JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Prediction of Employee Attrition and Customer Churn Using Machine Learning: A Survey

[1] **Kavita Sharma**, [2]**Dr. Sanjeev Kumar Sharma**
[1]M. Tech Scholar, [2]Professor(CSE)
[1]Technocrats Institute of Technology-CSE, Bhopal, India
[1]Technocrats institute of Technology-CSE, Bhopal, India

Abstract : Companies across all industries are currently most concerned about the problem of staff and customer churn. Given how costly personnel turnover is, reducing it is a primary issue for businesses. Likewise, reducing client churn is essential to avoiding substantial financial losses. This issue can be solved by calculating the probability that a customer or employee will leave a particular business. Machine learning classification algorithms are commonly used to forecast consumer churn. This investigation contrasts the typical accuracy of Decision Trees, Logistic Regression, Support Vector Machines, k-Nearest Neighbour, Random Forest, and Naive Bayes are further estimation techniques.

***IndexTerms* - Naive Bayes Model, Support Vector Machine, Decision Tree, Random Forest. Attrition, Customer Churn, Employee Churn, Prediction, Categorization, Churn Rate.**

1. INTRODUCTION

Employee churn, also known as attrition, is a significant problem for all organizations, but it is most acute in high-tech sectors (IT, telecom, manufacturing), service sectors (banking, finance, insurance), and support sectors (service desks, contact centre's, BPO). Infant attrition, in which a worker leaves within, say, six months of joining the company, is frequently distinguished from regular attrition. While managing and limiting attrition is a key duty of HR, attrition also has an impact on other organizational operations. However, the business operations of the organization, such as services, projects, manufacturing, delivery, and so forth, are where attrition has the greatest impact and effects. Understanding attrition is crucial for all HR responsibilities, including hiring and managing personnel, as shown in Figure 1.0.

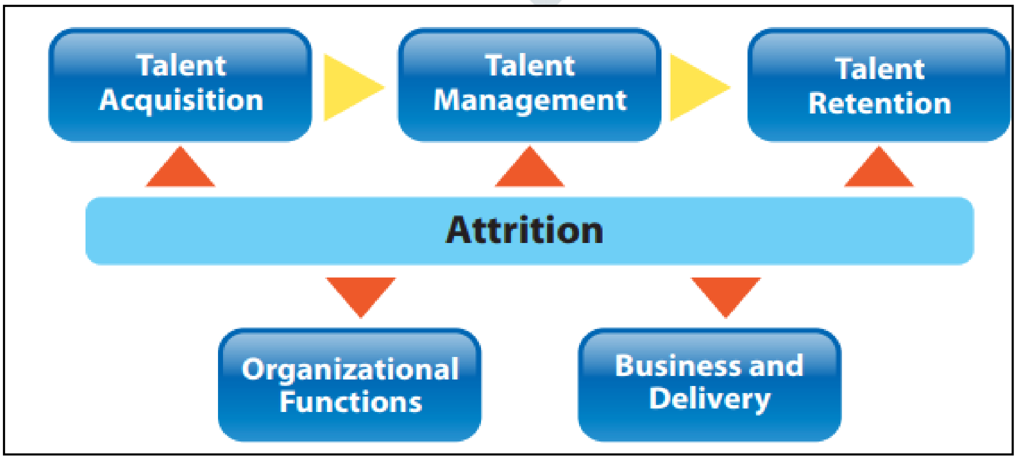


Fig.1: Attrition affects HR, other organizational functions as well as business and delivery

A big issue is when consumers and employees leave a business. Consumer churn occurs when a consumer discontinues using a company's product or service due to unmet expectations or discontent. This frequently results in switching to rival products. Staff turnover, on the other hand, refers to the rate of staff turnover inside an organization. This occurs when employees voluntarily choose to quit the company for a variety of reasons, such as poor performance or a discontent with their position. Organizations must identify trends and reasons for client churn in order to address these problems. Sample bias can be discovered by data analysis techniques like correlation matrices and histograms.

In the IT sector, employee turnover typically ranges between 12% and companies face a big issue from employee attrition since it causes workers to leave and forces employers to find and educate replacements. Officials from the company must pay attention to this process because it is expensive. The organization has financial losses as a result, which may limit their capacity to offer competitive pay and incentives, resulting in employee unhappiness and increased turnover. Organizations use a variety of strategies, including data mining to forecast and reduce staff and customer attrition. The classification models used in these techniques include Support Vector Machines (SVM), Linear and Logistic Regression, Decision Trees, Random Forests, Classification and Regression Trees (CART), K-Nearest Neighbour (KNN), Naive Bayes classifier, multi-criteria linear programming (MCLP), Neural Networks, Fuzzy Logic, Evolutionary Algorithms, Transfer Learning and others[1].

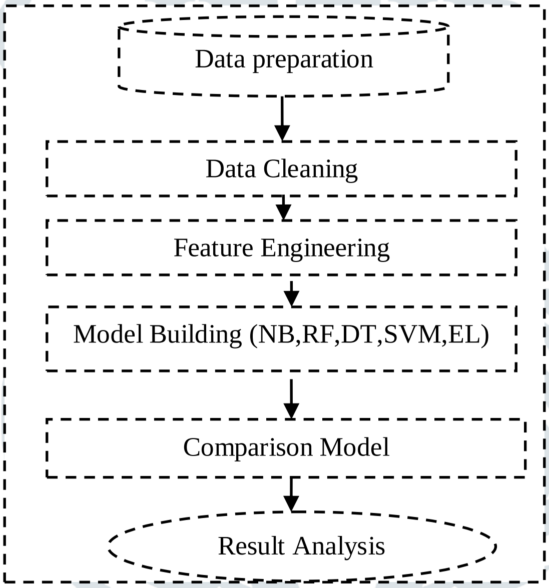


Fig.1: Basic methodology of prediction model

The objective of this paper is to evaluate and contrast various churn prediction methods. There are five important techniques will be evaluated on the basis of their effectiveness. In Section 2, we provide a thorough analysis of the different data mining categorization algorithms used for anticipating customer/employee churn. In Section 3, they gives a comparative analysis of the selected publications with information on the method used, dataset used, accuracy, research findings, and any limitations or future prospects, organized in a tabular format. Additionally, Section 4 assesses the outcomes of our survey and contrasts the average precision of the various methodologies. Finally, we draw meaningful conclusions from our findings in Section 5.

