
ISSN: 2321-2152
IJMECE
International Journal of modern
electronics and communication engineering
E-Mail
editor.ijmece@gmail.com
editor@ijmece.com
www.ijmece.com

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

Using Machine Learning to Forecast Employee Departure

[1]Ms. Y. Bindu Rajasri, [2]Sodasani Venkannababu,

[1]Assistant Professor, Department of MCA, Rajamahendri Institute of Engineering & Technology. Bhoopalapatnam, Near Pidimgoyyi,Rajahmundry,E.G.Dist.A.P. 533107.

[2]Student, Department of MCA, Rajamahendri Institute of Engineering & Technology. Bhoopalapatnam, Near Pidimgoyyi,Rajahmundry,E.G.Dist.A.P. 533107.

Abstract

Machine learning is becoming more popular among corporate decision-makers, therefore it's only natural that academics investigate its applications in corporate settings. The departure of brilliant workers is one of the biggest problems that company executives face today. The use of machine learning models to investigate employee turnover is the focus of this study. In order to forecast employee turnover, three primary experiments were carried out using synthetic data generated by IBM Watson. As a preliminary step, we trained different machine learning models on the original class-imbalanced dataset. These models included random forest, K closest neighbour, support vector machine (SVM) with several kernel functions, and another experiment. To address the issue of class imbalance, the second experiment employed an adaptive synthetic (ADASYN) technique. After that, the aforementioned machine learning models were retrained on a fresh dataset. In the third trial, we manually undersampled the data to ensure that the classes were balanced. Training an ADASYN balanced dataset using KNN (K = 3) yielded the best results, with an F1-score of 0.93. In the end, 12 features out of a total of 29 features were used to generate an F1-score of 0.909 utilizing feature selection and random forest. Machine learning, employee attrition, support vector machine, random forest, K closest neighbours, feature ranking, feature selection are all terms that may be used to describe this process.

INTRODUCTION

Staff attrition occurs when workers leave an organization for various reasons, such as dissatisfaction with their work, poor pay, or an unpleasant work environment. There are two main types of employee turnover: voluntary and involuntary. Employees experience involuntary attrition when their employers fire them for various reasons, such poor performance or company needs. In contrast, voluntary attrition occurs when high-performing workers voluntarily depart from the firm, even if the company has made efforts to keep them. Some causes of voluntary attrition include, but are not limited to, employment offers from competing companies or early retirement. Even while employee-focused businesses often put a lot of resources into training and creating a positive work environment, they nonetheless lose brilliant workers and experience voluntary attrition. Another problem is that it is expensive for the business to hire replacements, which includes the time and money spent on interviews, interviews, and training. Management may improve internal rules and tactics in response to employee turnover rate predictions. When good workers who are about to leave might be provided various incentives, such a pay raise or better training, to stay. Businesses may anticipate staff turnover with the use of machine learning models. formulate a machine learning algorithm capable of identifying departing workers. These models are taught to look for similarities and differences between characteristics of current and former workers.

RELATED WORK

Because of the far-reaching consequences, voluntary employee turnover is a key issue for any business. It may be difficult and time-consuming to replace talented people, who are a key component of a company's success [1]. Scientists have looked at what causes employees to voluntarily leave their jobs. Several variables may significantly contribute to employee attrition, as shown in the research study. Offering as remuneration is a key component influencing employee turnover and performance, as shown, for example, in [2] and [3]. The rate of

