

Employee Attrition Prediction using Machine Learning

Vivekanandan S J, Aruna A, Patrachalam C, Sathish Kumar M, Thamizharasan S

Department of Computer Science and Engineering
Dhanalakshmi College of Engineering, Manimangalam, Tambaram

Abstract: *This project uses the random forest algorithm to forecast whether or not a company's employee will leave. We use, among other things, the number of years an employee has worked for the company as well as the appraisal of employee performance on a monthly basis. Logistic regression, decision trees, and artificial neural networks (ANNs) are further methods for solving this issue. The dataset was divided, with 70% of it being used to train the algorithm and 30% of it being used to test it, yielding an accuracy of 89.73%. People analytics help organisations and their human resources (HR) personnel reduce attrition by modifying the tactics for attracting and retaining talent in the era of data science and big data analytics. Employee attrition poses a serious issue and a significant risk to firms in this situation since it has an impact on both productivity and the continuity of planning. The important contributions this study brought to the field are listed below. We first suggest an approach for people analytics that changes from a big data context to a deep data context by focusing on data quality rather than data quantity in order to anticipate employee attrition.*

Keywords: Deep people analytics, employee attrition, retention, prediction, interpretation, policies recommendation

I. INTRODUCTION

Every organisation must deal with employee resignations. However, if the circumstance fails to be addressed correctly, the departure of key employees may result in a decline in productivity. It may be necessary for the organisation to hire new staff members and teach them to understand the tool in use, which takes time. The majority of businesses want to know whether of their workers are most likely to leave.

Employee attrition or voluntary resignation is a significant problem for businesses since it impacts not merely their productivity and ability to continue doing the work, but also their long-term growth plans. Retention of staff is a significant difficulty for both companies and recruiters on this road since it results in an absence of business prospects in addition to skills, experiences, and persons.

First, we suggest an approach to people analytics that switches from a big data context to a deep data perspective by concentrating on data quality rather than data quantity in order to anticipate employee turnover. To build a useful employee attrition model and discover the main employee characteristics that affect an employee's attrition, this deep data-driven methodology actually relies on a hybrid strategy.

II. EXISTING METHODS

Employee attrition is the term used to describe the slow loss of workers over time. The majority of the research on employee attrition divides it into two categories: voluntary and involuntary. Involuntary attrition, which refers to an employee being fired by their employer for a variety of reasons, is viewed as the employee's fault. Employees that quit the company voluntarily do so of their own free will.

The subject of this essay is voluntary attrition. Age, salary, and job satisfaction were found to be the best predictors of voluntary job loss in a meta-analytic assessment of the subject. According to other studies, a number of additional factors, including working environment, satisfaction with work, and career opportunity, can affect voluntary attrition.

Employers use machine learning algorithms to forecast the likelihood that an employee will leave, and they then take proactive measures to avoid this happening.

2.1 Disadvantages

Only a few data mining techniques are used in the current systems to predict data. Effects of employee turnover on finances, productivity, and loss of effort for organisations. A trained and experienced person is expensive to replace, thus this is a significant problem.

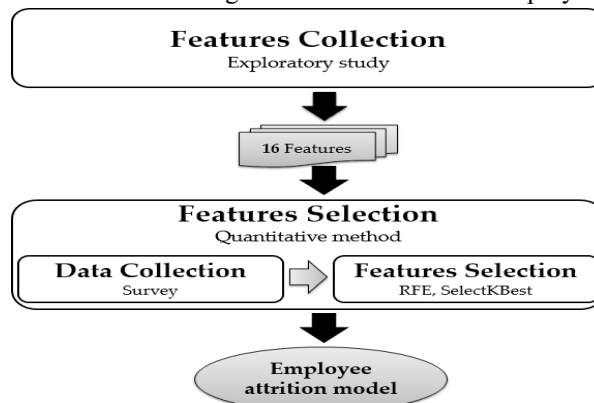
III. PROPOSED METHOD

The study's second section focuses on providing a method for predicting staff attrition. To do this, we'll begin this section with a summary of the relevant literature on attrition prediction strategies based on predictive models. Then, we'll concentrate on the specifics of our recommended predictive approach's steps.

