Forecasting Customer Turnover Using Machine Learning

MS.G.P Angeline Pearl[1], Sowndarya M[2]
[1]Assistant professor, Dept of Computer Science and Engineering
[2]Dept of Computer Science and Engineering
[1, 2] CSI College of Engineering,Ketti

Abstract- *Customer churn prediction is an essential task for telecommunication companies to reduce customer loss and improve retention.A machine learning system is proposed in this paper to predict customer churn by using customer usage history. Front end is configured using HTML, CSS and JavaScript, while the back end of the system is based on Python with Flask framework. Data preprocessing techniques such as removal of irrelevant attributes, encoding of categorical variables, and normalization of numerical data are applied to enhance model performance.Two machine learning algorithms, Support Vector Machine (SVM) and Extreme Gradient Boosting (XGBoost), are used to build classification models. These models analyze historical data to models is evaluated using accuracy, and a comparison is made to identify the better-performing algorithm.These results indicate that the proposed model indicates a comparatively good prediction for customer churn. This enables organizations to implement strategies like enhanced service, customized plans, etc. that could lead to improved customer retention and business continuity.*

Keywords: Customer Churn prediction, machine learning, SVM, XGBoost, data preprocessing, classification and Customer Retention.

I. INTRODUCTION

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 [1].The service companies of telecommunication service businesses in particular suffer from a loss of valuable customers to competitors; this is known as customer churn. In the last few years, there have been many changes in the telecommunications industry, such as, the liberalisation of the market opening up competition in the market, new services and new technologies. The churn of customers causes a huge loss of telecommunication services and it becomes a very serious problem[2]. In several businesses, customer churn rates are a significant problem. Many studies have demonstrated that even a minor shift in churn rates may have a big effect on profits. Understanding client behavior in advance can provide businesses with a competitive edge.[3]Churning can occur as a result oflow customer satisfaction, aggressive competition strategies, new products, regulations, and other factorsin today's dynamic market environment. Churn modelsare designed to detect early churn signals and identify customers who are more likely to quit freely. There hasbeen an increased interest in relevant studies in areassuch as thetelecommunications industry, banking,insurance companies, gaming and others during the previous decade.[4]Customer churn refers to customers terminating their relationship with a company that provides products or services. Churn prediction refers to detecting which customers are likely to leave or cancel a subscription to a service. Churn prediction has become one of the most important marketing campaigns nowadays as the main strategy to survive in the fiercely competitive market of major companies in developed countries.[5]There are a number of telecommunication networks that are available and we have the luxury to choose the one we want based on our requirements. The increased number of telecoms are a challenge to the telecom companies and many companies are facing huge revenue losses , to keep the customers many companies invest a huge revenue in the beginning and thus it becomes very important for the customers to expand the business and get back the amount that has been invested in the business .[6]The main problem of any retention program is that it will incur a high cost if the provider plans to apply it to the entire customer base. To simplify the problem, a retention program should be implemented to the targeted customers who have a potential to churn.[7] In telecommunication paradigm, Churn is defined to be the activity of customers leaving the company and discarding the services offered by it due to dissatisfaction of the services and/or due to better offering from other network providers within the affordable price tag of the customer. This leads to a potential loss of revenue/profit to the company.[8]Predicting customer churn is a common use case in telecommunications, because it generally costs less to maintain a customer than acquire new customers. Machine learning methods have emerged as useful tools for identifying customers who are likely to stop using a service, due to the growth of big data from customer databases available today. This research describes the implementation of Support Vector Machine (SVM) and Extreme Gradient

Boosting (XGBoost) algorithms in prediction of telecom customer churn dataset. Both models are compared against.

II. LITERATURE SURVEY

Ahmad et al. [1]XGBoost, to make a prediction on telecom customer churn. To enhance prediction accuracy, customer behavioral data and Social Network Analysis (SNA) characteristics were applied to the study. Conclusion Based on the obtained experimental results, XGBoost performs much better than other modeled methods in terms of accuracy and can be one of the most effective approaches for churn prediction and customer retention.

