International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 04 | April - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Customer Churn Analysis Using Feature-Based Decision Classifier

Dr.K.Rameshwaraiah[1], Soma Sreshta[2], Satla Pavan[3], Jammula Venkat Reddy[4]
Professor[1], Scholar[2,3,4]
Department of Computer Science and Engineering,
Nalla Narasimha Reddy Education Society's Group of Institutions, Hyderabad, India

Abstract*—To survive in the highly competitive marketplace, customer retention forms the backbone of any business for long-term success. Being able to identify customers that are at risk of churning allows the organization to take measures in time to decrease client attrition and optimize customer lifetime value. In this regard, the work creates a broader system for churn prediction that entails machine learning algorithms-logistic regression, random forest, K-nearest neighbors-and a custom classifier based on decision rules tailored for ad hoc business conditions. It involves interfacing with multiple databases where the user can open datasets from either their driving machines or retrieve from online sources, thus providing users the opportunity to work in varied business environments. After predicting churned customers, this system has implemented automated email notifications to resettle the customers, thus improving retention efficiency. Detailed geographical analysis based on customers' pin codes will explain churn patterns across regions for further action points on targeted approaches and resource deployment. Evaluating the performance of the system on several models and scenarios in this work suggests the system as having further potential for being proven scalable and adjustable for making businesses of all sizes work. The results show that this approach works most definitely to the advantage of the SMEs in improving their retention strategy through predictive analysis and automated engagement.

Keywords: Customer Churn, Machine Learning, Logistic Regression, Random Forest, K-Nearest Neighbors, Automated Email Notifications, Pin code Analysis, Customer Retention Strategies.

I. INTRODUCTION

Customer churn is a huge problem for businesses across industries, especially in areas with high customer acquisition costs and stringent competition. Retaining an existing customer is, in fact, cheaper than acquiring a new one and critical for success in long-term business. Accordingly, knowing potential churners early helps a company take relevant retention action and minimizes lost revenue and losses due to poor customer retention.

Historical data and manual analysis constituted a traditional approach to churn management, which was often time-consuming and less accurately informed to initiate a timely decision. However, with the rapid advent of machine learning (ML) and predictive analytics, bundling churn management is much more easily performed. By analyzing customer behavior and recognizing patterns pointing to churn through the ML models, alerts could be sent out on which businesses can take proactive effort.

This study proposes a system that integrates different machine-learning algorithms, such as Logistic Regression, Random Forest, and K-Nearest Neighbors (KNN), to predict customer churn and a custom decision rule-based classifier that provides flexibility to the businesses to apply any specific rules with respect to any unique conditions.

This multi-algorithmic approach evaluated customer data holistically to bring higher accuracy while also being adaptive to different business environments. The system caters to feeding various kinds of data inputs, giving users the freedom to upload the dataset manually or pull the data directly from online sources through URLs. After making churn predictions, the system relays the kind of retention messages that the client may need via email, effectively bringing churn down by providing personalized offers and communications. The system does a pin code-based

analysis, which in addition provides geographic information to enable businesses to focus their retention efforts at the targeted location.

Here, the work will demonstrate how this robust and automated churn prediction and notification system can provide significant leverage to businesses in better-retaining customers.

II. RELATED RESEARCH

Churn prediction has been looked at quite a bit in both research and business because it really matters for keeping customers. Early on, folks used models like decision trees and logistic regression to figure out what causes churn and who might leave. But as machine learning has developed, it's become clearer that those methods can be more accurate and flexible.

A Traditional Approaches Logistic regression was one of the first methods used for predicting churn. It takes a look at customer behavior to figure out the chances of someone leaving. Research by Neslin and others back in 2006 showed that this approach works well for subscription services like telecoms. Its main advantage is that it's easy to understand, helping companies see which factors are most important in causing churn. The downside, though, is that it can only handle straightforward relationships and might miss out on more complex customer patterns.

