International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 10 | Oct - 2021 ISSN: 2582-3930

Prediction of Employee Attrition through NSMs

Devansh Akhilesh Shukla , Vishal Rajendra Patil ,Nikhil Tiwari and Prof. Ranjita Gaonkar
Mumbai University , India

Abstract — Business leaders and decision makers have shown great interest in data and the sense it makes. There is an increase in demands that researchers explore its use within business organisations.Data has become a key asset for most businesses in a variety of industries, including those that deal with business operations. The use of new technology benefits all sorts of businesses, and data collection, administration, and analysis provide several advantages in terms of efficiency and competitive advantage. In fact, analysing massive amounts of data can result in better decision-making.In fact, analysing large amounts of data can lead to better decision-making processes, achievement of pre-established corporate goals, and increased business competitiveness. Different classification models were used to check the precision and accuracy for the dataset. The models which were used are logistic regression, Random forest ,KNN model, XGBoost, Light GBM and Categorical Boost. Different Machine learning techniques can give a better idea of which algorithm is the best fit.

Keywords—XGBoost,Light GBM,Categorical Boost.

1. Introduction

Employee management comes at a cost; a company must invest a substantial amount of time and money in training employees to align with the organizational requirements. When an employee leaves an organization, the company not only loses one of its most valuable employees, but it also loses the money it spent to recruit, select, and train those employees for their respective jobs. On the other hand, the organisation must continue to invest in the recruitment, training, and development of new employees in order to fill vacant positions. Because of these factors, every organisation wishes to reduce attrition and retain employees by implementing more beneficial company policies and work environments. In this paper, machine learning techniques are applied in the context of employee recruitment. The present research work would be helpful to most companies in gaining knowledge about their employees' satisfaction levels and acquiring useful information that would aid in controlling the attrition rate. Predictive analysis techniques have been applied to Human Resource data in this case. Data was cleaned, prepared, and pre-processed using the first exploration techniques. The data generated was then made subject to predictive models such as logistic regression, Random forest, KNN model, XGBoost,Light GBM, and Categorical Boost. Finally, the results of various machine learning models were examined by comparing which model is best-fitted and provides accurate results for a given problem.

2. Literature Survey

2.1 Early Prediction of Employee Attrition using Data Mining Techniques

A computer scientist was quoted in this paper as saying, "You remove our prime twenty workers, and we [Microsoft] become a mediocre company." This statement by a computer scientist drew our attention to one of the most serious issues in the workplace: worker attrition. Worker attrition (turnover) is a considerable expense to any organisation, and it can have a deleterious impact on its overall efficiency. According to CompData Survey data, total turnover has increased from 15.1% to 18.5 percent over the last five years.Finding a well-trained and experienced worker is indeed a challenging task for any organisation; however, switching such workers is even more challenging. This not only tends to raise many Human Resource (HR) costs, but it also has an effect on a company's market value. Despite these facts and ground facts, there has been little attention being paid to the literature that has seeded several misconceptions about time units and workers. As a result, the aim of this paper is to provide a framework for predicting worker churn by trying to analyse the employee's precise behaviour patterns and mistreatment classification techniques.

2.2 Prediction of Employee Attrition using Data Mining

As per this paper, today's worker attrition prediction has become a serious drawback within organisations. Worker attrition can be a significant issue for organisations, especially when trained, technical, and key personnel leave for a better opportunity somewhere else. This results in the loss of a trained worker. As a result, we tend to use this and previous worker knowledge to investigate the most common reasons for worker attrition or attrition. We usually use standard classification methods to avoid worker attrition, such as call tree, supply Regression SVM, KNN, Random Forest, and Naive mathematician methodologies. To achieve this, we hire feature selection techniques on the data and analyse the results in order to reduce worker attrition. This can assist companies predict worker attrition also and assist their economic process by reducing their human resource costs.

