, similar to the process in traditional decision trees.

Once the trees are built, predictions are made by each tree independently for a new instance. In classification tasks, for example, each tree "votes" on the class label. The final prediction is determined by aggregating the votes across all trees, often using a majority voting scheme. For regression tasks, predictions from all trees are averaged to produce the final prediction.

Random Forest mitigates overfitting issues commonly associated with individual decision trees by combining predictions from multiple trees. It provides better generalization by capturing various patterns present in the dataset while minimizing the impact of noise or outliers in individual trees. The ensemble nature of Random Forest improves robustness and predictive accuracy, making it a versatile and widely used machine learning algorithm across various domains.



c. Balanced Random Forest (BRF):

BRF employs a balanced sampling strategy which serves enormous datasets like telecom to classify the area of interest with more accurately. BRF randomly selects a subset of instances called bootstrap from the majority class while retaining all instances from the minority class.

BRF generates multiple balanced subsets of data by combining the selected instances from the majority class with all instances from the minority class. So, each subset contains an equivalent number of instances of both classes, ensuring a balanced representation.

Decision trees are constructed from these balanced subsets. Each tree is generated on these balanced subsets, adding variety to the ensemble.

BRF builds ensemble of decision trees, these trees can be range from hundreds to thousands, each of the Decision tree is trained on a different balanced subset that were generated. During training, each tree "votes" on the class label for a new instance.

When making predictions for a new instance, Balanced Random Forest sums the individual predictions from all trees in the ensemble. The final prediction that either customer will be churner or non-churner is determined by majority voting among the trees. BRF handle predictions of target variable by weighing the votes. Trees that give more accurate results or contribute better towards predicting the minority class) are given more weights in contrast to other to counter the imbalance and improve overall performance. Further, the performance measures for Balanced Random Forest are typically evaluated using metrics like Specificity, Sensitivity, F1-score, and area under the ROC curve (AUC-ROC).

d. Shallowed Random Forest (SRF):

Shallowed Random Forest (SRF) is a variant of the traditional Random Forest algorithm that usually concentrates on restriction of depth of individual decision trees within the ensemble.

Unlike traditional Random Forests that enable trees to grow deeper, it put a limitation on maximum depth of every decision trees. SRF tries to prevent the overfitting by striking a constraint on the tree depth and create shallower trees. SRF also constructs an ensemble of decision trees same like in Traditional Random Forest. However, during the construction of each tree, SRF chooses a randomized subset of features and instances from the dataset to create bootstrap data, ensuring diversity among trees. The decision trees in SRF are constrained to grow only up to a certain depth so by this it allows to captures more generalized patterns in the data and the trees from becoming overly complex, reducing the risk of overfitting. Like BRF and RF, SRF also builds multiple decision trees, each trained on a different randomly chosen subset of the data with restricted tree depth.

For the predictions of class label SRF aggregated the predications of each generated shallow trees in ensemble and then the final prediction is typically determined by voting for majority class. By directing the depth of the trees, it aims to prevent each tree from fitting too closely to the training data. Shallower trees are less prone to overfitting so they capture more generalizable patter. The performance of Shallowed Random Forest is evaluated using standard evaluation metrics like Specificity, Sensitivity, F1-score, and area under the ROC curve (AUC-ROC).

Table shows Parameter used for each model:

TABLE I. Parameters for tree based models

Models	n_estimators	random_states	max_depth
Decision tree	None	42	none
Random Forest	400	42	none
Balanced Random Forest	400	42	20
Shallowed Random Forest	400	42	10

V. Data Analysis

The "cell2cell" an open-source dataset seems to be an extensive compilation of telecom customer-level data. It contains a vast amount of data—99,999 instances and 226 features—that is essential for comprehending the behavior and preferences of customers. The information appears to contain a variety of elements, including call logs, consumption trends for data usage on 2G and 3G networks, roaming behavior, and other relevant telecom service indicators for a range of months. This abundance of data appears to offer a fertile ground for investigation and understanding of customer churn prediction—a crucial component for telecom firms hoping to hold onto their most valuable clients. Comprehending the patterns and signs included in the data could be crucial in creating predictive models that identify consumers who are susceptible to churn.

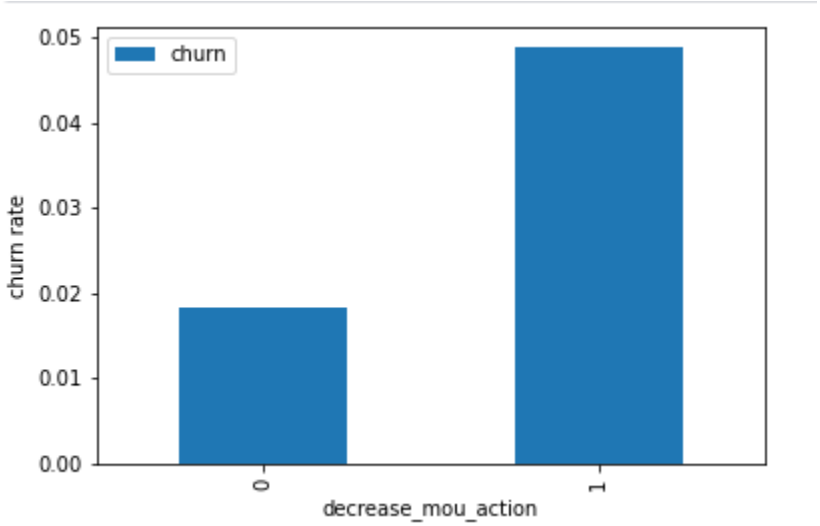
The other "Orange" dataset offers deep insights into telecom business consumer behavior. Predicting customer churn—the possibility that a client would leave the service—is the main goal of the analysis, which uses 230 variables and 44461 instances to capture a wide range of consumer interactions and preferences. After preprocessing and feature selection, this dataset revealed correlations that inform customer retention strategies for high-risk clients. Telecom firms must have a thorough understanding of turnover in order to customize targeted approaches and use predictive analytics to engage and retain their client base proactively.

