Proceedings of the International Conference on Industrial Engineering and Operations Management
Riyadh, Saudi Arabia, November 26-28, 2019

Predicting Employee Attrition Using Machine Learning

Imtinan F. Alsuahim, Feddah A. Alotaibi, Malak A. AlAsiri, Munirah S. Alkharji, and Shahad A. Alharthi

Emtnanf.is@gmail.com, FeddahAlotaibi@gmail.com, malak.abdullah.alasiri@gmail.com , munira.alkharji0@gmail.com, and shahad95alharthi@hotmail.com

School of Computing
Dublin City University
Dublin, Ireland

Abstract

One of the major issues facing business leaders within companies is the loss of talented employees. Several factors that might employee attrition are studied to predict whether a certain employee is leaving or not. Predicting Employee Attrition using Learning is the aim of this research. This research studies employee attrition using machine learning models with public data IB-world datasets may include a large number of features, that's why this research used three different methods for features selection: using all features, Weka techniques, and features used in literature. After preprocessing data and performing label encoding, dataset was splatted into training and test set. Several classifiers include K-nearest neighbors (KNN), Naive Bayes, Random Forest and Support Vector Machines (SVM) have been trained on 80% of data and they were validated using cross validation. Models that gave a better performance on training were used to test on unseen data. To evaluate those trained models on test set, several evaluation metrics were used. For the highest performing model SVM with accuracy 85.3% and f1-score 81%. To investigate more in the performance of SVM, an evaluation metric known as AUC ROC (Area Under the Receiver Operating Characteristics) were used. **Keywords:**

Data Analytics, Human Resources, Employee Attrition, Supervised Learning, and Classification.

1. Problem Definition

The employee attrition is defined as a process to reduce the strength of a particular object and thus reduce effectiveness. For an organization, employees are invaluable assets and attrition of employee does particularly affect the balance of enterprise resources and growth strategies. Reduced opportunities, the working environment, not satisfaction with the job profile, and also the challenges faced with the management can in turn lead to employee attrition. These and many other problems also hamper the status of the organization and thus to find a solution to the employee attrition, various methods have been implemented to predict the employee attrition. In this research, several factors that might lead to employee attrition are studied to predict whether a certain employee is leaving or not.

2. Related Work

HR analytics is an interesting research area, various research papers trying to solve the problem of losing high skilled employees which can result in a negative impact on the working of the organization. That is by studying the factors that affect employees' attrition.

2.1 Predicting Employee Attrition using Machine Learning

The growing interest in machine learning among business leaders and decision makers demands that researchers explore its use within business organizations. As argued by Alduayj S. and Rajpp̲ot K. (2018), one of the major issues facing business leaders within companies

is the loss of talented employees. This research studies employee attrition using machine learning models such as SVM, random forest, and KNN classification models. Using a synthetic data created by IBM Watson. From the dataset, two features were removed: 'Employee count' and 'Standard hours'. Mentioned classifiers were evaluated by ranking and selecting the important subset features only. SVMs were able to capture more than 0.70 using all the features, and more than 0.60 with just two features. All trained models were evaluated by measuring their accuracy, precision, recall, and F1 score. Adaptive synthetic (ADASYN) approach used overcome class imbalance as well as under sampling data to balance. Their best algorithm was KNN K=3 with 93% accuracy. Highest f1-score resulted from using quadratic SVM was only 53%. When balanced data techniques were used the performance of these models increased. However, using under sampling resulted in f1-score bellow 74%.

2.2 Evaluation of Employee Attrition by Effective Feature Selection using Hybrid Model of Ensemble Methods

According to Jain (2017) Employees are leaving organization's pre-maturely that results in high losses for the organization which cannot be predicted by HR. This paper contributes to HR predictive analytics (HRPA) that helps in a predicting the employees who will leave the organization in a certain period of time by using a hybrid model of machine learning techniques. This research explains how predicted accuracy, sensitivity and specificity can be enhanced by the use of ensemble methods in determining employee attrition with the feature selection method using efficient feature engineering, data wrangling, visualizing and analyzing results from previous models to increase the accuracy.

