BUSINESS GROWTH THROUGH CUSTOMER RETENTION

A PROJECT REPORT

Submitted by

S AJANEESHWAR (REG no:811721243005)

A KAMALESHWAR(REG no:8117212430023)

R SIDDHARTHAN(REG no:811721243052)

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

in

ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

K. RAMAKRISHNAN COLLEGE OF TECHNOLOGY

(An Autonomous Institution, affiliated to Anna University Chennai and Approved by AICTE, New Delhi)

SAMAYAPURAM – 621 112

DECEMBER, 2023

K. RAMAKRISHNAN COLLEGE OF TECHNOLOGY

(AUTONOMOUS)

SAMAYAPURAM – 621 112

BONAFIDE CERTIFICATE

Certified that this project report titled **“BUSINESS GROWTH THROUGH CUSTOMER RETENTION”** is the bonafide work of **S AJANEESHWAR (REG no:811721243005) , A KAMALESHWAR (REG no :811721243023) , R SIDDHARTHAN (REG no:811721243052)** who carried out the project under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form part of any other project report or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Dr.T.AVUDAIAPPAN M.E., Ph.D.,

HEAD OF THE DEPARTMENT,

Department of AI,
K.Ramakrishnan College of Technology
(Autonomous),
Samayapuram – 621 112.

SIGNATURE

Dr.T.AVUDAIAPPAN M.E., Ph.D.,

HEAD OF THE DEPARTMENT,

Department of AI,
K.Ramakrishnan College of Technology
(Autonomous),
Samayapuram – 621 112.

Submitted for the viva-voce examination held on ………………

INTERNAL EXAMINER **EXTERNAL EXAMINER**

DECLARATION

We jointly declare that the project report on **“BUSINESS GROWTH THROUGH CUSTOMER RETENTION”** is the result of original work done by us and best of our knowledge, similar work has not been submitted to **“ANNA UNIVERSITY CHENNAI”** for the requirement of Degree of **BACHELOR OF TECHNOLOGY**. This project report is submitted on the partial fulfilment of the requirement of the award of **DEGREE OF BACHELOR OF TECHNOLOGY**.

Signature

____

AJANEESHWAR S

____

KAMALESHWAR A

____

SIDDHARTHAN R

Place : Samayapuram

Date :

ACKNOWLEDGEMENT

It is with great pride that we express our gratitude and in-debt to our institution "**K.Ramakrishnan College of Technology (Autonomous)**", for providing us with the opportunity to do this project.

We are glad to credit honourable chairman **Dr. K.RAMAKRISHNAN**, **B.E.,** for having provided for the facilities during the course of our study in college.

We would like to express our sincere thanks to our beloved Executive Director **Dr. S. KUPPUSAMY, MBA, Ph.D.,** for forwarding to our project and offering adequate duration in completing our project.

We would like to thank **Dr. N. VASUDEVAN, M.E., Ph.D.,** Principal, who gave opportunity to frame the project the full satisfaction.

We whole heartily thanks to **Dr.T. AVUDAIAPPAN, M.E., Ph.D.,** HEAD OF THE DEPARTMENT, **ARTIFICAL INTELLIGENCE** for providing his encourage pursuing this project.

I express my deep and sincere gratitude to my project guide **Dr.T. AVUDAIAPPAN, M.E.,Ph.D.,** HEAD OF THE DEPARTMENT, **ARTIFICAL INTELLIGENCE** for his incalculable suggestions, creativity, assistance and patience which motivated me to carry out the project successfully.

I render my sincere thanks to my Project Coordinator **Mr.R.ROSHAN JOSHUA M.E.,** and other staff members for providing valuable information during the course.I wish to express my special thanks to theofficials and Lab Technicians of our departments who rendered their help duringthe period of the work progress.

ABSTRACT

Customer churn is a major problem and one of the most important concerns for large companies. Due to the direct effect on the revenues of the companies, especially in the telecom field, companies are seeking to develop means to predict potential customer to churn. Therefore, finding factors that increase customer churn is important to take necessary actions to reduce this churn. The main contribution of our work is to develop a churn prediction model which assists telecom operators to predict customers who are most likely subject to churn. The model developed in this work uses machine learning techniques on big data platform and builds a new way of features' engineering and selection. In order to measure the performance of the model, the Area Under Curve (AUC) standard measure is adopted, and the AUC value obtained is 93.3%. Another main contribution is to use customer social network in the prediction model by extracting Social Network Analysis (SNA) features. The use of SNA enhanced the performance of the model from 84 to 93.3% against AUC standard. The model was prepared and tested through Spark environment by working on a large dataset created by transforming big raw data provided by SyriaTel telecom company. The dataset contained all customers' information over 9 months, and was used to train, test, and evaluate 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.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

CP	Churn prediction
EGBM	Enhanced gradient boosting machine
AUC-ROC	Area under the receiver operating characteristic curve
CNN	Convolutional neural network
CRM	Customer relationship management
CV	Cross validation
DT	Decision tree
GBM	Gradient boosting machine
GBT	Gradient boosted tree
GWO	Gray wolf optimizer
HEOMGA	Heterogeneous euclidean-overlap metric genetic algorithm
KNN	K-nearest neighbours
LR	Logistic regression
MH	Meta-heuristic
ML	Machine learning
MPSO	Modified particle swarm optimization
MVO	Multi-verse optimizer
NB	Naive Bayes
PSO	Particle swarm optimization
RBF	Radial basis functions
RF	Random forest
SVM	Support vector machine
SVMRBF	Support vector machine with RBF kernel

CHAPTER 1

INTRODUCTION

Telecommunications sector has become one of the main industries in developed countries. Technical progress and the increasing number of operators raised the level of competition. Companies are working hard to survive in this competitive market depending on multiple strategies. Tree main strategies have been proposed to generate more revenues: acquire new customers, upsell the existing customers, and increase the retention period of customers. However, comparing these strategies taking the value of return on investment (Roll) of each into account has shown that the third strategy is the most profitable strategy, proves that retaining an existing customer costs much lower than acquiring a new one, in addition to being considered much easier than the upselling strategy. To apply the third strategy, companies have to decrease the potential of customer's churn, known as "the customer movement from one provider to another". Customers' churn is a considerable concern in service sectors with high competitive services. On the other hand, predicting the customers who are likely to leave the company will represent potentially large additional revenue source if it is done in the early phase. Many research confirmed that machine learning technology is highly efficient to predict this situation. This technique is applied through learning from previous data. data used in this research contains all customers' information throughout nine months before baseline. volume of this dataset is about 70 Terabyte on HDFS "Hadoop Distributed File System", and has different data formats which are structured, semi-structured, and unstructured. data also comes very fast and needs a suitable big data platform to handle it dataset is aggregated to extract features for each customer. We built the social network of all the customers and calculated features like degree centrality measures, similarity values, and customer's network connectivity for each customer. SNA features made good enhancement in AUC results and that is due to the contribution of these features in giving more different information about the customers. We focused on evaluating and analysing the performance of a set of tree-based machine learning methods and algorithms for predicting churn in telecommunications companies. We have experimented a number of algorithms such as Decision Tree