TABLE II. Characteristics of used telecom datasets

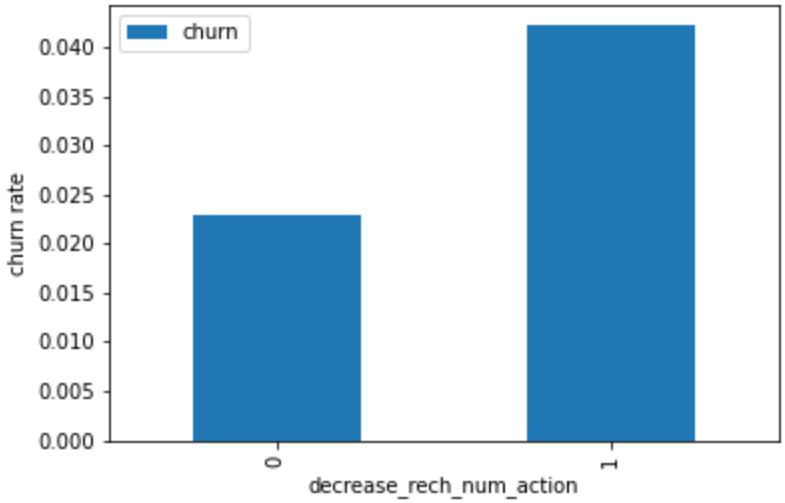
Characteristics	Cell2Cell	Orange
Total features	226	244
Total instances	99999	44461
Data distribution	Imbalanced	Imbalanced
No: of churners	7132	3380
No: of Non churners	62867	41081

a. Univariate analysis

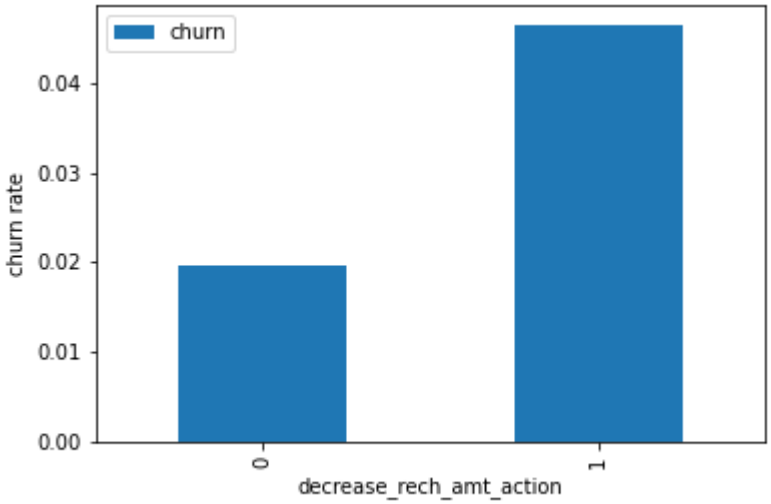
Univariate analysis refers to the statistical analysis that deals with the examination of a single variable in isolation. It helps in understanding the characteristics and properties of that specific variable without considering any relationship with other variables. Below is the univariate analysis of cell2cell. Churn rate on the basis whether the customer decreased her/his MOU in action month examining and analyzing a single variable in isolation to understand its characteristics, distribution, and behavior within a dataset.



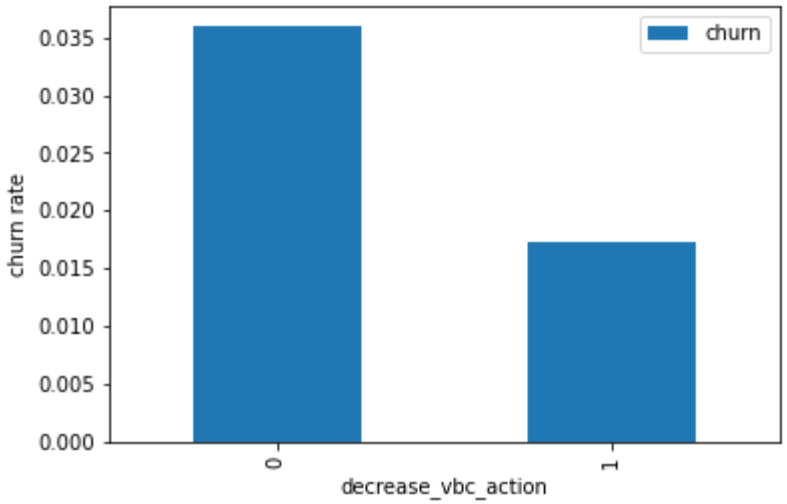
We can see that the churn rate is more for the customers, whose minutes of usage(mou) decreased in the action phase than the good phase. Churn rate on the basis whether the customer decreased her/his number of re charge in action month



As expected, the churn rate is more for the customers, whose number of re charge in the action phase is lesser than the number in good phase. Churn rate on the basis whether the customer decreased her/his amount of recharge in action month.

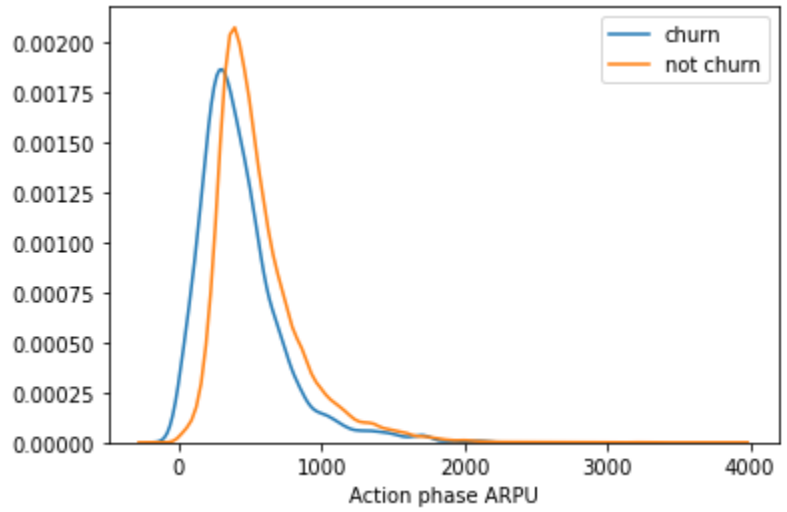


Here also we see the same behavior. The churn rate is more for the customers, whose amount of recharge in the action phase is lesser than the amount in good phase. Churn rate on the basis whether the customer decreased her/his volume-based cost in action month