Different types of ensemble methods like stacking, boosting and bagging were tested for classification, however boosting algorithms performed the best but Adaptive boosting gave the best evaluation score with getting up to 88.8% accuracy. The CRISP-DM method using three different ensemble methods was used (1-stacking) GLM, SVM, Decision Trees, KNN, (2-bagging) Random Forest and (3-boosting) GBM, Adaptive boosting (ADA). This achieved 88% accuracy by these techniques from which HR can place a sound strategy to raise employee retention. This research is limited to a small data- set which lacks to train the model well that might give low results. The second drawback is with the model is limited to only supervised machine learning that requires a lot of computation time.

2.3 Proactive Intervention to Downtrend Employee Attrition using Artificial Intelligence Techniques

Barvey et al. (2018) mentioned in their research that to predict the employee attrition beforehand and to enable management to take individualized preventive action. Using Ensemble classification modeling techniques and Linear Regression as well as feature engineering. Model could predict over 91% accurate employee prediction. The scope of model is restricted within the variables feuded in the algorithm and purity of available data.

2.4 Employee Attrition Prediction Using Data Mining Techniques

In the research conducted by: Sukhadiya et al. (2018) tries to predict employee attrition. Authors have used random forest and extreme gradient boosting for features selection. The models that were developed are: SVM, Random Forest, Logistic Regression and extreme Gradient Boosting. Ensemble average had the highest accuracy of 0.855127. However, the top 5 factors that are highly responsible for employee attrition include overtime, monthly income, daily rate, age, and total working years. The classification techniques used give good accuracies but amongst all the techniques implemented in this paper, Extreme Gradient Boosting proves to have a upper hand on the attrition prediction task. This paper did not use any domain knowledge for the purpose of features selection.

2.5 An Approach for Predicting Employee Churn by Using Data Mining

In the research conducted by: Yigit and Shourabizadeh (2017) well-known classification methods including, Decision Tree, Logistic Regression, SVM, KNN, Random Forest, and Naive Bayes methods were used on the HR data. Then, they analyzed the results by calculating the accuracy, precision, recall, and F-measure values of the results. Moreover, they implemented a feature selection method on the data and analyze the results with previous ones. SVM is their best model in terms of accuracy and f-measure, where its accuracy is 0.857 and its F-score is 0.28. The results will lead companies to predict their employees' churn status and consequently help them to reduce their human resource costs. The problem of churn prediction is not just to identify churners from no churners. It would be useful to

build a comprehensive and universal model that the organization can use for the better of the employees, cost effectiveness and future prospects to get the most advantages of this approach.

3. Proposed Method

Various supervised machine learning algorithms and data mining techniques have been implemented to predict employee attrition. The techniques used are K-nearest neighbors (KNN), Naïve Bayes, Random Forest, and Support Vector Machines (SVM). Each method was discussed in detail in the experiment section.

4. Dataset Description

Dataset was downloaded from IBM website in comma separated value (.CSV) format. Which has 1470 rows and 35 columns. Some of the features are numeric such as: Age, Daily Rate, Distance from home, Hourly Rate, Monthly income, Monthly Rate, Number of Companies Worked, Percent Salary Hike, Total working years, Years at Company, Years in current Role and Years with current Manager. The rest of the features are of categorical data type such as: Business Travel, Department, Education, Gender, Environment satisfaction, Job level, Job Role, Marital Status, Over time, Performance Rating, Stock option level, education field and Work life balance. In addition to Job Involvement, Job Satisfaction, Relationship Satisfaction, Training Times Last Year, Years Since Last Promotion and Attrition.
Attrition is the main (target) column which is to be predicted whether employee leaves the company or not (yes or no) and other columns are the independent variables which are taken into consideration to build the models. The Dataset contains factors like age, marital status, gender, job satisfaction, monthly rate and work-life balance, ethnicity, education were mostly used to predict the employee attrition in many researches (Ajit; 2016) and (Kursa et al.; 2010) but this research extends these researches by considering 25 attributes and dropping 10 columns which were of no use.