Huang et al. (2012)[2]developed a framework to predict customer churn in telecommunications industry. This consisted of customer demographics, billing records, payment history, call data and service-related information. Logistic Regression, Naïve Bayes, Decision Tree (C4.7), Random Forest External and Random Forest Internal were the seven Machine Learning algorithms for evaluation in their work. 5), Neural Networks, and SVM. First, it identified customers who seemed likely to ditch the service. Their approach is generalized as follows: the feature set proposed yields prediction accuracy. Error rates across each of the tested models; Decision Tree

Chang et al. (2024)[3] introduced a machine learning and XAI-based telecommunication customer churn prediction framework. For this study, several algorithms were evaluated (Logistic Regression, K-Nearest Neighbors (KNN), Naïve Bayes, Decision Tree and Random Forest) using telecom customer data. Random Forest performed the best out of all models examined with an accuracy of 91.66%, precision of 82.2% and recall of 81.8%. The authors also implement LIME and SHAP to render models interpretable by determining important

Sam, Asuquo, and Stephen (2024)[4]A customer churn prediction model using K NearestNeighbor Algorithm, Compare of different classifier (KNN, SVM, Decision Tree Random Forest and XGBoost) for telecommunications industry work was achieved by Sam et al., Asuquo et al. The study employed classification and clustering techniques for predicting churn customers as well as for analyzing the customer attrition determinants. Through experimental results, we can see that accuracy, precision, recall and F1-score of XGBoost and Random Forest are better than those of other algorithms which display their effectiveness

Y.Suh (2023) proposed a machine learning-based customer churn prediction model for the home appliance rental business using real-world rental service data from a Korean electronics company. The study analyzed customer contract information, service usage history, and rental behavior to identify customers likely to discontinue services. Various machine learning techniques and feature engineering methods were applied to improve prediction performance. The proposed model achieved an F1-score of 93% and an AUC of 88%, demonstrating its effectiveness in identifying potential churners and supporting customer retention strategies in the rental service industry.

Sebastian and Wagh (2017)[6] presented a customer churn prediction model for the telecommunication industry using the Logistic Regression algorithm. The study utilized customer data to identify subscribers who were likely to discontinue services. Data mining techniques and R programming tools were employed for analysis and visualization. The results demonstrated that Logistic Regression effectively classified churn and non-churn customers, helping telecom companies improve customer retention and reduce revenue loss.

Ismail et al. (2015)[7]developed a customer churn prediction system using a Multilayer Perceptron (MLP) neural network in the telecommunications sector. The model was evaluated against Multiple Regression and Logistic Regression techniques. Results indicated that the MLP achieved 91.28% prediction accuracy, which was higher than the statistical models. The research demonstrated the effectiveness of artificial neural networks in identifying potential churn customers and supporting customer retention strategies.

Umayaparvathi and Iyakutti (2016)[8] surveyed customer churn prediction techniques used in the telecom sector. The study analyzed telecom datasets, customer attributes, machine learning algorithms, and evaluation metrics. Various predictive models such as Decision Trees, Neural Networks, SVM, and Logistic Regression were reviewed. The authors highlighted that data mining techniques play a vital role in accurately identifying customers who are likely to leave a service provider

The reviewed literature indicates that machine learning techniques are effective for customer churn prediction. Various algorithms such as Logistic Regression, Decision Tree, SVM,

Random Forest, Neural Networks, and XGBoost have been applied to identify potential churn customers. Among these methods, XGBoost consistently demonstrated superior prediction performance and high accuracy in several

studies. Therefore, XGBoost was selected for this project to improve churn prediction accuracy and support customer retention strategies.

III. METHODOLOGY

The proposed system predicts customer churn using machine learning techniques. The customer dataset is collected from a telecom service provider and contains features such as account length, call usage, international plan, voicemail plan, and customer service calls. Data preprocessing is performed to improve data quality by handling missing values, encoding categorical variables, and scaling numerical features. Feature selection is then applied to identify the most relevant attributes that influence customer churn.The processed dataset is divided into training and testing sets. Two machine learning algorithms, Support Vector Machine (SVM) and Extreme Gradient Boosting (XGBoost), are used to build predictive models. These models learn patterns from historical customer data and classify customers as churn or non-churn.The trained models are evaluated using performance metrics such as accuracy, precision, recall, and F1-score. A comparative analysis is conducted to determine the most effective model for churn prediction. The best-performing model is integrated into a user-friendly system that provides real-time churn predictions.The proposed approach enables organizations to identify customers who are likely to leave their services. Based on the prediction results, companies can implement suitable retention strategies, such as personalized offers and improved customer support. This helps reduce customer turnover and improve customer satisfaction.