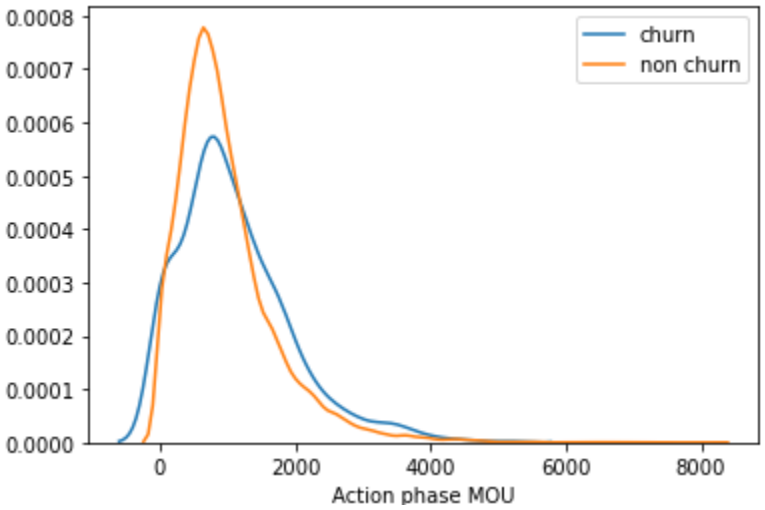


Here we see the expected result. The churn rate is more for the customers, whose volume-based cost in action month is increased. That means the customers do not do the monthly recharge more when they are in the action phase.

Below is analysis of the average revenue per customer (churn and not churn) in the action phase



Average revenue per user (ARPU) for the churned customers is mostly densed on the 0 to 900. The higher ARPU customers are less likely to be churned. ARPU for the not churned customers is mostly densed on the 0 to 1000. Analysis of the minutes of usage MOU (churn and not churn) in the action phase.

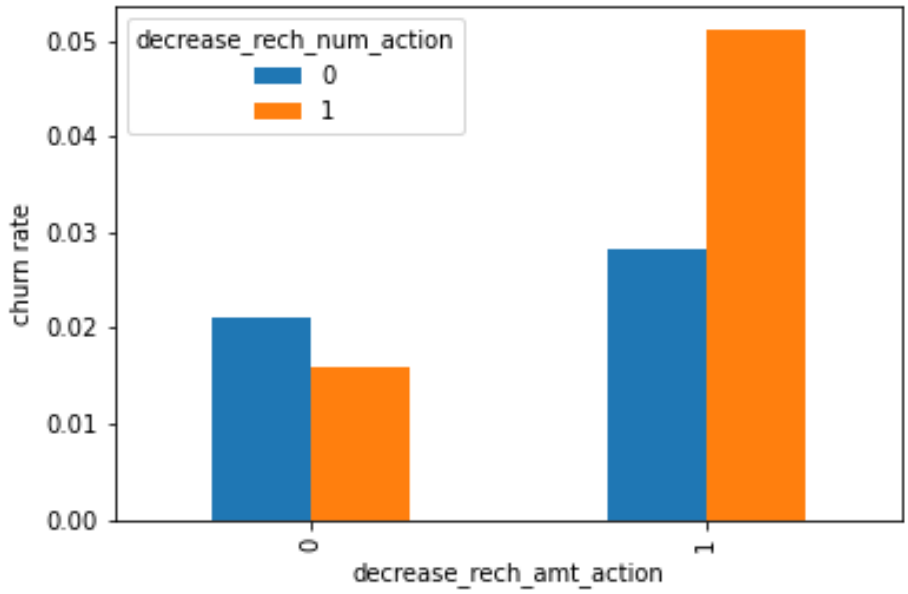


Minutes of usage (MOU) of the churn customers is mostly populated on the 0 to 2500 range. Higher the MOU, lesser the churn probability.

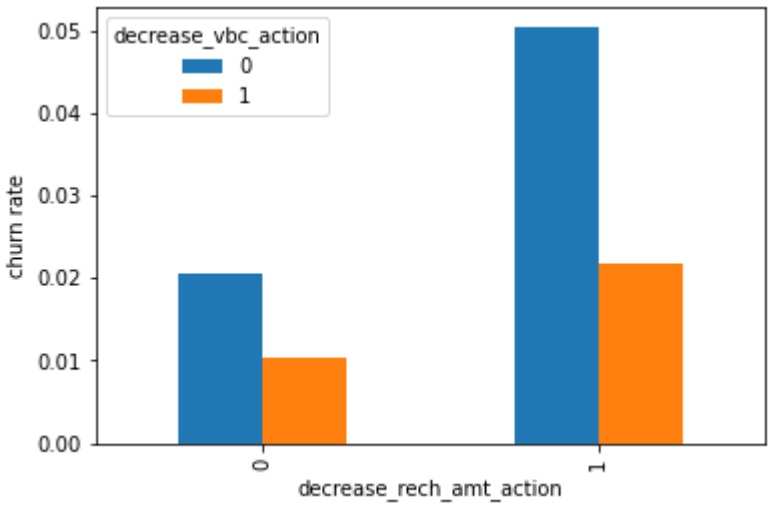
b. Bivariate analysis:

Bivariate analysis involves the simultaneous analysis of two variables to determine if there is any relationship between them. Unlike univariate analysis, which focuses on a single variable, bivariate analysis explores the connection, association, or correlation between two variables. In the implementation for data analysis of variables bivariate analysis on cell2cell was performed.

