Document Information

Analyzed document	BE_Major_Project_2023_24_Final_Report.pdf (D178356105)
Submitted	2023-11-10 02:31:00 UTC+01:00
Submitted by	Manoj Dhande
Submitter email	manoj.dhande@sakec.ac.in
Similarity	4%
Analysis address	manoj.dhande.sakec@analysis.urkund.com

Sources included in the report

W	URL: https://arxiv.org/pdf/1904.00690 Fetched: 12/8/2020 11:02:52 AM	3
W	URL: https://www.hindawi.com/journals/mpe/2022/8534739/ Fetched: 8/5/2022 12:39:22 PM	5
W	URL: https://www.researchgate.net/publication/344582269_Customer_Churn_Prediction_in_Telecom_Using_Machine_Learning_in_Big_Data_Platform Fetched: 12/11/2020 7:01:24 PM	1
W	URL: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0191-6 Fetched: 12/20/2019 11:26:16 PM	1

Entire Document

A Project Report on Advanced Customer Churn Prediction Submitted in partial fulfillment of the requirements of the degree of Bachelor in Engineering by Viraj Shah(20UF15911CM050) Vinit Damania (20UF15737CM008) Vansh Barot (20UF15733CM002) Devang Khopkar (20UF15585CM024) Under the guidance of Prof. Manoj Dhande Department of Computer Engineering Shah and Anchor Kutchhi Engineering College Chembur, Mumbai – 400088. 2023 – 2024

CERTIFICATE This is to certify that the report of the project entitled Advanced Customer Churn Prediction is a bonafide work of Viraj Shah(20UF15911CM050) Vinit Damania (20UF15737CM008) Vansh Barot (20UF15733CM002) Devang Khopkar (20UF15585CM024) submitted to the UNIVERSITY OF MUMBAI during semester VII in partial fulfilment of the requirement for the award of the degree of BACHELOR OF ENGINEERING in COMPUTER ENGINEERING. Prof. Manoj Dhande Guide Prof. Uday Bhave Dr. Bhavesh Patel Head of the Department Principal

Attendance Certificate Date: 10/11/23 To, The Principal, Shah and Anchor Kutchhi Engineering College, Chembur, Mumbai-88 Subject: Confirmation of Attendance Respected Sir, This is to certify that Final year students Viraj Shah, Vinit Damania , Vansh Barot , De- vang Khopkar have duly attended the sessions on the day allotted to them during the period from 02/08/2023 to 31/10/2023 for performing the Project titled Advanced Customer Churn Prediction . They were punctual and regular in their attendance. Following is the detailed record of the student's attendance. Attendance Record: Date Viraj Shah Vinit Damania Vansh Barot Devang Khopkar Present/Absent Present/Absent Present/Absent Present/Absent 02/08/2023 Present Present Present Present 16/08/2023 Present Present Present Present 30/08/2023 Present Present Present Present 13/09/2023 Present Present Present Present 04/10/2023 Present Present Present Present 11/10/2023 Present Present Present Present 18/10/2023 Present Present Present Present 31/10/2023 Present Present Present Present Prof. Manoj Dhande

Approval for Project Report for B. E. Semester VII This project report entitled Advanced Customer Churn Prediction by Viraj Shah, Vinit Da- mania , Vansh Barot and Devang Khopkar is approved for semester VII in partial fulfilment of the requirement for the award of the degree of Bachelor of Engineering. Examiners 1. 2. Guide 1. 2. Date: 10/11/23 Place: Mumbai iv

Declaration We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the orig- inal sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed. Name of the Student Roll No. Signature Viraj Shah 20UF15911CM050 Vinit Damania 20UF15737CM008 Vansh Barot 20UF15733CM002 Devang Khopkar 20UF15585CM024 Date: 10/11/23 Place: Mumbai v

Acknowledgement I would like to express my sincere gratitude to all those who have supported and guided me throughout the process of conducting this report on Advanced Customer Churn Predic- tion . This endeavor would not have been possible without their valuable contributions and assistance. We are thankful to our college Shah and Anchor Kutchhi Engineering College for consider- ing our project and extending help at all stages needed during our work of collecting infor- mation regarding the project. We are deeply indebted to our Principal Dr. Bhavesh Patel and Head of the Computer Engineering Department Prof. Uday Bhave for giving us this valuable opportunity to do this project. We express our hearty thanks to them for their assistance without which it would have been difficult in finishing this project synopsis and project review successfully. We take this opportunity to express our profound gratitude and deep regards to our guide Prof. Manoj Dhande for her exemplary guidance, monitoring and constant encouragement throughout the course of this project. This work would not have been possible without the collective efforts of these individuals and organizations. While any shortcomings in this report are solely our responsibility, their contributions have significantly enriched its content. vi