5. Data Preparation:

IBM dataset as mentioned, contains 35 columns and 1470. However, some of these columns are not providing any useful or meaningful information. They also affecting negatively on models' performance. These columns are: Employee count where all the column has a value of 1, Standard hours where all values of this column are 80 which doesn't distinguish any employee from the others. The other columns are: Employee number and Over 18 where obviously most the employees are over 18 and the Employee number column is nothing but ascending census from 1 to 1470. As a result, we decided to drop those columns to allow the models to get more reasonable results. The data has no missing values.
As per industry practice we split the dataset into (80% train and 20% test) using python. Before building the models in Python, we had to convert all categorical data into numeric data. We used some libraries to help us with that, these libraries are: from sklearn preprocessing and LabelEncoder.

6. Feature Selection

Real-world datasets may include a large number of features. Some of these features are considered noise and might not have a positive influence on training machine learning algorithms. Using all available features will increase model complexity, hence affecting model performance and training time. There are different methods that can be used to evaluate and rank all features.
Feature Selection is the process where user automatically or manually select from available features which contribute most to the right prediction variable or output. Having irrelevant features in the dataset can decrease the accuracy of the models and make the model learn based on irrelevant features.

There are several methods for features selection, some of them are used in this report. Three features selection techniques have been used in our research. One way is to using feature selection techniques in Weka. Using Weka is helpful to decide which features has higher effect on the target feature. Weka is a combination of algorithm used in machine learning for data analysis and there are some methods to select features such as: Info gain (ranker) and Cfs subsetEval (best first). Another way to select features is using domain knowledge which is done by using same set of features used by researchers in current literature who are using same dataset. Last way to select feature is when all features are considered. The following table (Table 1.) describes different methods used and their associated subset of features as well as their evaluation accuracy using SVM algorithm:

Table 2. Methods and SVM algorithm accuracy

Preprocessing way	Feature	Method	Percentage (SVM)
Using domain knowledge and features in literature	environment satisfaction, years with current manager, marital status, monthly income(salary), Job level, Overtime, Stock option level, job role , Job satisfaction, and total working year.	Based on current literature	83% - 84%
Using all Features	All feature are considered	No method is used	87% - 88%
Using feature selection techniques	job role, total work years, age, monthly income, job level, overtime, years with current manager, years at company, years in current role, Stock option level, marital status, business travel, environment satisfaction, education field, job involvement, work life balance, department, job satisfaction, gender.	Info gain (ranker)	86%
	Age, business travel, environment satisfaction, job involvement, job level, job satisfaction, monthly income, overtime, Stock option level, total work years, work life balance, years at company, years with current manager.	Cfs subsetEval (best first)	84%

As a result of using different feature selection techniques and domain knowledge. It is clear that these subsets of features are not helping, and all features have to be considered in the experiment.

7. Experiment:

After preprocessing data and doing the label encoding, the training set was splatted into train data which include all features except the last column and train target which contains the class. The class is attrition column that may take yes or no as possible values. To predict employee attrition, several classifiers were trained such as KNN, Naive Bayes, Random forest and Support Vector Machine to solve a binary classification problem. Since our data is imbalanced, there has to be some trade-off between variance and bias to avoid any overfit or underfit. So, we choose the settings that lead to balance those. And in our experiments gave us a low variance meaning no overfitting of data.
The first classifier trained was KNN which is non-parametric algorithm that make no assumptions about the data. Thus, KNN is a good choice when the study does not include any prior knowledge. Since our problem is a binary classification, KNN could deliver a good performance, however it might be affected by class imbalance. For the purpose of training the model on the data, the hyper parameter K (number of neighbors to consider when assigning class label) was changed each time K=3, K=5, K=7. For evaluating each KNN classifier on the training set, cross validation technique was used in which data splits into number of folds and all entries in the data are used for training and validation.

There are few points to mention about cross validation that was taken into account when choosing the number of folds in cross validation. As number of folds increases the difference between data decreases hence the bias becomes smaller. So, we decided to use 10-fold cross validation to evaluate the model on training set. Each KNN classifier were evaluated and the mean accuracy is calculated for each. For KNN where K=3 the mean accuracy was 81%, K=5 the mean accuracy was 82% and for K=7 mean accuracy was 83.6%. It is clear that as we increase the number of neighbors in KNN the performance of the model increases. As a result, a KNN algorithm with K=7 will be tested on unseen data which is described in the next section.