Random Forest, Gradient Boost Machine Tree and XGBoost tree to build the predictive model of customer Churn after developing our data preparation, feature engineering, and feature selection methods. There are two telecom companies in Syria which are SyriaTel and MTN. SyriaTel company was interested in this fields of study because acquiring a new customer costs six times higher than the cost of retaining the customer likely to churn. The dataset provided by SyriaTel had many challenges, one of them was unbalance challenge, where the churn customers' class was very small compared to the active customers' class. We experimented three scenarios to deal with the unbalance problem which are oversampling, under sampling and without re-balancing. evaluation was performed using the Area under receiver operating characteristic curve "AUC" because it is generic and used in case of unbalanced datasets. Many previous attempts using the Data Warehouse system to decrease the churn rate in SyriaTel were applied. Data Warehouse aggregated some kind of telecom data like billing data, Calls/SMS/Internet, and complaints.

Data Mining techniques were applied on top of the Data Warehouse system, but the model failed to give high results using this data. In contrast, the data sources that are huge in size were ignored due to the complexity in dealing with them. Data Warehouse was not able to acquire, store, and process that huge amount of data at the same time. In addition, the data sources were from different types, and gathering them in Data Warehouse was a very hard process so that adding new features for Data Mining algorithms required a long time, high processing power, and more storage capacity. On the other hand, all these difficult processes in Data Warehouse are done easily using distributed processing provided by big data platform.

We believe that big data facilitated the process of feature engineering which is one of the most difficult and complex processes in building predictive models. By using the big data platform, we give the power to SyriaTel company to go farther with big data sources. In addition, the company becomes able to extract the Social Network Analysis features from a big scale social graph which is built from billions of edges (transactions) that connect millions of nodes (customers).

The hardware and the design of the big data platform illustrated in “Proposed churn method” section fit the need to compute these features regardless of their complexity on this big scale graph.The model also was evaluated using a new dataset and the impact of this system to thedecision to churn was tested. The model gave good results and was deployed to production.

1.1 PROBLEM STATEMENT:

Background:

In the highly competitive landscape of [industry], retaining customers is paramount for sustained business growth and profitability. Customer churn, the phenomenon where customers cease their association with a product or service, poses a significant challenge. Identifying and predicting potential churners in advance can empower the business to take proactive measures to retain customers, enhance customer satisfaction, and optimize marketing strategies.

Problem Definition:

Develop a robust churn prediction model to anticipate customer churn in [company name]'s [product/service] based on historical customer data. The goal is to identify patterns and indicators that precede customer churn, allowing for timely intervention and targeted retention efforts

KEY OBJECTIVES:

Data Collection and Exploration:

- ❖ Gather and preprocess relevant historical customer data, including but notlimited to usage patterns, transaction history, customer support interactions, and demographic information.
- ❖ Explore and analyze the data to identify trends, correlations, and potential features that may influence customer churn.

Feature Selection and Engineering:

- ❖ Select key features that are likely to have a significant impact on churn prediction. Engineer additional relevant features that could enhance the model's predictive capabilities.

Model Development:

- ❖ Choose appropriate machine learning algorithms for churn prediction (e.g., logistic regression, decision trees, random forests, or neural networks).
- ❖ Train the model on a subset of the data and validate its performance using another subset.

Model Evaluation:

- ❖ Evaluate the model's performance using metrics such as accuracy, precision, recall, and F1-score.
- ❖ Employ cross-validation techniques to ensure the model's generalizability.

Deployment and Integration:

- ❖ Deploy the trained model into [company name]'s existing infrastructure for real-time predictions.
- ❖ Integrate the model outputs into business processes to enable timely and targeted intervention strategies.

Monitoring and Optimization:

- Implement a monitoring system to track the model's performance over time.Continuously optimize the model based on new data and changing customer behavior patterns.

Success Criteria:

- Achieve a high level of accuracy in predicting customer churn.
- Demonstrate the model's effectiveness in reducing actual churn rates through proactive retention strategies.
- Ensure seamless integration of the model into [company name]'s operations.

Significance:

A successful churn prediction model will not only enhance customer retention efforts but also contribute to a more data-driven and customer-centric approach, ultimately fostering sustained growth and competitiveness in the market.

1.2 AIM:

This study focuses on the development of an advanced customer churn prediction model tailored for telecom operators, with specific emphasis on SyriaTel. By leveraging sophisticated machine learning techniques on a big data platform, our aim is to significantly improve the accuracy of churn prediction and offer actionable insights for reducing customer attrition.

A key aspect of our approach involves innovative feature engineering methods, designed to identify and incorporate pertinent factors that influence customer churn. Moreover, the incorporation of Social Network Analysis (SNA) features aims to enrich the predictive capabilities of the model.

Creates online communities where customers can connect, share information, and interact with the brand. Fosters a sense of belonging and loyalty among customers. Provides valuable insights into customer needs and preferences.

A central platform for storing and managing all customer data, including contact information, purchase history, interactions, and preferences. Enables segmentation and personalization efforts based on customer data. Provides a comprehensive view of each customer relationship, facilitating better communication and engagement.

1.3 OBJECTIVE:

This study endeavours to revolutionize the approach to customer churn prediction within the telecommunications sector, with a specific focus on SyriaTel. Employing cutting-edge machine learning techniques on a big data platform, our primary goal is to elevate the accuracy of churn prediction models. A central component of our methodology involves pioneering feature engineering methods, designed to identify and incorporate pivotal factors influencing customer churn. Moreover, we aim to push the boundaries of predictive modeling by integrating Social Network Analysis (SNA) features, recognizing the impact of social relationships on customer behaviour.

Through a rigorous comparison of four machine learning algorithms—Decision Tree, Random Forest, Gradient Boosted Machine Tree "GBM," and Extreme Gradient Boosting "XGBOOST"—we aim to pinpoint the most effective algorithm for precise churn prediction and subsequently optimize its parameters. Our model's performance will be evaluated using the Area Under Curve (AUC) standard, providing a robust metric for accuracy assessment.