Abstract Customer attrition poses a significant challenge and ranks among the top concerns for large corporations. Its impact on a company's revenue is particularly pronounced in the telecom- munications sector. Consequently, businesses are actively searching for ways to predict potential customer churn. Identifying the key factors that contribute to customer churn is of utmost importance in taking proactive measures to mitigate this problem. Our primary contribution lies in the development of a churn prediction model designed to aid telecom op- erators in identifying customers most likely to churn. The model we've constructed utilizes machine learning techniques within a robust big data framework, and it introduces innova- tive methods for feature engineering and selection. To assess the model's performance, we employ the widely accepted Area Under Curve (AUC) metric, achieving an impressive AUC score of 93.3 percent.Another notable innovation is our incorporation of customer social network data into the prediction model through the extraction of Social Network Analysis (SNA) features. This integration significantly enhances the model's performance, elevating the AUC score from 84percent to 93.3percent . The model was meticulously crafted and rig- orously tested within the Spark environment, utilizing a substantial dataset derived from the transformation of raw data provided by SyriaTel, a prominent telecom company. This dataset spans a comprehensive nine-month period and encompasses detailed customer information. It served as the foundation for training, testing,

and evaluating

55% **MATCHING BLOCK 1/10** W

the system at SyriaTel. Our experimentation involved four distinct algorithms: Decision Tree, Random Forest, Gradient Boosted Machine Tree (GBM), and Extreme Gradient Boosting (XGBOOST). The results clearly favored the XGBOOST algorithm, which was subsequently employed for classifica- tion within this churn prediction model. Keywords: Customer churn prediction, Churn in telecom, Machine learning, Feature selec- tion, Classifcation, Mobile Social Network Analysis, Big data

vii

Chapter 1 Introduction The telecommunications industry has emerged as a prominent sector in developed nations. With the advancement of technology and a growing number of operators, competition has intensified [1]. Companies are diligently devising strategies to thrive in this competitive landscape. Among the three primary strategies proposed to boost revenues [2] – (1) acquiring new customers, (2) upselling to existing customers, and (3) extending customer retention – it has been empirically demonstrated

62% **MATCHING BLOCK 2/10** W

that the third strategy is the most financially rewarding [2]. This underscores that retaining an existing customer

is not only more cost-effective than acquiring a new one [3] but also generally simpler than implementing the upselling strategy [4]. To implement this third strategy successfully, companies must reduce the risk of customer churn, which refers to customers shifting from one provider to another [5]. The issue of customer churn is particularly significant in service sectors characterized by intense competition. Additionally, forecasting which customers are likely to depart from a company can potentially open up a substantial additional revenue stream, especially when done early in the process [3]. Numerous research studies have corroborated the high efficacy of machine learning technology in predicting this scenario. This technique is applied by leveraging insights gleaned from historical data [6,7] 1.1 Background Telecommunication companies (telco's) have a significant problem with customer churn, which is the loss of customers who stop using their services. To solve this problem, telco's need to identify the customers who are likely to churn and take preemptive measures to retain them. Machine learning models can help telco's in predicting the customers who are most likely to churn, based on various factors such as customer usage patterns, payment history, and demographics. 1

Chapter 1. Introduction 1.2 Motivation The motivation behind the project "Telecom Customer Churn Prediction" likely stems from several compelling factors: 1. Business Imperative: The telecommunications industry is highly competitive, and com- panies in this sector are driven by the need to maximize their customer base and revenue. Customer churn can have a direct and significant negative impact on revenue, making it imperative for telecom companies to address this issue. 2. Revenue Enhancement: Telecom companies continuously seek strategies to boost their revenues. One of the most profitable approaches identified is extending the retention period of existing customers. Retaining customers is often more cost-effective and less challenging than acquiring new ones or upselling to existing ones. As such, predicting and preventing customer churn becomes a key focus for revenue enhancement. 3. Customer Retention: Maintaining a loyal customer base is essential in the long-term success of telecom companies. Reducing customer churn through predictive methods en- sures that customers stay with the company, fostering a more stable and sustainable business model. 4. Competitive Advantage: Companies that can successfully predict and prevent customer churn gain a competitive advantage. They can offer better service and incentives to at-risk customers, leading to higher customer satisfaction and loyalty. In summary, the project "Telecom Customer Churn Prediction" is motivated by the need to address a critical business challenge, enhance revenues, improve customer retention, and leverage advanced technologies to gain a competitive edge in the telecommunications indus- try. Computer Engineering 2

Chapter 2 Literature Review 2.1 Survey of Existing system 3

Chapter 2. Literature Review Sr No Title Author Remarks 1.