With the aid of our prior research studies and information gathered from the employee survey, we identified the key factors that have the greatest impact on employee attrition and will aid in our ability to accurately anticipate this attrition. The gathered and chosen data will be utilised as an input to our three-step prediction methodology.

Pre-processing of the data is the initial step.

In the second, attrition prediction using machine, ensemble, and deep learning models is covered. The third one addresses interpretation in order to inform HR managers of the causes of this employee loss.



3.1 Advantages

To anticipate future attritions and investigate the causes of employee turnover, we attempt to analyse previous and present employee data. The study's findings show that data extraction algorithms can be used to build trustworthy and precise forecast models for employee attrition.

A. DATA PREPROCESSING

Data pre-processing is one of the crucial processes for improving the training of predictive models. This is accomplished by transforming and encoding the data provided by respondents so that it is suitable for processing and training using the library methods offered and implemented in Python's library scikit-learn [27]. To put all the features on a similar scale, categorical features, for example, were One-Hot Encoded, which involved converting each unique value in the categorical fields to a numerical value. This method prevents outliers from influencing the predictions by normalising data to ranges from -1 to 1.

B. ATTRITION PREDICTION MODEL

The prediction of employee attrition is approached as a supervised learning problem, and more specifically, as a binary classification problem. In other words, we are interested in determining if the employee's intention to quit actually exists or not. In order to do this, we tested various supervised machine learning and deep learning approaches, as well as the implementations offered by the scikit-learn module for Python [29]. We have specifically used the following classifiers:

Random Forest, XGBoost, Vote Classifier, Decision Tree, Logistic Regression, and Support Vector Machine (as machine learning models), as well as three deep learning models (DNN, LSTM, and CNN). Each classifier's hyperparameters were tuned using a gridsearch technique, and the dataset was divided into validation, training, and test sets at a ratio of 10:70:20. The various models were then trained on the training dataset using their optimal configuration

IV. MACHINE LEARNING BASED PREDICTION MODEL

1. Paths from the root to the leaf of a decision tree represent categorization rules during a recursive partitioning procedure [28]. Each internal node denotes a "test" on an attribute, each branch the partitioned result of the test, and each leaf a class label in the case of classification or a numeric value in the case of regression.
2. A supervised learning algorithm called the Support Vector Machine is utilised to solve both linear and nonlinear classification issues. It makes use of a hyperplane or collection of hyperplanes in higher dimensional space to create class separation. The hyper-plane with the greatest distance to the nearest training data points of any class achieves a satisfactory separation, according to the intuition behind this statistical learning-based technique [29].
3. Using the logistic function to describe categorical or binary dependent variables, logistic regression is a straightforward statistical approach and one of the fundamental linear models for classification. It's frequently used in conjunction with regularisation in the form of fines based on the L1-norm or L2-norm to prevent over-fitting [30].

V. DEEP LEARNING BASED PREDICTIVE MODELS

1. The term "deep neural networks" (DNN) refers to artificial neural networks (ANNs) that are deep and have a number of hidden layers—at least two—through which data is processed from the input to the output layers. Each layer in a standard DNN is made up of a group of neurons and an activation function and is fully connected. Each neuron is impacted by a series of weights, each of which is multiplied by one input. After being supplied via the activation function, they are then added together to create the neuron's output.
2. Recurrent neural networks (RNNs) can handle sequential and temporal data and predict time series. Long short-term memory networks (LSTMs) are an improvement on RNNs [32]. In order to develop a more stable RNN for time series prediction by identifying and remembering the long-term dependencies present in the time series, a cell state is added to LSTM to store long-term states.
3. Input, convolutional, pooling, and fully connected layers make up the structure of convolutional neural networks (CNNs) [33], which typically include four different layer types. The input will be convoluted with several filters in the convolutional layer, which is the most significant CNN component, where each filter is viewed as a smaller matrix. Following the convolution operation, the corresponding feature maps will then be produced. The size is decreased while maintaining key characteristics during the pooling operation. As a result, the network's effectiveness is increased and overfitting is prevented.