Practical implementation on a nine-month dataset from SyriaTel will showcase the model's real-world applicability. By achieving these objectives, our study aspires to contribute a transformative predictive tool, empowering telecom operators to proactively address customer churn, optimize retention efforts, and fortify overall business sustainability.

CHAPTER 2

LITERATURE SURVEY

2.1 TITLE: CHURN PREDICTION IN TELECOM: A RANDOM FOREST MODEL ANALYSIS.

AUTHOR: Irfan Ullah, Basit Raza ,Saif Ul Islam ,and Sung Won Kim.

YEAR OF PUBLICATION : IEEE May 6, 2019.

ALGORITHM USED: Logistic Regression and K-Nearest Neighbour algorithm.

ABSTRACT:

In the telecom sector, a huge volume of data is being generated on a daily basis due to a vast client base. Decision makers and business analysts emphasized that attaining new customers is costlier than retaining the existing ones. Business analysts and customer relationship management (CRM) analyser need to know the reasons for churn customers, as well as, behaviour patterns from the existing churn customers data.This paper proposes a churn prediction model that uses classification, as well as, clustering techniques to identify the churn customers and provides the factors behind the churning of customers in the telecom sector. Feature selection is performed by using information gain and correlation attribute ranking filter .The proposed model first classifies churn customers data using classification algorithms, in which the Random Forest (RF) algorithm performed well with 88.63% correctly classified instances. Creating effective retention policies is an essential task of the CRM to prevent churners. After classification, the proposed model segments the churning customer's data by categorizing the churn customers in groups using cosine similarity to provide group-based retention offers. This paper also identified churn factors that are essential in determining the root causes of churn.

2.2 TITLE: BUSINESS INTELLIGENCE AND ANALYTICS:

FROM BIG DATA TO BIG IMPACT.

AUTHOR : Veda C. Storey ,Roger H. L. Chiang ,Hsinchun Chen.

ALGORITHM : Linear Regression and K-Nearest Neighbouring Algorithm.

YEAR OF PUBLICATION: IEEE December 2012.

ABSTRACT:

Business intelligence and analytics (BI&A) and the related field of big data analytics have become increasingly important in both the academic and the business communities over the past two decades. Industry studies have highlighted this significant development.

For example, based on a survey of over 4,000 information technology (IT) professionals from 93 countries and 25 industries, the IBM Tech Trends Report (2011) identified business analytics as one of the four major technology trends in the 2010s. In a survey of the state of business analytics by Bloomberg Businessweek (2011), 97 percent of companies with revenues exceeding $100 million were found to use some form of business analytics.

A report by the McKinsey Global Institute (Manyika et al. 2011) predicted that by 2018, the United States alone will face a short-age of 140,000 to 190,000 people with deep analytical skills, as well as a shortfall of 1.5 million data-savvy managers with the know-how to analyse big data to make effective decisions.

2.3 TITLE: PROFIT OPTIMIZING CHURN PREDICTION FOR LONG-TERM LOYAL CUSTOMER IN ONLINE GAMES.
AUTHOR : Eunjo Lee , Boram Kim , Sungwook Kang , Byungsoo Kang Yoonjae Jang ,and Huy Kang Kim.

ALGORITHM: Bayes algorithm , linear discriminant analysis , support vector machine.

YEAR OF PUBLICATION:IEEE October 8, 2018**.**

ABSTRACT:

To successfully operate online games, gaming companies are introducing the systematic customer relationship management model. Particularly, churn analysis is one of the most important issues, because preventing a customer from churning is often more cost-efficient than acquiring a new customer. Churn prediction models should, thus, consider maximizing not only accuracy but also the expected profit derived from the churn prevention. We thus, propose a churn prediction method for optimizing profit consisting of two main steps: first, selecting prediction target, second, tuning threshold of the model. In online games, the distribution of a user’s customer lifetime value is very biased that a few users contribute to most of the sales, and most of the churners are no-paying users. Consequently, it is cost-effective to focus on churn prediction to loyal customers who have sufficient benefits. Furthermore, it is more profitable to adjust the threshold of the prediction model so that the expected profit is maximized rather than maximizing the accuracy. We applied the proposed method to real-world online game service, Aion, one of the most popular online games in South Korea, and then show that our method has more cost-effectiveness than the prediction model for total users when the campaign cost and the conversion rate are considered.

CHAPTER 3

SYSTEM SPECIFICATION

3.1 SOFTWARE SYSTEM CONFIGURATION:

- ❖ Seamless integration between various software systems, such as CRM, marketing automation, and customer service platforms, ensures data consistency and eliminates data.
- ❖ This facilitates personalized communication and targeted marketing campaigns based on a complete customer profile. Automated data synchronization across systems reduces manual effort and improves operational efficiency.
- ❖ Advanced analytics and data segmentation tools enable businesses to segment customers based on demographics, behaviors, preferences, and purchase history.
- ❖ This allows for personalized communication, targeted promotions, and product recommendations tailored to individual customer needs, leading to higher engagement and satisfaction.
- ❖ AI-powered chatbots and virtual assistants provide personalized customer support and answer questions 24/7, enhancing customer experience and reducing churn.

3.1.1 AUTOMATION AND WORKFLOW:

- ❖ Automating repetitive tasks, such as email marketing, order confirmations, and appointment reminders, frees up employee time for focusing on high-value activities.
- ❖ Automated workflows improve efficiency and reduce manual errors, ensuring smooth customer service interactions and timely responses.
- ❖ A well-configured software system allows for customization to adapt to specific business needs and workflows.
- ❖ This ensures the system can accommodate changing customer expectations and support the implementation of new customer retention strategies.

- Flexible integrations with third-party applications further enhance functionality and enable businesses to leverage specialized tools for specific needs.
- Self-service portals empower customers to manage their accounts, track orders, and access support resources independently, increasing convenience and satisfaction.

3.2 SOFTWARE ANALYTICS:

- Login frequency, feature usage, engagement metrics, support tickets. Purchase history, subscription status, payment failures.Demographics, psychographics, communication preferences.
- Industry trends, competitor analysis, economic indicators. Integrate data from various sources and prepare it for analysis.
- Logistic regression, decision trees, neural networks to identify. patterns in churn behaviour.
- Create dashboards and reports to understand churn trends and risk factors. Proactive interventions like targeted offers, personalized communication, and improved customer service.
- Retaining existing customers is cheaper than acquiring new ones .Focus marketing and customer success efforts on high-risk customers.Insights into churn drivers inform product development, pricing strategies, and customer service initiatives.
- Accurate and complete data is essential for effective models. Choosing the right algorithm and optimizing its performance. Translating predictions into concrete retention strategies.Avoiding bias and discrimination in churn prediction models.