81% **MATCHING BLOCK 3/10** W

Customer churn prediction in telecom using machine learn- ing in big data platform[1] Abdelrahim Kasem Ahmad, Assef Jafar Kadan Aljoumaa The

authors developed a new way of feature engineering and selec- tion to improve the per- formance of the model.. 2.

100% **MATCHING BLOCK 4/10** W

Developing a prediction model for customer churn from electronic banking services using data mining[2]

Abbas Keramati, Hajar Gha- neei Seyed Mohammad Mir- mohammadi The authors conclude that their model can be used by banks to iden- tify customers who are at risk of churning and take proactive measures to retain them. 3.

76% **MATCHING BLOCK 5/10** W

The use of knowledge extrac- tion in predicting customer churn in B2B[3]

Arwa A. Jamjoom The study concludes that data mining tech- niques can be effec- tively used to predict customer churn in B2B settings.. 4. Customer churning Analy- sis using machine learning algorithms[4] B. Prabhadevi , R. Shalini, B. R. Kavitha They recommend that businesses use machine learning to identify cus- tomers who are at risk of churn and take steps to retain them. 5.

71% **MATCHING BLOCK 6/10** W

Survey on Customer Churn Prediction Using Machine Learning techniques [5] Saran Kumar A. Chandrakala D.

The authors discuss the different data pre- processing and feature engineering techniques that can be used to improve the accuracy of churn prediction models. 6. Predicting Customer Churn in Telecom Industry using neu- ral network[6] Omar Adwan, Ossam Faris,Khalid Jaradat This paper explores the application of neural networks to predict customer churn in the telecommunications industry.. Computer Engineering 4

Chapter 2. Literature Review 2.2 Limitation of Existing system or research gap The mentioned research papers provide valuable insights into churn prediction in the telecom industry using various machine learning and data mining approaches. However, there are limitations to these proposed systems, and there is a research gap that can be addressed: 1. Data Size and Representativeness: Many of the studies mentioned in the text use relatively small datasets, and some are based on specific telecom companies' data, which may not be representative of the broader industry. 2. Feature Engineering: The paper highlights that most previous research did not perform feature engineering but relied on ready-made features provided by telecom companies. 3. Class Imbalance: Unbalanced datasets,

55% **MATCHING BLOCK 7/10** W

where the churned customer class is smaller than the active customer class, is a significant challenge in churn prediction.

While some papers address this issue with oversampling and undersampling techniques, it's important to men- tion that these techniques can introduce biases or overfitting, and their effectiveness may vary depending on the dataset. 4. Model Evaluation Metrics: The mentioned papers primarily use AUC (Area Under the Curve) to evaluate model performance. While AUC is a useful metric, it's not the only metric that should be considered. In practice, other metrics like precision, recall, F1-score, and cost- sensitive metrics should be evaluated to provide a more comprehensive understanding of a model's performance.[?] 2.3 Problem Statement and Objective Problem Statement To create a predictive application for telecom companies to address the challenge of customer churn, focusing on both retaining and acquiring users. As the telecom industry experiences rapid growth, retaining customers has become crucial, as the loss of subscribers can ad- versely affect a company's profitability. Churn prediction helps in identifying potential cus- tomer defection to competing providers. Telecom companies grapple with an ever-increasing churn rate, and this study employs machine learning algorithms to develop effective churn- reduction strategies. Silent churn, a particularly challenging type to predict, involves users who may leave in the near future. Decision-makers and advertisers should prioritize reduc- ing churn rates, as existing customers are more valuable assets for companies compared to acquiring new ones. Computer Engineering 5

Chapter 2. Literature Review Objective • Reduce Customer Churn: The main objective is to reduce the number of customers who leave the telecom service, thus increasing customer retention and revenue. • Business Insights: Interpret the model results to understand which features are most influential in predicting customer churn. This can provide valuable business insights for decision-making. • Measuring Success: Define key performance indicators (KPIs) to measure the success of the churn prediction project, such as a reduction in churn rate, increased customer retention, or improved profitability. 2.4 Scope The project "Telecom Customer Churn Prediction" holds significant potential as it addresses a critical challenge in the telecom industry. By utilizing advanced machine learning and data analysis techniques, it aims to accurately forecast customer churn, allowing telecom companies to proactively retain customers and reduce revenue losses. This project can en- hance customer satisfaction, optimize marketing strategies, and improve overall business performance. With the ever-growing competition in the telecommunications sector, effective churn prediction offers a valuable competitive edge and ensures better resource allocation, making it a promising and impactful endeavor. Computer Engineering 6

