Ensemble method based architecture using random forest importance to predict employee's turnover

Md. Anwar Hossen\textsuperscript{1*}, Emran Hossain\textsuperscript{1}, Zahereel Ishwar Abdul Khalib\textsuperscript{2}, and Fatema Siddika\textsuperscript{3}

\textsuperscript{1} Department Software Engineering, Daffodil International University, Bangladesh
\textsuperscript{2} School of Computer and Communication Engineering, Universiti Malaysia Perlis, Malaysia
\textsuperscript{3} Department of Computer Science and Engineering, Jagannath University, Bangladesh

Email: *anwar.swe@diu.edu.bd

Abstract. The departure of a skilled employee can create a problem for a company and this incident is increasing globally. Employee turnover has become an important issue these days due to the heavy workload, low pay, low job satisfaction, poor working environment. Companies face problems as their budget will increase, losing skilled manpower and employees’ trust. It’s taking time to adjust for a new employee and bring risk and increase the cost for the company. It is necessary to bring appropriate solutions to the problem. The main purpose of this paper is to predict the turnover of employees with the help of state of the art machine learning classifier. We have determined employee turnover selection factors using some prediction models. We first pre-processed the dataset by removing correlative attributes. Then, we have scaled the attributes. Secondly, a Sequential selection algorithm (SBS) has been used to reduce features from a high number to a relatively small signal-canton. Then use Chi-square and Random Forest important algorithms to determine the most significant shared key features. Then we get \textit{average\_monthly\_hours}, \textit{satisfaction\_level}, \textit{time\_spend\_company} are responsible for the employee’s departure. Then, we have applied different state of the art machine learning algorithm to measure the accuracy. We have achieved the highest accuracy of 99.4\% using the reduced feature with 10-Fold Cross-validation by applied the Random Forest classifier and which is higher than the mentioned reference work.

Keywords: Employee Turnover, Feature Selection, SBS, Chi2-Square, Random Forest

1. Introduction

Recently, Employee turnover rate is the measurement of the number of employees within typically consider the one-year time frame who leave the institution [1]. An organization considers the sum of employees who left, turn over can also apply to subdivisions within an industry like independent departments. Moreover, employees turnover is very mundane occurrence for the industries and it creates a problem for them. Lack of satisfactory salary, large numbers project assigned to the employees, lack of satisfactory working environment is the prime reasons for employee turnover [2]. The skilled employees are the wealth for the company when they leave the company is facing major problems to run smoothly [1]. A higher rate of employee attrition is bad for company’s reputations and hamper their reputations [2]. Therefore, identifying the key reasons behind the employee's attrition is a prime concern for any company to take the necessary step to retain the position and accomplishment [3]. Numerous numbers of authors tried to do research and find out the proper reason for employee turnover. Researchers tried
to establish a predictive model to predict whether an employee will leave the company or continue where they worked. H. Liu and R. Setiono [4] has shown compassion between two well-known algorithms Decision Tree (DT) and Artificial neural network (ANN) to detect the particular employee's turnover factor by data mining approach. Hall and Holmes [5] have analyzed 15 different software datasets from the UCI dataset repository and observe that six major attributes are the prime factor for employee turnover. In this work, we have proposed an architecture based on the state of the art machine learning approach that contributes acumen into employee's turnover by discovering out the key factors which can be applied to any company. Our main objective was to bring out the most appropriate attributes which were the prime reason for turnover and compared our model accuracy with an existing model. Firstly, we have used two Chi-Square and Random forest importance feature selection algorithm to bring out the most principal attributes. After that, we have applied the state of art machine learning algorithms. Also, we applied 10-fold cross-validation to overcome over-fitting problems. We have executed five different models to get better accuracy named as with all features, with reduced features, with 10-fold cross-validation, and with bagging, booting. Therefore, we have found that the model consists of random forest importance as feature selection, Decision Tree (DT) with 10-fold cross-validation as the learning algorithm has achieved an accuracy of 99.4% and outperform than reference work [6].

The remainder of the paper is organized as follows. Section II contains the literature review and state of the artwork on employee turnover. Section III describes the system model including data preprocessing, feature selection, and algorithm selection. Section IV discusses the experimental implementation details, the results, analysis, performance evaluation, and comparison and finally, we concluded the paper in Section V.

2. Literature Review

There are major five reasons to which employees leave the organization according to different researchers. Dissatisfaction at work is one of the main causes of absenteeism. Robbins and Decenzo [7] state that a person with high job satisfaction has positive attitudes towards work. According to [8] the work environment is also a major cause of employee turnover. The important reasons for turnover are the exit rate and employment stability [9]. Another reason may be that employees are not satisfied with the organization's culture. Because of this, the employees leave the company. Chen and Huang [10] have investigated the effect of improving job satisfaction among employees. Human's nature is to seek growth. If you feel that there is no room to grow in a particular organization, they will look for a place where they can grow more.