2.3 Predicting Employee Attrition using XGBoost Machine Learning Approach

This paper discusses Given the global competitive state of affairs, there is an ocean of opportunities for hot as well as gifted people all over the globe, and given an honest chance, workers half from one organisation to another. Turnover is currently considered as the most significant issue for all organisations, owing to its negative impacts on work productivity and meeting structure objectives on time. To overcome this drawback, organisations are currently relying on machine learning techniques to predict worker turnover. Organizations will take necessary actions for worker retention or continuation in course of time, with high precision in prediction. The majority of the data comes from basic time unit primarily based data systems, which aren't very cost effective in prediction and modelling, and these models aren't terribly correct in knowledge models and can't assist organisations to make effective decision. The primary objective of this analysis paper is to predict worker attrition, or if a worker is about to leave or is still continuing to work within the organisation. In this paper, we propose a completely new model for predicting worker attrition mistreatment. XGBoost, a machine learning-based approach, is especially powerful. To validate the accuracy of the system projected for worker Attrition, the information set is non inheritable via online data fetched to the the system, and extremely beautiful exactitude results with respect to voluntary turnover

2.4 Predicting Employee Attrition using Machine Learning

The increased interest in machine learning among business leaders and call manufacturers necessarily requires that researchers investigate its application within business organisations, according to this paper. One of the most serious issues confronting business leaders today is the loss of talented employees. This study looks at machine learning techniques for worker attrition and mistreatment. Three key experiments were conducted to predict worker attrition using IBM Watson's artificial knowledge. The primary experiment includes coaching the initial class-imbalanced dataset with the following machine learning models: support vector machine (SVM) with many kernel functions, random forest, and Knearest neighbour (KNN). The second experiment centred on a mistreatment adaptational artificial (ADASYN) approach to solve category imbalance, accompanied by preparation on the new dataset mistreatment of the above- mentioned machine learning models. The third experiment involved manual undersampling of information to balance between categories. As a result, the coaching associate ADASYN balanced the dataset with KNN (K = 3) achieved the highest F1- score of 0.93. Finally, by leveraging feature selection and random forest, an F1-score of 0.909 was acquired by mistreating twelve options out of a total of twenty nine options.

3. Proposed Work

We accomplished three goals in this study. We first analysed the data, then developed many models, and then, using a confusion matrix, determined which models are the best. In addition, each experiment required the training and validation of a collection of machine learning classifiers to predict employee attrition from an unknown dataset. All classifiers were validated using 5-fold cross validation. Furthermore, feature selection approaches were devised to reduce the complexity of trained models and improve their performance. Each classifier was iteratively trained and evaluated, with the number of features growing with each iteration..The proposed

approaches are listed below, along with further information.

3.1 System Architecture

The system architecture is given in Figure 1. Each block is described in this Section.

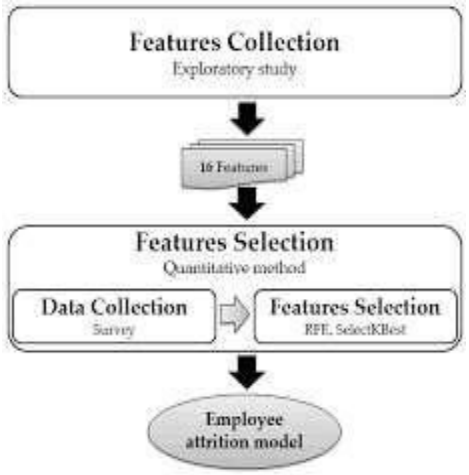

Fig. 1 Proposed system architecture

3.1.1. Training Set: The training set is the material that the computer uses to learn how to analyse data. Machine learning employs algorithms to simulate the human brain's ability to take in a variety of inputs and weigh them in order to generate activations in the brain's individual neurons.

3.1.2. Test Set: In machine learning, a test set is a secondary data set used to evaluate a machine learning programme after it has been trained on an initial training data set. A test set, also known as a test data set or test data, is a collection of data that is used to create a test.

3.1.3. Modelling

The modelling phase entails choosing models based on a variety of machine learning approaches to be employed in the testing. Various predictive models are used in prediction, including logistic regression, Random forest, KNN model, XGBoost, Light GBM, and Categorical Boost. Our goal is to find the most appropriate classifier for our situation. Each classifier can be trailed on the feature set for this, with the classifier with the best classification results being utilised for prediction.

