INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 5, May 2026 | ISSN: 2456-4184 | IJNRD.ORG

Customer Retention Analysis Through Machine Learning-Based Churn Prediction

1st Saranya.R
Department Of Advanced Computing and Analytics
Vels Institute of Science, Technology & Advanced Studies
Chennai,India
Saranyasaranya2816@gmail.com

2nd B.Kamatchy,Assistant Professor
Department Of Advanced Computing and Analytics
Vels Institute of Science, Technology & Advanced Studies
Chennai,India
Kamatchi6282@gmail.com

ABSTRACT:Many businesses are struggling with customer defection/turnover; this problem is widespread amongst competing organisations due to the currently increasing level of competition which adds an extra challenge for organisations attempting to deal with these problems. The purpose of this research paper is to examine how machine learning classification algorithms can be used to classify and predict the likelihood of customer turnover based on a variety of customer attributes. The initial step toward reaching the goals of this research paper was to find the best possible classification algorithm that could be used to predict service levels for customers in the future; thus providing information about which classification algorithm will give the most accurate predictions of classifications.

Several classification algorithms were reviewed in the course of this research paper and three classification algorithms were selected as the main focus of this research paper through further review (i.e., Logistic Regression (LR), Decision Tree (DT) and Random Forest (RF)), and to evaluate these algorithms at three different levels to compare the performance of the classification algorithms based upon common classification prediction measurement (i.e., accuracy, precision and total errors) to determine which classification algorithm produced the most accurate predicted classifications.

I.INTRODUTION

Today's customers are just as important to keep as they were to get in the first place. All the money and time spent to find someone who buys your product means less than keeping that person happy and coming back for more. Customer retention is cheaper than acquiring new customers since you wouldn't need to spend as much money advertising; existing customers will buy from you again and provide you with long-term value (CLV).The major issue faced by companies is customer churn. The term "customer churn" refers to the loss of all or a portion of a company's customers, meaning these customers have chosen not to continue using the company's services/products. The reasons for customer churn may vary widely, but include things such as poor service, high pricing, customers finding better opportunities elsewhere, or just general dissatisfaction with a service they have been provided by the company. If companies fail to identify and correct any of these issues early enough in the process when they can still keep their customers, they will keep losing customers and continue trying to retain them, while at the same time continue losing out on the expected profits from the current customers.Employ Machine Learning for Customer Loss Prevention.Telecom companies are applying machine learning techniques to gain insight into when their customers will complete purchasing of telecom services/products. Machine learning techniques enable telecom companies to identify ways of retaining their customers and preventing customer attrition by analyzing past trends in customer purchasing data, behavior, and product usage.Accurate prediction of customer retention is necessary for ongoing business growth and profitability in the telecommunications, financial services, insurance and e-commerce industries. Machine learning algorithms predict customer loss, based on client historical activity to predict future customer losses. In many cases, there are predictive models that demonstrate a deliberate correlation between the former customer's death and certain traits and behaviors of the new customer.Using existing data related, to historical, client loss, machine learning algorithms will create an accurate expectation of how many of their customers they will lose in the future.

II. LITERATURE SURVEY

Predicting current churn has been studied through various methods for a number of years. While there are many different combinations of modeling techniques that have been used to predict customer churn (from logistic regression to support vector machines), the majority of previous researchers agree that no single predictive technique is appropriate for every organization, and they also agree that a well-designed and implemented ensemble model can produce good results. The entire predictive model prediction process is a multi-step process that includes both the preparation of data and the identification of certain attributes to be included in the predictive model. There have been numerous articles written reviewing the technical approaches to customer churn prediction at many different industries, and the consensus among all of the articles reviewed is that using an approach based on machine learning in creating the predictive model is one way to forecast which customers will churn and take appropriate steps to decrease churn in the future.