Howley and other [11] authors have examined the use of Principal Component Analysis (PCA) and Random Forest algorithm to reduce high dimensional phantom data and to improves the imminent performance of some state of the art machine learning methods. M M Alam [6] has also accomplished the most beneficial attributes by applying feature selection techniques Chi-square and Random Forest Importance. Nevertheless, Random Forest Importance has been performed better than chi-square to determine the weight of each attribute. Fan and Chan [12] had found out that company management issues, lack of good leadership, internal faithfulness identification, were the prime cause of employee turnover from a company. They have used Hybrid ANN and clustering analysis that found out the above mentioned reason. In our procedure, we have estimated out the actual problem of employee turnover are average monthly hours, satisfaction level, and last evaluation.
3. System Model

Our system model consisted of numerous different segments. Firstly, we have collected a Human Resources Analytics dataset from the Kaggle [13]. From the raw data, we have done a few pre-processing steps. We have removed the null values, scaling the value of the feature, find out correlative features. With the pre-processed dataset, we tried to find out the key features using Sequential Backward Selection (SBS), Chi-square and Random Forest Importance as all the features have not the same effect on the label attributes. We found out which group of features has a separate value than other attributes from the correlation matrix based on correlation value. Then we were putting each group of attributes into an atomic feature set after multiplying them with their respective Random Forest Feature Importance value. After our dataset was ready to apply different classification algorithms. We have applied different state of the art classifier including SVM, DT, GNB, RF, MLP, KNN. We have also applied 10-fold cross-validation to overcome the over-fitting problem. The meta-algorithms that combine different learning techniques in a predictive model to reduce variance and distortion or predictions. We have used bagging and boosting for reducing variance and boosting respectively. We have tried five different approaches to find out the appropriate ways to predict the employee’s turnover.

The proposed system model was given in Fig 1.

![System Model Diagram]

**Figure 1. System Model**
3.1. Dataset Description
We have collected a Human Resources Analytics data set from the Kaggle [13]. The primary purpose to collect the data and make a repository to find out the prime factors why the potential employees are leaving institution. The data set has nine attributes named satisfaction_level, last_evaluation, number_project, average_monthly_hours, time_spend_company, work_accident, promotion_last_5years, type, salary and one label attribute named left. The left column is dependent on the other nine attributes. The data set has a total of 15000 samples where only 3572 employees left the company.

3.2. Data Pre-processing
Our data set didn't have any null values among 15000 instances. This data set has two categorical attributes named 'type' and 'salary'. We didn’t consider the 'type' attribute as the job category does not affect more on the label column. The salary attribute has three different values named high, medium, and low. We have to encode the salary attributes as 3, 2, and 1 for the high, medium, and low value respectively. After that, we scaled some of the particular features because of their erratic variety. The value of the attributes 'Satisfaction Level' and 'Last Evaluation' was between 0 to 1. For feature scaling, we have few known scaler methods in python Scikit-learn Libraries such as MinMax Scaler, Robust Scaler, or Standard Scaler [14]. But we didn't apply these methods as Standard Scaler and Robust Scaler generate negative values that are not permissible for the Chi-square feature selection algorithm. That's why we decided to multiply those values with 100.

3.3. Feature Reduction using Sequential Backward Selection
Sequential Backward Selection (SBS) is a feature selection algorithm where a subset of features is selected based on the requirements. We have used a large data set with many features. Therefore, an appropriate learning algorithm needs to be select to avoid over-fitting problems which will also remove model complexity and reduce the dimensionality [15]. If a data set has N number of features, then SBS methods select M number of features among them where N is greater than M. We have applied a KNN learning classifier and achieve 96.75% accuracy with seven features which were mentioned in Table 4.

3.4. Feature Reduction using Sequential Backward Selection
The important weight for each attribute was found and those values were multiplied with some threshold valued to combine all the features. For each categorical feature, we have created a few dummy columns in our data set, when we combine those dummy columns value with the actual value by finding the mean of those dummy columns [16]. After applying the Random Forest Importance to the data set, we have achieved the weights for all the attributes which were mentioned in Table 1. The data are shown as decreasing order. From Table 1, we have seen that the satisfaction_level, number_project, average_monthly_hours, average_monthly_hours and time_spend_company were the most significant features among the nine features.