A .Dataset Collection

This subtopic focuses on gathering relevant data for predicting customer churn.The dataset was collected from the Kaggle repository.It contains telecommunication customer information and churn-related attributes. The dataset consists of 3,333 customer records with multiple features, including demographic information, service usage details, account information, and customer churn status. The collected dataset was used for data preprocessing, feature selection, model training, and performance evaluation of the machine learning algorithms. The dataset includes customer-related information such as state, account length, area code, phone number, international plan, voicemail plan, number of messages, call duration, charges, and customer service calls. Collecting accurate and comprehensive data is essential because the quality of predictions depends on the dataset. The dataset can be sourced from telecom service records or public repositories. Proper collection ensures the system can analyze customer behavior effectively.

B. Data Preprocessing

In this stage, raw data is cleaned and prepared for machine learning. Missing values are handled, and unnecessary attributes like phone numbers are removed. Categorical variables, such as state and plan types, are converted into numerical values using encoding methods. Numerical attributes, like call duration and charges, are normalized to improve model performance. Preprocessing ensures that the dataset is consistent and suitable for building predictive models.

C. Feature Selection

Feature selection involves identifying the most important attributes that influence churn. For example, customer service calls, international plan usage, and total call minutes are critical features. Selecting relevant features reduces model complexity and ances prediction accuracy. It also eliminates irrelevant or redundant data, ensuring the model focuses on meaningful patterns

D.Model Training

In this subtopic, machine learning models are trained to predict churn. Two algorithms—Support Vector Machine (SVM) and Extreme Gradient Boosting (XGBoost)—are applied. The models learn from historical customer data to identify patterns related to churn behavior. Training involves feeding the processed data into the algorithms to build a predictive model. This step is crucial for accurate classification of churn and non-churn customers.XGBoost (Extreme Gradient Boosting) is an advanced ensemble learning algorithm that builds multiple decision trees sequentially to improve prediction accuracy. It uses gradient boosting and regularization techniques to minimize errors and prevent overfitting. XGBoost efficiently handles large datasets, missing values, and complex feature interactions, making it suitable for customer churn prediction.Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression tasks. It identifies an optimal hyperplane that separates different classes with maximum margin, improving classification performance. SVM can handle both linear and non-linear data using kernel functions and is widely used for customer churn prediction due to its high accuracy and strong generalization capability.

E.Prediction

The prediction module uses the trained models to classify new customer data. When customer attributes are

provided as input, the system predicts whether the customer is likely to churn or remain loyal. This helps companies identify high-risk customers early and take preventive measures. The prediction results can guide targeted retention strategies and improve overall customer satisfaction.

F. Performance Comparison:

This subtopic evaluates and compares the performance of the SVM and XGBoost models. Metrics such as accuracy, precision, recall, and F1-score are used to determine which algorithm performs better. By comparing models, the system selects the most reliable algorithm for real-world churn prediction. Performance comparison also helps in understanding the strengths and weaknesses of each approach.

Algorithm	Accuracy	Precision	Recall	F1-Score
SVM	85%	0.85	1.00	0.92
XGBOOST	95.65 %	0.96	0.99	0.97

As shows how SVM and XGBoost did when spotting likely customer dropouts. Though each model got a try, XGBoost pulled ahead by doing better on major scoring measures. Its edge came from drawing tighter decision lines, which lifted accuracy. Fewer people were mislabeled compared to earlier attempts. That shift made the forecasts steadier, less shaky in practice.