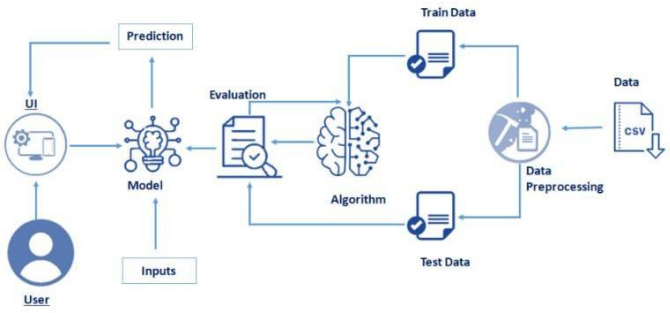
The recent trend in researching customer churn forecasting has been to develop models with an explanation about why someone is predicted to churn. Customer churn prediction has increasingly moved from traditional statistical methods to much more complicated methods, such as different forms of machine learning and deep learning algorithms. Numerous researchers have developed and validated many different models, and additional research is needed to further improve these methods. Some researchers used logistic regression, because it's easy to use and interpret. However, in addition to developing accurate predictions for customer churn, researchers are trying to develop models that can provide clearer explanation on why a customer is expected to churn. In

III. PROPOSED METHODOLOGY

It is essential for companies in today's economy not to only acquire new customers but also to ensure they retain existing customers. While it may take the company considerable time and money to acquire new customers, it will only cost the company a small fraction of that cost to retain its current customers. As such, continuing to provide value to a company's current customers creates future value for the company. Customer churn is one of the most basic problems many companies experience in relation to the retention of customers; customer churn occurs when an existing customer has ceased purchasing products and/or services from a supplier.There are numerous reasons why customers experience customer churn, including: failure to receive satisfactory service from a vendor; receiving a higher rate than offered by another supplier; having easier access to competitive offerings through a competitor; or dissatisfaction with the quality of the products and/or services offered. If companies do not take steps immediately to resolve the customer churn issue, they will continue to have losses associated with customer churn and low profitability.

Once we've gathered our customer data, we'll begin the cleansing process for those records. The purpose of this process is to remove any errors or incorrect entries in the customer records that may limit the potential for successful predictive models for predicting customer churn. Cleansing Activities will consist of (but not limited to): correcting any errors in the customer data (such as spelling mistakes), formatting the variables stored in the customer agreement records for processing with the customer churn prediction software (e.g. gender, service type), and standardizing all variables for comparative purposes. After completing the predictive models of the customer churn prediction software, the models will be validated to verify proper performance and functionality.

IV. ARCHITECTURE DIAGRAM



V. VARIOUS METHODOLOGY

1. Data Collection

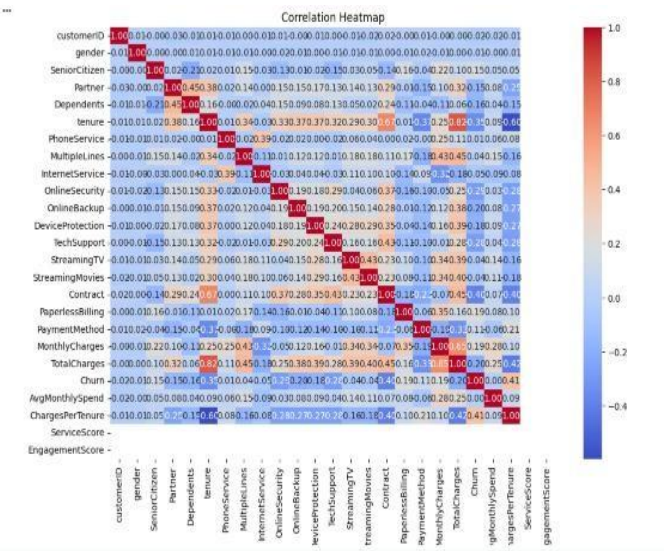
The data we have regarding your company's subscribers comes from various sources (employee input and data found on the internet). Gender, age and way of payment/subscription type are some of the many data points we collect on subscribers. Subscription usage is tracked monthly, which gives us an ongoing record of how subscribers use the subscription over time. Data collected on subscribers that contain errors typically result in more inferior performance; therefore, accuracy of your data is linked to the level of performance achieved by somebody (and therefore their success).