3.3 LIBRARIES:

- Scikit-learn: A versatile library for machine learning in Python, offering various tools for classification, regression.
- XGBoost: An efficient and scalable implementation of gradient boosting, often used for predictive modeling, including churn prediction.

- LightGBM: Another gradient boosting framework that is efficient and designed for distributed and efficient training.
- TensorFlow and Keras: Widely used for building neural networks, these libraries are powerful for complex models, including deep learningapproaches for churn prediction.
- PyTorch: Similar to TensorFlow, PyTorch is another deep learning library that provides flexibility for building and training neural networks.

3.3.1 EFFECTIVE COMMUNICATION:

- Clearly communicate to patrons how the library uses churn prediction to improve services and resources, emphasizing its focus on retention and engagement, not exclusion.
- Be transparent about data collection and usage: Inform patrons about what data is collected for churn prediction, how it's used, and who has access to it. Assure them of data privacy and security measures.
- Proactively address potential concerns about data privacy and algorithmic bias. Offer opt-out options for data collection or model inclusion.
- Use churn prediction models to segment patrons into different risk categories. This allows for targeted communication and their identified needs and interests.
- Don't just inform patrons about their churn risk; offer concrete steps they can take to re-engage with the library, such as recommending relevant resources, programs, or events.
- Continuously seek feedback from patrons on the effectiveness and appropriateness of communication methods and messages.
- Use feedback to refine communication strategies and ensure they remain relevant and resonate with patrons.intervention strategies.
- Tailor communication methods (email, SMS, in-person) and message content to individual patron preferences and risk levels.Highlight the specific benefits and value patrons receive by staying engaged with the library, addressing.

CHAPTER 4

SYSTEM ANALYSIS

4.1 EXISTING SYSTEM:

- Centralize customer data and interactions, allowing for personalized communication and targeted marketing campaigns.
- Track customer journey and identify areas for improvement in the customer experience.
- Automate routine tasks like follow-ups, reminders, and birthday greetings. Facilitate collaboration between different departments (sales, marketing, service) to deliver a unified customer experience.
- Reward repeat customers for their purchases, encouraging them to continue doing business with your company.Offer tiered rewards based on spending levels to incentivize increased engagement.Provide exclusive benefits like early access to new products, discounts, and special events.
- Utilize loyalty program data to gain valuable insights into customer preferences and purchase behaviour.
- Provide 24/7 support through chatbots and self-service portals, reducing customer wait times and improving satisfaction.Personalize responses based on customer history and preferences.Offer proactive support by sending notifications.
- About relevant updates, promotions, and offers.Analyze customer service interactions to identify areas for improvement and ensure consistent quality.
- Collect and analyze customer feedback through surveys, reviews, and social media interactions.Identify common pain points and address them proactively to improve customer experience.
- Communicate with customers who provide feedback, demonstrating that their input is valued .Use feedback data to inform product development, marketing campaigns, and customer service initiatives.

- Analyze customer data to gain valuable insights into their behaviour, preferences, and purchase history. Segment customers into different groups based on shared characteristics for targeted marketing campaigns.
- Predict future customer behavior and identify opportunities to increase engagement and retention.Measure the effectiveness of customer retention strategies and adjust them as needed.

4.1.1 DRAWBACKS:

- Implementing effective customer retention strategies requires significant investment in resources, technology, and training .Maintaining a consistent level of service and personalization across channels and touchpoints can be challenging .Building and maintaining strong customer relationships takes time and effort from employees at all levels.
- Focusing heavily on satisfying existing customers might lead to neglecting the needs of potential new customers .Acquisition costs might be lower than retention costs, impacting the overall marketing budget allocation .New customer acquisition can be essential for market expansion and revenue growth.
- Implementing excessive retention tactics, like over-personalized communication or frequent offers, can lead to customer fatigue and annoyance .Constant promotions and discounts can devalue the brand and diminish customer perceived value.Striking a balance between engagement and intrusiveness is crucial to maintaining customer loyalty.
- Quantifying the exact return on investment (ROI) of customer retention efforts can be challenging.Long-term benefits, like increased customer lifetime value, might not be readily apparent in short-term financial metrics.Attributing revenue to specific retention strategies can be difficult due to multiple influencing factors.

- Collecting and analyzing customer data for personalized experiences requires careful consideration of privacy concerns.Misusing customer data or failing to ensure its security can damage brand reputation and trust.Implementing ethical data practices is crucial to building long-term customer relationships.

4.2 PROPOSED SYSTEM:

The proposed system represents a groundbreaking approach to addressing the pervasive challenge of customer churn in the telecom sector. By amalgamating advanced machine learning techniques, notably the XGBOOST algorithm, with Social Network Analysis (SNA), this system aims to achieve unparalleled accuracy in predicting potential customer churn.

Leveraging a comprehensive dataset spanning 9 months from SyriaTel and operating within the Spark environment, the model undergoes rigorous testing and evaluation. The innovative feature engineering methods not only optimize traditional factors influencing churn but also introduce a novel dimension by incorporating insights from customer social networks.

The integration of SNA features significantly enhances the model's performance, elevating the Area Under Curve (AUC) value to an impressive 93.3%. This system holds the potential to transform how telecom operators approach customer retention, offering a holistic understanding of churn dynamics and empowering proactive, targeted retention strategies. As we implement this system in real-time, its seamless integration into existing operational frameworks ensures its practical applicability and heralds a new era of data-driven decision-making in the telecom industry.