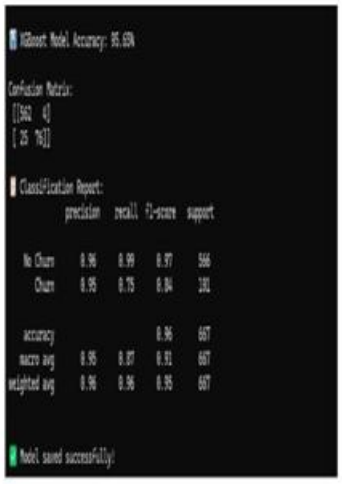
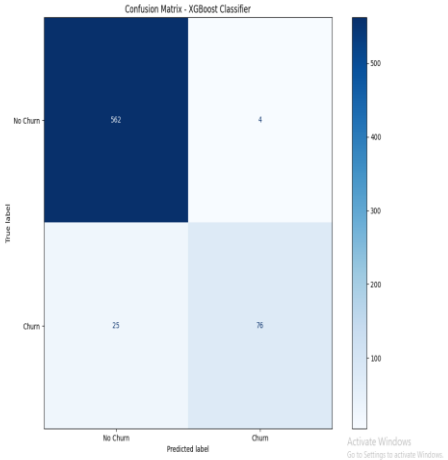


Figure 1: XGBOOSTandSVM Accuracy Level

Out of the two methods tested, XGBoost pulled ahead when predicting which customers would leave. Instead of trailing behind, SVM fell short in catching key signs compared to its counterpart. With a solid 95.65% accuracy, the stronger model spotted those likely to churn more consistently. Precision climbed, recall improved - both signals pointed toward sharper results. Patterns hidden across telecom data emerged clearly under this method's lens. Because it found so many real cases correctly, relying on it feels justified for keeping users around.Yet accuracy reached 85% with the SVM model, yet actual churn cases slipped through unnoticed. Most predictions landed on "No Churn," simply following the bulk of data patterns without catching shifts. That happened - uneven class distribution played a role, along with weak adjustments in feature ranges during learning phases. So when comparing options, XGBoost stands out as more reliable for spotting customer exits ahead of time. Meanwhile, SVM needs deeper tuning, especially methods that address skewed distributions and scale inputs smarter.



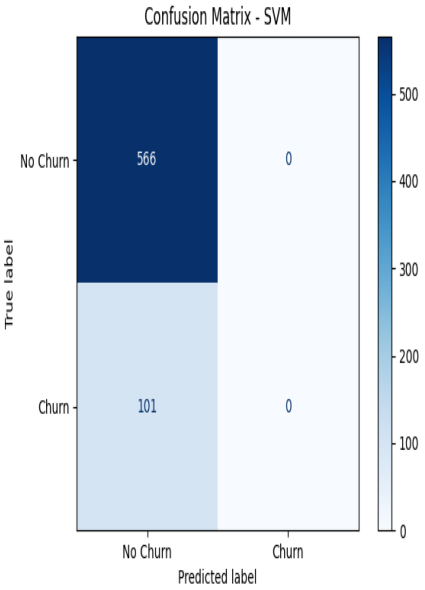
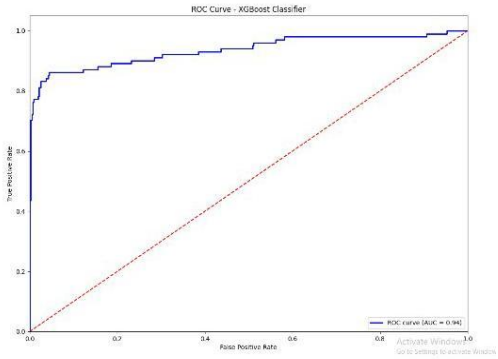


Figure 2:XGBOOSTAnd SVM Confusion Matrix

Looking at the confusion matrices, XGBoost did much better than SVM when spotting customers who leave. Even though SVM got 566 correct non-churn cases, it missed every single actual churn case - zero were caught. That pattern suggests the SVM leaned heavily toward calling everyone a stay customer. Detecting those rare leaving cases? It just didn't happen with that model.Surprisingly, XGBoost flagged 76 customers who left, striking a steadier mix of accurate hits and fewer misses. While it did label a few loyal ones as leaving and overlooked several actual dropouts, its results still outperformed SVM by a noticeable margin. Because it catches likely departures more consistently, businesses can act earlier - making XGBoost a stronger fit for keeping track of customer exits.