Chapter 3 Software Requirement Specification 3.1 Introduction Purpose The purpose of this project is to develop a telecom customer churn prediction system using machine learning techniques. Through exploratory data analysis (EDA) and the application of decision tree, random forest, and PCA algorithms, we aim to create a model that can effec- tively predict and mitigate customer churn by addressing class imbalance with SmoteENN. Document Conventions This document follows standard documentation conventions for clarity and consistency. Intended Audience and Reading Suggestions This project is tailored for data scientists, machine learning enthusiasts, and telecom profes- sionals seeking insights into customer churn prediction. It assumes a basic understanding of machine learning concepts and offers a comprehensive view of the process, from exploratory data analysis (EDA) to model selection, with a focus on decision trees, random forests, and PCA, while addressing sample imbalance using SMOTEENN. Product Scope The telecom customer churn prediction project aims to develop a machine learning solution for identifying potential customer churn. It includes exploratory data analysis, model training with decision tree, random forest, and PCA, along with sample balancing using SmoteENN, to enhance predictive accuracy and inform targeted retention strategies. 7

Chapter 3. Software Requirement Specification References List any other documents or Web addresses to which this SRS refers. These may include user interface style guides, contracts, standards, system requirements specifications, use case documents, or a vision and scope document. Provide enough information so that the reader could access a copy of each reference, including title, author, version number, date, and source or location. 3.2 Overall Description Product Perspective From a product perspective, our telecom customer churn prediction project employs a diverse range of machine learning algorithms, including decision trees, random forests, and PCA, to enhance predictive accuracy. With the integration of SMOTEENN for sample balance, we provide a robust solution for reducing customer churn, ensuring greater customer retention and improved service quality. Product Functions 1. Data Preprocessing: Perform data cleaning, feature engineering, and balancing using SMOTE-ENN to prepare the dataset for modeling. 2. Decision Tree Model: Implement a Decision Tree algorithm to predict customer churn based on input features. 3. Random Forest Model: Develop a Random Forest model for improved predictive accu- racy. 4. Principal Component Analysis (PCA): Implement PCA for dimensionality reduction and feature selection to enhance model performance. 5. Evaluation: Evaluate model performance using relevant metrics like accuracy, precision, recall, and F1-score. 6. Visualization: Create visualizations for data exploration and model results. User Classes and Characteristics Data Analysts: Require comprehensive exploratory data analysis (EDA) insights, model per- formance evaluation, and data preprocessing expertise. Machine Learning Engineers: Need Computer Engineering 8

Chapter 3. Software Requirement Specification to implement and fine-tune decision tree, random forest, and PCA-based models, ensuring optimal predictive accuracy. Business Stakeholders: Seek actionable insights on customer churn factors and model-driven recommendations for retention strategies. Data Scientists: Must possess proficiency in handling imbalanced datasets, utilizing SMOTEENN for effec- tive sample balancing. Telecom Executives: Rely on the project to make informed decisions and prioritize efforts to reduce customer churn, based on predictive results. Operating Environment Operating Environment: This telecom customer churn prediction project relies on Python 3.x and popular libraries like scikit-learn, pandas, and NumPy. It utilizes decision trees, random forests, and Principal Component Analysis (PCA) for modeling. Sample balancing is achieved using the SmoteENN technique, ensuring robustness in handling imbalanced data. Design and Implementation Constraints The system should be designed to run on standard desktop hardware and must not consume excessive system resources. User Documentation List the user documentation components (such as user manuals, on-line help, and tutori- als) that will be delivered along with the software. Identify any known user documentation delivery formats or standards. Assumptions and Dependencies Data Quality: Assumes that the input data is accurate, complete, and representative of the telecom customer population. Algorithm Suitability: Assumes that the chosen machine learning algorithms (Decision Tree, Random Forest, PCA) are appropriate for the prob- lem and dataset. Balanced Data: Depends on SMOTEENN for addressing class imbalance. Model Interpretability: Assumes the models provide interpretable insights into customer churn factors. External Factors: Assumes no major external factors affecting customer churn are missing. Deployment: Assumes successful deployment for real-time predictions. 3.3 External Interface Requirements User Interfaces Upload your data. Perform Exploratory Data Analysis (EDA). Choose your ML model: Decision Tree, Random Forest, or PCA. Apply SMOTEENN for sample balance. Predict Computer Engineering 9