A) **Logistic Regression** : is an approach similar to linear regression but with a discrete dependent variable (e.g., 0 or 1). While maximising a classification criterion, linear logistic regression estimates the coefficients of a linear model using the selected independent variables. This is an example of our data's logistic regression parameters. The logistic function, also known as the sigmoid function, was created by statisticians to characterise the characteristics of population increase in ecology, such as how it rises swiftly and eventually reaches the environment's carrying capacity. It's an S-shaped curve that can transfer any real-valued integer to a value between 0 and 1, but never exactly between those two points.

$$1 / (1 + e^{-value})$$

This is the sigmoid function's equation. Where e is the natural logarithms' base (Euler's number or the EXP() function in your spreadsheet) and value is the numerical value to be transformed. The values between -5 and 5 have been changed into the range 0 and 1 using the logistic function, as shown below.

B) **Random Forest** : Random Forest is also known as Random Decision Forest (RFA), and it is used for classification, regression, and other tasks that require many decision trees to be constructed. This Random Forest Algorithm is based on supervised learning, and it has the advantage of being able to perform both classification and regression. It's also known as a lazy learner algorithm since it doesn't learn from the training set right away; instead, it saves the dataset and performs an action on it when it comes time to classify it.

C) **KNN** : K-nearest neighbour analysis is a notion that has been employed in a number of anomaly detection systems. The k-nearest neighbour algorithm, which is a supervised learning technique in which the result of a new instance query is classified based on the majority of K-Nearest Neighbor category, is one of the finest classifier algorithms that has been utilised in credit card fraud detection.

D) **XGBoost** : XGBoost provides the ability to handle missing values by default. When XGBoost comes across nodes with a low value, it splits the left and right hands and learns all possible paths to the biggest loss. This is the point at which the test is run on the data. Extreme Gradient Boosting (XGBoost) is a supervised learning algorithm that uses synthesis. It has an objective function (written) that consists of a loss function (d) and a regularisation term ():

$$\Omega(\theta) = \underbrace{\sum_{i=1}^{n} d(y_i, \hat{y}_i)}_{Loss} + \underbrace{\sum_{k=1}^{K} \beta(f_k)}_{regularization} ,$$

Fig 2.XGBoost equation

Where yi is the predictive value, n is the number of instances in the training set, K is the number of trees formed, and fk is the synthesis from a tree. Regularization is defined as follows:

$$\beta(f_i) = \gamma T + \frac{1}{2}\left[\alpha \sum_{j=1}^{T} | c_j | + \lambda \sum_{j=1}^{T} c_j^2\right] ,$$

Fig3.Regularization equation

E) LGBM: Light GBM is a tree-based learning algorithm-based gradient boosting framework. Light GBM develops trees vertically, whereas other algorithms grow trees horizontally, implying that Light GBM grows trees leaf-by-leaf rather than level-by-level. It will grow the leaf with the greatest delta loss. Leaf-wise algorithms can decrease more loss than level-wise algorithms while expanding the same leaf.

F) CATBoost: CatBoost presents a novel technique for processing categorical features based on target encoding, a well-known preprocessing strategy. The encoded quantity is, in general, an estimate of the predicted goal value in each feature category.

$$\mathbb{E}(y|x^i = x_k^i)$$

Fig4.Formula of encoded quantity

Let's look at category I of the k-th training example in more detail. We wish to replace it with an approximation of the encoded quantity's formula. An example of a widely used estimator is

$$\hat{x}_k^i = \frac{\sum_{j=1}^{n} \mathbb{1}_{\{x_j^i = x_k^i\}} \cdot y_j + a\,p}{\sum_{j=1}^{n} \mathbb{1}_{\{x_j^i = x_k^i\}} + a}$$

Fig 5.Formula estimator for encoded quantity

This is just the average target value for samples in the same category as xi of sample k, smoothed by some prior p, and with weight a > 0. The parameter p is typically set to the sample mean of the desired value. CatBoost's Ordered Target Statistics (TS) approach attempts to address a frequent problem that occurs when employing such a target encoding: target leaking. The authors of the original work give a simple but powerful illustration of how a naïve target encoding may lead to large mistakes in the test set predictions.