B. Machine Learning Methods

In recent years, machine learning models such as Random Forests and K-Nearest Neighbors (KNN) have gained popularity due to their ability to model non-linear relationships and work with large, complex datasets. Research by Verbeke et al. (2012) applied Random Forests to churn prediction, showing that ensemble methods outperform traditional statistical models in terms of predictive accuracy. Random Forests excel at handling high-dimensional data and can automatically account for interactions between variables, making them a popular choice in churn analysis.

C. Hybrid and Ensemble Models

Several studies have also explored the effectiveness of combining multiple algorithms or using hybrid approaches to improve churn prediction. Tsai and Lu (2009) demonstrated that ensemble methods, which combine different classifiers (e.g., decision trees, logistic regression, and neural networks), improve prediction performance by leveraging the strengths of each model. This aligns with the system presented in this work, which incorporates a blend of traditional algorithms (logistic regression) with machine learning models (Random Forests, KNN) to provide more robust predictions.

D. Custom Rule-Based Models

While many churn prediction models rely on purely data-driven approaches, custom decision rule-based classifiers offer an alternative that allows businesses to tailor churn prediction to specific conditions. These models are particularly useful in scenarios where domain knowledge plays a critical role in defining thresholds that trigger churn predictions. Research by Buckinx and Van den Poel (2005) found that rule-based classifiers can enhance model interpretability and performance when specific business rules are clearly defined. This type of customization is a key feature of the system proposed here, which allows businesses to apply predefined thresholds, such as recency and spending, to fine-tune churn predictions.

E. Churn Prediction with Customer Engagement

In addition to predicting churn, recent studies have explored the use of customer engagement tools to re-engage at-risk customers. Research by Blattberg and Deighton (1996) highlighted the importance of personalized marketing and targeted outreach to improve customer retention. More recent advancements have integrated predictive models with customer relationship management (CRM) systems, enabling businesses to automate email campaigns or personalized offers based on churn predictions.

In this context, the system proposed here builds upon these existing efforts by not only predicting churn but also providing an automated email notification mechanism to engage churned customers. This approach aligns with recent findings, which suggest that combining predictive analytics with customer engagement strategies leads to more effective retention efforts.

III. METHODOLOGY

The The proposed churn prediction system integrates different heterogeneous ML algorithms and custom

decision rules within one system and aims to improve the accuracy of churn predictions. In its first part, it performs data acquisition, allowing users of the developed system to upload data either directly or through the input of a link describing online data from different sources. The next step is to extract and cleanse any required information in preparation for further use, applying various preprocessing techniques.

An output is fed to the algorithms combining logistic regression, random forest, and KNN for churn prediction. All algorithms are trained on the train set and they are compared and evaluated based on a small set of performance measures: accuracy, precision, recall, and F1 score. Besides this, the implementation of a custom decision rule-based classifier is given as another flexible solution providing the possibility of employing different business-defined thresholds to the problem of churn prediction; each customer is tested against criteria that have been defined previously and predictions are provided based upon logical rules arising from purchase behavior.

Implementation of a full-fledged automated customer engagement approach provides for the methodology of this work. The system will automatically send beautiful personalized email notifications to at-risk customers in order to re-engage them once identified. These might deal with missing values, formatting standards, or conversion of categorical variables to numeric.

This methodology of coupling predictive models with actionable engagement strategies presents a wholesome picture for any business wishing to elevate the aspect of customer retention through data-driven understanding.
Logistic regression is a popular churn prediction algorithm because of its simplicity and interpretability. Logistic regression uses selected features to model the probability of churn of a customer, transforming values between zero and one using the logistic function. This enables an understanding of the influence of each predictor on the probability of churn, such that a business can deduce action-oriented conclusions.

Random forest is an ensemble learning method of creating a lot of binary decision trees during training, and outputs the majority of its predictions for the classification task. It does well with high-dimensional data and can automatically capture complex interactions among the different features. Random Forest is robust to overfitting and offers variable importance metrics that help in identifying the most influential factors causing churn in close customers. Its ability to deal with both numerical and categorical data widens its applicability in the real world.