employee turnover decreases as remuneration improves. Despite this, additional variables, including workload, performance compensation, and a lack of a strong career plan, have contributed to the high turnover rate in the retail sector ([1]). Using machine learning to foretell how employees would act has been the subject of many investigations. When predicting employee performance, the authors of [4] used decision trees (ID3 C4.5) and the Naïve Bayes classifier. They discovered that job title was the most significant factor, whereas age did not seem to have any discernible impact. The authors of [5] used a dataset with 1575 records and 25 attributes to investigate several data mining methods for the purpose of predicting staff turnover (or attrition). The following machines learning methods were employed: naïve Bayes, support vector machines, logistic regression, decision trees, and random forests. An SVM, with an accuracy of 84.12%, is suggested by the study's findings. Classification and regression trees (CART), C4.5, and REPTree were among the decision tree algorithms investigated in [6]. The researchers used a dataset comprising 309 employee records (out of 4,326) and six attributes to train and evaluate the decision trees. Consequently, out of all the decision trees tested, the C5 decision tree had the best accuracy rate of 74%. Salary and duration of service were also identified as significant factors in the dataset of the examined organization. The authors of [7] forecasted the employee turnover rate for small-west manufacturing firm using neural networks. Thus, they came up with the neural network simultaneous optimization method (NNSOA) and 10-fold cross validation, which together produced a 94% accurate turnover rate prediction. In addition, by using a tweaked genetic algorithm, they determined which "Tenure of employee on January 1" was the most crucial and pertinent. To forecast employee turnover, the authors of [8] combed over 6,909,746 online employee profiles. Included in the profiles were details on the individuals' schooling and job history, as well as their

do an SVM model evaluation. Clearly, the model's accuracy was low, with an average of 55%. The study's author suggested enhancing the trained model by supplementing the dataset with additional personal characteristics, such as employees' ages, genders, and workplaces. [9] forecasted staff turnover for an American branch of a multinational store. The dataset had 73,115 observations and 33 characteristics. With an area under the curve (AUC) of 0.88, XGBoost emerged as the best accurate model out of seven machine learning algorithms studied by the researchers. When compared to the other models, it also fared better in terms of memory use. A model for predicting staff turnover at Swedbank was created by the author in [10]. With an accuracy of 98.6%, a random forest model surpassed SVM and MLP models in this investigation. In their use of diverse datasets and machine learning models, prior research has offered a variety of accuracy metrics. Consequently, settling on the optimal model to use is somewhat challenging. Furthermore, the class imbalance issue that is present in real-world attrition data has not been addressed in earlier research. Consequently, we investigated a number of approaches to address class imbalance, which greatly improved the training procedure. What follows is an outline of the rest of the paper. Methods that were considered for this study are detailed in Section 3. The investigation is concluded in Section 5, after which Section 4 describes the experimental setup and outcomes. Part III: Advised Procedures We have looked at three primary experiments to forecast staff turnover in this study. We started by trying to forecast staff turnover using the initial unbalanced dataset (section IV provides specifics of the data). To address the issue of class imbalance, we used the adaptive synthetic sampling strategy in the second trial. The "yes" class, which was a minority in this instance, was oversampled using this method. The third trial included a random undersampling of the data, in which we chose an equal number of subjects from each category. Additionally, in order to forecast an unknown dataset of employee turnover, each experiment trained and validated a set of machine learning classifiers. A 5-fold cross-validation procedure was used to verify all classifiers. To further reduce the complexity of the trained models and improve their performance, we have also devised a feature selection approach. Increasing the amount of features for each cycle allowed us to train and test each classifier repeatedly. Detailed descriptions of the recommended approaches follow. We shall present the classifiers utilized in this study below. A. Classification In this article, we have employed various current machine learning classification models to categorize unseen data. Classification and regression are two applications of support vector machines (SVMs), a kind of non-probabilistic supervised machine learning. Using a decision boundary, often called a hyperplane, SVMs will train algorithms with specified classes [11], [12]. When the decision boundary is not easily discernible, we say that the issue is nonlinear. Nevertheless, a kernel function—sometimes called a kernel trick—can resolve this issue. After assigning the result of the dot product to one vector and one high-dimensional space, this method returns the result of the product. In addition, many kernel functions, including linear, Gaussian, and polynomial kernels, are available [13],