4.2.1 ADVANTAGES:

- Retained customers spend more money over time. Studies show that existing customers are 50% more likely to try new products and spend 31% more than new customers.Retaining customers reduces acquisition costs. It costs 5-25 times more to acquire a new customer than to retain an existing one.
- Customer retention increases customer lifetime value (CLTV), which is the total revenue a customer generates throughout their relationship with the business. 5% increase in customer retention can lead to a 25-95% increase in profit.
- Satisfied customers are more likely to become loyal advocates for your brand.They will recommend your products or services to their friends, family, and colleagues, which can help you acquire new customers.Positive word-of-mouth recommendations can have a powerful impact on your brand reputation and sales.
- Customer churn is the rate at which customers stop doing business with you.High churn rates can significantly impact your revenue and profitability.Customer retention strategies help to reduce churn by improving customer satisfaction and loyalty.
- Retained customers provide valuable data and insights about your products, services, and brand. This data can help you improve your offerings, marketing campaigns, and customer service. By understanding your customers better, you can tailor your business to their needs and preferences.
- Retaining customers can lead to greater operational efficiency. Satisfied customers are less likely to contact customer service, which can free up resources for other tasks.Additionally, loyal customers are more likely to forgive mistakes and disruptions, which can improve overall operational effect.

CHAPTER 5

ARCHITECTURE

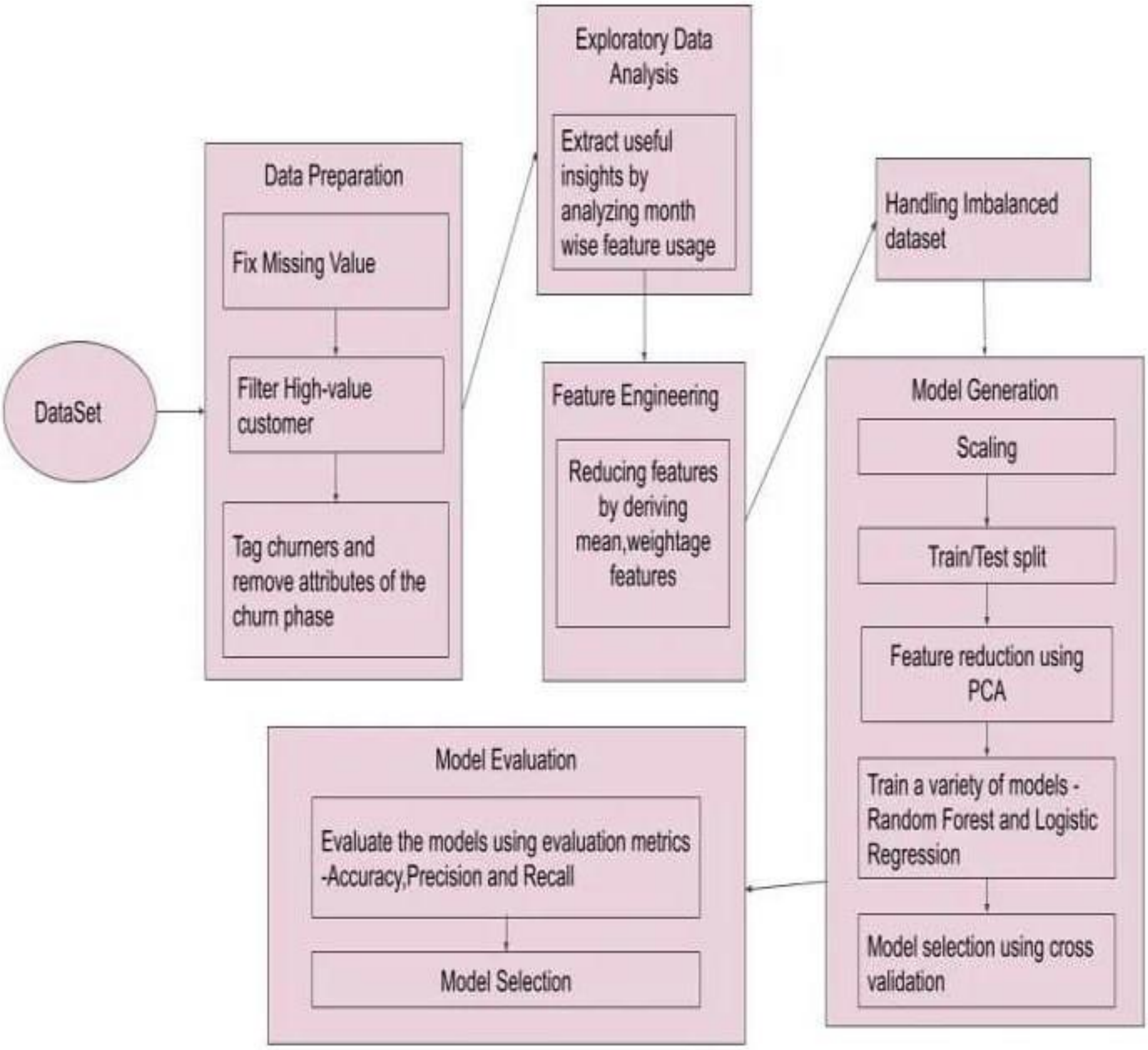

DataSet
Data Preparation
Fix Missing Value
Filter High-value customer
Tag churners and remove attributes of the churn phase
Exploratory Data Analysis
Extract useful insights by analyzing month wise feature usage
Feature Engineering
Reducing features by deriving mean,weightage features
Handling Imbalanced dataset
Model Generation
Scaling
Train/Test split
Feature reduction using PCA
Train a variety of models - Random Forest and Logistic Regression
Model selection using cross validation
Model Evaluation
Evaluate the models using evaluation metrics -Accuracy,Precision and Recall
Model Selection

FIG 5.1 WORKING PROCEDURE.

APPENDIX (SAMPLE CODE)

LOGISTIC REGRESSION:

```
log = LogisticRegression()
log.fit(X_train, y_train)
log_y_pred = log.predict(X_test)
log_y_pred_train = log.predict(X_train)
log_test_as = metrics.accuracy_score(log_y_pred, y_test)
log_train_as = metrics.accuracy_score(log_y_pred_train, y_train)
```

SUPPORT VECTOR CLASSIFIER:

```
y.pred_svc = svc.predict(X_test)
y_pred_train = svc.predict(X_train)
svc_train_as = metrics.accuracy_score(y_train, y_pred_train)
svc_as = metrics.accuracy_score(y_test, y_pred_svc)
from sklearn.metrics import roc_curve, auc
y_scores = svc_new.decision_function(X_train_sc)
fpr, tpr, thresholds = roc_curve(y_train, y_scores)
# Compute the area under the ROC curve (AUC)
roc_auc = auc(fpr, tpr)
plt.figure(figsize=(8, 8))
plt.plot(fpr, tpr, color='darkorange', lw=2, label=f'AUC = {roc_auc:.2f}')
plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic (ROC) Curve')
plt.legend(loc='lower right')
plt.show()
```