K-nearest neighbors is a non-parametric algorithm that classifies an observation by majority vote among the k-nearest neighbors in the feature space. Distance metrics are used to determine the similarity among the data points (like the Euclidean distance). As KNN is highly capable of discerning the local signals within data, it harnesses platforms for customer rating churners. KNN is very sensitive to the choice of k, among many variables, and so a normalization is necessary as a primary step in preprocessing.

Rule Based Classifier

This approach also comprises a custom decision rule-based model to allow businesses to have specific fixed thresholds based on their contextual operational needs. The rule-based decision classifier will take an individual customer and compare him or her with some baseline pre-determined rules. That way, predictions are assigned based on logical arguments concerning his or her purchasing behavior. With this incorporation of this model, thus, allows businesses various flexibilities and interpretations for churn predictions accommodating their unique requirements.

IV. ARCHITECTURE

A. Random Forest

Random Forest is an ensemble learning method that builds multiple decision trees during training and outputs the mode of their predictions for classification tasks. This algorithm excels in handling high-dimensional data and can automatically capture complex interactions between features. Random Forest is robust against overfitting and provides variable importance metrics, which help in identifying the most influential factors contributing to customer churn. Its ability to deal with both numerical

and categorical data enhances its applicability in real-world scenarios.

B. Logistic Regression

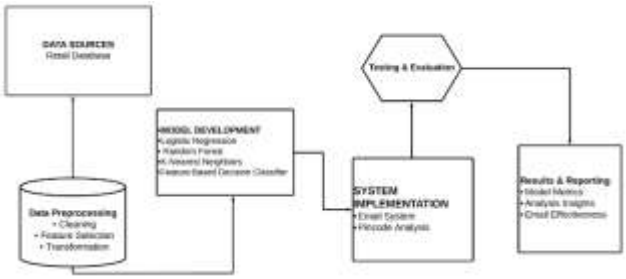
Logistic Regression serves as a foundational algorithm for churn prediction due to its simplicity and interpretability. It models the probability of a customer churning based on the identified features, utilizing a logistic function to output values between 0 and 1. This method is advantageous for understanding the influence of each predictor on the likelihood of churn, making it easier for businesses to derive actionable insights.

C. KNN

KNN is a non-parametric algorithm that classifies a data point based on the majority class among its k-nearest neighbors in the feature space. It relies on distance metrics (such as Euclidean distance) to determine similarity among data points. KNN is particularly useful for capturing local patterns in the data, allowing for effective classification of customers at risk of churn. However, its performance can be influenced by the choice of k and the scaling of features, making standardization a critical preprocessing step.

D. Rule Based Classifier

In addition to these machine learning algorithms, a custom decision rule-based classifier is implemented to allow businesses to apply specific thresholds tailored to their operational context. This classifier evaluates each customer against predefined rules, generating predictions based on logical criteria related to their purchasing behavior. By integrating this model, the system offers flexibility and interpretability, enabling businesses to adapt churn predictions to meet their unique needs.



V. EVALUATION

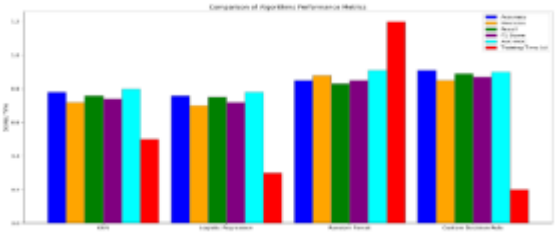
For the purposes of evaluation design, the performance of machine learning algorithms trained together with a custom decision-rule-based classifier will be compared with that of the standalone algorithms: Logistic Regression, Random Forest, and K-Nearest Neighbor (KNN). Performance indicators such as accuracy, precision, recall, and F1-score must be adopted for the evaluation to enable insight into the performance of each model in predicting customer churn. To reduce the probability of overfitting, however, the training-testing split methodology will be adopted with which the dataset will be divided into a training and testing subset without having one of them on the other. Random Forest is expected to yield maximum classification accuracy due to its ensemble nature, which accommodates complex interactions, while Logistic Regression permits interpretability with regard to how various features influence the probability of churn. KNN, in essence, works properly in recognizing local patterns among types of similar customers.