Chapter 3. Software Requirement Specification customer churn likelihood. Get actionable insights and recommendations. Hardware Interfaces Hardware Interfaces for your telecom customer churn prediction project include a standard computer with a minimum of 8GB RAM, a multi-core CPU for faster model training, and a storage capacity of at least 100GB to store datasets and model checkpoints. Ensure GPU support for accelerated machine learning tasks, if available. Software Interfaces Software Interfaces: The project offers a user-friendly interface with options for data explo- ration, model training, and testing. Users can select from Decision Tree, Random Forest, and PCA models, and choose SMOTE-ENN for sample balancing. The interface simplifies the entire churn prediction process for telecom customers. Communications Interfaces The software may require internet connectivity for updates and cloud-based services. 3.4 System Features EDA (Exploratory Data Analysis) for data understanding. Utilized Decision Tree, Random Forest, and PCA algorithms for prediction. Employed SMOTEENN for sample balancing. Provides insights into customer churn likelihood. Helps telecom companies optimize re- tention strategies. Scalable for larger datasets. User-friendly interface for input and results visualization. Robust and accurate predictions for informed decision-making. 3.4.1 EDA (Exploratory Data Analysis) for data understanding. Description and Priority Description: Exploratory Data Analysis (EDA) is a crucial phase in the telecom customer churn prediction project. It involves data cleaning, visualization, and statistical analysis to understand data patterns, identify outliers, and gain insights into feature importance. Priority: EDA is of high priority as it lays the foundation for model building. A well- executed EDA helps in selecting relevant features, addressing data imbalances, and making informed decisions about the choice of machine learning algorithms, leading to more accu- rate churn predictions. Computer Engineering 10

Chapter 3. Software Requirement Specification Stimulus/Response Sequences Response: Explored data intricacies through EDA, harnessed decision trees' interpretability, enhanced accuracy with random forests, and optimized features via PCA. Addressed imbal- ance with SmoteENN, ensuring robust model training for precise telecom customer churn predictions. Functional Requirements • REQ-1:Implement exploratory data analysis for telecom customer churn prediction. • REQ-2:Train and test decision tree, random forest, and PCA models. • REQ-3:Integrate SmoteENN for sample size balancing. • REQ-4:Ensure seamless execution and compatibility across machine learning algo- rithms. 3.4.2 Provides insights Provides insights into customer churn likelihood. Helps telecom companies optimize reten- tion strategies. 3.5 Other Nonfunctional Requirements Performance Requirements Ensure accurate customer churn prediction with a minimum accuracy of 85 percent. Opti- mize model training time to under 5 minutes. Maintain a precision of at least 80 percent to minimize false positives. Strive for a recall of 85 percent to capture most churn instances. Validate the model's stability with consistent performance across diverse datasets. Safety Requirements Ensure data privacy compliance throughout the project, implement secure data handling practices, and regularly update model training to reflect evolving patterns. Monitor and mit- igate bias in predictive outcomes. Adhere to ethical AI principles, promoting transparency and fairness in the use of customer data. Security Requirements To enhance the security of your telecom customer churn prediction project, implement ro- bust encryption protocols for sensitive data during EDA and model training. Employ access controls to restrict system entry. Regularly update algorithms and conduct security audits. Ensure compliance with data protection regulations, prioritizing customer privacy and con- fidentiality. Computer Engineering 11

Chapter 3. Software Requirement Specification Software Quality Attributes Efficiency: Leveraging Decision Trees and Random Forests enhances predictive accuracy, while Principal Component Analysis (PCA) streamlines feature dimensionality. Robustness: Employing SmoteENN ensures a balanced dataset, mitigating biases. Maintainability: Clear EDA documentation facilitates ongoing project understanding. Overall, your project priori- tizes efficiency, robustness, and maintainability for impactful telecom churn prediction. Business Rules • 1. Algorithmic Diversity: Utilized decision tree, random forest, and PCA for robust model comparison. • 2. Data Balance: Addressed sample size imbalance through SmoteENN to enhance predictive accuracy. • 3. Iterative Evaluation: Conducted thorough EDA and iterative model training/testing for data-driven insights. • 4. Ensemble Strength: Leveraged the power of random forests for improved predictive performance. 3.6 Other Requirements Other essential requirements for the system to run smoothly are mentioned in the require- ments section Computer Engineering 12