G) SVC: The techniques SVC and NuSVC are similar, but they take slightly different sets of parameters and have somewhat different mathematical formulations (see section Mathematical formulation). LinearSVC, on the other hand, is another (faster) Support Vector Classification implementation for the situation of a linear kernel. The parameter kernel is not accepted by LinearSVC since it is believed to be linear. It also lacks several of the SVC and NuSVC characteristics.

3.1.4 Deployed Model: Deployment is the process of integrating a machine learning model into an existing production environment in order to make data-driven business decisions. It's one of the last steps in the machine learning process, and it's also one of the most time-consuming.

4. Results and Evaluation

4.1 Analysis of Data:. The number of frauds and the percentage were calculated after examining the dataset and doing feature extraction, and the results are as follows:

Figure 5.Analysis of Employee Attrition

After doing a quick study, we can determine what elements contribute to attrition, which aids in the development of the model.

4.2 Evaluation:

After building several classification models for the dataset, accuracy ,precision,recall and the f1-score for various models were calculated. This gave us a brief idea on which model is best for the dataset and working on different models gave a better understanding on how they work and how exactly we can find the best fit.

Model Type	Accuracy	Precision	Recall	F1 Score
Logistic regression	0.969	0.966	0.896	0.93
Random Forest	0.998	0.999	0.955	0.977
KNN	0.994	0.973	0.975	0.974
XGBoost	1.000	0.996	0.997	0.996
LGBM	1.000	0.996	0.996	0.996
CatBoost	1.000	0.995	0.995	0.995
SVC	0.982	0.982	0.94	0.961

Fig 6. Accuracy precision recall and f1-score for different models

A matrix below gives a more clear idea on which model is the best for the dataset we have worked on.A confusion matrix lets us decide on which classification model is the best.

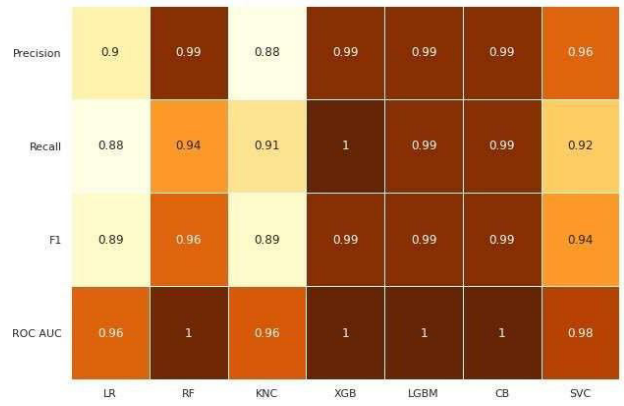

Fig 7.Matrix for different models with their accuracy,precision and F1-scores

After a brief data analysis and model building and trying different models we came to a conclusion that XGBoost, Light GBM and Categorical Boost gave us the best results

Acknowledgment

It is our privilege to express our sincerest regards to our supervisor Prof. Ranjita Gaonkar of Pillai College of Engineering, New Panvel, for the valuable inputs, able guidance, encouragement, whole-hearted cooperation and constructive criticism throughout the duration of this work.

References

1. Sandeep Yadav, Aman Jain, Deepti Singh, "Early Prediction of Employee Attrition using Data Mining Techniques" in IEEE 2018.
2. R Shiva Shankar, J Rajanikanth, V.V .Siva Rama Raju, K VSSR Murthy, " Prediction of employee attrition using data mining", in IEEE 2018.
3. Rachna Jain, Anand Nayyar," Predicting Employee Attrition using XGBoost Machine Learning Approach", in IEEE 2018.
4. Sarah S. Alduayj, Kashif Rajpoot, "Predicting Employee Attrition using Machine Learning",in IEEE 2018.
5. Explaining and predicting employees' attrition: a machine learning approach Praphula Kumar Jain, Madhur Jain & Rajendra Pamula .
6. Predicting Employee Attrition using XGBoost Machine Learning Approach Rachna Jain1 Rachna Jain and Anand Nayyar.
7. M. Maisuradze, Predictive Analysis On The Example Of Employee Turnover (Master's thesis), Tallinn: Tallinn University of Technology, 2017.
8. S. Kaur and R. Vijay, "Job Satisfaction – A Major Factor Behind Attrition or Retention in Retail Industry," IJIR, vol. 2, no. 8, 2016.