The second algorithm trained is Naive Bayes, which assumes independency between features. At first Gaussian Naive Bayes was trained on training data and the training mean accuracy using cross validation was 78%. However, since our data is not normally distributed and has so many binary values; another special type of Naive Bayes was used which is Bernoulli. Bernoulli naive Bayes was built with

alpha=1 as smoothing parameter. The training result using cross validation was 82.4% with a very low value of variance which indicates a low variance in the data. A Bernoulli Naive Bayes is more suitable to be tested on unseen data which is discussed in next section.

The third algorithm trained is Random Forest, ensemble learning algorithms that constructs multiple trees and compute the majority vote when performing prediction. We trained the model using 300 as number of estimators (trees) and Gini as creation method. After that, the algorithm evaluated on training set using cross validation with 10-fold. The mean accuracy was 86%, which shows improvements over the previous classifiers. Testing this model on unseen data is discussed in the next section. Also, the standard deviation between those resulted accuracies was calculated, and it gave us 0.014 that is 1.4 which means a very low variance. This indicates that the model generated was not obtained by chance.

The fourth and last algorithm trained is SVM, which is defined by a separating plane or a separating line. In 2D space, this algorithm generates an output which is a line that divides the data into two parts with different classes on the either side of the line (say for example class 0 and class 1). The implementation of the SVM algorithms is done using a kernel in practice. In linear SVM, the plane learns by transforming the problem using linear algebra. SVM proved to be the most used algorithm in current literature. The idea behind SVM is to optimize the decision boundary that separate classes for prediction. It has several parameters that plays an important role in building a classifier that best understands the data. There parameters include kernel, gamma and the regularization parameter known as C. The values to give these parameters cannot be easily determined, hence, GridSearchCV in Scikit Learn used for parameter tuning. GridSearchCV was imported and its parameters were given. GridSearchCV takes the model and range of values the C and gamma can take as well as types of kernels. Then the grid is fitted to training data and it gives best score and best estimators kernel, C and gamma. The best estimators from GridSearchCV are when C=100, gamma is auto deprecated, and kernel is linear. What those values of parameters mean is a large number of C indicates low bias, it tells how much we want to avoid misclassification, allowing fewer outliers.

After that, this output is used to build SVM classifier with those resulted parameters that best understands the data. The accuracy of the model on training set was 84%. Testing it on unseen data is shown in the next section.

8. Result:

After training classifiers and tuning parameters and hyperparameters, those models were tested on unseen data. The testing set (20% of data) were saved into a separate file in the preparation phase. In order to evaluate how good is the model performance when used on test data, various performance metrics available. The choice of metrics depends on the problem to be solved and on business objective. Choosing a metric with relevant variable like false positive and false negative is suitable for our highly imbalanced data.
To evaluate those trained models on test set, accuracy metric and classification report were used. Classification report includes various metrics. Such as recall, precision and f1 score. Recall measures sensitivity, positives that were correctly classified. On the other hand precision is same as accuracy, which measures specificity. In most problems there has to be some trade-off between precision and recall. That is sensitivity and specificity are proportional to each other. In which when sensitivity increase, specificity decreases, and vice versa. Hence, a more general metric could be used known as f1-score which takes into account both precision and recall. It is solely the mean of precision and recall. In this study we aim to maximize accuracy and weighted average f1-score that are used to evaluate the algorithms on unseen data. The weighted average of f-score combines micro and macro averages. The following table (Table 2.) shows each tested algorithm and its associated result.

Table 2. Tested algorithms and result

Model	Accuracy	F1-score (weighted Avg)
KNN (K=7)	83%	77%
Naive Bayes	82.4%	80%
Random Forest	84%	80%
SVM	85.3%	81%

The metrics were somehow closer to each other's, however several implications about the algorithms must be taken into account. To illustrate, KNN classifier is not a good model for our data because it is affected by class imbalance. In which it assigns labels for new points based on a majority vote amongst the k-nearest neighbors of the new point. Which might be very much misleading when class is imbalanced. As it was shown in the experiment when K increased the model performance increased as well. However, data points that

belongs to the class with lower instances might get misclassified. And when algorithm tries to label all of those points belong to the class which has most instances, the model might suffer from overfitting trying to learn most of the data points. Since KNN is sensitive to class imbalance, it is believed that it not useful even if it yields high performance.