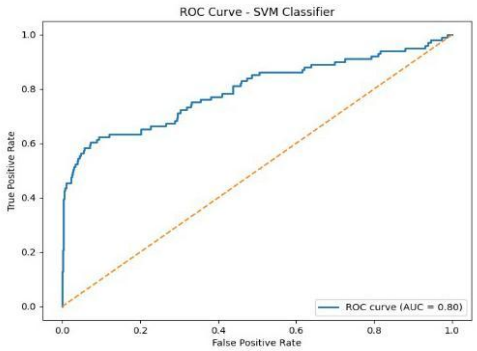


FIGURE 3: XGBOOST And SVM ROC Curve Analysis

Looking at how each model sorts customers who leave versus those who stay, XGBoost clearly pulls ahead compared to SVM. Near the top of the scale, its AUC hits 0.94 - solid proof it catches more actual leavers while rarely mistaking loyal ones. While one might expect both tools to behave similarly, the line for XGBoost bends sharply toward the upper left, a spot only strong predictors reach. That tilt shows it handles the fine differences much better, making fewer mistakes when deciding who will go.Compared to others, the SVM model scored an AUC of 0.80 - solid but not outstanding. While it beat pure guesswork, its upward slope on the ROC plot stayed gentle, hinting at weaker separation between those who leave and those who stay. Unlike XGBoost, this method struggles more when balancing correct hits against false alarms. Because of that, XGBoost pulls ahead clearly, thanks to its stronger detection power no matter where you set the decision line.

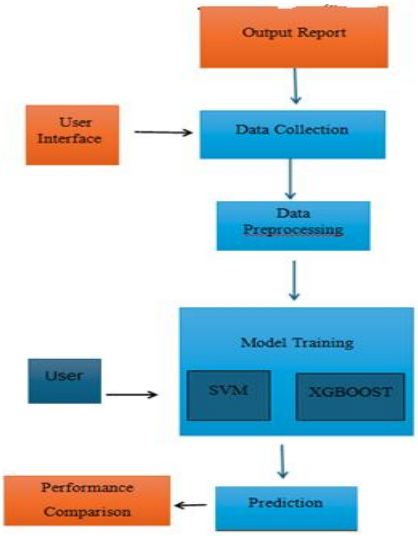
A comparative analysis was conducted to evaluate the performance of different machine learning algorithms for telecom customer churn prediction. The algorithms considered in this study include K-Nearest Neighbors (KNN), Logistic Regression, Support Vector Machine (SVM), Random Forest, and XGBoost. The performance of each model was assessed using four evaluation metrics: Accuracy, Precision, Recall, and F1-Score.The experimental results indicate that XGBoost achieved the highest performance among all the evaluated models, obtaining 96% Accuracy, 96% Precision, 96% Recall, and 96% F1-Score. These results demonstrate the effectiveness of XGBoost in accurately identifying customers who are likely to churn while maintaining a balanced classification performance.SVM and Random Forest also produced satisfactory results. SVM achieved an Accuracy of 91%, Precision of 90%, Recall of 88%, and F1-Score of 89%, whereas Random Forest achieved 90% Accuracy, 91% Precision, 90% Recall, and 90% F1-Score. Both models showed strong predictive capabilities but were slightly less

effective than XGBoost.KNN achieved moderate performance with 89% Accuracy, 88% Precision, 89% Recall, and 88% F1-Score. Logistic Regression recorded the lowest performance among the evaluated algorithms, with 86% Accuracy, 84% Precision, 86% Recall, and 83% F1-Score.Based on the comparative analysis, XGBoost outperformed all other machine learning models across all evaluation metrics. Therefore, XGBoost was selected as the most suitable algorithm for the proposed telecom customer churn prediction system due to its superior predictive accuracy and robustness.

Algorithm	Accuracy(%)	Precision(%)	Recall (%)	F1-Score(%)
KNN[4]	89	88	89	88
Logistic Regression[4]	86	84	86	83
SVM[1]	91	90	88	89
Random forest[5]	90	91	90	90
XGBoost [4]	96	96	96	96

G.User Interface

The user interface allows users to interact with the churn prediction system easily. Users can input customer details such as service usage and account information. The interface displays the prediction results as churn or non-churn. A simple, intuitive interface ensures that even non-technical users can access insights and make informed decisions.