VI. A ENSEMBLE LEARNING BASED PREDICTIVE MODELS

1. A prominent tree-based ensemble learning technique called Random Forest uses a bagging process in which consecutive trees are built using various bootstrap samples from the dataset. A simple majority vote is finally taken for the prediction. As opposed to regular trees, random forests are robust to over-fitting since each node is split using the best among a selection of predictors randomly selected at that node [35].
2. The three deep learning models (DNN, LSTM, and CNN) that have been selected for this model are stacked together to form a new dataset that includes the real expected value for each row. This dataset will be used to train a new DNN learning model dubbed the meta-learner. To find the optimum hyperparameters for each model, we used GridSearch to validate 10% of the dataset (for example, the decision criterion and maximum depth for DT, the number of hidden layers and units or neurons in each layer for DNN, LSTM, and CNN).
3. A series of weak learners are fitted using the gradient boosted tree algorithm XGBoost, and the combined predictions from all of them are used to construct the final prediction through a weighted majority vote (or sum). This boosting approach is highly robust and performs well because it is based on the usage of a regularized model formalisation to control over-fitting [36].

VII. EXPERIMENTATION RESULTS

We are now prepared to move on with building our models and evaluating their performance after completing an exploratory and thorough data analysis, finding all model settings (parameters and hyper-parameters), and then analysing the data. In fact, we shall discuss the experimental outcomes of machine, ensemble, and deep learning predictive models in this section. The large Kaggle HR simulated dataset (15000 samples), the medium IBM HR simulated dataset (1470 samples), and our tiny HR real dataset (450 samples) are utilised to best evaluate the performance of these prediction models in a variety of scenarios. In order to help the HR manager find retention strategies, the primary contribution of these models will be presented towards the conclusion of this experiment. This will allow them to not only anticipate attrition but also to understand why it occurs. In the parts that follow, evaluation standards for these models and a comparison of their outcomes are explained.

RESULTS OF PREDICTIVE MODELS FOR TWO SIMULATED HR DATASETS

The two simulated human resources datasets are used to assess our predictive algorithms. The first is a sizable dataset that Kaggle supplied, with 15000 samples and a target variable of "left" and 9 features including "satisfaction level," "last evaluation," "number of projects," "average monthly hours," "time spent at company," "work accident," "promotion within the past five years," "sales," and "salary." The second simulated human resources analytics dataset includes an attrition target variable that can be written as "Yes" (employee departed) or "No" (employee did not depart). IBM created this medium-sized dataset, which has 1470 samples and 34 attributes. We will utilise the entire IBM dataset with all 34 characteristics to assess how well our predictors function because our 11 selected features are contained in the 34 features of this second simulated dataset. Then, using the same dataset, we will assess their performance, focusing only on the 11 features of our employee attrition model that we have specifically chosen (marriage status, age, tenure, grade, rewards, job involvement, training, business travel, job satisfaction, job performance, and environment satisfaction). Table 4 presents the results using the two simulated datasets in terms of accuracy (defined as the model's percentage of correctly classified data and a measure of the proportion of all valid predictions) and F1-score.

RESULTS OF PREDICTIVE MODELS FOR OUR DATASETS

Using our actual dataset, we evaluate our classification predictors in this part to determine which predictor is most useful for classifying churners and non-churners.

Models accuracies are measured before and after feature selection algorithms which means that for the first time we use the entire real dataset with its 16 features. Next, models are evaluated using only the 11 features selected after applying the feature selection process by combining RFE and SelectKbest.