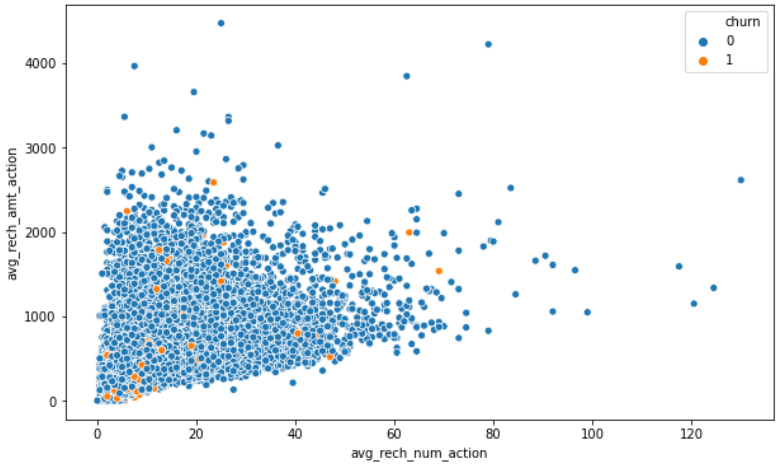
Analysis of churn rate by the decreasing recharge amount and number of re charge in the action phase



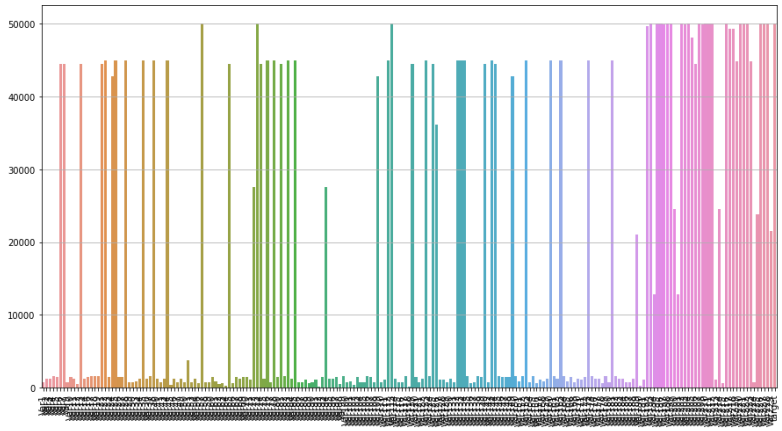
We can see from the above plot, that the churn rate is more for the customers, whose recharge amount as well as number of recharge have decreased in the action phase than the good phase. Analysis of churn rate by the decreasing recharge amount and volume based cost in the action phase.



Here, also we can see that the churn rate is more for the customers, whose recharge amount is decreased along with the volume-based cost is increased in the action month. Analysis of recharge amount and number of re charge in action month.



We can see from the above pattern that the recharge number and the recharge amount are mostly proportional. More the number of re charge, more the amount of the re charge. Dropping few derived columns, which are not required in further analysis



This graph shows about orange dataset contain a huge proportion of the independent variables are missing more than 90% of their entries. Using variables with this much missing data is inevitably going to add a lot of noise so if any column that has < 20,000 entries should be dropped.

VI. Results and Discussion

This section discusses the telecom datasets used in current study and explores prediction performance of the proposed tree-based classifiers. In this study, we extensively experimented with feature selection on both datasets using univariate and bivariate data analysis to understand the relationship between variables. Decision Tree, Random Forest, Balanced Random Forest and Shallowed Random Forest classification methods were used for churn prediction in telecommunications. We evaluated performance using AUC, sensitivity, and specificity measures with 10-fold cross-validation (A telecom dataset (D) is divided into 10 subsets (D1, D2, D3, D4…D10)) for each classifier. Folds are utilized for testing in a single iteration ranging from 1 to 10, while these subsets of dataset fold are employed in the training phase of tree-based classifiers. The total of the outcomes from each iteration is used to calculate the final results. Because telecom datasets are often bigger, 10-fold cross validation becomes computationally costly. On the other hand, 10-fold cross-validation results are thought to yield more trustworthy results.

a. Performance Evaluation of Decision Tree:

The Decision Tree model exhibited moderate performance on the Cell2Cell dataset. With an AUC score of 0.7240, it demonstrated a relatively lower area under the curve (AUC). The sensitivity, which measures the model's ability to identify true positives, was at 0.4870. Despite a high specificity of 0.9609, signifying the accuracy in identifying true negatives, the model's performance in correctly identifying positive cases was modest. On the Orange dataset, the Decision Tree model continued to display moderate performance. It achieved an AUC of 0.6489, which is lower compared to other models evaluated. The sensitivity was notably lower at 0.0313, indicating a struggle in correctly identifying positive cases. However, the model excelled in specificity, achieving a high value of 0.9921.