1.1 Impact of Attrition

A difficulty arises when an employee leaves for a number of reasons, including:

1. Finding acceptable replacements for lost personnel can be challenging, especially for individuals with extensive experience and specialized expertise.
2. It takes time, effort, and money to find new personnel with the necessary knowledge and experience, train them, and assist them in achieving performance and product levels equivalent to those of the departing staff.

3. The termination of an employee has a negative impact on current services and projects, which makes customers and other stakeholders unhappy.
4. New hires need time and effort to develop the same levels of knowledge and productivity.
5. The loosing of cost employee is incurred when hiring, training, and paying replacement hires.
6. Annual employee turnover rates might reach 12–15%.
7. In addition, attrition occurs all year long. As a result, "fire-fighting" attrition is a constant in high churn organizations.

1.2 Equipment and Technology

1. **Programming Languages**: This poll will determine whether Python, R, or MATLAB are the programming languages used to predict customer attrition. Large libraries and frameworks are available in these languages for statistical modeling, machine learning, and data analysis.
2. **Machine Learning Libraries**: Include well-known machine learning frameworks and libraries like Tensor Flow, Scikit-Learn, or PyTorch. These libraries offer a broad array of tools and algorithms for feature selection, model training, and evaluation.
3. **Data Analysis and Visualization Tools**: To discuss tools like Pandas, NumPy, and Matplotlib in Python that is used for exploratory data analysis and visualization. These tools assist with comprehending the data, spotting trends, and visualizing the outcomes.
4. **Feature Engineering Tools**: Scikit-Learn's Feature Selection module tools are used in feature engineering. To help churn prediction models perform better, these technologies aid in translating and producing pertinent features from raw data.
5. **Model Evaluation and Interpretability:-**Cross-validation methods and performance indicators like recall, accuracy, and F1 score are some of the instruments used to evaluate models. Mention any additional tools or methods that are utilized to improve model interpretability.
6. **Cloud Platforms**: The purpose of creating and machine learning models are developed by cloud platforms like Amazon Web Services, Google Cloud Platform (GCP), or Microsoft Azure provide scalable infrastructure and services. These technologies can be used for large-scale customer churn prediction programmers.
7. **Big Data Processing**: This refers to the appropriate technologies, such as Apache Spark or Hadoop, that permit distributed processing and analysis of massive datasets, which can be helpful for managing sizable customer churn prediction projects.

1.3 Attrition Cost – a ballpark

The HR industry is aware that finding suitable replacements for lost staff may be expensive and time-consuming, which makes it costly to lose good personnel. To give you a sense of the attrition costs, consider the following hypothetical situation: With 5000 cases of attrition each year and an average replacement cost of $10,000 per person, the annual cost of attrition-related expenses comes to $50 million. In addition to raising customer satisfaction and project success rates, cutting attrition by 20% will result in savings of around $10 million.

1.4 Analysis of Attrition

Due to data limitations (many aspects pertinent to attrition may not be assessed) and attrition-related human factors, there will always be some percentage of inevitable and essentially random attrition. When the attrition data is seen over a sufficient amount of time, however, there are typically some broad, plainly discernible patterns. In addition, attrition frequently has well-known root causes. It now becomes possible to create focused analytics that automatically extract new, practical in the form of data attrition. Consider this:-

(a) Systematically determining the organization's "as-is" understanding of the attrition issue.

(b) Developing data-driven prediction models for attrition in order to forecast future attrition extra time, for instance, location- or designation-based and for identifying people at high risk of attrition.

(c) Choosing the most suitable corrective actions and incentives to keep the employee by determining the most likely underlying causes for each anticipated occurrence (instance) of attrition.

(d) Creating the best retention strategy for each employee whose attrition is anticipated to be at a high level.

(e) Creating a thorough and effective attrition-handling plan to lessen the effects of anticipated attrition. Such a strategy calls for the hiring of suitable replacement personnel for lost or at-risk workers.

The attrition phenomenon is covered in this paper, and tasks (a) and (b), which deal with understanding and forecasting attrition, are given more weight than tasks (c), (d), and (e), which deal with retention. Understanding attrition can be useful for HR professionals in a variety of ways, especially when combined with recent, data-driven insights about attrition derived from targeted analytics. It can help to

(i) Lower attrition,

(ii) Keep good people,

(iii) Lower attrition expenses, and

(iv) Make hr systems sensitive to and proactive towards attrition reasons.

(v) Reduce attrition's negative effects (for example, by maintaining team quality and having
The least amount of an impact on projects and customer satisfaction

Fortunately, attrition is largely predictable, but because there are so many complicated human elements at play, there will always be some "unpredictable" and even "unexplained" occurrences of attrition. Business intelligence (BI) software today includes a variety of statistical and machine learning techniques that can automatically find a predictive model from historical data. These methods can be applied to analyze detailed historical data on both resigned and non-resigned employees to find a forecast attrition model. For tasks like data cleaning, feature selection, dimensionality reduction, exploratory analysis, algorithm selection, parameter tuning, and application of domain knowledge, these BI techniques require analytics expertise and extensive experimentation. Once mastered, Using this technique, one can foresee (for instance, for the upcoming quarter) which employees are most likely to leave their jobs. The HR can assist with the aforementioned activities by having a very accurate predictive attrition model at their disposal. For instance, HR management can proactively spot and solve problems with candidates whose high tendency for attrition is anticipated by the model (or "red flags"). For these workers, they can also develop a retention strategy that is unique to them.