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

[14]. When it comes to supervised machine learning algorithms, random forest (RF) ranks high in terms of power for producing regressions and classifications. Data is trained using RF using multiple decision trees [15]. The RF model takes into consideration the number of votes cast by each decision tree in order to determine the most popular class for a given dataset [16]. When it comes to classification and regression, one of the easiest machine learning methods to utilize is K-nearest neighbours (KNN). To make KNN function, one must set the value of K, which represents the number of nearest training points for a particular data point. The majority of the votes from each data point's neighbors will be used to classify each new data point [13]. B. A method for adaptive synthetic sampling By using the minority class's density distribution to generate new synthetic instances, the ADASYN method eliminates class imbalance [20]. To do this, ADASYN will modify the weights for instances belonging to the minority class using adaptive learning. Consequently, the decision border will move, facilitating the acquisition of knowledge from challenging situations. Part C: Choosing Features Feature counts in real-world datasets could be really high. It is possible that training machine learning algorithms may be negatively affected by some of these characteristics as they are deemed noise. Model performance and training duration will be impacted by the increased complexity caused by using all accessible features [21]. It is possible to rank all characteristics using a variety of approaches. In this study, we compute the means and standard deviations of the training data points' binary class labels using the t-test technique. Here is one way to express the t-test formula:

$$t(x)=(\bar{y}_1(x)-\bar{y}_2(x))/\sqrt{(s_1^2(x)/n_1+s_2^2(x)/n_2)} \quad (1)$$

in where are the class means, and are the class labels' standard deviations, divided by the sample size.

.

DATASET AND TOOLS

The This study made use of a dataset that is available to the public via IBM Watson Analytics1. Artificial data generated by data scientists at IBM makes up the bulk of the dataset. There are a total of 1470 workers' HR records in the dataset, which has 32 different attributes. Furthermore, 1233 active workers were in the "No" attrition category, meaning that

Of the 237 departing workers, 207 fell into the "Yes" attrition group. Two elements were eliminated from this research: "Standard hours" since all workers have the same standard hours and "Employee count" because it is a series of numbers (1, 2, 3..). Also, for processing, all non-numerical data were given numerical values, such Sales=1, R&D=2, and HR= 3. Additionally, the study used MATLAB R2017b for training and evaluating the ML models. Section V: Lab Work Here, we provide the outcomes of the three primary dataset trials. With each trial, several ML models were trained. We used the F1 score, recall, accuracy, and precision to rank the models. The subsections that follow elaborate on this. Part A: Assessing Work Quality F1 score, recall, accuracy, and precision were the metrics used to assess each trained model.

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FN+FP} \quad (2)$$

$$\text{Precision}=\frac{TP}{TP+FP} \quad (3)$$

$$\text{Recall}=\frac{TP}{TP+FN} \quad (4)$$

$$\text{F1 Score}=2*\frac{\text{Precesion}*\text{Recall}}{\text{Precesion}+\text{Recall}} \quad (5)$$

where TP stands for true positives, TN for true negatives, FP for false positives, and FN for false negatives. Experiments with Unbalanced Data (B) In this part, we use the initial class-imbalanced dataset to forecast employee turnover. Classification models such as SVM, random forest, and KNN were examined in this study. To start, we used all of the characteristics in the dataset to test how well each classifier performed. The next step in evaluating classifiers was to rank and pick the most significant subset characteristics. You can see how different categorization models fared in Table I. Although the F1 score was very low, the accuracy was 86.9% after training using linear SVM. The majority of the minority group is being incorrectly classified, as this shows. Various kernel types, including quadratic, cubic, and Gaussian, were used to train SVM for future analysis. However, F1 performances remained poor. Quadratic support vector machine (SVM) produced an F1 score of 0.503. Random forest and KNN work in the same way. Training KNN with varied K values (1, 3, 5, and 10) did not improve its

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

performance compared to SVM. In order to rank all features in the unbalanced dataset, feature ranking was used. After calculating the two-sample t-test, it produced an ordered index of the most significant features that were used during training. Among them, monthly salary, work level, and overtime ranked highest. As a further step in our investigation, we trained and evaluated a linear SVM algorithm with only the two most important features—monthly income and overtime—and achieved an accuracy rate of 83.9%. But it failed to mark any data point as "Yes" for attrition, therefore F1 got a zero. The F1 scores features significantly while employing SVM with multiple kinds of kernels. In addition, the F1 score was poor while training the random forest with just two features. Even after adjusting for overtime, monthly salary, and work level, its accuracy was remained poor in comparison to other characteristics (F1). The top two and three characteristics were also used to train KNN. In both trials, when K=1 and 5, KNN produced no F1 outcomes. There was a statistically significant drop in KNN performance when K = 3. The findings were negligible, even after continuing feature selection up to 12 characteristics in this investigation. Consequently, there was no discernible uptick in model performance when utilizing feature selection on data that was unbalanced.

