

Customer Churn Prediction at Scale for Telecom

Arpitha S¹, Dr. Geetha M²

P.G. Student, Department of Master of Computer Applications, BIET, Davangere-577004.

Arpithas693@gmail.com, Davangere, Karnataka, India¹

Associate professor Department of Master of Computer Applications, BIET, Davangere-577004.

bmgeethamallad@gmail.com, Davangere, Karnataka, India²

ABSTRACT

The phenomenon of customer churn, wherein customers decide to use services from other telecom providers, is one of the main problems that cause huge losses for organizations. Nowadays, acquiring new customers costs significantly more than retaining existing ones. Consequently, customer churn prediction and prevention have become crucial factors in effective CRM for the telecommunications industry.

The project “Telecom Customer Churn Detection Using Machine Learning” is aimed at detecting customers with a high probability of churn based on analysis of previous information including data about the use of services, demographics, payment history, complaints, and interactions. This problem is solved with machine learning models such as Logistic Regression, Random Forest, Gradient Boosting, and XGBoost.

This whole process consists of steps such as data preparation, feature creation, model training, evaluating the model, and deploying the model. This process involves the use of interactive dashboards for visualizing churn trends, segmenting customers, and making sound decisions based on analysis.

The objective of this project is to develop an intelligent, scalable, and automated solution that enables telecom companies to prevent customer attrition and thus save their income.

Keywords: Customer Churn Prediction, Machine Learning, Telecom Analytics, Classification Models, Customer Retention, Predictive Analytics

I. INTRODUCTION

Customer churn remains one of the major issues facing telecommunications companies in the current digital age. The competition from various telecommunication service providers, flexible payment methods, and ease in transferring telephone numbers from one company to another have increased the need for retention. Being able to predict and identify such customers before they defect gives companies an upper hand as they can try to entice them into remaining loyal through different offers.

Predicting customer churn entails the analysis of previous customer records in order to find trends that may result in customer churn. Such trends include low consumption levels, customer complaints about billing, dissatisfaction with the customer care department, network problems, or desire to use other competitors' services. Machine learning allows this process to be more convenient, as it identifies customer trends without involving any physical effort on the part of analysts.

In general, telecom databases contain details such as calls, messages sent, data usage, plans subscribed to, roaming services, customer complaints, customer tenure, bill payments, and demographics. Machine learning algorithms can effectively predict customer churn based on such parameters by analysing all of them. As part of the current assignment, data preprocessing is done carefully, which includes treating missing data, data normalization, variable encoding, and feature engineering.

The importance of predicting customer churn cannot be understated since there are many reasons for that. It costs significantly more money to attract a new customer than to retain an old one. On the other hand,

the telecommunications sector is highly competitive without much difference between providers. Customers' expectations nowadays are higher, as they require fast and reliable connections, clear bills, and prompt customer support.

In addition, this project can also assist telecommunications companies in making strategic decisions. In particular, by detecting potential customers that may churn in the future, telecom companies can target their marketing campaigns, optimize their loyalty schemes, improve their customer service, and even plan for investments in infrastructure. Questions such as who will be the churning customers? Why do they churn? How should the company deal with the problem? will be answered by this system.

In addition, the capability to process live data is another feature of this model. By incorporating Apache Spark and Kafka in our system, we can use real-time data such as customer behavior on the network.

II. LITERATURE REVIEW

There is a substantial body of literature that examines customer churn prediction through the application of statistical techniques and machine learning algorithms. In the earlier days, classical approaches were widely used such as survival analysis and basic classifiers to comprehend the customer lifecycle and detect customer churn. For example, Buckinx & Van den Poel examined the retention periods of customers, whereas Coussement & Van den Poel stressed the significance of predictive modeling in subscription-based industries. On the other hand, Huang et al., and Verbeke et al. employed data mining and predictive modeling techniques on telecommunication industry data to conclude that historical customer data could play a crucial role in predicting customer churn.

With advancements in machine learning, sophisticated and more precise models were created. Machine learning techniques like SVM, Gradient Boosting, and ensemble modeling were found to work far better when dealing with highly complex data. Several researches, including Idris et al., Amin et al., and Xiao et al., demonstrated that integrating more than one model could help in achieving greater prediction accuracy by identifying intricate patterns in customer

behavior. Deep learning and big data technologies have recently been added to telecom data analytics to enable the effective management of voluminous data sets. The contributions of several authors, including Goodfellow et al. and Friedman, have been crucial for building advanced algorithms, such as artificial neural networks and boosting algorithms, that are currently being utilized in predicting churn in telecommunication companies.