K-NEAREST NEIGHBOUR:

```
import numpy as np
from sklearn.neighbors import KNeighborsClassifier
from sklearn import metrics
X_train_np = np.array(X_train)
X_test_np = np.array(X_test)
y_train_np = np.array(y_train)
y_test_np = np.array(y_test)
testscores = []
trainscores = []
for i in range(1, 10):
 model = KNeighborsClassifier(i)
model.fit(X_train_np, y_train_np)
test_pred = model.predict(X_test_np)
train_pred = model.predict(X_train_np)
testscores.append(metrics.accuracy_score(y_test_np, test_pred))
trainscores.append(metrics.accuracy_score(y_train_np, train_pred))
knn = KNeighborsClassifier(8)
knn.fit(X_train, y_train)
X_test_array = X_test.values
X_train_array = X_train.values
y_pred_knn = knn.predict(X_test_array)
y_pred_knn_train = knn.predict(X_train_array)
```

DECISION TREE CLASSIFIER:

```
decision_tree = DecisionTreeClassifier()
decision_tree.fit(X_train, y_train)
y_pred_dt = decision_tree.predict(X_test)
y_pred_train_dt = decision_tree.predict(X_train)
t_as = metrics.accuracy_score(y_test, y_pred_dt)
dt_as_train = metrics.accuracy_score(y_train, y_pred_train_dt)
```

RANDOM FOREST CLASSIFIER:

```
decision_tree.predict random_forest =RandomForestClassifier()
random_forest.fit(X_train, y_train)
y_pred_rf = random_forest.predict(X_test)
y_pred_train_rf = random_forest.predict(X_train)
random_forest = RandomForestClassifier()
random_forest.fit(X_train, y_train)
y_pred_rf = random_forest.predict(X_test)
y_pred_train_rf = random_forest.predict(X_train)
```

XG BOOST:

```
xg = XGBClassifier()
xg.fit(X_train_sc, y_train)
y_pred_xg = xg.predict(X_test_sc)
y_pred_xg_train = xg.predict(X_train_sc)
xg_as = metrics.accuracy_score(y_test, y_pred_xg)
xg_as_train = metrics.accuracy_score(y_train, y_pred_xg_train)
```

GRADIENT BOOSTING CLASSIFIER:

```
grad_boost = GradientBoostingClassifier()

grad_boost.fit(X_train_sc, y_train)

y_pred_grad = grad_boost.predict(X_test_sc)

y_pred_grad_train = grad_boost.predict(X_train_sc)

grad_as = metrics.accuracy_score(y_test, y_pred_grad)

grad_as_train = metrics.accuracy_score(y_train, y_pred_grad_train)
```

5.1 CODE ACCURACY:

Logistic Regression results :

Accuracy score of test data : 0.8009478672985783.

Accuracy score of train data : 0.8038632986627043.

KNN results :

Accuracy score for test data : 0.7781128823782852.

Accuracy score for train data : 0.8106559116960306.

SVC result without parameter tunning :

Accuracy score for test data : 0.7970702283498492.

Accuracy score for train data : 0.8193589471449798.

SVC results with parameter tunning :

Accuracy score for test data : 0.7987936234381732.

Accuracy score for train data : 0.8102313733814477.

LGBM results without parameter importance :

LGBM accuracy score for test data 0.7850064627315813.

LGBM accuracy score for train data 0.9006580343876035.

LGBM result with feature importance :

LGBM accuracy score for test data 0.7819905213270142.

LGBM accuracy score for train data 0.855656973041817.

Decision Tree Classifier results with parameter importance :

Accuracy score for test data : 0.7242567858681602.

Accuracy score for train data : 0.9987263850562513.

Random Forest Classifier without parameter tunning :

Accuracy score for test data : 0.7841447651874193.

Accuracy score for train data : 0.9987263850562513.

Random Forest Classifier with parameter tunning :

Accuracy score for test data : 0.8035329599310642.

Accuracy score for train data : 0.8342177881553811.

XGBoost results without parameter tunning :

Accuracy score of test data : 0.7694959069366653.

Accuracy score of train data : 0.9622160900021227.

XGBoost results with parameter tunning :

Accuracy score of test data : 0.8065489013356312.

Accuracy score of train data : 0.8098068350668648.

Gradient Boosting Classifier results without parameter tunning :

Accuracy score of test data : 0.7983627746660922.

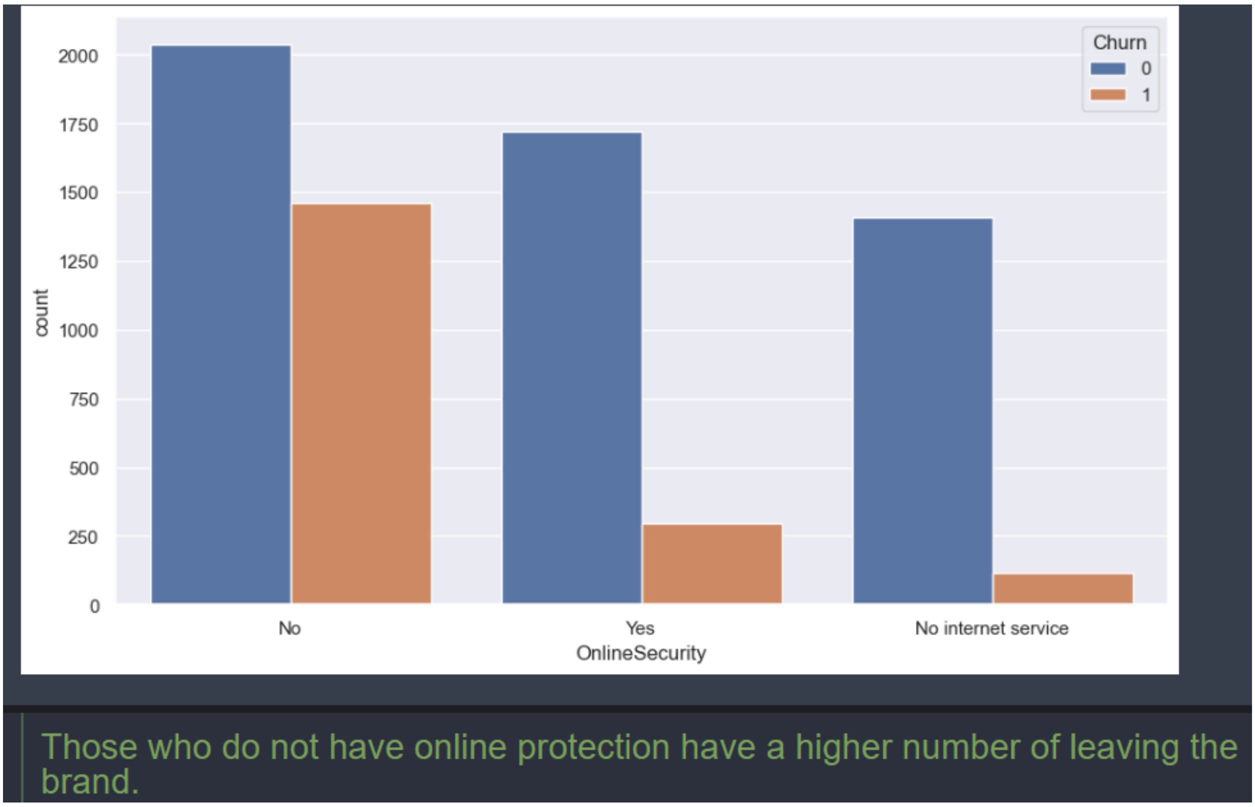
Accuracy score of train data : 0.8282742517512206.