TABLE I CLASSIFIER PERFORMANCE WITH IMBALANCED DATA

Model Type	Accuracy	Precision	Recall	F1 Score[a]
Linear SVM	0.869	0.814	0.240	0.371
Quadratic SVM	0.871	0.662	0.405	**0.503**
Cubic SVM	0.841	0.508	0.418	0.458
Gaussian SVM	0.865	0.788	0.219	0.343
Random Forest	0.856	0.75	0.164	0.269
KNN (K=1)	0.827	0.275	0.046	0.079
KNN (K=3)	0.8374	0.25	0.004	0.008

[a] Bold values indicate highest F1 score

Using Oversampling to Balance Data This section uses a dataset that has been artificially balanced to forecast staff turnover. To get to this point, we used the ADASYN approach after we scaled the dataset. Consequently, additional synthetic data points were created in order to oversample the 'Yes' minority class. As a result, there was an increase of 1152 observations in the "Yes" class, while the 1233 observations in the "No" class remained unchanged. The overall performance for all prediction models was greatly increased when trained with balanced classes, as shown in Table II, which compares the performance of numerous classification models when trained with all characteristics. The F1 score was improved to 0.779 after training using linear SVM. Using quadratic, cubic, and Gaussian kernels to train SVM resulted in even better F1 scores: 0.881 for quadratic SVM, 0.927 for cubic SVM, and 0.912 for Gaussian SVM. Kernels may be used to transfer data to higher dimensions, which aids in defining the ideal boundary, and thus proves that the newly balanced dataset is nonlinearly separable. Furthermore, random forest was used for both training and evaluation on the balanced dataset. Random forest, in contrast to the unbalanced dataset, managed to reach F1 scores of 0.921. In addition, KNN was trained using a range of K values, including 1, 3, 5, and 10. Overfitting may have been the cause of KNN's very high scores when K = 1. During this time, KNN was able to reach F1 scores of 0.931 and.909 with K = 3 and K = 5, respectively. Last but not least, KNN's performance suffered when K = 10, resulting in F1 scores of 0.88.

TABLE II. CLASSIFIER PERFORMANCE WITH SYNTHETIC BALANCED DATA

Model Type	Accuracy	Precision	Recall	F1 Score[b]
Linear SVM	0.782	0.763	0.795	0.779
Quadratic SVM	0.879	0.839	0.927	0.881
Cubic SVM	0.926	0.879	0.981	**0.927**
Gaussian SVM	0.912	0.885	0.941	**0.912**
Random Forest	0.926	0.950	0.893	**0.921**
KNN (K=1)	0.967	0.939	0.997	**0.967**
KNN (K=3)	0.929	0.877	0.992	**0.931**
KNN (K=5)	0.904	0.843	0.987	**0.909**
KNN (K=10)	0.872	0.804	0.970	0.880

Once the synthetic data points were generated, the top features contributing to the training process were ranked using the feature ranking algorithm. Overtime, total working years, and job level were determined to

ISSN 2321-2152
www.ijmece.com
Vol 13, Issue 2, 2025

be the top three characteristics. Table III shows that out of all the models, random forest had the best outcomes with an F1 score of 0.829. After being trained with the two characteristics, the remaining prediction models performed quite poorly. While random forest achieved 0.806 with only three features, similar results were seen when trained using just two features. To include the 12 top traits, the tests continued. Here are the top 12 traits that were employed in the training: Table IV. Random forest achieved F1 scores of 0.909 with only 12 subset features. In addition, KNN achieved scores of 0.882,.861, and 0.839 for K = 3, 5, and 10, respectively. More than 0.83 F1 scores were achieved by cubic and Gaussian SVMs as well. D. Employing Undersampling to Maintain Data Balance Here, to address the issue of class imbalance, we forecast staff attrition by manually undersampling the dataset. Each class contained 237 observations total, and this was achieved by randomly picking an equal amount of observations for each. The total number of observations in the new dataset was 474. When trained with all features, Table V compares the performance of many categorization models. Support vector machines (SVMs) allowed us to get the highest possible F1 score; our quadratic SVM achieved 0.74, while our linear and Gaussian SVMs also achieved 0.73. Random forest and cubic SVM both achieved F1 scores of 0.69. Lastly, at K=10, KNN produced poor results (up to a value of 0.59). According to these findings, manual undersampling might cause crucial data that could be useful for forecasting attrition to be lost.