DESIGN AND IMPLEMENTATION

The Customer Churn Prediction is a systematic and structured process to make it reliable and productive. It begins with the gathering of telecom data that contains customers' demographics, services, and billing records. These are cleaned and processed by addressing missing values, eliminating irrelevant variables, and transforming categorical variables into numeric variables. Steps such as normalizing numeric values and balancing the data set are undertaken as well. In addition, significant features such as spending habits, usage behavior, and billing habits of customers are generated. Different machine learning algorithms, such as logistic regression, decision trees, random forest, support vector machines, gradient boosting, XGBoost, and neural networks, are utilized and analysed for their effectiveness. The data is split into training and testing samples, and the efficiency of the algorithms is tested based on performance parameters, such as accuracy, precision, recall, F1-score, and ROC-AUC. Once the most successful algorithm is determined, the model is implemented to forecast customer churn.

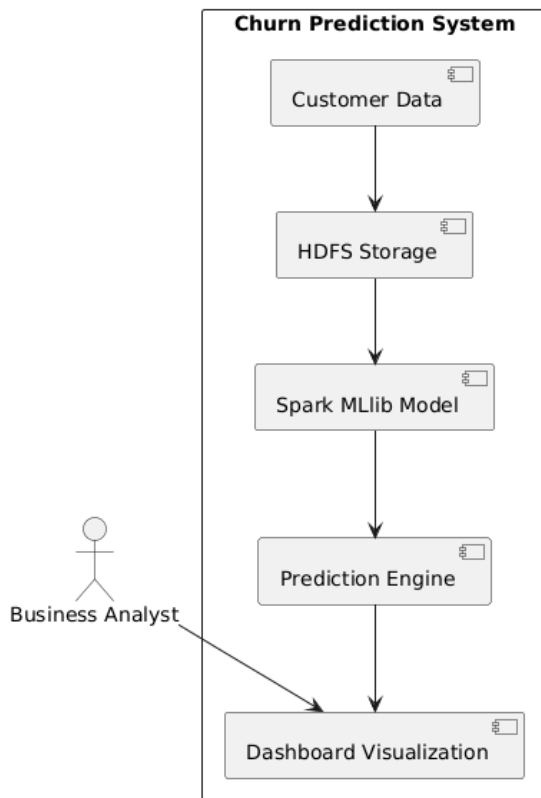


Figure: Architecture of Customer Churn

III. PROPOSED SYSTEM

The project, “Customer Churn Prediction”, the system involves following a step-by-step process that comprises data collection, data preprocessing, feature extraction, modeling, testing, and implementation. For this research, publicly available telecom data sets such as the IBM Telco Customer Churn dataset can be used. This data set provides details about demographic information, services subscribed by customers, billing data, duration of service tenure, monthly/total charges, and churn status denoting whether a customer has stopped using the service. In the data preprocessing stage, any missing data is managed, irrelevant features are eliminated, and the data is encoded to convert categorical data into numerical form. Numerical data is also scaled or normalized for better modeling results. Churn data sets are usually unbalanced; therefore, balancing techniques such as oversampling are used.

Advantages of Proposed System

The advantages of the new telecom churn prediction solution include a range of aspects making it an efficient approach to solve the problem. First of all, the high accuracy of predictions is provided thanks to the advanced techniques of machine learning used by the

system, including XGBoost, Random Forest, and Gradient Boosting models, which outperform conventional approaches. Moreover, the algorithm incorporates automated feature engineering, whereby the most informative features are generated automatically to make the prediction process more accurate. These factors include the average number of GB per month consumed by a user, call activity, payment delay, communication with the customer service, and other elements.

Among other key strengths of the proposed telecom churn prediction model is the ability to predict churn in real time. Thanks to the integration of Kafka and Spark technologies, the tool allows constantly processing streaming data on telecom customers and notifying the administrators about those who are at the highest risk. Furthermore, the solution is scalable enough to process huge amounts of data in relation to a large number of millions of users. Additionally, there are convenient dashboards for analyzing results of the work. Explainability is yet another important aspect that has been given consideration in this model through such approaches as SHAP values, explaining to the customers why they are predicted to leave the service, which helps in creating a more transparent environment for all. The final stage of the project deals with automation of the process in general, starting from data acquisition and processing up to prediction and visualizing.

IV. EXPERIMENTAL RESULT

Experimental evaluation demonstrates that ensemble methods such as Random Forest and Gradient Boosting outperform traditional Logistic Regression in terms of predictive accuracy and robustness. Artificial Neural Networks show competitive performance, especially when feature interactions are complex.

Evaluation results indicate that recall is particularly important in churn prediction since identifying potential churners is critical for retention strategies. Models optimized for higher recall achieve better business value even if overall accuracy is slightly lower.