Gradient Boosting Classifier results with parameter tunning :

Accuracy score of test data : 0.8048255062473072.

Accuracy score of train data : 0.8098068350668648.

5.2 SAMPLE OUTPUT:



Those who do not have online protection have a higher number of leaving the brand.

FIG NO 5.1.1: COUNT OF LEAVING THE BRAND.

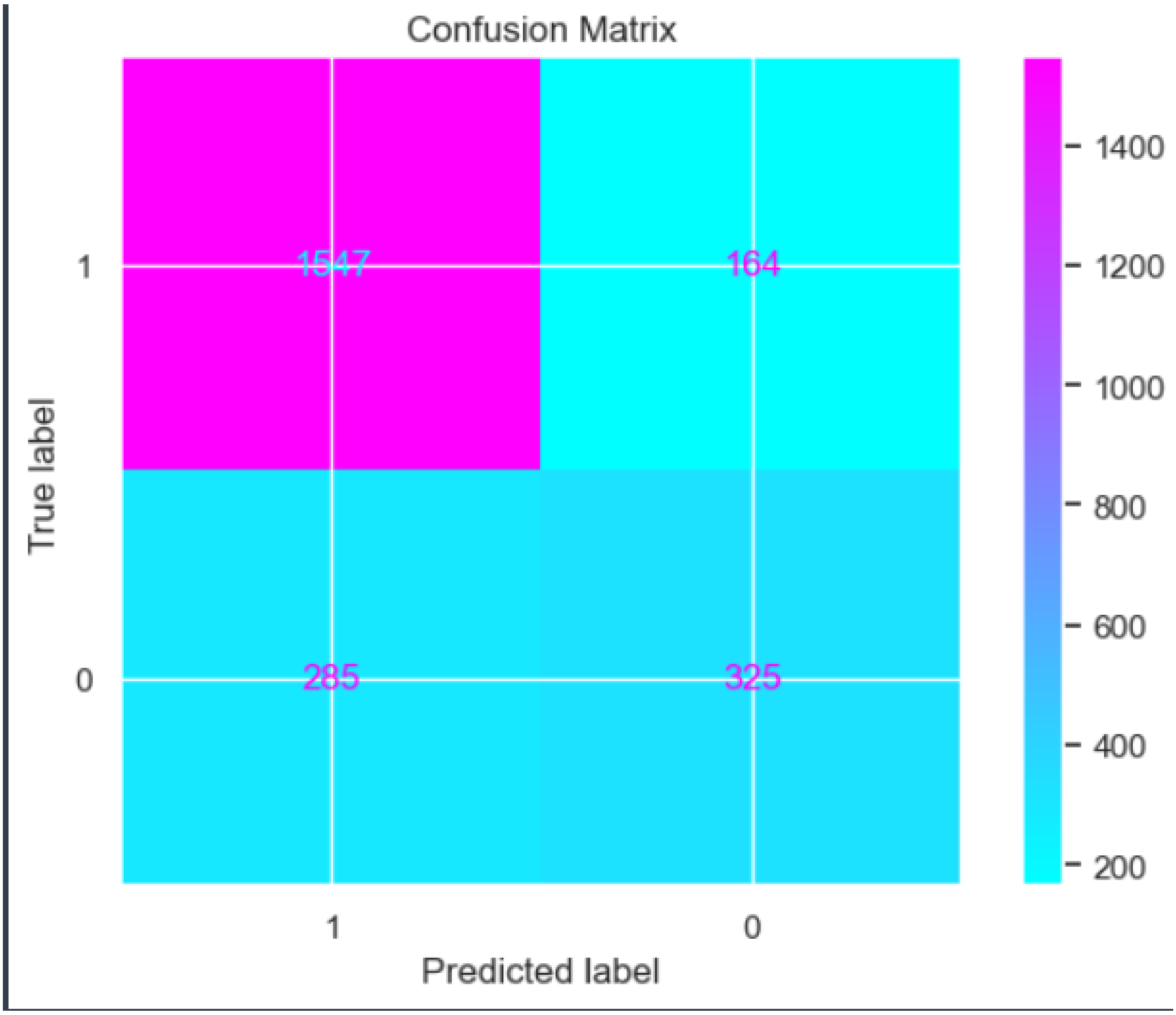


FIG NO 5.1.2: PREDICTION LABEL IN CONFUSION MATRIX.