2. Data Preprocessing

Sometimes data comes messy, maybe missing bits here and there. Errors pop up too. At this stage clarity matters most - cleaning begins right away. Fixing wrong entries happens first, then gaps get filled if answers exist somewhere. Unchecked pieces? They get set aside completely. Things like gender labels or agreement categories shift into numbers quietly behind the scenes. Scaling adjusts everything evenly so one part doesn't overpower another later.

3. Exploratory Data Analysis (EDA)

Looking closely at a dataset lets us spot connections and oddities hiding inside. Patterns start to show up once you begin asking questions of the numbers. Instead of jumping to conclusions, we watch how pieces move together across time or groups. Visuals such as bar charts or color-coded grids turn long rows into clear views. Important clues about why customers leave often come from these early glimpses. Sometimes a quiet signal appears only after viewing data from several angles first.



4. Feature Selection

In many instances, a model does not need to be precise surrounding all possible details. Rather, there are certain characteristics that are significantly more important for having real value. Characteristics that individuals value will comprise the basis for determining projected outcomes. Which characteristic(s) will be the deciding factor in establishing an association between how strongly a characteristic correlates with people leaving as customers? The determination of which important characteristics to use will be evaluated based on how the significant characteristics correlate with turnover trends.

5. Model Training

This document outlines the process for developing training data using an automated approach to train a machine learning model with historical records. The initial step will involve dividing the dataset into two equal parts, 50% for training and 50% for subsequent validation of the trained model. Logistic Regression and Decision Tree analysis results will be presented randomly resulting in Random Forest representations being generated randomly from each respective tree in the overall sample.

6. Model Evaluation

Having trained the model we now need to evaluate the model for its ability to accurately predict and especially for its ability to identify intent of our competitors

There are several different parameters which can be used to evaluate the model,these are

- Accuracy
- Precision
- Recall
- F1 Score

8. Deployment

After it runs, out pops a web page or real-time screen where folks can tap in facts. Built using things like Flask or Streamlit, the outside piece takes questions. People drop info there, answers come right back. Hidden under the surface, the ready-made model acts like software on the internet. Each time someone reaches out, help shows up without delay.

VI.PSEUDOCODE AND IMPLEMENTATION

Input:

A single row stands for each individual when reviewing the figures. That detail arrives through corporate documents or official sources.A single fact fits into each box, shaped by lines that cross. Rows stretch left to right while columns climb top to bottom. Information settles where these meet, placed without clutter. Each cell holds just enough for clarity.Age shows up right away. Sex appears alongside it. Sometimes details come in pieces. Other times they group together. What you see depends on where you look. Information splits into parts that make sense later.Profile creation day shows up too.Information about money and payments is nearby.Each line of data is about a person.One listing per file shows the person's sex. File entries include this detail just a single time. Sex appears exactly once in every record kept here.Fragments of the information flow out from within the company's network systems.Facts sometimes arrive through channels beyond our control.At times, a part keeps count of login moments. Sometimes it notes when users come back again. A corner might show how often people return after logging in. Each visit could be marked by this feature quietly working. It may record every entry without making noise about it.

Neatly arranged, row after row stretches forward without a break. Each line follows the one before, staying straight and close.Beside every customer's ID sits a remark showing their agreement type. Though small, it reveals whether terms were fixed or flexible. A dash appears where none was recorded. Where forms differ, so does the wording - sometimes brief, sometimes longer. Each line holds just one clue, nothing more. Each month, their payments are clearly listed in the documents.Still, the numbers show exactly what provider they picked.That one shows their strategy as well.Fresh each call arrives, a record forms. Notes appear when voices reach out. A log builds with every request made. Each contact leaves its mark on file. Paper trails follow whenever help is asked. When dialogue starts, things will fall into detail; each time a ring occurs, something is written down. Arranging all of the pieces allows for their arrangement to help you visualize everything together as well as see each component in relation to the complete image. As everything comes into alignment, you'll begin to see a complete image. Once a person has exited your business, they become Churn.