TABLE V. CLASSIFIER PERFORMANCE FOR UNDERSAMPLED DATA

Model Type	Accuracy	Precision	Recall	F1 Score[d]
Linear SVM	0.745	0.754	0.725	**0.739**
Quadratic SVM	0.747	0.760	0.722	**0.740**
Cubic SVM	0.707	0.733	0.650	0.689
Gaussian SVM	0.751	0.779	0.700	**0.738**
Random Forest	0.717	0.756	0.641	0.694
KNN (K=1)	0.589	0.595	0.552	0.573
KNN (K=3)	0.573	0.572	0.586	0.579
KNN (K=5)	0.565	0.562	0.586	0.574
KNN (K=10)	0.588	0.584	0.611	0.597

[d] Bold values indicate highest F1 score

Feature ranking and selection were implemented in this part despite the poor undersampling findings. For the purpose of training, the most important characteristics were ranked using the feature ranking tool. The results showed that total working years, years with present management, and overtime were the top three attributes. When feature selection is used, all prediction models perform as shown in Table VI. The results obtained by Gaussian SVM, random forest, and KNN were quite similar, ranging from 0.66 to 0.68. Actually, the majority of observations were labeled as 'Yes' by KNN. Using all three characteristics during training also yielded fairly similar outcomes.

TABLE VI. CLASSIFIES PERFORMANCE WITH FEATURE SELECTION FOR UNDERSAMPLED DATA

Model Type	No. Features	Accuracy	Precision	Recall	F1 Score[e]
Linear SVM	2	0.652	0.698	0.536	0.606
Cubic SVM	2	0.631	0.661	0.536	0.592
Gaussian SVM	2	0.681	0.676	0.696	**0.686**
Random Forest	2	0.679	0.682	0.671	**0.677**
KNN (K=1)	2	0.523	0.511	0.995	**0.676**
KNN (K=3)	2	0.506	0.503	0.995	**0.668**
KNN (K=5)	2	0.5	0.5	1	**0.666**
Linear SVM	3	0.652	0.698	0.536	0.606
Cubic SVM	3	0.515	0.524	0.316	0.395
Gaussian SVM	3	0.67%	0.687	0.620	**0.652**
Random Forest	3	0.618	0.612	0.646	0.628
KNN (K=1)	3	0.5253	0.513	0.953	**0.667**
KNN (K=3)	3	0.5464	0.525	0.945	**0.675**
KNN (K=5)	3	0.5	0.527	0.919	**0.670**

[e] Bold values indicate highest F1 score

CONCLUSION

Businesses suffer greatly when their staff turnover rate is high. For firms that put resources into their personnel, losing top performers is like losing a tooth. Because of how hard it is to find suitable successors, the business may end up spending more time and money than necessary. Predicting employee turnover using feature-based machine learning models was the

primary goal of this study. Signs backed by machine learning technologies will be sent to the company's management. Management will be able to take swifter action as a consequence, lowering the probability that brilliant personnel will leave. In order to build prediction models, this study employed three different experimental methods on the dataset. To begin, many prediction models were used to train the initial unbalanced data; the best of these models was quadratic SVM, which achieved an F1 score of 0.50. The second point is that the ADASYN method allowed for parity between the two groups. All four models—cubic, Gaussian, random forest, and KNN (K = 3)—noticedably improved their performance, with F1 values ranging from 0.91 to 0.93. Feature selection also yielded very similar results: random forest got 0.92 F1 scores with only two features and 0.90 with the top twelve features. Manually undersampling the dataset to achieve equal classes was the last approach. Performance suffered because crucial data was lost in the process. But with all the characteristics, SVMs captured over 0.70, and with just two features, they captured over 0.60.