Alternately, more sophisticated data mining approaches like subgroup discovery can find logically connected groups of workers who have an extremely high attrition rate. Such a fascinating set of workers might be described as "candidates with BE and MBA degrees AND age between 28-30 AND gender = male AND number of job changes > 3 AND major skill = Microsoft Technologies," to give an example. A 19% attrition rate could be seen in this group of 170 employees, compared to a 7% attrition rate across the board For such a team of workers target analytics might be created to also identify the main causes of attrition For instance, (a) role; (b) compensation; or (c) supervisor may be the main underlying causes. Then, by taking into account the core causes of attrition within this particular group, HR executives may build an ideal retention plan to lower the attrition levels within that group.

Attrition data insights can also be useful for recruitment:

- Priorities hiring people who have a high propensity for stability (also known as "green flags for attrition").
- Identify applicants who have a high tendency for attrition (also known as "red flags for attrition") and address them (with corrective procedures). Such applicants are likely to have problems that could soon lead to attrition.
- In order to find these kinds of insights from historical data, one might create focused analytics.

2. LITERATURE REVIEW AND RELATED WORK

Ibrahim Onramp and colleagues look into many data mining methods to forecast voluntary employee attrition. While the decision tree technique performs best in recall, the support vector machine performs better in terms of accuracy and precision. The study also emphasizes how crucial it is to select essential features from an IBM dataset, which comprises 1470 entries and 34 attributes and is cited in [2].

Andry Alamsyah and colleagues' research focuses on a dataset they acquired from an Indonesian telecoms provider. They adopt a four-stage process to forecast employee turnover, starting with data collection from the 12 attributes-containing human resource information system. The researchers develop a categorization model using the algorithms Naive Bayes, Random Forest and Decision Tree.. Each approach is given its own confusion matrix, which demonstrates that Naive Bayes produces the most real positive outcomes. But according to [1], Random Forest has a 97.5% accuracy rate, making it the most trustworthy prediction technique.

Sepideh Hassankhani Dolatabadi and colleagues offer a method for developing a prediction model for customer and employee turnover using data mining techniques and neural network algorithms. They employ a 1.5-year-old dataset with 21 attributes. Churning cases account for 24.2% of all the data in the analysis, which involved 9239 records. To forecast client behaviours, 15 carefully chosen criteria are used. Decision Tree, Naive Bayes, SVM, and neural network models are all utilized for prediction. Naive Bayes demonstrates the fastest information processing speed, according to reference [8]. SVM, on the other hand, comes in first for accuracy and True Positive rate, closely followed by the Bayesian approach. In [4], V. Vijaya Saradhi and colleagues forecast employee churn using three classification methods: SVM, Naive Bayes, and Random Forest. The dataset includes in-depth information about the employees of a certain customer unit of an organization.

The dataset comprises three distinct output classifications, including resigned, released, and kept, along with 25 attributes. After removing unimportant traits, more qualities were produced. RFC, SVM, and NB score highest in the output for accuracy. Due to its greater true positive rates, SVM, nevertheless, outperformed the other two algorithms. This advantage was brought about by adding penalties for particular classes, which enhanced SVM's capacity to handle the dataset's class imbalance issue. Customers with bad credit were discovered to be frequent churners. Plotting lift curves was done in a variety of ways, and the curve was then compared for each of the several Random Forest methods. The results demonstrated that IBRF (Improved Random Forest) provided much better performance in terms of speed, training rate, and scalability, as described in [5].

A technique for analyzing user attrition on a website or in a business email is described by Aihua Li and colleagues in [7]. Both decision trees and penalized multi-criteria linear programming are employed by them. These methods enable the distinction between loyal and churn-prone clients. In order to achieve this, ideal values for each class—both good and bad—are established, and regret measures are computed by comparing the actual values of samples with the ideal values. Huanqi Wang and colleagues recommend the in CART model in order to reduce the rate of misclassification to a minimum. To divide the dataset into blocks, the rate of idea partition attribute (CPA) is supplied. Based on the CPA for each block, they produce a cost matrix.

Classification of Math Subjects Using Machine Learning-Based Churn Prediction, 14 February 2021 Manas Kumar, Praveen Lalwani, Jasroop Singh, Chadha, and Pratyush Seth. Boosting and ensemble approaches are used in the prediction process along with logistic regression, naive bayes, support vector machines, random forests, decision trees, etc. to assess the impact on model accuracy. In order to fine-tune the hyper parameters and avoid model overstating, the train set is subjected to K-fold cross validation. Confusion matrix and AUC curve analysis was performed on the test set results. With accuracies of 81.71% and 80.8%, Ad boost and GBoost Classifier are the most accurate. The highest AUC of 84% is achieved by the Ad boost and GBoost Classifiers.

Analysis of Machine Learning Methods for Predicting Churn and Identifying Factors in Telecom Ahmad Kamran Malik, Muhammad Imra, Saif Ul Islam, IRFAN ULLAH1, Basit Raza, and Sung Won Kim are also mentioned. IEEE, 4/30/19. This article identified churn traits that point to its causes. CRM may increase productivity, provide pertinent incentives to consumers who are prone to churn based on similar behaviour patterns, and optimize company marketing operations by identifying key churn determinants from customer data. The churn prediction model is assessed using accuracy, precision, recall, f-measure, and ROC area. The RF churn classification and k-means clustering customer profiles were both improved by our churn prediction technique. It also provides reasons why customers leave businesses using the attribute-selected classifier technique.