A. Pseudocode

1. Begin
- Load Dataset
- Import competitors data from CSV or API
2. Data Processing
- Eliminate duplicate records
- Find and eliminate records with missing data
3. Extraction of Characteristics
- Extract characteristics of your competitors such as number of mentions in news sources, hiring activity, published research and product launches.
4. Normalization- Normalizing all features using the formula Xnorm = (X mins(X)) / (max(X) mins(X)).
5 Split Data- Splitting total data into training and testing sets.
6. Train Model (Random Forest)
- Instantiate random forest classifier.
- Train random forest model using training set.
7. Making Predictions
Make predictions regarding competitive intelligence based on test set.
8. Evaluating the Model
- Calculate: £) accuracy, precision, recall and F1 score of model.
9. Visualizing Results
- Create visual representation of competitors' activities and threats.
10. Output
- Finalize the dashboard output.

Output:

The information comes in, gets sorted - either "Churn" or "No Churn." A group goes, another stays, pulled by the unseen strings the machine picked up on earlier. Each one has a number next to their name - a churn score, quietly glowing. Large numbers mean they're leaving; small ones mean they're staying. The unknown has become known. Results like these? A form pops up - perhaps bars, lines, boxes - nothing is untested, nothing is vague. Vision dictates movement, movement dictates decisions. What is seen shapes what is done. A face appears on the screen, much like journalists do when they're breaking the news. This quick look is a hint to businesses about those likely to leave them soon, and help is at hand early, whether it is to cut costs, smooth out rough edges, or simply to make contact. Behavior is a tell-all sign without needing to guess or wonder, and people naturally

segregate without much effort, divided according to type into those leaving and those holding on..

B. Implementation

The answer to the question "who leaves?" starts here. This program is based on using the programming language Python to enforce machine learning rather than rely upon gut feelings. The program will alert corporations to potential problems (red flags) before a customer leaves. All of the information must be obtained prior to any further analysis; Pandas handles all aspects of obtaining the information without any disruption or problems. After acquiring the necessary data, the program creates a data set for its models; this happens immediately upon execution of the program. While all of the data are organized as if they were in a spreadsheet format, in reality, they were specifically designed to allow the program to accurately predict future changes to the information set. Pandas has the ability to move data into and out of data cells; if a data cell has become full or has become cluttered due to a large number of records occurring in the cell, Pandas can easily move or reduce the number of records within the cell. The result will be that if there is something missing or incorrectly grouped, the customer churn model will have an inaccurate prediction. Therefore, it will be imperative that a complete set of recorded data is available before generating a customer churn prediction. The input determines the outcome of the program; focusing on correcting and organizing the customer churn predictions will allow Pandas to clean up the data, and applying the encoding will allow Pandas to convert the labeled data into numerical coding.

VII RESULT AND DISCUSSION

A. Model Performance Analysis

Most recently, algorithms attempted to identify those who may cease the service. Logistic Regression was the first, followed by Decision Tree, and finally, Random Forest. Each of the models was given its time to work through the examples in the practice phase. Following the identification of trends, the models were evaluated based on correctness, exactness, completeness, and the balance between the two. Eventually, it was Random Forest that won with the least margin when the results were aggregated. The Random Forest model was precise.

B. Evaluation Results

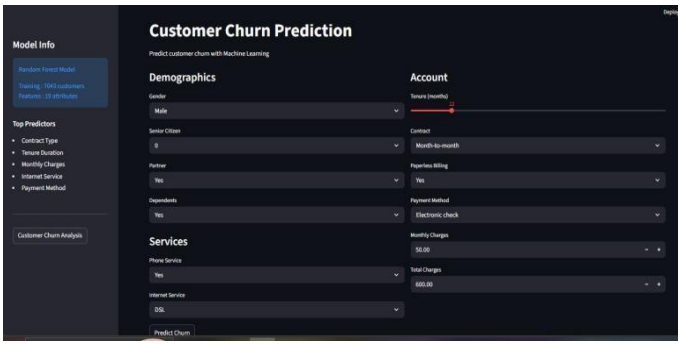
The first thing evaluated was the overall performance of the model itself, and then the accuracy of each individual guess was measured using a measuring device designed to determine how accurate and precise an answer was.Other behaviors of the buyers that were observed did not conform to any sort of regularity in order for the models to accurately calculate what behaviors should be expected, and thus created a large amount of variability between the two models' predictions about which buyers will stay and which buyers will leave, but there was still a general trend of agreement between them.