VIII. LIST OF MODULES

- Data Collection and Pre-processing
- Data cleaning
- Data transformation
- Data selection
- Training and Testing
- Algorithm

DATA COLLECTION AND PREPROCESSING

One of the most important tasks in creating an AI model is probably information collection. It is a social affair where data about errands is dependant on some targeted factors to research and produce some major outcome. In any event, some of the material can be outrageous, for instance, it might contain false, insufficient, or wrong information. Therefore, handling the data is essential before dissecting it and moving on to the conclusions. Information cleansing, information modification, and information determination need to make information pre-handling practicable.

DATA CLEANING

Cleaning up the information includes adding missing details, taming turbulent material, identifying or getting rid of outliers, and fixing irregularities. Information can alter through smoothing, collection, conjecture, and other processes that enhance its quality. Information selection includes a number of methods or skills that enable us to select the pertinent information for our framework.

DATA TRANSFORMATION

Information transition refers to the organisation and transformation of information from one arrangement to the next. For instance, XML data can be transformed from XML records valid for one XML Schema to another XML record valid for a different XML Schema. Different models take into account the information transition from non-XML to XML.

DATA SELECTION

Information determination is defined as the most popular method of selecting the best information kind, source, and instruments for gathering information. Information selection is done before actual information gathering. Information uprightness can be influenced by the most popular method of selecting reliable information for a research project.

TRAINING AND TESTING

To determine how precise your model is, use the train/test method. The term "train/test" refers to the method's division of the data set into two sets, one for training and the other for testing. 70% of training, 30% of testing. The model is trained using the training set.

ALGORITHM

SUPPORT VECTOR MACHINE

Both regression as well as classification are performed using supervised machine learning techniques known as Support Vector Machines (SVM). Although we also talk about regression issues, classification is the better word. The SVM method aims to find the hyperplane in a space of N dimensions that clearly separates the data points.

RANDOM FOREST

Step 1: Pick K data points at random from the initial data set.

Step 2: Create the decision trees linked to the chosen data subsets.

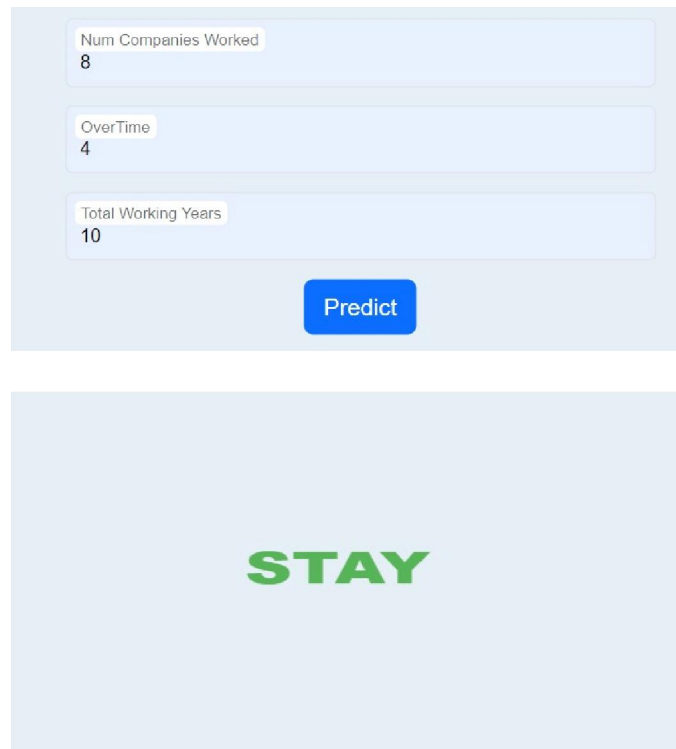
Step 3: Select N to determine the size of the decision trees you wish to construct.

Step 4: Repetition of steps 1 and 2.