Chapter 4 Project Scheduling and Planning • Phase 1. Research of existing systems and gap identification. Researched papers by reputation data scientists and software developers in advantage of profiting businesses and industries Realised and identified the requirement of "cus- tomer retention" Further studied customer retention techniques and came across cus- tomer churning prediction and went through more papers over the same topic. • Phase 2 selection of appropriate data set. Researched through kaggle as mainstream companies and businesses do not release their customer information and data Found sufficient amount of high variant low biased data on kaggle. • Phase 3. Data Pre-Processing and Data-Cleaning. Eliminating the records with high null values or empyty data. Also eliminating features that does not contribute to the end result like Customer IDs and tenure etc. Converting categorical data into 0s and 1s for the ML models to work better. • Phase 4. Selection of algorithm Studied various ML algorithms for deploying the best possible algorithm applicable for the topic. Through ensemble learning method, realised that PCA, Random Forest And Desicion Trees were best suitable and most complimenting to each other and were devices to return confident and accurate output. • Phase 5 Balancing the dataset using SMOOTE-ENN Balancing the highly imbalanced data using SMOOTENN so that the algorithms can work more efficiently and can yield more accurate results. This resulted in an enhanced accuracy for each algorithm and we were able to get better results. • Phase 6 Finalising the Algorithm Choosing model with the best accuracy and f-score and saving it as a model for the API and UI connections that we can display the Churn Confidence and Insights for the particular record. 13

Chapter 5 Proposed System 5.1 Algorithms and Frameworks • 1) Decision Tree: Classification bushes are tree fashions wherein the centered variable can take a discrete set of values; in those tree structures, leaves suggest magnificence labels and branches constitute function conjunctions that cause the one's magnificence labels. Regression bushes are choice bushes wherein the goal variable can take non- stop values (typically actual numbers). To create a prediction, this set of rules divides an information pattern into or extra homogeneous units primarily based totally on the maximum sizeable differentiator in entering variables. The tree is created by using department of every split. As a result, a tree containing choice nodes and leaf nodes (that are connected) is formed. • 2) Random Forest: The random woodland is a category set of rules product of several choice bushes. We use Random Forest to forecast whether or not or now no longer the client will terminate his membership. Random Forest makes use of Decision bushes to are expecting whether or not a client might cancel his subscription. A choice tree specializes in one precise elegance. A elegance with the maximum votes may be the classifier for a selected client. Decision bushes are particularly touchy to the records on which they're trained. We use Bagging to keep away from this. Bagging is a way wherein we take a random pattern from a dataset to educate choice bushes. • 3) XGBoost: XGBoost has become widely used model amongst Kaggle competition. It is effortlessly on hand as open-supply software, and it is able to be used on lots of structures and interfaces. XGBoost stands for excessive Gradient Boosting. The key purpose for the use of XGBoost is its execution pace and version performance. XGBoost employs ensemble studying methods, this means that it employs a group of more than one algorithm to supply output. XGBoost gives parallel and disbursed computing at the same time as offering reasonably-priced reminiscence use. 14

Chapter 5. Proposed System 5.2 Details of Hardware & Software –Hardware Requirements– 1. Processor (CPU): A quad-core or higher CPU is recommended. 2. Memory (RAM): The more RAM you have, the better, as it allows you to work with larger datasets and run complex models. At least 16GB of RAM is advisable. 3. Graphics Processing Unit (GPU): NVIDIA GPUs, such as the GeForce is preferable 4. Storage: Need sufficient storage space to store dataset, code, and the models. An SSD is preferable for faster data access. 5. Internet Connection: A stable and reasonably fast internet connection is necessary for downloading datasets, libraries, and updates. –Software Requirements– 1. Operating System: Use any major operating system such as Windows, macOS, or Linux. 2. Python: Most popular programming language for machine learning. Install Python 3.x on your system. Official Python website (https://www.python.org/). 3. Integrated Development Environment (IDE): Google Colab, Visual Studio Code. 4. Machine Learning Libraries: Various Python libraries for machine learning, including: • NumPy: For numerical operations. • pandas: For data manipulation. • scikit-learn: For machine learning algorithms. • Matplotlib and Seaborn: For data visualization. • imblearn(imbalanced-learn). • Importing functions like SmoteENN (Synthetic Minority Over-sampling Technique (SMOTE) and Edited Nearest Neighbors (ENN)), Decision Trees, PCA etc. 5. Data Visualization Tools: Power BI can be useful for creating interactive visualizations to present findings. Computer Engineering 15

Chapter 5. Proposed System 5.3 Design Details 5.4 Methodology (your approach to solve the problem) We've used previous data for predicting future customer churn. We examine data from con- sumers who have already churned (respond) as well as their attributes/behavior (predictors) prior to the churn. Customers' demographic information, total charges, and the sort of ser- vice they receive from the company are all included in the dataset. It is made up of churn data from over a thousand consumers divided across 21 parameters gathered from Kaggle. We will try to predict the reaction for existing customers by fitting statistical models that relate the predictors to the response. Computer Engineering 16