C. System Implementation Result (Dashboard)

System Implementation Results Dashboard is where the model will be deployed, and it's also where the way the model is deployed provides clarity of insight. The Web Framework provides clean and familiar interfaces, which generally allow you to navigate naturally between points. As a result, retrieving data from the system is nearly effortless and is similar to discovery as if you knew about it beforehand.

D. Prediction Output

Predictions show up the minute data enters the system. As data enters the system - the way people behave, what they purchased in the past, and how they have used services - the future starts to materialize on its own. No further action is needed. Who will leave is detected well in advance. People who stay send different signals all along. Before people leave, signals show up. Decisions happen through behaviors that determine the outcome.



E. Discussion

To help companies identify who will likely leave them (customers) and therefore predict customer churn, use a predictive capability. By using predictive models, the company will have a better understanding of what the future holds.

Traditional predictive models have consistently performed well; however, since Random Forests were introduced, these models have greatly improved their accuracy by using their ability to handle small and fragmented records quickly.

Traditionally, companies would first sort or clean records before identifying the most significant variables to include in their final predictive model. This approach made producing a predictive model more difficult. However, with the

introduction of the Random Forest method, the aforementioned processes have become significantly easier.

The most significant influence of our system will reduce the number of customers that a company loses due to a high customer churn. High customer churn is a costly problem for a company to solve. By enhancing the company's overall customer service performance, it will be able to keep the existing customer, who otherwise may have left. Customers have a greater willingness to return to a company after improvements have been made. Additionally, customers will desire to do more business with that company.

VIII.CONCLUSION

Developing a solid understanding of when someone may leave is achieved through observing their behavior before they leave. Individuals engaging in certain activities on a daily basis as well as understanding their previous spending habits are two of the biggest generators of indicators of potential leaving. Understanding contract details is another key indicator. These previous methods provided us with ways to create indicators of the ways in which individuals may leave and generated theories on what has caused them to stop using our services, while ensuring privacy was protected. The data gained through all stages was entered into computer algorithm programs, which developed patterns based on the length of time until discontinuation of the use of our services. We created educated guesses regarding employees who would potentially leave the company, but we also tried multiple validation methods to substantiate these educated guesses. Many of these validation methods confirmed our educated guess were incorrect however there was one validation method, Random Forest, that did help substantiate a portion of our educated guess. Our individual data helped us in setting the future for all of the individuals involved. Now that we had accomplished collecting the data, we began to realize that we needed other ways of predicting employee turnover from the workplace. By utilizing alternate methods companies will have improved employee turnover resulting in lowered overall company costs. Looyalty is created by the use of products and services offered by a business to those who utilize them. Customer loyalty is achieved through ongoing use of products and/or services.

X.REFERENCES

1. I.H., Frank, E., and Hall, M.A. (2011) 'Data Mining: Practical Machine Learning Tools and Techniques', 3rd edition (Morgan Kaufmann);
2. Han, J.Kamber, M.andPei, J. (2012) 'Data Mining: Concepts & Techniques', 3rd edition, (Elsevier);
3. Pedregosa, et al. (2011) 'Scikit-Learn: Machine Learning in Python, The Journal of Machine Learning Research, Vol.12, 2825–2829;
4. Verbraken, T,Verbeke,W.andBaesens,B. (2014) 'Profit Optimisation by Churn Prediction using Bayesian Network Classifiers' published in (IEEE Transactions on knowledge and data engineering)), 2(8)2774-2786;
5. Amin, S. et al. (2019) 'Prediction of Customer Churn in the Telecommunications Sector Using Certainty data', published in The Journal of Business Research Volume 94, 290-301.