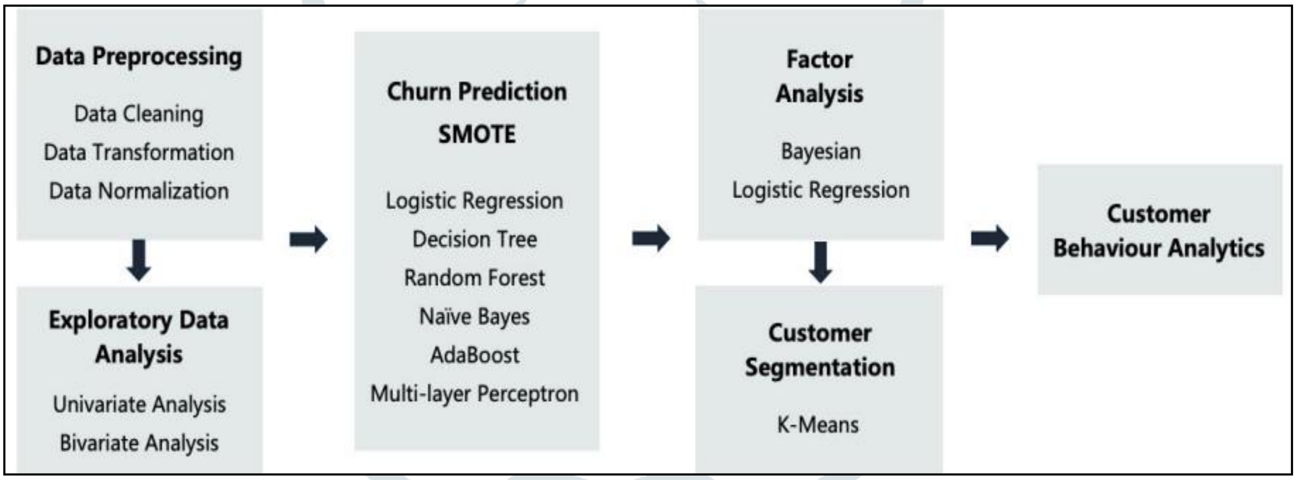
Machine learning was used by a B2B SaaS provider to forecast customer churn. Second-cycle physics degree project by Marie Sergue, 30 credits Stockholm2020, This thesis examines actual data from a SaaS firm called Air call, which offers a cutting-edge cloud-based business phone solution. The monthly customer data collection for this use case has an uneven target distribution because most customers remain loyal. The influence of the imbalance is attempted to be lessened in a number of ways while still adhering to reality and the chronological structure. Cross-validation of time series, under sampling, and oversampling are all employed. Logistic regression and random forest models are used to forecast and explain churn. For our use case, non-linear models performed better than logistic regression. Precision and recall are improved by oversampling with under sampling. Cross-validating time series results in better model performance. Instead of forecasting churn, the model is better at explaining it. It placed emphasis on churn-influencing factors related to product use.

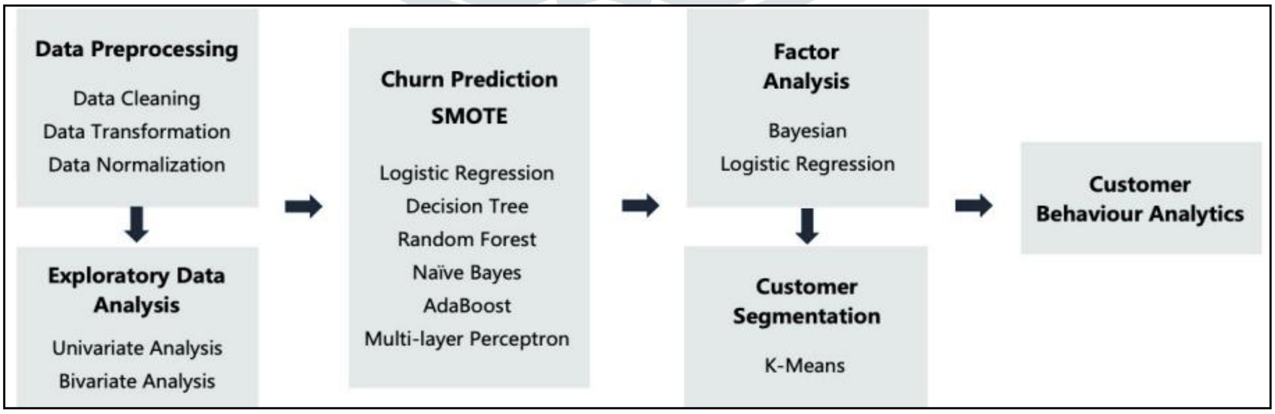
Measurement and understanding of customer churn Models for Defect Detection Scott Neslin and others. 2019 April, JMR AMA ISSN 43 This article explains how methodological considerations impact the accuracy of customer churn models. Researchers and professionals collected information from a public website, calculated a model, and made predictions using two validation databases.

The results are important. Methods matter. Variations in predictive accuracy could reduce the profitability of a churn control programme by hundreds of thousands of dollars. Models endure. "Using Data Mining To Anticipate Telecom Churn," by Wei and Chiu, in 23 (Aug. 2020). This technique allows for the identification of contract-level churners for a particular projection period. The suggested solution employs a multi-classifier class-combiner to handle the skewed class distribution between churners and non-churners. The results of the empirical evaluation show that the proposed call-behavior-based churn-prediction technique performs better when more current call details are used. The suggested technique has acceptable or respectable predictive power over the month-long period between the model's inception and the predicted rate of churn. Our suggested solution achieves acceptable lift factors when compared to a previous demographics-based churn-prediction system.

3. METHODS FOR CHURN PREDICTION

An integrated framework for consumer analytics is suggested in this study, as seen in Figure 1. In the pre-processing stage, data cleaning, data transformation, and data normalization are performed. The exploratory data analysis (EDA) step, which includes uni- and bi-variate analysis, comes next. Before supplying each feature to machine learning models, the goal of doing EDA is to make sense of the data and assist us better comprehend it, making the modeling process more effective. Secondly, to forecast customer turnover, six machine learning classifiers—Logistic Regression, Decision Tree, Random Forest, Naive Bayes, AdaBoost, and Multi-layer Perceptron—are utilized. These classifiers are used in the research for performance comparison since they are common and have done well in previous churn prediction studies, which is supported by the literature review [1], [5], and [7]. In addition, SMOTE is an oversampling technique in figure 3.