Chapter 6 Implementation Plan for Next Semester In the upcoming semester, our focus will be on enhancing the accuracy and f-score of the telecom customer churn prediction project. We'll kick off by delving into hyperparameter tuning, fine-tuning the existing algorithms to extract optimal performance. This step is cru- cial for maximizing the predictive power of our models and improving overall effectiveness. The next significant addition to our arsenal will be XGBoost, a powerful and efficient al- gorithm known for its performance in structured/tabular data scenarios. By incorporating XGBoost into our ensemble, we aim to elevate the predictive capabilities of our model, po- tentially outperforming the previous algorithms. Addressing the imbalances in our dataset will remain a priority. While SmoteENN has been effective, we'll explore other advanced techniques to further enhance our model's ability to handle class imbalances. This includes experimenting with different resampling methods and evaluating their impact on model performance. To ensure our findings are accessible and comprehensible for stakeholders, we plan to de- velop both an API and a user interface (UI). The API will facilitate seamless integration of our model into existing systems, enabling real-time predictions. Simultaneously, the UI will provide an intuitive platform for stakeholders to interact with the model's predictions and gain valuable insights. The UI design will prioritize user-friendly visualizations, making complex machine learning outputs understandable for non-technical stakeholders. This approach fosters better collab- oration and decision-making by ensuring that insights derived from the model are easily digestible. Regular model monitoring and updates will be implemented to maintain relevance and accu- 17

Chapter 6. Implementation Plan for Next Semester racy. Continuous evaluation of model performance against real-world data will inform any necessary adjustments or retraining. This iterative process ensures the model remains robust and adaptable to changing patterns in the telecom industry. In conclusion, our next semester's implementation plan revolves around hyperparameter tun- ing, integrating XGBoost, addressing class imbalances, and developing an API/UI for en- hanced accessibility. This comprehensive approach aims to not only boost the predictive power of our model but also make it a valuable tool for stakeholders, fostering informed decision-making in the dynamic landscape of telecom customer churn prediction. Computer Engineering 18

Chapter 7 Summary In our telecom customer churn prediction project, We embarked on a comprehensive journey. Beginning with exploratory data analysis (EDA), we delved into the intricacies of the dataset. To fortify my predictive prowess, we harnessed the power of machine learning algorithms, starting with the decision tree and advancing to the robust random forest. The finale featured Principal Component Analysis (PCA), offering a nuanced perspective. Mindful of data im- balances, we strategically employed SmoteENN to harmonize sample sizes, enhancing the models' efficacy. This multifaceted approach not only bolstered predictive accuracy but also illuminated the subtle dynamics influencing customer churn. The decision tree provided in- terpretability, the random forest brought ensemble strength, and PCA distilled complex fea- tures. Through meticulous training and testing, we sculpted a predictive framework poised to discern customer churn patterns, an invaluable asset for preemptive retention strategies in the dynamic realm of telecom services. This project exemplifies the synergy between ex- ploratory analysis, algorithmic diversity, and strategic sampling techniques, converging to fortify the predictive arsenal in the ever-evolving landscape of customer behavior analysis. 19
Bibliography [1]

85% **MATCHING BLOCK 8/10** W

A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom us- ing machine learning in big data platform,"

in Proceedings of the IEEE International Conference on Big Data, 2019. [2]

100% **MATCHING BLOCK 9/10** W

A. Keramati, H. Ghaneei, and S. M. Mirmohammadi, "Developing a prediction model for customer churn from electronic banking services using data mining,"

in Proceedings of the IEEE International Conference on Data Mining, 2019. [3]

100% **MATCHING BLOCK 10/10** W

A. A. Jamjoom, "The use of knowledge extraction in predicting customer churn in b2b,"

in Proceedings of the IEEE International Conference on Knowledge Extraction and Data Mining, 2021. [4] B. Prabhadevi, R. Shalini, and B. R. Kavitha, "Customer churning analysis using ma- chine learning algorithms," in Proceedings of the IEEE International Conference on Machine Learning and Data Analysis, 2023. [5] S. K. A. and C. D., "Survey on customer churn prediction using machine learning tech- niques," in Proceedings of the IEEE International Workshop on Machine Learning Ap- plications in Customer Churn Prediction, 2022. [6] O. Adwan, O. Faris, and K. Jaradat, "Predicting customer churn in telecom industry using neural network," in Proceedings of the IEEE International Conference on Neural Networks, 2023. 20
Appendix A Appendices A.1 Plagiarism Report 21

Hit and source - focused comparison, Side by Side

Submitted text	As student entered the text in the submitted document.
Matching text	As the text appears in the source.