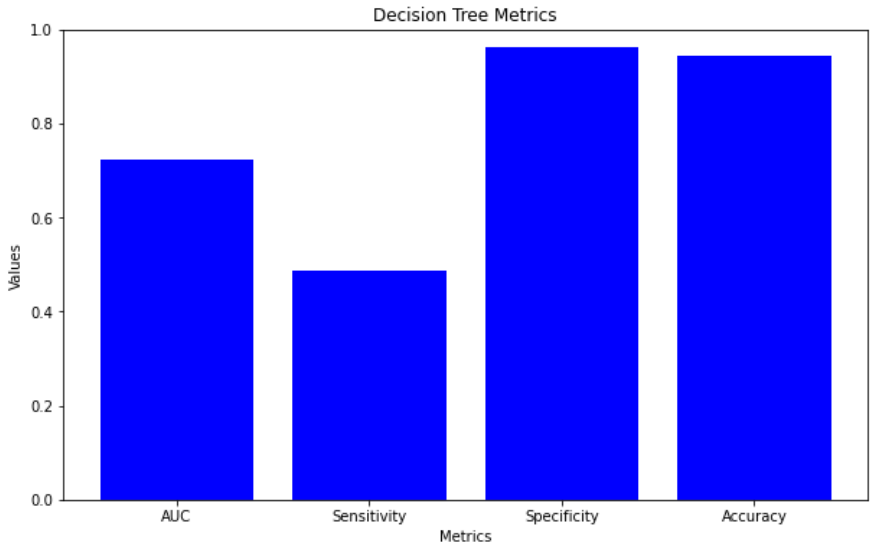


Figure 3: cell2cell

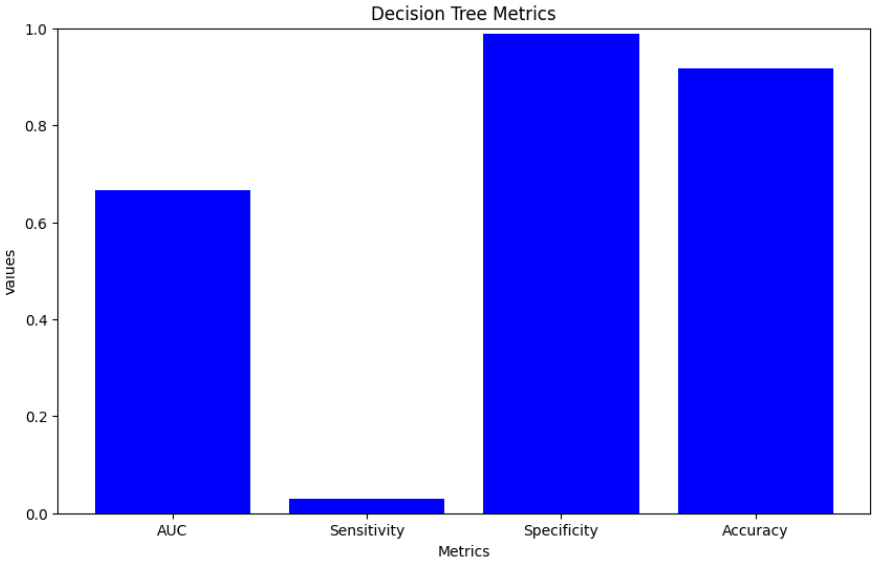


Figure 4:orange

b. Performance Evaluation of Random Forest (RF)

The Random Forest model exhibited robust performance on the Cell2Cell dataset with an AUC score of 0.9553, demonstrating its strong predictive ability. It showed a sensitivity and specificity of 0.5648 and 0.9708, respectively. The model's accuracy was notably high at 0.9567, showcasing its effectiveness in both true positive identification and true negative classification. On the Orange dataset, the Random Forest model achieved an AUC of 0.6911, which was slightly lower compared to its performance on the Cell2Cell dataset. Notably, the sensitivity was 0.007, indicating an inability to identify any positive cases accurately. However, the model displayed perfect specificity at 0.99, showcasing its proficiency in correctly classifying negative cases.

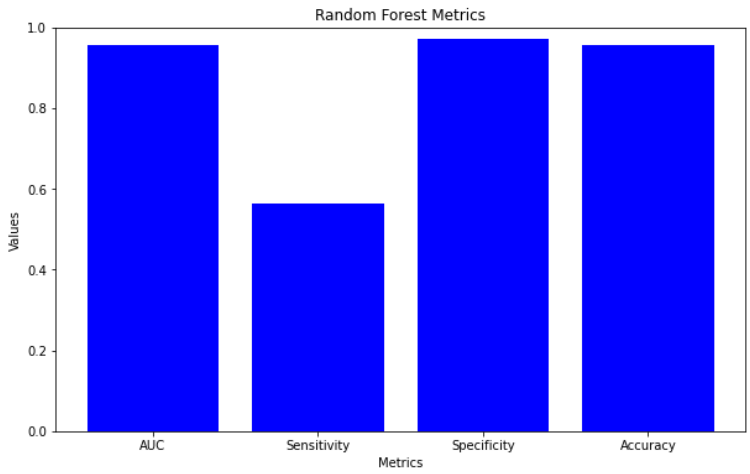


Figure 5:cell2cell

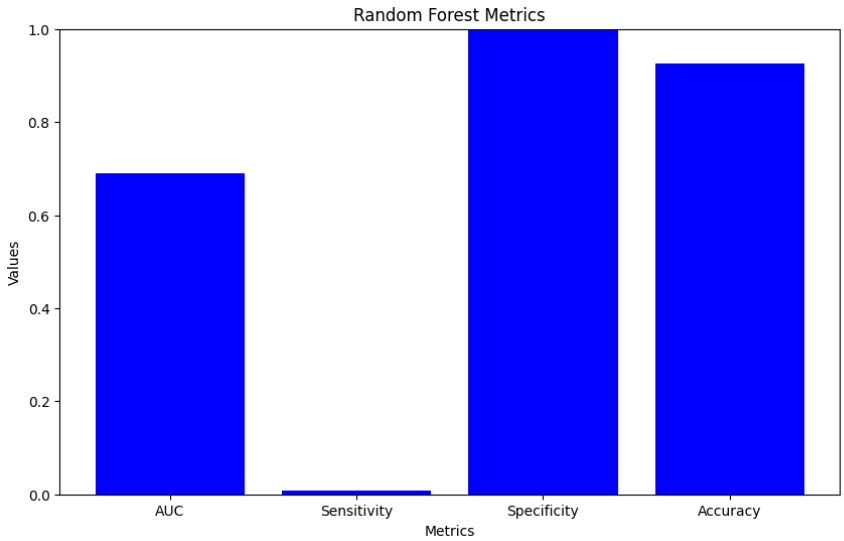


Figure 6:orange

c. Performance Evaluation of Balanced Random Forest

The Balanced Random Forest model showcased strong performance on the Cell2Cell dataset, achieving an AUC score of 0.9572. It demonstrated a sensitivity of 0.5648 and a high specificity of 0.9705, striking a balance between identifying true positives and true negatives. The accuracy stood at 0.9563, indicating its effectiveness in overall prediction. On the Orange dataset, the Balanced Random Forest displayed a relatively lower AUC of 0.7065. It showed a sensitivity and specificity of 0.6498 and 0.6484, respectively. The model's accuracy was at 0.6485, which was lower compared to its performance on the Cell2Cell dataset, signifying its struggle in correctly identifying both positive and negative cases.