The training set will help the classifier perform better. The k-Nearest Neighbour samples are obtained for each sample x in the minority class after SMOTE uses the Euclidean distance to determine its distance to all samples in the minority class in the sample space. K has a default value of 5. A new sample is created using (1) for each randomly chosen neighbour ex, where x stands for the original sample, ex for the neighbour sample, and $_{x}$new for the synthetic sample. Different SMOTEs, SMOTE-ENN and SMOTE Tomek, for example, feature an extra cleaning process that removes overlapping synthetic samples that are hard to identify from samples from the majority class. Additionally, some of the additional hybrid sampling techniques outlined in [23] outperformed

SMOTE in their study. But different approaches behave in various datasets in different ways. SMOTE and its variants are the main subjects of our study, and SMOTE consistently outperforms the others in the datasets we use. SMOTE is therefore chosen in order to balance the instances of the two classes.

$$\mathbf{x_{new} = x + random(0, 1) \times (ex - x)\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ (1)}$$

Thirdly, to identify some key causes of the churning, factor analysis is carried out using Bayesian Logistic Regression. Bayesian Logistic Regression modeling is the procedure used to create helpful features for more accurate consumer segmentation. Since the aim of this research is to carry out churn management, the customer segmentation is solely performed on the churn data. In this study, the churn clients are segmented using the segmentation technique known as K-means clustering. Then, each cluster's properties can be discovered. Following client segmentation and churn prediction, customer behaviour analytics are then carried out. Churn unpredictability is shown visually. We compare various machine learning classifiers. Additionally, each churn segment's characteristics are summarized, which can help marketers and strategists come up with unique retention strategies for each segment.

In order to smoothly combine attrition prediction and client segmentation, this article uses Bayesian analysis as an intermediary mechanism. Following the churn prediction, factor analysis is carried out using Bayesian Logistic Regression with the goal of determining the cause of churn and providing some key criteria for churn customer segmentation. In contrast to conventional probability theory, Bayesian thinks that people's perceptions of probability, which express how much they feel something will happen, are subjective. By amassing evidence, Bayesian analysis refers to a method of determining an event's likelihood. It is clear that when making predictions,

First, a prior probability must be inferred based on prior experience and knowledge, and when fresh data mounts, this likelihood must be modified. The Bayesian theorem is used with the Logistic Regression model in Bayesian Logistic Regression. Probability is used by Bayesian approaches to measure uncertainty [24].

Custom priors are first selected before Bayesian Analysis is performed. For each parameter, the previous distribution is defined as a normal distribution with a mean and standard deviation. Then, various parameters from the prior distribution are taken and added to the logistic regression model to produce simulated data, which are then contrasted with actual data. And those filters eliminate the factors that provide data that is contradictory with the actual data. The posterior can finally be obtained [25], [26], [27]. The most frequent parameters in the posterior probability distribution can be observed, and the probability of the uncertainty can also be stated. (2) is a valid expression for the posterior probability.

$$\mathbf{P(\theta|D) = P(\theta) \cdot P(D|\theta)\ PP(\theta) \cdot P(D|\theta)\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ (2)}$$

$$P = \frac{g(z)}{n} = \frac{\dfrac{1}{1+e^{-z}}}{n} \qquad (3)$$

$$z = x_0 + w_1x_1 + w_2x_2 + \ldots + w_mx_m \qquad (4)$$

To determine the cause of churning, the probability of churn can be visualized for each attribute after modeling. By using odds ratios, the most significant causes of churning are chosen in order to further divide the churning customers into various groups. In order to help the operator manage customer churn more effectively, the characteristics of each cluster are examined. Additionally, the overall probability of each cluster is determined using (3) based on the findings of the Bayesian Analysis and customer segmentation. Where z is the linear combination of a set of features given by (4) and n is the number of clients in a cluster. Here, the offset is represented by x0, the customer features are represented by x0, xm, and the weight of each feature is represented by w0, wm.

An earlier survey study reveals that information from the Human Resource Information System (HRIS) is frequently extracted via data mining for human resource management objectives. As a result, we discuss the theoretical foundation of the study's approach in this chapter. A brief summary of similar works using data mining methods used in HRIS is also provided.

3.1 Naïve Bayes: - A well-liked classification model in supervised machine learning is naive Bayes [16]. Because of its effectiveness and simplicity, Nave Bayes captures interest. For each class, this algorithm calculates an a posteriori probability. These are the probabilities of witnessing churn and non-churn given an employee record in the context of the employee churn problem. The posterior probability calculated for each class of a given employee record using the Bayes method and naive Bayesian assumption.

Assigning new employee records to the class with the highest posteriori probability is the goal of the Bayes decision rule. In the field of text categorization, where naive Bayes technique is frequently and effectively utilized, its success rate has been shown to be significant [15]. However, the use of Nave Bayes is still restricted in forecasting churner. In their paper, Saradhi et al. [15] propose Nave Bayes as a possible model to forecast staff churns.

3.2 Decision Tree:- When participants are divided into a few groups, a decision tree is created as a prediction diagram of the decision-making rules [17]. Every internal node (non-leaf node), in a decision tree, represents a test on an attribute. It is a type of flowchart that resembles a tree structure. Each leaf node (or terminal node) stores a class label, and each branch reflects an outcome of the test. The root node is the topmost node in a tree [18]. The generation of the prediction model uses the decision tree C4.5 classifier because it can decompose complex decision-making into considerably simpler steps [19]. The decision tree model is suggested by Jantan et al. [7] as a suitable model for prediction in HRIS.