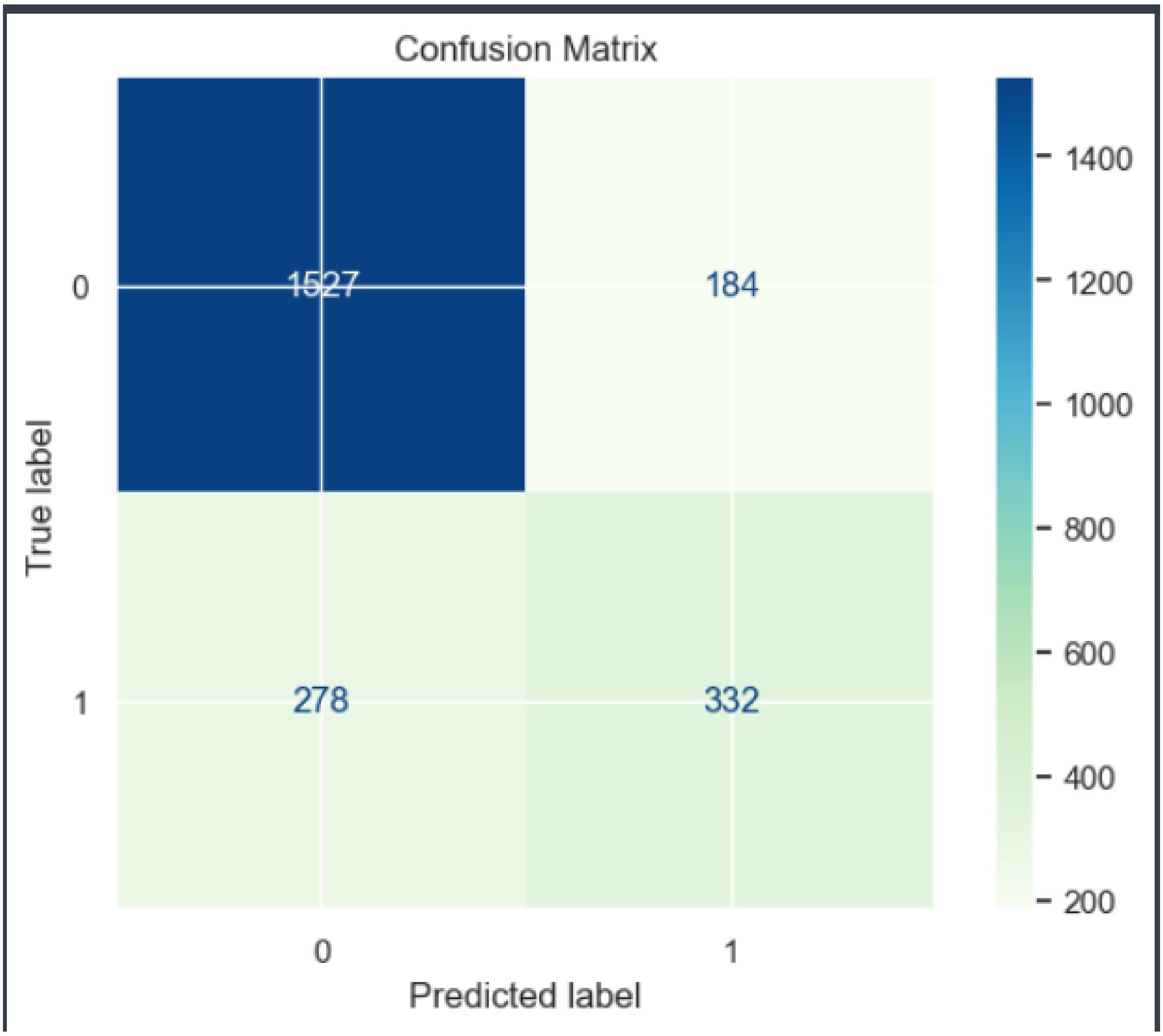


FIG NO 5.1.3: PREDICTION LABEL IN CONFUSION MATRIX.

CHAPTER 6

CONCLUSION

Telecom importance of this type of research in the telecom market is to help companies make more profit. It has become known that predicting churn is one of the most important sources of income to telecom companies. Hence, this research aimed to build a system that predicts the churn of customers in SyriaTel telecom company .These prediction models need to achieve high AUC values. To test and train the model, the sample data is divided into 70% for training and 30% for testing. We choose to perform cross-validation with 10-folds for validation and hyperparameter optimization. We have applied feature engineering, effective feature transformation and selection approach to make the features ready for machine learning algorithms. In addition, we encountered another problem: the data was not balanced. Only about 5%of the entries represent customers' churn. Tis problem was solved by under sampling or using trees algorithms not affected by this problem. Four tree based algorithms were chosen because of their diversity and applicability in this type of prediction. These algorithms are Decision Tree, Random Forest, GBM tree algorithm, and XGBOOST algorithm. The method of preparation and selection of features and entering the mobile social network features had the biggest impact on the success of this model, since the value of AUC in SyriaTel reached 93.301%. XGBOOST tree model achieved the best results in all measurements. The AUC value was 93.301%. The GBM algorithm comes in the second place and the random forest and Decision Tree came third and fourth regarding AUC values. We have evaluated the models by fitting a new dataset related to different periods and without any proactive action from marketing, XGBOOST also gave the best result with 89% AUC. The decrease in result could be due to the non-stationary data model phenomenon, so the model needs training each period of time. The use of the Social Network Analysis features enhance the results of predicting the churn in telecom.

REFERENCE

1.Irfan ullah,Basitraza,Saif ul islam and Sung won kim(2019)Churn Prediction in Telecom: A Random Forest Model Analysis.pg:1-17.

2.Veda C. storey,Roger H.L.Chiang ,Hsinchun chen(2012)Business intelligence and analytics: from big data to big impact.pg:5-12.

3.Eunjo lee , Boram kim , Sungwook kang , Byungsoo kang Yoonjae jang ,and Huy kang kim(2018)Profit Optimizing Churn Prediction for Long-Term Loyal Customers in Online Games.pg:2-16.

4.Jaehyun ahn,Junsik hwang, Doyoung kim, Hyukgeun choi and Shinjin kang (2020)A Survey on Churn Analysis in Various Business Domains.pg:4-8.

5.Nien ting lee,Hau chen lee,Joseph hsin and Shih hau fang(2022)Prediction of customer behavior changing via a hybrid approach.pg:3-16.

6.Taek Lee, Jaechang Nam, Donggyun Han, Sunghun Kim and Hoh Peter(2016) Developer micro interaction metrics for software defect prediction.pg:5-19.

7.Noman ahmad,Mazhar javed awan, Haitham nobanee,Azlan mohd zain,Ansar naseem,Amena Mahmoud(2017)Customer Personality Analysis for Churn Prediction Using Hybrid Ensemble Models and Class Balancing Techniques.pg:12-18.

8.Soumi De and P.Prabu(2022)A Sampling-Based Stack Framework for Imbalanced Learning in Churn Prediction.pg:10-22.

9.Soumi De and P. Prabu(2021)A Representation-Based Query Strategy to Derive Qualitative Features for Improved Churn Prediction.pg:1-15.

10.Adnan amin,Sajid anwar, Awais adnan,Muhammad Nawaz,Newton howard, Junaid qadir,Ahmad hawalah and Amir hussain(2016)Comparing Oversampling Techniques to Handle the Class Imbalance Problem:A Customer Churn Prediction Case Study.pg:10-27.