References

[1]. S. Kaur and R. Vijay, "Job Satisfaction – A Major Factor Behind Attrition or Retention in Retail Industry," Imperial Journal of Interdisciplinary Research, vol. 2, no. 8, 2016.

[2]. D. G. Gardner, L. V. Dyne and J. L. Pierce, "The effects of pay level on organization-based self-esteem and performance: a field study," Journal of Occupational and Organizational Psychology, vol. 77, no. 3, pp. 307-322, 2004.

[3]. E. Moncarz, J. Zhao and C. Kay, "An exploratory study of US lodging properties' organizational practices on employee turnover and retention," International Journal of Contemporary Hospitality Management, vol. 21, no. 4, pp. 437-458, 2009.

[4]. Q. A. Al-Radaideh and E. A. Nagi, "Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance," nternational Journal of Advanced Computer Science and Applications, vol. 3, no. 2, p. 144–151 , 2012.

[5]. G. K. P. V. Vijaya Saradhi, "Employee churn prediction," Expert Systems with Applications,, vol. 38, no. 3, pp. 1999-2006, 2011.

[6]. D. A. B. A. Alao, "Analyzing employee attrition using decision tree algorithms," Computing, Information Systems, Development Informatics and Allied Research Journal, no. 4, 2013.

[7]. R. S. Sexton, S. McMurtrey, J. O. Michalopoulos and A. M. Smith, "Employee turnover: a neural network solution," Computers & Operations Research, vol. 32, no. 10, pp. 2635-2651, 2005.

[8]. Z. Ö. KISAO˘GLU, Employee Turnover Prediction Using Machine Learning Based Methods (Thesis), MIDDLE EAST TECHNICAL UNIVERSITY, 2014.

[9]. R. Punnoose and P. Ajit, "Prediction of Employee Turnover in Organizations using Machine Learning Algorithms," International

[10]. Journal of Advanced Research in Artificial Intelligence, vol. 5, no. 9, 2016.

[11]. M. Maisuradze, Predictive Analysis On The Example Of Employee Turnover (Master's thesis), Tallinn: Tallinn University of Technology, 2017.

[12]. K.-B. Duan and S. S. Keerthi, "Which is the best multiclass SVM method? An empirical study," International workshop on multiple classifier systems, 2005.

[13]. K. P. Bennett and C. Campbell, "Support vector machines: hype or hallelujah?," Acm Sigkdd Explorations Newsletter, vol. 2, no. 2, pp. 1-13, 2000.

[14]. S. Rogers and M. Girolami, A first course in machine learning, CRC Press, 2016.

[15]. N. Cristianini and B. Scholkopf, "Support vector machines and kernel methods: the new generation of learning machines," Ai Magazine, vol. 23, no. 3, p. 31, 2002.

[16]. T. K. Ho, "Random decision forests," in proceedings of the third international conference on Document Analysis and Recognition, 1995. [16] L. Breiman, "Random forests," Machine learning, vol. 45, no. 1, pp. 5-32, 2001.

[17]. X. Zhu, Knowledge Discovery and Data Mining: Challenges and Realities: Challenges and Realities, Igi Global, 2007.

[18]. D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC,

informedness, markedness and correlation," Journal of Machine, 2011.

[19]. H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Transactions on knowledge and data engineering, vol. 21, no. 9, pp. 1263-1284, 2009.

[20]. H. He, Y. Bai, E. A. Garcia and S. Li, "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning," in IEEE International Joint Conference on Neural Networks, 2008.

[21]. I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," Journal of machine learning research, vol. 3, pp. 1157-1182, 2003.

[22]. W. Zhu, X. Wang, Y. Ma, M. Rao, J. Glimm and J. S. Kovach, "Detection of cancer-specific markers amid massive mass spectral data," Proceedings of the National Academy of Sciences, vol. 100, no. 25, pp. 14666-14671, 2003.