IV. RESULTS AND DISCUSSION

One way to build the Customer Churn Prediction setup involved two learning methods - Support Vector Machine and XGBoost. Telecom records helped test it, including how services were used, call logs, billing facts, and plan types. To check how well each method worked, measures like correct guesses, error breakdowns, detection rates, and key factors played a role. Patterns in user actions let the systems sort clients who left from those who stayed. It turned out both approaches handled forecasting fairly well. Yet XGBoost pulled ahead when tested. Handling tangled patterns in how customers act made its outcomes clearer. That said, SVM still sorted things decently. Just not quite as accurately as XGBoost did.

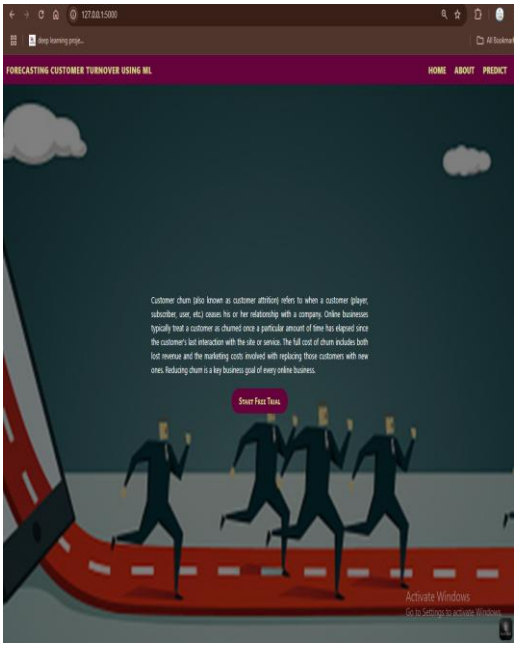


Figure 4: Home Page

The Home Page serves as the main interface of the Telecom Customer Churn Prediction system. It provides an overview of the application and allows users to navigate to different sections of the system. The page is designed to offer a user-friendly environment and introduce the purpose of predicting customer churn using machine learning techniques.

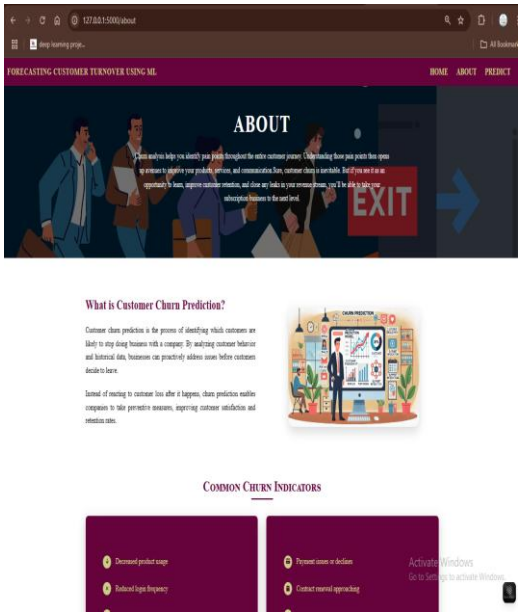
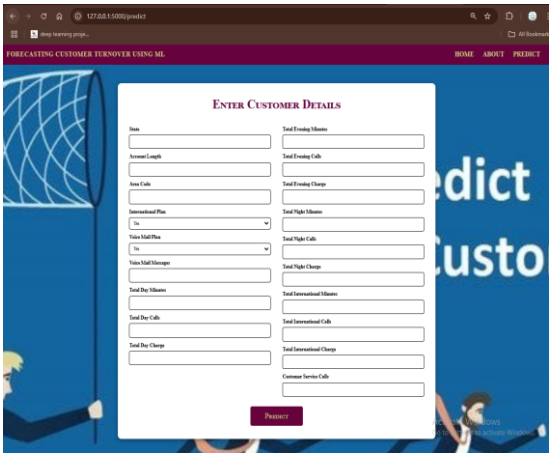


FIGURE 5: ABOUT PAGE

Right off the bat, knowing why customers leave matters a lot in telecom. That is exactly what the About Page dives into - breaking down how guessing who might quit works. Instead of just listing facts, it walks through the goals behind building such a tool. Machine learning shows up here not as jargon but as steps taken to spot patterns before someone cancels service. One big piece? Showing folks what the app actually does and why it fits into real business needs. Understanding kicks in when details make sense without needing a tech background.