3.3 Random Forest: - Combining aggregation and bootstrap principles introduces random forest, which is based on decision trees [20][21]. The trees in the random forest are independently constructed using a unique bootstrap sampling of the data set. Using only a small selection of qualities, random forest aims to build several decision trees based on sampling data [1]. In a random forest, a simple majority decision is made regarding the prediction. The best split among a selection of predictors is used to split each node in a random forest. At that node, the predictors are selected at random, which makes it robust against over-fitting [22].

3.4 Artificial Bee Colony (ABC): Every outcome from the swarm knowledge-based algorithm ABC is known as a bumble bee-nourishment hot spot since it is inspired by the bees' intellectual behaviour of seeking out food. The nature of the food supply is taken into consideration while determining the condition. The three types of honey bees are observers, returners, and scouts**.**

Employed bees are thought of as observers who seek for the food supply and gather information. Based on the information gathered by the hired honey bees, observer honey bees stay in the hive and hunt for food sources. Scout bees randomly take advantage of fresh food supplies in deserted areas. ABC uses iteration and focuses on three phases in order to arrive at a solid solution.

3.5 **Machine learning methods:** Open source software classification models have been utilized for prediction using R, as shown in figure 3.5. The model is described in detail as follows: 1. Decision trees (C5.0): These algorithms are an expanded version of Quinlan's C45 classification methods. 2. AdaBoost: This algorithm is a straightforward, effective, and user-friendly method for creating models. 3. Random forest (RF): This uses random inputs and is based on trees in a forest.

”

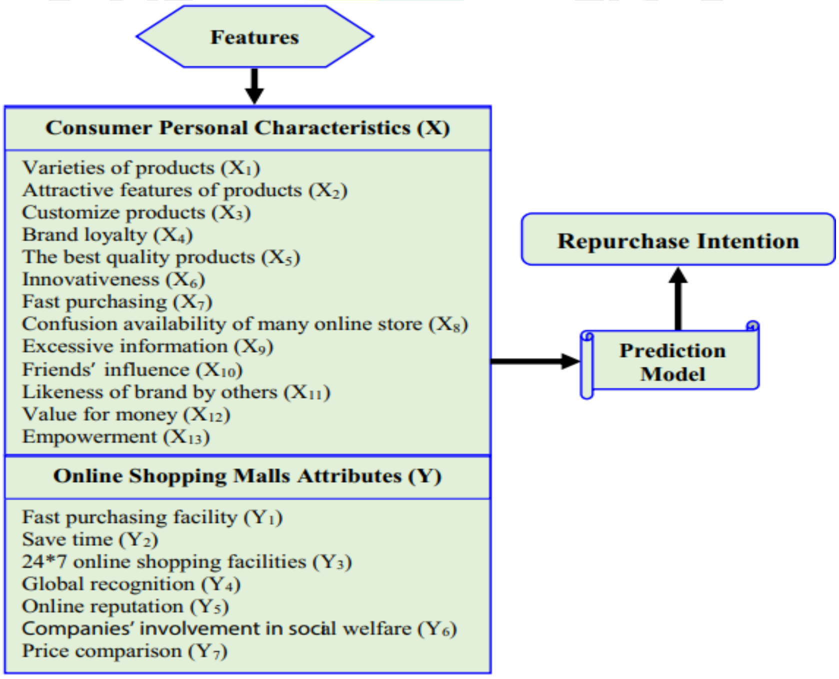

Features
Consumer Personal Characteristics (X)
Varieties of products (X_1)
Attractive features of products (X_2)
Customize products (X_3)
Brand loyalty (X_4)
The best quality products (X_5)
Innovativeness (X_6)
Fast purchasing (X_7)
Confusion availability of many online store (X_8)
Excessive information (X_9)
Friends' influence (X_{10})
Likeness of brand by others (X_{11})
Value for money (X_{12})
Empowerment (X_{13})
Online Shopping Malls Attributes (Y)
Fast purchasing facility (Y_1)
Save time (Y_2)
24*7 online shopping facilities (Y_3)
Global recognition (Y_4)
Online reputation (Y_5)
Companies' involvement in social welfare (Y_6)
Price comparison (Y_7)
Repurchase Intention
Prediction Model

Fig.3.5 : Prediction Method

3.6 **Evaluation Measurement**: - Four scores are produced by the classification model, which is based on the confusion matrix. Utilizing the confusion matrix displayed in table I, we evaluate the classification models [19]. The comprehensive justifications are:

1) TP (True Positives): The quantity of not transferring that has been accurately labeled as such.

2) TN (True Negatives): The quantity of data that have been accurately categorized as transfers.

3) FP (False Positives): The proportion of those who actually transfer but were mistakenly classified as not transferring by the algorithm.

4) FN (False Negatives): The amount of employees that aren't transferring but were mistakenly classified as such by the algorithm [14].

4. CAUSES OF ATTRITION

1. Root-causes for Attrition:

HR executives are aware of a wide range of factors (sometimes called root causes) for employee churn as part of their organization- and domain-specific knowledge. These recognized underlying reasons can be categorized broadly into personal, professional, facility-related, and supervisor-related issues. There can be additional categories in which to group these basic causes. Additionally, in order to account for any other instances of attrition, a catch-all root-cause such as "unknown" must be included. One or more of these well-known root reasons is frequently used by HR executives to "explain" why a specific person departed their position. There is a need to acquire additional evidence that supports or contradicts the root-cause claimed by the employee upon departing, including information from analysis of past work history, even though exit interviews can help identify the accurate root-cause of why an employee is leaving. Interviews with coworkers and superiors are another method HR executives can use to acquire further proof.