Step 5: Assign the extra data to the category that receives the most votes by pulling up the forecasts for every tree of decisions for the data points that were added.

IX. OUTPUT SCREENSHOTS





X. FINDINGS AND DISCUSSION

First, in terms of a quantitative evaluation of the effectiveness of our predictors, Tables 4 and 5's results demonstrate that the ensemble learning model Voting Classifier VC outperforms the others for both simulated and real data. In fact, VC performs better in terms of accuracy than all the other classifiers, especially when using our real dataset as opposed to the simulated ones. In comparison to other machine learning classifiers, ensemble learning VC produces more accurate results for both simulated and real datasets, regardless of the use of feature selection. In particular, for our final dataset, ensemble learning VC gives the best results with an accuracy of 0.99. This can be explained by the fact that the ensemble learning aims to combine (weak) learners in one method by taking advantage of their complementarity to output best accurate results. Additionally, our ensemble learning VC outperforms deep learning predictors in both simulated and real-world data. The volume of data presented may help to explain this outcome. "Relatively" huge datasets are actually necessary for deep learning algorithms to function properly and produce superior results, as well as the infrastructure to train them quickly. Deep learning algorithms are also more effective when dealing with complicated problems and genuine huge data with a larger number of features, but they also require a lot more experience.

Additionally, we present various results in Table 6 to compare the proposed models' accuracy to recent studies that made use of the simulated HR datasets. We point out that our ensemble learning VC delivers the best results with an accuracy of 0.93 for the IBM HR simulated dataset. Ensemble learning VC equally provides the best results for the Kaggle HR simulated dataset, with an accuracy of 0.98.

The main contributions of this study, aside from the presented predictive models and their combination to obtain more precise predictions of employee attrition, mainly focus on two issues. The first is about the suggestions for a deep data-driven predictive methodology. In actuality, rather than using all of the data gathered, our method concentrates on the utilisation of pertinent data and the choice of influential features. It is important to note that feature selection offers a practical method for reducing the complexity of classification issues by eliminating redundant and irrelevant data. This can speed up computation, increase learning accuracy, and enable a better understanding of the learning model. These substantiations were, in accordance with the findings, experimentally demonstrated here as indicated in tables 4 and 5.

In reality, using feature selection results in a noticeable improvement of accuracy measures for the majority of classifiers. We also observe an increase in the F1-score following feature selection. This demonstrates the efficacy of the employee attrition model we used for this study, and the positive outcomes of numerous classifiers following feature selection support the notion that the attributes we chose successfully contribute to voluntary attrition. Reducing the amount of existing features and maintaining only our 11 chosen features has improved predictors' performance even for the human resources IBM simulated dataset. In particular, ensemble method VC accuracy has marginally increased from 0.93 prior to feature selection to 0.94. Additionally, ensemble learning with an accuracy of 0.99, VC applied to our final dataset following feature selection produces the best results. This further affirms that the two feature selection algorithms SelectKBest and RFE are a viable choice to enhance and validate our employee attrition model. Therefore, this in-depth study adds to earlier findings about the features that have an impact on employee attrition reported in the literature and only confirms the necessity of the 11 features that were chosen.

XI. CONCLUSION AND FUTURE WORKS

The major objective of this study is to assist HR managers in preventing attrition by assisting them in applying predictive analytics methodologies to identify an employee's intention to depart as soon as possible. Three things can be said about the contributions:

The suggestion of a novel employee attrition model that uses a mixed research technique and only 11 features that are both essential and sufficient to detect intention to leave and predict positive attrition.

The suggestion of machine, deep, and ensemble learning prediction models, as well as their experimentation in various settings (large-sized simulated dataset, medium-sized simulated dataset, and small-sized real dataset) to better assess their performance.

The explanation and interpretation that help HR managers adopt crucial retention policies by helping them comprehend what causes an employee to desire to quit.