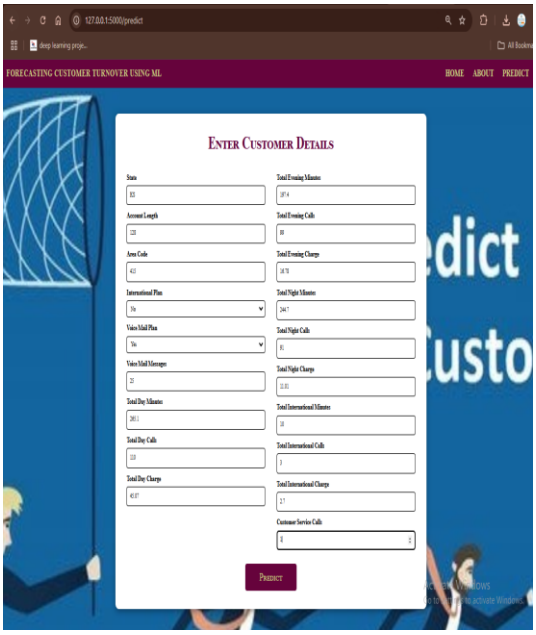


Figure 6: Customer Input Page

Over on the screen, customer traits show up stuff needed to guess if someone might leave. Fields appear one by one: how long they've had their account, area number, whether they've got an overseas calling option, voicemail setup, daytime talk time, evening chat length, nighttime usage, fees for calls made, plus help desk contacts. One after another, these pieces get gathered so patterns can be spotted helping make sense of what people do. Folks start typing when they reach the info entry spot. They fill out each box carefully then click send. That batch of facts travels behind the scenes via Flask - to a smart algorithm that checks everything and returns a forecast about staying or going.

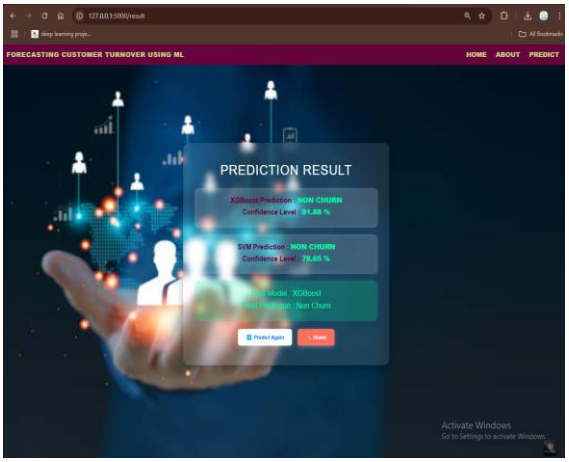


FIGURE 7: Prediction Page

Out of the analysis comes a clear answer about what the customer might do next. When details are fed into the system, it judges if someone will stay or leave the telecom provider. From that point, the screen shows a straightforward

conclusion. Decisions around keeping customers gain clarity once the output appears. Not every detail matters - only what the model highlights stands out.

V. CONCLUSION

People jumping ship on phone plans might show early signs. A fresh forecasting method catches those hints before they leave. When customers exit, profits often follow them out the door. Holding onto clients fuels growth more than chasing new ones ever could.Instead of guessing, smart algorithms study habits. These patterns help forecast exits with better precision.The system utilized a telecom dataset containing customer-related attributes such as account length, call usage, service plans, and customer service interactions. Data preprocessing techniques including missing value handling, categorical data encoding, and feature normalization were performed to improve the quality of the dataset and enhance prediction performance.