Naive Bayes performance improved in the training phase when normalizing data as solution to the model's sensitivity towards the distribution of data among classes. However, the algorithm might not perfectly describe the data as its accuracy were below 83%, even if it is good at generalizing on held-out data. Random forest performance was good. However random forest works like a black box, and to improve the model further, its parameters like the maximum depth of the tree and maximum leaf nodes may be modified to yield a better performance**.**

Lastly, the model that gave us the highest performance on both metrics is SVM with C=100 and linear kernel, the reason behind that is that its parameters were tuned unlike some work in current literature as in (The related work paper shows the performance of SVM equal to 53%. Using recall increases the performance to 80%). With accuracy of 85.3% and 81% average weighted f1-score.

To investigate more in the performance of SVM, an evaluation method known as AUC ROC (Area Under the Receiver Operating Characteristics) were used. Which is a probability curve that indicates measure of separability. It demonstrated how well the model can distinguish between classes. The higher the roc the better the model. The ROC curve is shown below with true positive rate against false positive rate (sensitivity and specificity). The value of AUC= 0.76 which means that there is 76% chance that the model is able to distinguish between positive and negative classes. (See Figure 1.)

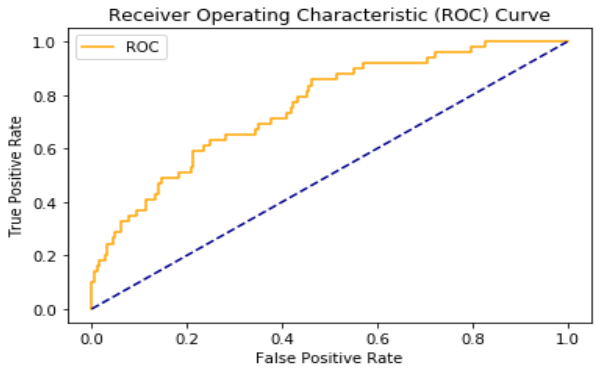

Figure 1. Receiver Operating Characteristic (ROC) Curve

In the research conducted by (Alduayj S. and Rajppot K. 2018), their under-sampling method was not capturing important information which led to lower performance. Using all features their SVM model captured only 70%. In contrast, our proposed methodology resulted in SVM with 85.3% accuracy and f1-score 81%. In addition, auc roc = 76% in SVM developed in our study, which is higher than that of work proposed by (Aasheesh Barvey et.al 2018), their SVM AUC score was only 72%. The work performed by attained a similar accuracy of SVM model with using all features. By using subset of features their algorithm yielded a better accuracy. However, accuracy might be misleading when all of their f1-scores were below 53%. For future work, understanding the hidden causes that may lead an employee to leave is an interesting area of research. Our proposed methodology could be improved by using techniques that deal with imbalanced class. Like over sampling of minority class in our case (Attrition = yes). Different classifiers could be used, and their parameters can be tuned to yield a better performance. Also, it may be interesting to gather more features and to do feature engineering.

9. Conclusion:

To sum up, the main idea in this research is to predict attrition of employees. The focus was on finding the best way to do features selection that gives the highest accuracy to the proposed data set. However, all feature selection techniques were not helping, so all features considered in the experiment. Data cleaning and processing performed, and several classifiers include KNN, Naive Bayes, Random Forest and SVM have been trained on 80% of data and they were validated using cross validation. Models that gave a better performance on training set were used to test on unseen data. To evaluate the models, accuracy was not the only metric used, a more

general metric was used, f1-score that led to better understanding of imbalanced class. For the highest performing model SVM with accuracy 85.3% and f1-score 81%, a probability curve was used to show how well the model distinguish between negative and positive classes. Which gave us a good score to discriminate between classes around 76% even when our data is imbalanced. As a feature direction, a model that operates based on historical data of employees who left can be developed to effectively predict employee attrition and to make proactive approaches in transferring knowledge of employees who will leave. As well as reducing HR costs of training new employees.