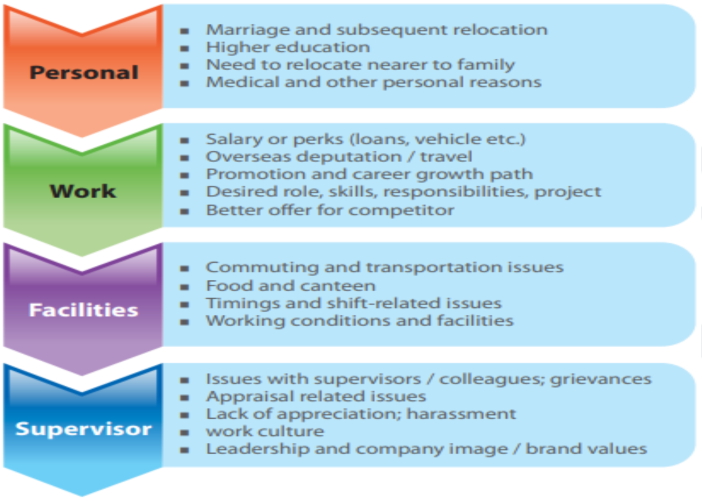

Personal
▪ Marriage and subsequent relocation
▪ Higher education
▪ Need to relocate nearer to family
▪ Medical and other personal reasons
Work
▪ Salary or perks (loans, vehicle etc.)
▪ Overseas deputation / travel
▪ Promotion and career growth path
▪ Desired role, skills, responsibilities, project
▪ Better offer for competitor
Facilities
▪ Commuting and transportation issues
▪ Food and canteen
▪ Timings and shift-related issues
▪ Working conditions and facilities
Supervisor
▪ Issues with supervisors / colleagues; grievances
▪ Appraisal related issues
▪ Lack of appreciation; harassment
▪ work culture
▪ Leadership and company image / brand values

Fig.4: Several common causes ("root causes") of attrition

Therefore, determining the most likely root cause for each occurrence of attrition using the specific employee's work history is a crucial task for HR analytics. Each root cause can be conceptualized as a binary variable that, in the simplest example, either applies ("explains") or does not apply ("does not explain") to a particular case of attrition. How can we determine whether a specific root cause applies or does not apply in a given example of attrition? This is the crucial question. It should be emphasized that, as is frequently the case with human factors, one can also rank each root cause for elucidating the causes in a specific example of attrition in figure 4.0.

4.1 Retention Strategy

The HR executives can select from a variety of retention techniques when an employee formally signals an intention to retire in order to stop that resignation. They can provide:

1. A greater pay cheque
2. An advancement
3. International deputation

4. Continue to a chosen destination
5. A new position or one with more responsibilities Loans or aid with finances
6. A project change
7. Initiatives for competency building and training
8. Redress of any particular complaints
9. In addition to these focused (individually-specific) retention initiatives, HR has other, the general tools at its disposal to lessen the impact of attrition.

Among them are:

a. Pro-active identification (in advance) of employees at high risk of attrition
b. Training and deployment for "back-up" team members for critical tasks and core employees
c. Improved and effective knowledge transfer mechanisms
d. Creation and implementation of a succession plan for leadership positions etc.
e. Creating a thorough and effective attrition-handling plan to lessen the effects of anticipated attrition.

5. PERFORMANCE ANALYSIS

Use this checklist to identify and address problems. It has been found that the random forest distribution, or DT, is the best distribution for prediction as a consequence. It is also reasonable to classify SVM, Naive Bayes, regression, max-min, balanced and unbalanced data, over fitting, AdaBoost, bagging, boosting, pruning, etc. as machine learning algorithms. Because they are less accurate than other models, decision trees are not the greatest choice for mystifying or perplexing forecasts. The accuracy score outcomes, monitored as the suggested SVM model underwent training and testing. The unseen data models accuracy is 90% high and navie bias accuracy is 89% and lowest accuracy of random forest classifier 84% .After that, we made sure that our concept could be applied broadly.

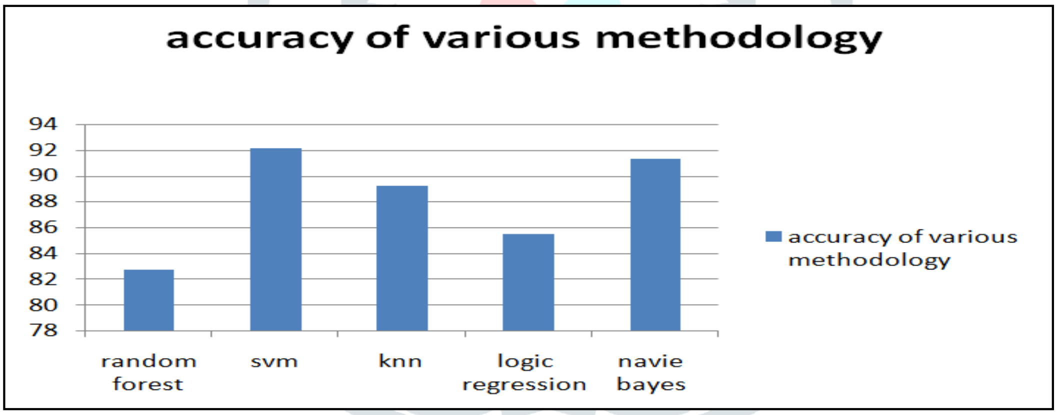


Fig.5: Performance analysis

6. CONCLUSION

Organizations are already experiencing serious challenges as a result of the enormous financial losses brought on by both customer and personnel loss. In order to address this issue, a precise estimation of the probability of losing customers and employees is required. Using a number of data mining-based categorization techniques, customer churn can be anticipated consistently and precisely. In this study, we thoroughly examine a range of models and compare how well they perform in terms of average accuracy. Observations showed that Random Forest, with a score of 93, had the highest prediction accuracy. The 16% is the greatest choice for developing exact models to forecast employee or customer turnover. Other methods, such as NB and SVM, show the prediction's average accuracy as well, making them a good choice for developing predictive models. Compared to other strategies, Decision Trees did the best, with an accuracy rate of only 82.74%.