3.5. Chi-Square Feature Selection
The highest valued features can be calculated from the chi-square statistic from the training and test set [17]. Our training data set only contains only the positive value in the form of pattern or Boolean form. As chi-square only can be applied to those situations, therefore we applied the chi-square in our data set. We have applied the chi-square feature selection algorithm and found the weights of the individual features. Top five features with weights mentioned in II. We have seen from the Table 2 that
average_monthly_hours, satisfaction_level, time_spend_company, promotion_last_5 years and number_project were the most significant features among nine features.

Table 1. Random Forest Importance Scores

| Feature Name           | Importance |
|------------------------|------------|
| satisfaction_level     | 0.345032   |
| last_evaluation        | 0.182678   |
| number_project         | 0.180379   |
| average_monthly_hours  | 0.143579   |
| time_spend_company     | 0.121260   |
| work_accident          | 0.012071   |
| left                   | 0.008336   |
| promotion_last_5years  | 0.005729   |
| department             | 0.000937   |

Table 2. Chi-square scores

| Feature Name                   | Importance |
|-------------------------------|------------|
| promotion_last_5years         | 145.008319 |
| Department                    | 40.070240  |
| time_spend_company            | 22.665090  |
| average_monthly_hours         | 11.374087  |
| satisfaction_level            | 3.816073   |
| Work_accident                 | 1.292191   |
| last_evaluation               | 0.261557   |
| number_project                | 0.623815   |
| salary                        | 0.264293   |

3.6. System model with state of the art learning algorithms

We have applied four different approaches to build a system model. After data pre-processing and selecting the most significant six features named average_monthly_hours, satisfaction_level, time_spend_company, promotion_last_5 years and number_project, we have split our data set into a train and test set.

We have applied state of the art classifiers including Support Vector Machine (SVM), Decision Tree (DT), Gaussian Naive Bayes (GNB), Random Forest (RF), Multi-layer Perception (MLP) and K-Nearest Neighbor (KNN). The five different approaches have mentioned below:

- Applied classifier algorithms to measure performance with all features
- Applied classifier algorithms to measure performance with SBS feature selection
- Applied classifier algorithms to measure performance with selected most significant five features which were mentioned above
- Applied classifier algorithms to measure performance with 10-fold Cross-validation to circumvent over fitting problem and get better-trained model
- Applied ensemble method named bagging and boosting to combine multiple learning methods.
3.7. K-fold cross-validation
The original sample is randomly divided into subsets of equal size. Thus between the subfamilies K, a single sub-undercut model is retained as validation data for the test and the remaining K-1 samples were used as training data. The process is repeated circularly several time, with the subset K used precisely one time as validation data. The advantage cross validation of K folds through circular repeated random sampling is that all the sample are used for both training and validation, and each sample was used once for validation. In the cross-validated stratified K folds, the partitions are selected so that the average response value between the partitions is approximately equal.

3.8. Ensemble method
The methods set are meta-algorithms that combine different machine learning techniques in a predictive model to reduce variance (bagging), distortion (reinforcement), or improve predictions (stacking). The basic motivation of parallel methods is to exploit independence among the basic students since the error can be drastically reduced by the average.

3.8.1. Bagging
We used a bagging classifier, a combined meta-estimate that adapted each base classifier to a random satellite from the original database and then combined them into separate forecasts (by grade or average) to form the final forecast. This national meta-estimator is generally used as a way to reduce the variance of a black box estimate by introducing randomization into a compilation set method and then creating the representation. Fit many large trees to bootstrap-resampled versions of the training data, and classify by majority vote [18].

3.8.2. Boosting
The increase is a sequential process, in which each successive model attempts to correct the errors in the previous model. Later models depend on the previous model. Next, we will discuss how Boosting works. Fit many large or small trees to re-weighted versions of the training data. Classify by weighted majority vote [19].

A subset is created from the original data.
- Initially, all data points received the same weight
- A basic model was created in this subset
- This model was used to predict the entire data set

4. Experimental Results
Each machine learning algorithm has a set of variables that learn from provided training data. A classifier has a diverse set of learning parameters which value is important for better learning. Every learning algorithm has not used all parameters. However, there are a set of parameters whose values are set before the learning process begins, this variable is called hyper-parameter. The Python Hyper-parameter is passed as an argument to the Model Constructor class in the Scikit-Learn Machine Learning Library.

Here are the hyper parameters that were followed in these research models:
- RF Classifier: random_state = 42, n_estimators = 100, max_features = 'auto', class_weight = none, n_jobs = none, verbose = 0, warm_start = False.
- DT Classifier: random_state = 0, criterion = "gini".
- KNN Classifier: n_neighbors = 1.
Now, we discussed the experimental result and analysis of our five different approaches in the following sections.