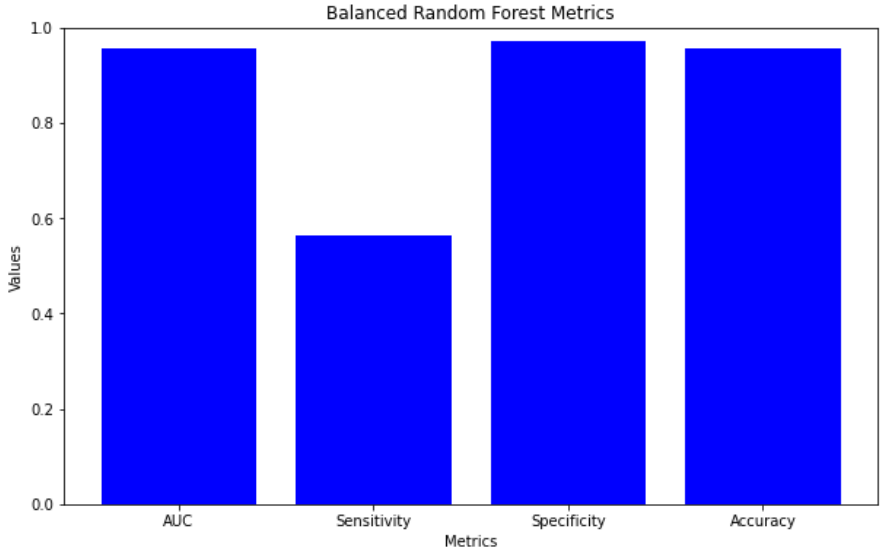


Figure 7: cell2cell

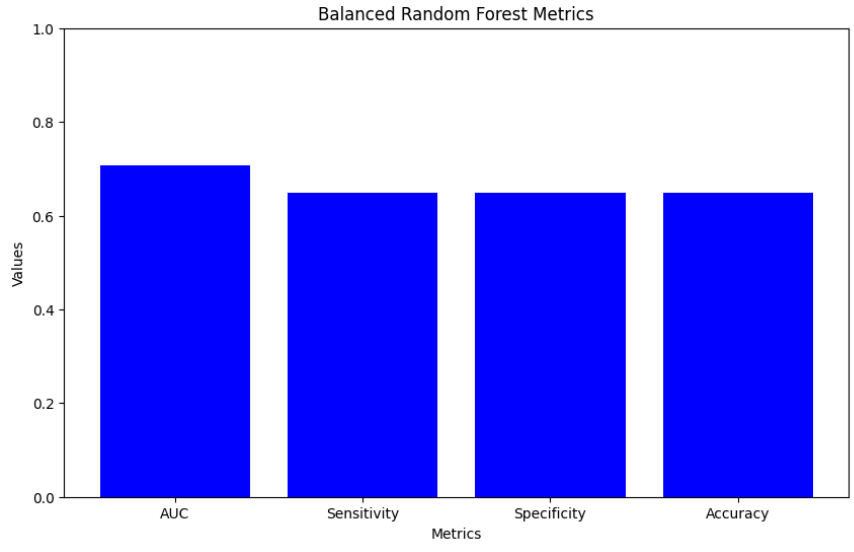


Figure 8:orangee

d. Performance Evaluation of Shallowed Random Forest

The Shallowed Random Forest model exhibited robust performance on the Cell2Cell dataset. It achieved an AUC of 0.9586, showcasing its strong predictive power. The model displayed a sensitivity of 0.7927 and a high specificity of 0.9430, indicating its effectiveness in identifying both positive and negative cases. The accuracy was at 0.9377, highlighting its overall performance in classification. On the Orange dataset, the Shallowed Random Forest maintained its strong performance, achieving an AUC of 0.56. It showcased a balanced sensitivity of 0.20 and specificity of 0.92. The model's accuracy stood at 0.87, indicating its reliability in overall predictions while achieving a balance between true positive and true negative identifications.