The accuracy of the predictions could be affected by the usage of unneeded or undesirable dataset attributes in any of the machine learning approaches mentioned above that are used to forecast staff attrition or customer churn. In order to add the most salient traits and remove unneeded and unnecessary features, a new strategy of feature engineering is therefore offered. After that, the accuracy of the f1-measure, precision, and recall characteristics will be used to evaluate the performance of the overall strategy.

REFERENCES

[1] A. ALAMSYAH, N. SALMA: *A comparative study of employee churn prediction model*, 4th International Conference on Science and Technology (ICST), (2018), 1–4. IEEE, 2018.

[2] I. O. YIGIT, H. SHOURABIZADEH: *An approach for predicting employee churn by using data mining*, 2017 International Artificial Intelligence and Data Processing Symposium(IDAP), (2017), 1–4.

[3] D. S. SISODIA, S. VISHWAKARMA, A. PUJAHARI: *Evaluation of machine learning mod-els for employee churn prediction*, International Conference on Inventive Computing and Informatics (ICICI), (2017), 1016–1020.

[4] V. V. SARADHI, G. K. PALSHIKAR: *Employee churn prediction*, Expert Systems with Ap- plications, **38**(3) (2011), 1999–2006.

[5] W. YING, X. LI, Y. XIE, E. JOHNSON: *Preventing customer churn by using random forests modeling*, IEEE International Conference on Information Reuse and Integration, (2008), 429–434.

[6] C. WANG, R. LI, P. WANG, Z. CHEN: *Partition cost-sensitive cart based on customer value for telecom customer churn prediction*, 36th Chinese Control Conference (CCC), (2017), 5680–5684.

[7] A. LI, Z. LIN: *Email users churn analysis based on pmclp and decision tree*, Sixth Interna- tional Conference on Fuzzy Systems and Knowledge Discovery, **7** (2009), 348–350.

[8] S. H. DOLATABADI, F. KEYNIA: *Designing of customer and employee churn prediction model based on data mining method and neural predictor*, 2nd International Conference on Computer and Communication Systems (ICCCS), (2017), 74–77.

[9] L. XIE, D. LI, J. XIAO: *Feature selection based transfer ensemble model for customer churn prediction*, International Conference on System science, Engineering design and Manufac- turing informatization, **2** (2011), 134–137.

[10] E. M. PETERS, G. DEDENE, J. POELMANS: *Understanding service quality and customer churn by process discovery for a multi-national banking contact center*, IEEE 13th Interna- tional Conference on Data Mining Workshops, (2013), 228–233.

[11] R. Bååth. (2017). Introduction to Bayesian Data Analysis—Part 1: What is Bayes?. [Online]. Available: https://www.youtube.com/ watch?v=3OJEae7Qb_o

[12] R. Bååth. (2017). Introduction to Bayesian Data Analysis—Part 2: Why use Bayes? [Online]. Available: https://www.youtube.com/watc h?v=mAUwjSo5TJE

[13] R. Bååth. (2017). Introduction to Bayesian Data Analysis—Part 3: How to do Bayes? [Online]. Available: https://www.youtube.com/watch?v=Ie6H_r7I5A.

[14] S. Khodabandehlou and M. Rahman, "Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior," Journal of Systems and Information Technology, vol. 19, no. 1/2, pp. 65-93, 2017.

[15] 1] V. V. Saradhi and G. K. Palshikar, "Employee Churn Prediction," Expert System with Application, vol. 38, no. 3, pp. 1999-2006, 2011.

[16]] A. Alamsyah and N. F. Zuhri, "Measuring Public Sentiment Towards Services Level in Online Forum using Naive Bayes Classifer and Word Cloud," CRS-ForMIND International Conference and Workshop, 2017.

[17] E.-B. Lee, K. Jinhwa and S.-G. Lee, "Predicting customer churn in mobile industry using data mining technology," Industrian Management & Data Systems, vol. 117, no. 1, pp. 90-109, 2017.

[18] J. Han, M. Kamber and P. Jian, Data Mining: Concepts and Techniques, San Fransisco: Morgan Kaufmann Publisher, 2006.

[19] H. M. Setiadi, C. Ariandika and A. Alamsyah, "Prediction Models Based on Flight Tickets and Hotel Rooms Data Sales for Recommendation System in Online Travel Agent Business," The 7 h Smart Collaboration for Business in Technology and Information Industry (SCBTII), vol. 6, 2016.

[20] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5- 32, 2001.

[21] R. Genuer, J.-M. Poggi, C. Tuleau-Malot and N. Villa-Vialaneix, "Random Forests for Big Data," Big Data Research , vol. 9, pp. 28-46, 2017.

[22] R. Punnoose and P. Ajit, "Prediction of Employee Turnover in Organizations using Machine Learning Algorithms (A case for Extreme Gradient Boosting)," International Journal of Advanced Research in Artificial Intelligence (IJARAI), vol. 5, no. 9, pp. 22-26, 2016.

[23] A. Amin, S. Anwar, A. Adnan, M. Nawaz, N. Howard, J. Qadir, A. Hawalah, and A. Hussain, ''Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study,'' IEEE Access, vol. 4, pp. 7940–7957, 2016.

[24] J. O. Berger, Statistical Decision Theory and Bayesian Analysis, 2nd ed. New York, NY, USA: Springer-Verlag, 1985.

[25] R. Bååth. (2017). Introduction to Bayesian Data Analysis—Part 1: What is Bayes?. [Online]. Available: https://www.youtube.com/ watch?v=3OJEae7Qb_o.

[26] R. Bååth. (2017). Introduction to Bayesian Data Analysis—Part 2: Why use Bayes? [Online]. Available: https://www.youtube.com/watc h?v=mAUwjSo5TJE.

[27] R. Bååth. (2017). Introduction to Bayesian Data Analysis—Part 3: How to do Bayes? [Online]. Available: https://www.youtube.com/watch?v=Ie6H_r7I5A.