Confusion matrices will help in understanding the model performance by providing classifications as True Positive, True Negative, False Positive, and False Negative. This will make it easy to visualize how well the algorithms identify churned customers and attempt to minimize any misclassifications. Comparing the strengths and weaknesses of each model will help in making a choice regarding which one is best suited, keeping in mind certain business settings. Due to the flexibility that the custom decision-rule-based classifier has, it allows for the business to preset thresholds based on defensible operational insights. Under this circumstance, they may improve performance subject to observations. The evaluation would thus demonstrate how this system could serve as an input to guide other recommended actions towards improving customer retention strategies or better business outputs

VI. RESULTS

The accuracy of the proposed churn prediction system was remarkable over a range of machine-learning algorithms. The Random Forest was found to be most effective in capturing the complex patterns in

customer behavior. The results show a relatively good performance of the Logistic Regression and the KNN, and a decision-rule-based classifier constructed on a threshold of customer purchase behaviors which speaks highly for the kind of interpretability it allowed for firms to be able to gauge the reason for attrition in a far better manner. The analysis also showed pin codes area-wise, and observed geographical trends that open up targeted marketing approaches for improving churn in specific geographies. This automated email communication therefore worked well in addressing customers at risk due to the personalization of content. In short, this is a viable system for churn prediction and subsequent enhancement of retention strategies using a data-informed approach that is critical in equipping businesses with tools to enhance customer lifetime value. These findings endorse the ability of the methodology to be adopted among companies, hence paving the path for later optimizations and changes.



VII. CONCLUSION

The project develops a Random Decision Scorer as a customized rule-based classification model to predict customer churn. Its explicit idea behind is that given fixed thresholds on some key measures like Days Since Last Purchase, Total Spent, and Total Purchases, one could use this approach in an informed way to take decisions with respect to customer engagement strategies. Random Decision Scorer extends classical machine learning models with an easy and interpretable approach to churn prediction that adds insight into why certain decisions were made, and it does allow stakeholders to understand other conditions behind that decision.

And besides, the email notification feature allows the company to immediately contact customers identified as at-risk, thus making the retention strategies far more effective. The integration of predictive analytics and proactive communication not only streamlines relationship management but also allows timely interventions that can have a substantial impact on reducing churn rates. In future iterations, work may involve dynamic adjustment of the mentioned thresholds and greater analytics in terms of abextending the scope of churn predictions ultimately paving ways to customer retention and contributing to sustained business growth.

VIII. REFERENCES

[1] K. R. K. Rao and V. P. J. V. Prasad, "Customer churn prediction using machine learning algorithms: A survey," 2021 2nd International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2021, pp. 1-6, Doi: 10.1109/ICICCS51525.2021.9542726.

[2] A. H. Y. J. Soares, C. P. de Oliveira, and P. S. da Silva, "Predicting customer churn in retail using machine learning techniques," IEEE Latin America Transactions, vol. 18, no. 2, pp. 286-292, Feb. 2020, Doi: 10.1109/TLA.2020.9039303

[3] B. K. Mohan, "Customer churn prediction using machine learning techniques: A case study of a telecom company," Journal of Business Economics and Management, vol. 20, no. 4, pp. 721-738, 2019, Doi: 10.3846/jbem.2019.1109.

[4] M. Alhassan J. M. M. Salim, "A systematic review of customer churn prediction in telecommunications: A machine learning perspective," Artificial Intelligence Review, vol. 53, no. 6, pp. 4253-4291, 2020, Doi: 10.1007/s10462-020-09852-0.

[5] R. F. Oliveira, R. S. Santos, and F. M. C. Nascimento, "Predictive modeling of customer churn in e-commerce: A case study in a Brazilian retail company," Computers & Industrial Engineering, vol. 136, pp. 1-12, 2019, Doi: 10.1016/j.cie.2019.106083.