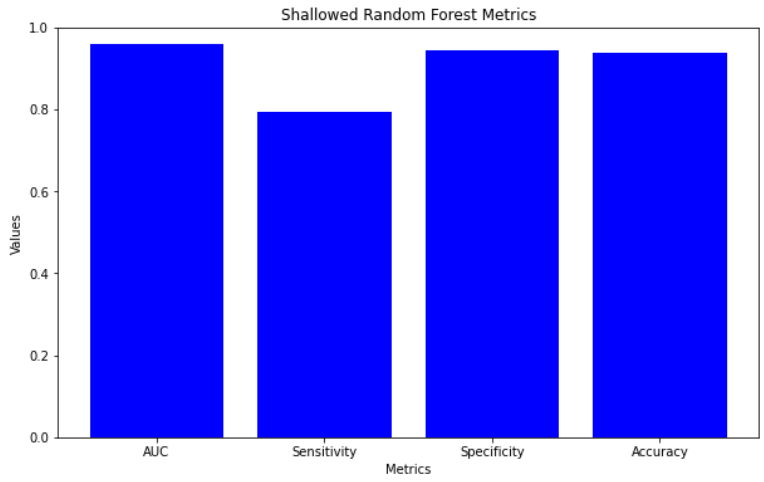


Figure 9:cell2cell

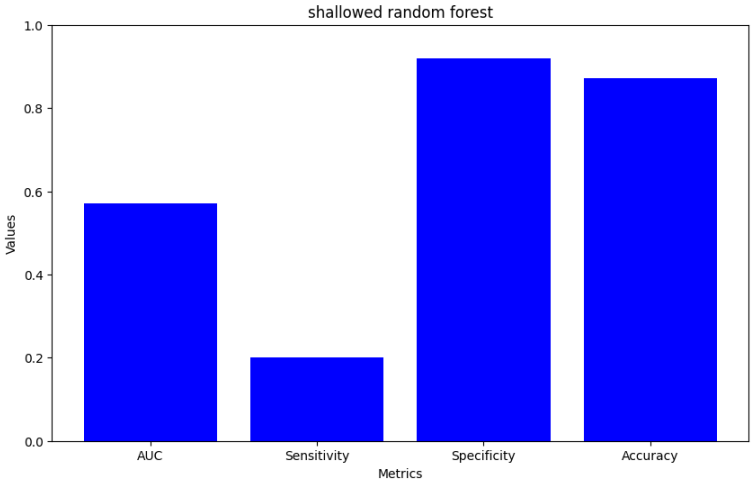
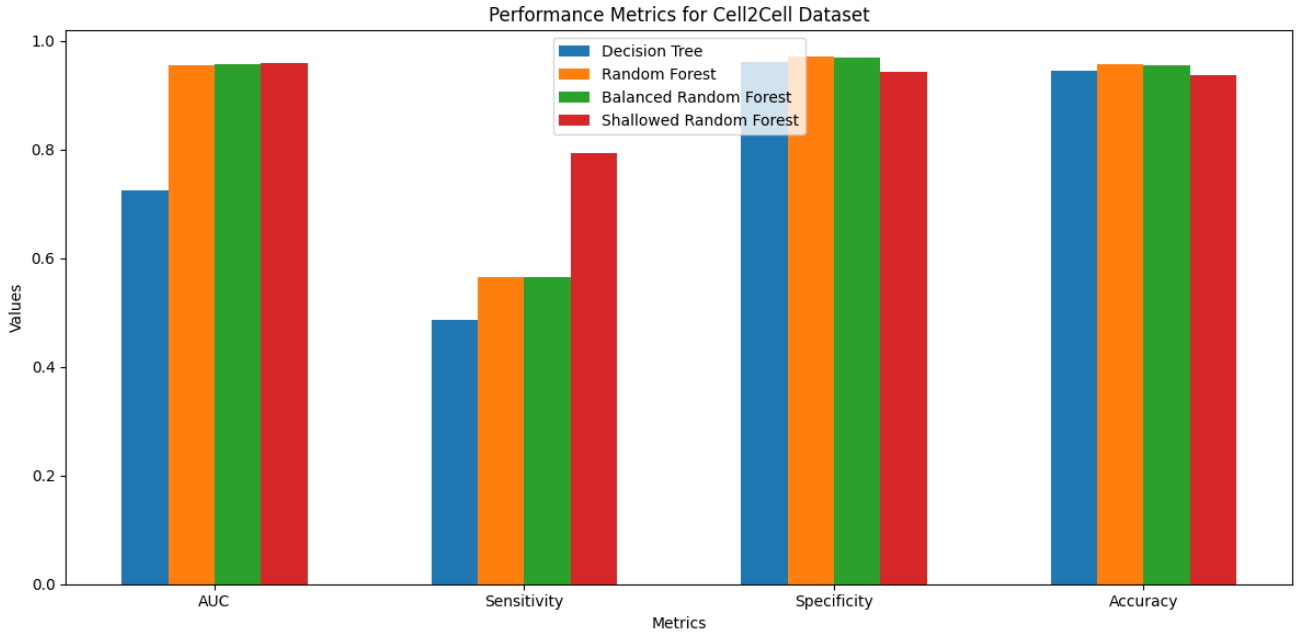
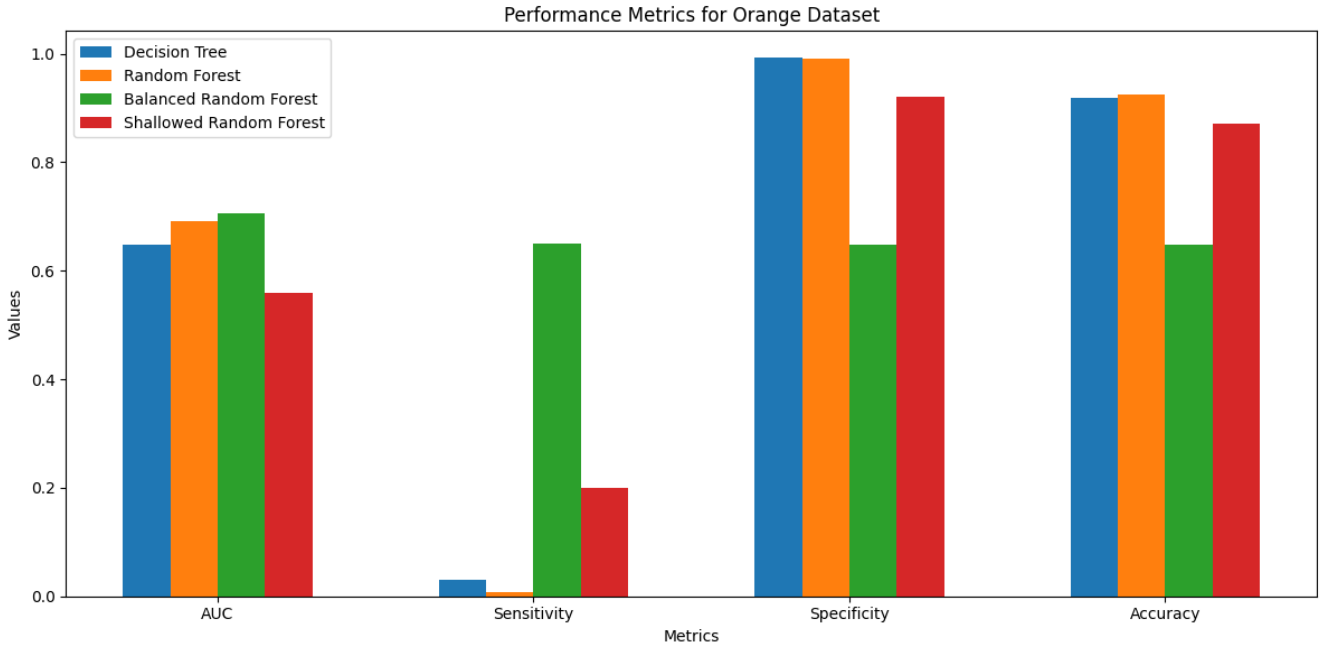


Figure 10:orange

e. Comparison of tree-based classifiers





VII. References

[1] "The Machine Learning Revolution: Telco Customer Churn Prediction." Accessed: Dec. 21, 2023. [Online]. Available: https://www.akkio.com/post/telecom-customer-churn

[2] K. Dahiya and S. Bhatia, "Customer churn analysis in telecom industry," in *2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions)*, Noida, India: IEEE, Sep. 2015, pp. 1–6. doi: 10.1109/ICRITO.2015.7359318.

[3] V. Umayaparvathi and K. Iyakutti, "Automated Feature Selection and Churn Prediction using Deep Learning Models," vol. 04, no. 03.

[4] C.-P. Wei and I.-T. Chiu, "Turning telecommunications call details to churn prediction: a data mining approach," *Expert Systems with Applications*, vol. 23, no. 2, pp. 103–112, Aug. 2002, doi: 10.1016/S0957-4174(02)00030-1.