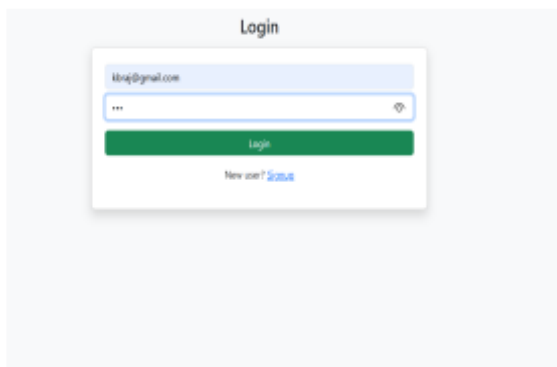
Feature importance analysis reveals that contract type, tenure, monthly charges, and customer service interactions are among the most significant predictors of churn. These insights provide actionable

intelligence for telecom companies to design targeted marketing campaigns and improve service offerings.

Moreover, feature engineering is done to generate valuable attributes like the average expenditure per month, contract length attributes, and the proportion of services used. Relevant features that have an impact on churn are discovered via correlation analysis and feature importance techniques. Several machine learning algorithms are then trained and evaluated, including logistic regression as a benchmark model, decision tree and random forest algorithms to learn intricate patterns in the data, support vector machine algorithms for high-dimensional spaces, gradient boosting methods such as XGBoost for better accuracy, and artificial neural networks for identifying hidden connections among the attributes. The data set is divided into train and test samples, and the effectiveness of each model is measured based on various performance metrics like accuracy, precision, recall, F1-score, and AUC. Hyperparameter tuning using cross-validation is carried out to optimize model performance.



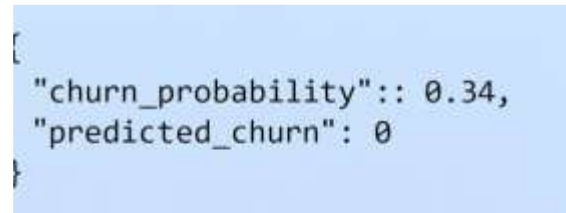
(a)



(b)



(c)



(d)

V. CONCLUSION

This paper shows how machine learning can help effectively minimize the problem of customer attrition in the telecom sector. With the use of data collected about user behavior, services used, as well as interactions, the model is capable of predicting users' intentions to churn. It turned out that algorithms such as XGBoost and Random Forest are very efficient in recognizing hidden patterns which other approaches cannot detect. In addition to increased accuracy of prediction, the created system offers valuable insights and enables visualization of results through dashboards. Ultimately, such an approach can assist telecommunication firms in minimizing their losses due to customer attrition.

The solution can be made more powerful by leveraging more sophisticated technologies and information resources. Thus, by including sentiment analysis based on customer opinions, emails, or conversations with representatives of the company, one can gain additional insight into the customers' discontent. The application of the deep learning approach may increase the accuracy of predictions by accounting for more complex dependencies in data. Moreover, the system can be supplemented with a feature to suggest customized methods of retaining each particular customer. Periodic updates of the model, integration of fresh data, and learning from experience will keep the solution efficient and up-to-date.

VI. REFERENCES

- [1] E. Coussement and K. Van den Poel, "Churn Prediction in Subscription Services," IEEE Transactions on Systems, Man, and Cybernetics, 2008.
- [2] B. Huang, M. Kechadi, and B. Buckley, "Customer Churn Prediction in Telecommunications," Expert Systems with Applications, 2012.
- [3] T. Verbeke et al., "Predictive Modeling for Churn Prediction," IEEE International Conference on Data Mining Workshops, 2011.
- [4] S. Idris, A. Khan, and Y. S. Lee, "Intelligent Churn Prediction in Telecom," IEEE Access, 2019.
- [5] H. Amin et al., "Customer Churn Prediction Using Data Mining Techniques," IEEE International Conference on Computer and Information Technology, 2017.
- [6] J. Buckinx and D. Van den Poel, "Customer Base Analysis Using Survival Analysis," IEEE Transactions on Knowledge and Data Engineering, 2005.
- [7] L. Huang et al., "Applying Machine Learning to Predict Customer Churn," IEEE Big Data Conference, 2018.
- [8] C. Wei and I. Chiu, "Turning Telecommunications Call Details to Churn Prediction," Expert Systems with Applications, 2002.
- [9] X. Xiao et al., "Ensemble Learning for Telecom Churn Prediction," IEEE Access, 2020.
- [10] J. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," Annals of Statistics, 2001.
- [11] C. Cortes and V. Vapnik, "Support-Vector Networks," Machine Learning, 1995.
- [12] I. Goodfellow, Y. Bengio, and A. Courville, "Deep Learning," MIT Press, 2016.