Acknowledgements

This work was supported by College of Computer Science and Information Systems at Princess Nourah Bint Abdulrahman University and School of Computing at Dublin City University. Also, we would like to thank IBM for the public dataset used in this research.

References

Barvey, A., Kapila, J., and Pathak, K., Proactive Intervention to Downtrend Employee Attrition using Artificial Intelligence Techniques, 2018.

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

IBM Sample Data: HR Employee Attrition and Performance:. (n.d.). *https://www.ibm.com/communities/analytics/watson-analytics-blog/hr-employee-attrition/*.

Jain, D., Evaluation of Employee Attrition by Effective Feature Selection using Hybrid Model of Ensemble Methods, 2017.

Sukhadiya, J., Kapadia, H., and D'silva, M., Employee Attrition Prediction using Data Mining Techniques. *International Journal of Management, Technology And Engineering*, 2018.

Yigit, I., and Shourabizadeh, H., An Approach for Predicting Employee Churn by Using Data Mining, 2017.

Understanding Support Vector Machine algorithm:. (n.d.). *https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/* .

AUC - ROC Curve:. (n.d.). *https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5*.

Confusion Matrix: . (n.d.). *https://machinelearningmastery.com/classification-accuracy-is-not-enoughmore-performance-measures-you-can-use/*.

Cross Validation:. (n.d.). *https://chrisalbon.com/machine_learning/model_evaluation/cross_validation_parameter_tuning_gri d_search/*.

IBM Sample Data: HR Employee Attrition and Performance:. (n.d.). *https://www.ibm.com/communities/analytics/watson-analytics-blog/hr-employee-attrition/*.

Model evaluation:. (n.d.). *https://scikit-learn.org/stable/modules/model_evaluation.html* .

Parameter estimation using grid search with cross-validation:. (n.d.). *https://scikitlearn.org/stable/auto_examples/model_selection/plot_grid_search_digits.html*.

Biographies

Imtinan Alsuahim, Ms. Imtnan is Data Analytics master's student in DCU@PNU. She holds a Bachelor of Science in Information Systems from Princess Nourah Bint Abdulrahman University. She works on many researches focused on Data Analysis and Data Visualization**.**

Feddah Alotaibi, Ms. Feddah studies Masters of Computing in Data Analytics in DCU@PNU. She holds a Bachelor of Science in Information Systems from Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia. Ms. Feddah's Undergraduate thesis focused on Mobile App Development and Human Computer Interaction. Her previous research focused on Sentiment Analysis and NLP. Ms. She is interested in Machine Learning and Deep Learning.

Malak AlAsiri, Ms. Malak currently studying Masters of Computing in Data Analytics in DCU@PNU. She graduated from Princess Nourah Bint Abdulrahman University with a bachelor's degree Science in Information Systems. Her undergraduate work was related to building web-based systems that provide remote communications. She is interested in Data Analytics and visualization.

Munirah Alkharji, Ms.Munirah studies Masters of Computing in Data Analytics at Dublin City University (DCU) and Princess Nourah Bint AbdulRahman University (PNU), she holds a Bachelor's Degree in Computing Information System from Prince Sattam Bin Abdulaziz University (PSAU), she earned the first place in the field of Scientific Research from The Second Scientific Forum for Excellence, She has written and published a book specialized in the Equivalence Systems Using Information Retrieval Methods, earned two certified certificates from Research & Translation Center (RTC), and earned DELL EMC certified Cloud Computing Associate. Ms.Munirah is interested in Big Data Analytics, Machine Learning, and Artificial Intelligence.

Shahad Alharthi, Ms. Shahad a master's student that is specialized in Data Analytics in DCU@PNU. Prior to her bachelor's degree in computer information systems from Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia. Ms. Shahad hopes to contribute to Data science and Data Analytics in addition to Big Data field. Ms. Shahad is also interested in Data Manipulation & Data Visualization.