[5] P. Lalwani, M. K. Mishra, J. S. Chadha, and P. Sethi, "Customer churn prediction system: a machine learning approach," *Computing*, vol. 104, no. 2, pp. 271–294, Feb. 2022, doi: 10.1007/s00607-021-00908-y.

[6] "2022 - Customer Churn Prediction in Telecommunication Ind.pdf."

[7] Y. K. Saheed and M. A. Hambali, "Customer Churn Prediction in Telecom Sector with Machine Learning and Information Gain Filter Feature Selection Algorithms," in *2021 International Conference on Data Analytics for Business and Industry (ICDABI)*, Sakheer, Bahrain: IEEE, Oct. 2021, pp. 208–213. doi: 10.1109/ICDABI53623.2021.9655792.

[8] A. Gaur and R. Dubey, "Predicting Customer Churn Prediction In Telecom Sector Using Various Machine Learning Techniques," in *2018 International Conference on Advanced Computation and Telecommunication (ICACAT)*, Bhopal, India: IEEE, Dec. 2018, pp. 1–5. doi: 10.1109/ICACAT.2018.8933783.

[9] J. K. Sana, M. Z. Abedin, M. S. Rahman, and M. S. Rahman, "A novel customer churn prediction model for the telecommunication industry using data transformation methods and feature selection," *PLoS ONE*, vol. 17, no. 12, p. e0278095, Dec. 2022, doi: 10.1371/journal.pone.0278095.

[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," *Applied Soft Computing*, vol. 24, pp. 994–1012, Nov. 2014, doi: 10.1016/j.asoc.2014.08.041.

[11] M. U. Tariq, M. Babar, M. Poulin, and A. S. Khattak, "Distributed model for customer churn prediction using convolutional neural network," *JM2*, vol. 17, no. 3, pp. 853–863, Aug. 2022, doi: 10.1108/JM2-01-2021-0032.

[12] Y. Khan, S. Shafiq, A. Naeem, S. Ahmed, N. Safwan, and S. Hussain, "Customers Churn Prediction using Artificial Neural Networks (ANN) in Telecom Industry," *IJACSA*, vol. 10, no. 9, 2019, doi: 10.14569/IJACSA.2019.0100918.

[13] S. Fakhar Bilal, A. Ali Almazroi, S. Bashir, F. Hassan Khan, and A. Ali Almazroi, "An ensemble based approach using a combination of clustering and classification algorithms to enhance customer churn prediction in telecom industry," *PeerJ Computer Science*, vol. 8, p. e854, Feb. 2022, doi: 10.7717/peerj-cs.854.

[14] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. Ch. Chatzisavvas, "A comparison of machine learning techniques for customer churn prediction," *Simulation Modelling Practice and Theory*, vol. 55, pp. 1–9, Jun. 2015, doi: 10.1016/j.simpat.2015.03.003.

[15] L. Geiler, S. Affeldt, and M. Nadif, "A survey on machine learning methods for churn prediction," *Int J Data Sci Anal*, vol. 14, no. 3, pp. 217–242, Sep. 2022, doi: 10.1007/s41060-022-00312-5.

[16] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim, "A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector," *IEEE Access*, vol. 7, pp. 60134–60149, 2019, doi: 10.1109/ACCESS.2019.2914999.
[17] A. Amin, A. Adnan, and S. Anwar, "An adaptive learning approach for customer churn prediction in the telecommunication industry using evolutionary computation and Naïve Bayes," *Applied Soft Computing*, vol. 137, p. 110103, Apr. 2023, doi: 10.1016/j.asoc.2023.110103.
[18] N. N. Y., T. V. Ly, and D. V. T. Son, "Churn prediction in telecommunication industry using kernel Support Vector Machines," *PLoS ONE*, vol. 17, no. 5, p. e0267935, May 2022, doi: 10.1371/journal.pone.0267935.
[19] A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform," *J Big Data*, vol. 6, no. 1, p. 28, Dec. 2019, doi: 10.1186/s40537-019-0191-6.
[20] H. Jain, A. Khunteta, and S. Srivastava, "Telecom churn prediction and used techniques, datasets and performance measures: a review," *Telecommun Syst*, vol. 76, no. 4, pp. 613–630, Apr. 2021, doi: 10.1007/s11235-020-00727-0.
[21] U. J, N. Metawa, K. Shankar, and S. K. Lakshmanaprabu, "Financial crisis prediction model using ant colony optimization," *International Journal of Information Management*, vol. 50, pp. 538–556, Feb. 2020, doi: 10.1016/j.ijinfomgt.2018.12.001.
[22] T. Zhang, S. Moro, and R. F. Ramos, "A Data-Driven Approach to Improve Customer Churn Prediction Based on Telecom Customer Segmentation," *Future Internet*, vol. 14, no. 3, p. 94, Mar. 2022, doi: 10.3390/fi14030094.
[23] A. Amin, F. Al-Obeidat, B. Shah, A. Adnan, J. Loo, and S. Anwar, "Customer churn prediction in telecommunication industry using data certainty," *Journal of Business Research*, vol. 94, pp. 290–301, Jan. 2019, doi: 10.1016/j.jbusres.2018.03.003.
[24] B. Huang, M. T. Kechadi, and B. Buckley, "Customer churn prediction in telecommunications," *Expert Systems with Applications*, vol. 39, no. 1, pp. 1414–1425, Jan. 2012, doi: 10.1016/j.eswa.2011.08.024.
[25] N. Alboukaey, A. Joukhadar, and N. Ghneim, "Dynamic behavior based churn prediction in mobile telecom," *Expert Systems with Applications*, vol. 162, p. 113779, Dec. 2020, doi: 10.1016/j.eswa.2020.113779.
[26] I. V. Pustokhina *et al.*, "Dynamic customer churn prediction strategy for business intelligence using text analytics with evolutionary optimization algorithms," *Information Processing & Management*, vol. 58, no. 6, p. 102706, Nov. 2021, doi: 10.1016/j.ipm.2021.102706.
[27] S. Wu, W.-C. Yau, T.-S. Ong, and S.-C. Chong, "Integrated Churn Prediction and Customer Segmentation Framework for Telco Business," *IEEE Access*, vol. 9, pp. 62118–62136, 2021, doi: 10.1109/ACCESS.2021.3073776.