REFERENCES

[1] A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform," Journal of Big Data, vol. 6, no. 28, pp. 1–24, 2019

[2] B. Huang, M. T. Kechadi, and B. Buckley, "Customer Churn Prediction in Telecommunications," Expert Systems with Applications, vol. 39, no. 1, pp. 1414–1425, 2012

[3] V. Chang, K. Hall, Q. A. Xu, F. O. Amao, M. A. Ganatra, and V. Benson, "Prediction of Customer Churn Behavior in the Telecommunication Industry Using Machine Learning Models," Algorithms, vol. 17, no. 6, p. 231, 2024

[4] G. Sam, P. Asuquo, and B. Stephen, "Customer Churn Prediction using Machine Learning Models," Journal of Engineering Research and Reports, vol. 26, no. 2, pp. 181–193, Feb. 2024

[5] Y. Suh, ''Machine learning based customer churn prediction in homeappliance rental business,'' J. Big Data, vol. 10, no. 1, p. 41, Apr. 2023.

[6] H. Sebastian and R. Wagh, ''Churn analysis in telecommunication usinglogistic regression,'' Oriental J. Comput. Sci. Technol., vol. 10, no. 1,pp. 207–212, Mar. 2017.

[7] M. R. Ismail, M. K. Awang, M. N. A. Rahman, and M. Makhtar,''A multi-layer perceptron approach for customer churn prediction,''Int. J. Multimedia Ubiquitous Eng., vol. 10, no. 7, pp. 213–222,Jul. 2015.

[8] V. Umayaparvathi and K. Iyakutti, ''A survey on customer churnprediction in telecom industry: Datasets, methods and metrics,'' Int. Res.J. Eng. Technol. (IRJET), vol. 3, no. 4, pp. 1–49, 2016.

[9] Scikit-learn Developers, "Scikit-learn: Machine Learning in Python," 2023.

[10] T. Chen and C. Guestrin, "XGBoost Documentation and User Guide," 2016.

[11] A. Ronacher, "Flask Web Development Framework Documentation," 2010.

[12] K. Khadka and S. Maharjan, ''Customer satisfaction and customerloyalty,'' Centria Univ. Appl. Sci. Pietarsaari, vol. 1, no. 10, pp. 58–64,2017.

[13] J. Ahn, J. Hwang, D. Kim, H. Choi, and S. Kang, ''A surveyon churn analysis in various Bus. Domains,'' IEEE Access, vol. 8,pp. 220816–220839, 2020.

[14] J. N. Sheth and C. Uslay, ''Creating enduring customer value,'' J. Creating Value, vol. 8, no. 2, pp. 241–252, Nov. 2022.

[15] T. Dierkes, M. Bichler, and R. Krishnan, ''Estimating the effect ofword of mouth on churn and cross-buying in the mobile phone marketwith Markov logic networks,'' Decis. Support Syst., vol. 51, no. 3,pp. 361–371, Jun. 2011.

[16] M. Alizadeh, D. S. Zadeh, B. Moshiri, and A. Montazeri, ''Developmentof a customer churn model for banking industry based on hard and softdata fusion,'' IEEE Access, vol. 11, pp. 29759–29768, 2023.

[17] N. Edwine, W. Wang, W. Song, and D. Ssebuggwawo, ''Detecting therisk of customer churn in telecom sector: A comparative study,'' Math.Problems Eng., vol. 2022, pp. 1–16, Jul. 2022.

[18] L. C. Cheng, C.-C. Wu, and C.-Y. Chen, ''Behavior analysis of customerchurn for a customer relationship system: An empirical case study,''J. Global Inf. Manage., vol. 27, no. 1, pp. 111–127, Jan. 2019.

[19] A. Somosi, A. Stiassny, K. Kolos, and L. Warlop, ''Customer defectiondue to service elimination and post-elimination customer behavior: Anempirical investigation in telecommunications,'' Int. J. Res. Marketing,vol. 38, no. 4, pp. 915–934, Dec. 2021.

[20] W. Soliman and T. Rinta-Kahila, ''Toward a refined conceptualizationof is discontinuance: Reflection on the past and a way forward,'' Inf.Manage., vol. 57, no. 2, 2020, Art. no. 103167