Regarding study restrictions, it will be fruitful to investigate the influence of dynamic elements that deal with employees' behaviour and emotional states on employee attrition. Predictive model training in this situation must be done online because the data will be dynamic and can be updated as needed. We also appreciate that the people who answered our questionnaire provided additional considerations that could lead to voluntary turnover and that could be incorporated into our upcoming research. In fact, they have suggested that the corporation take into account health concerns, employment stability, and the utilisation of new technologies. Finally, because the adopted predictive models are experimentally unsuitable for unbalanced data, taking into consideration unbalanced data in future research will be a real challenge, particularly for organisations and businesses with high turnover rates.

REFERENCES

- [1]. G. King and L. Zeng, "Logistic regression in rare events data," *Political Anal.*, vol. 9, no.2, pp. 137–163, 2001.
- [2]. G.E.Hinton, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.
- [3]. A.Liaw and M.Wiener, "Classification and regression by random Forest," *R News*, vol.2, pp.18–22, Dec.2002.
- [4]. J.H.Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol.29, no. 5, pp.1189–1232, Oct. 2001.
- [5]. J.V.Beaverstock, B.Derudder, J.R.Faulconbridge, and F.Witlox, "International business travel: Some explorations," *Geografiska Annaler, B, Hum. Geogr.*, vol. 91, no. 3, pp. 193–202, Sep. 2009.
- [6]. P. Runeson and M. Höst, "Guidelines for conducting and reporting case study research in software engineering," *Empirical Softw. Eng.*, vol. 14, no.2, pp.131–164, Apr.2009.
- [7]. B.K.Goswami and S.Jha, "Attrition issues and retention challenges of employees," *Int.J.Sci.Eng.Res.*, vol.3, no.4, pp. 1–6, Apr.2012.
- [8]. G. Brown, "Ensemble learning," in *Encyclopedia of Machine Learning*, vol. 312. 2010, pp. 15–19.
- [9]. F.Pedregosa, G.Varoquaux, A.Gramfort, V.Michel, B.Thirion, O.Grisel,

- [10]. M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and J. Van der Plas, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.
- [11]. S. V. Kalinin, B. G. Sumpter, and R. K. Archibald, "Big-deep-smart data in imaging for guiding materials design," *Nature Mater.*, vol. 14, no. 10, pp. 973–980, Oct. 2015.
- [12]. N. Shah, Z. Irani, and A. M. Sharif, "Big data in an HR context: Exploring organizational change readiness, employee attitudes and behaviors," *J. Bus. Res.*, vol. 70, pp. 366–378, Jan. 2017.
- [13]. A. Tursunbayeva, S. D. Lauro, and C. Pagliari, "People analytics—A scoping review of conceptual boundaries and value propositions," *Int. J. Inf. Manage.*, vol. 43, pp. 224–247, Dec. 2018.
- [14]. S. Shah, S. Alatekar, Y. Bhangare, B. Kasar, and R. Patil, "Analysis of employee attrition and implementing a decision support system providing personalized feedback and observations," *J. Crit. Rev.*, vol. 7, no. 19, pp. 2372–2380, 2020.
- [15]. F. Fallucchi, M. Coladangelo, R. Giuliano, and E. W. DeLuca, "Predicting employee attrition using machine learning techniques," *Computers*, vol. 9, no. 4, p. 86, Nov. 2020.
- [16]. S. Kakad, R. Kadam, P. Deshpande, S. Karde, and R. Lalwani, "Employee attrition prediction system," *Int. J. Innov. Sci. Eng. Technol.*, vol. 7, no. 9, p. 7, 2020.
- [17]. S. R. Ponnuru, "Employee attrition prediction using logistic regression," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 8, no. 5, pp. 2871–2875, May 2020.
- [18]. N. Jain, A. Tomar, and P. K. Jana, "A novel scheme for employee churn problem using multi-attribute decision making approach and machine learning," *J. Intell. Inf. Syst.*, vol. 56, no. 2, pp. 279–302, Apr. 2021.