1/10 **SUBMITTED TEXT** 56 WORDS	**55%** **MATCHING TEXT** 56 WORDS
the system at SyriaTel. Our experimentation involved four distinct algorithms: Decision Tree, Random Forest, Gradient Boosted Machine Tree (GBM), and Extreme Gradient Boosting (XGBOOST). The results clearly favored the XGBOOST algorithm, which was subsequently employed for classifica- tion within this churn prediction model. Keywords: Customer churn prediction, Churn in telecom, Machine learning, Feature selec- tion, Classifcation, Mobile Social Network Analysis, Big data	the system at SyriaTel. The model experimented four algorithms: Decision Tree, Random Forest, Gradient Boosted Machine Tree "GBM" and Extreme Gradient Boosting "XGBOOST". However, the best results were obtained by applying XGBOOST algorithm. This algorithm was used for classification in this churn predictive model. Keywords: Customer churn prediction, Churn in telecom, Machine learning, Feature selection, Classification, Mobile Social Network Analysis, Big data
W https://arxiv.org/pdf/1904.00690	

2/10 **SUBMITTED TEXT** 17 WORDS	**62%** **MATCHING TEXT** 17 WORDS
that the third strategy is the most financially rewarding [2]. This underscores that retaining an existing customer	that the third strategy is the most profitable strategy [2], proves that retaining an existing customer
W https://arxiv.org/pdf/1904.00690	

3/10 **SUBMITTED TEXT** 17 WORDS	**81%** **MATCHING TEXT** 17 WORDS
Customer churn prediction in telecom using machine learn- ing in big data platform[1] Abdelrahim Kasem Ahmad, Assef Jafar Kadan Aljoumaa The	Customer churn prediction in telecom using machine learning in big data platform Abdelrahim Kasem Ahmad * , Assef Jafar and Kadan Aljoumaa The
W https://arxiv.org/pdf/1904.00690	

4/10 **SUBMITTED TEXT** 11 WORDS	**100%** **MATCHING TEXT** 11 WORDS
Developing a prediction model for customer churn from electronic banking services using data mining[2]	Developing a prediction model for customer churn from electronic banking services using data mining,"
W https://www.hindawi.com/journals/mpe/2022/8534739/	

5/10 SUBMITTED TEXT 10 WORDS	76% MATCHING TEXT 10 WORDS
The use of knowledge extrac- tion in predicting customer churn in B2B[3]	The use of knowledge extraction in predicting customer churn in B2B,"

W https://www.hindawi.com/journals/mpe/2022/8534739/

6/10 SUBMITTED TEXT 11 WORDS	71% MATCHING TEXT 11 WORDS
Survey on Customer Churn Prediction Using Machine Learning techniques [5] Saran Kumar A. Chandrakala D.	Survey on Customer Churn Prediction using Machine Learning Techniques Article Full-text available • Nov 2016 • Saran Kumar A.Saran Kumar A. • Chandrakala D.

W https://www.researchgate.net/publication/344582269_Customer_Churn_Prediction_in_Telecom_Using_Mac ...

7/10 SUBMITTED TEXT 19 WORDS	55% MATCHING TEXT 19 WORDS
where the churned customer class is smaller than the active customer class, is a significant challenge in churn prediction.	where the churned customer classes are smaller than the active customer classes, as it is a major issue in churn prediction

W https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0191-6

8/10 SUBMITTED TEXT 21 WORDS	85% MATCHING TEXT 21 WORDS
A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform,"	A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform,"

W https://www.hindawi.com/journals/mpe/2022/8534739/

9/10 SUBMITTED TEXT 22 WORDS	100% MATCHING TEXT 22 WORDS
A. Keramati, H. Ghaneei, and S. M. Mirmohammadi, "Developing a prediction model for customer churn from electronic banking services using data mining,"	A. Keramati, H. Ghaneei, and S. M. Mirmohammadi, "Developing a prediction model for customer churn from electronic banking services using data mining,"

W https://www.hindawi.com/journals/mpe/2022/8534739/

10/10 SUBMITTED TEXT 14 WORDS	100% MATCHING TEXT 14 WORDS
A. A. Jamjoom, "The use of knowledge extraction in predicting customer churn in b2b,"	A. A. Jamjoom, "The use of knowledge extraction in predicting customer churn in B2B,"

W https://www.hindawi.com/journals/mpe/2022/8534739/