4.1. Applied classifier with all features

After data pre-processing, I have trained the model with all the features and applied six different classifiers with all necessary hyper-parameter.

From the Table 3 we observed that RF, DT, SVM, LR, GNB and KNN have achieved accuracy 98.49%, 97.44%, 94.45%, 75.78%, 79.91% and 92.80% respectively and seen that DT has the height accuracy with 98.49%. The Fig 2 showed the ROC curve for the classifier with all features and we observed that RF, DT, LR and KNN classifier curve converge very quickly whereas SVM and GNB curve converge slowly.

|                | RF  | DT  | SVM | LR  | GNB | KNN |
|----------------|-----|-----|-----|-----|-----|-----|
| Training accuracy | 100% | 100% | 96.11% | 93.61% | 77.97% | 100% |
| Test accuracy    | 98.49% | 97.44% | 94.45% | 75.78% | 79.91% | 92.80% |

![Receiver Operating Comparison](image)

**Figure 2.** ROC curve with all feature

4.2. Applied classifier with SBS feature selection

SBS feature selection method applied and all the observation listed in the Fig. 3 and Table 4 From this image and table, we have seen that the least and highest number of the feature were not good for the classifier whereas a middle number of feature bring good classification accuracy. With 1 or 2 features bring less than 95 percent accuracy and with all the feature accuracy fallen down to less than 90 percent. The ideal number of features were 4 to 7. These last_evaluation, number_project, average_monthly_hours, Work_accident, promotion_last_5years, salary features bring highest accuracy 96.75%.

4.3. Applied classifier with reduced features by chi-square and Random forest importance

We picked one more stage to reduce the feature and also applied the importance of Chi2 and Random Forest to perform. By doing this, we can access their prominent features and their respective scores. The results of Chi2 and Random Forest are given in Table 1 and Table 2 for calculating the feature importance. From Tables 1 and Table 2 we can clearly see the importance of random forest features and
score of Chi2. According to Chi2, average monthly hours, satisfaction level, Time Spend company, promotion last 5 years, and number of project were the five highest impacting factors. Using Random Forest Importance feature selection, five most important factors are select satisfaction level, last evaluation, number of project, average monthly hours and Time Spend Company. We experimented with Chi 2 and Random Forest.

So, we have combined the most important features from these two methods and final features are average monthly hours, satisfaction level, Time Spend company, promotion last 5 years, and number of project. Subsequently, it was run with different classifications built into the selected algorithms. We visualize the ROC curve after the second phase characteristic reduction, which is provided in Figure 4 after applying the second step feature reduction, it is randomized that the random forest classifier offers the best accuracy in both cases only and it is 99.00%, SVM 97.00% and decision tree 97.00%. Further, the KNN model did not change much 97.00%. Comparing with Table 5 we can detect the unrecognized accuracy of the MLP model 81.00% after applying the feature reduction which previously had an accuracy of 97.00%. The accuracy of the Gaussian NB has not changed much, though it has been reduced to 81.00%, which was 83.00% earlier.

Figure 3. SBS Feature Reduction

| Selected Features                                                                 | Accuracy |
|----------------------------------------------------------------------------------|----------|
| last_evaluation, number_project, average_monthly_hours, time_spend_company,     | 88.55%   |
| Work_accident, promotion_last_5years, Department, salary                         |          |
| last_evaluation, number_project, average_monthly_hours, time_spend_company, Work | 95.00%   |
| accident, promotion_last_5years, salary                                          |          |
| last_evaluation, number_project, average_monthly_hours, Work_accident, promotion | 96.75%   |
| last_5years, salary                                                              |          |
| last_evaluation, number_project, average_monthly_hours, Work_accident, promotion | 96.68%   |
| last_5years, salary                                                              |          |
| last_evaluation, number_project, average_monthly_hours, Work_accident            | 96.57%   |
| last_evaluation, number_project, Work_accident                                   | 96.17%   |
| last_evaluation, Work_accident                                                   | 94.57%   |
| last_evaluation                                                                   | 93.64%   |
In the ROC curve, we try to evaluate the best-classified model according to the highest true-positive value and the low-false positive value, and we also compared all the classified models that are controlled by the ROC curve. Table 5 shows the results of the accuracy of the different classifiers after the second-level feature reduction process.

![ROC curve after feature selection](image)

**Figure 4. ROC curve after feature selection**

| Path | RF   | DT     | SVM | MLP | GNB | KNN   |
|------|------|--------|-----|-----|-----|-------|
| Training accuracy (Reduced variables) | 99.75 | 99.98  | 96.35 | 93.37 | 80.32 | 99.03 |
| Test accuracy(Reduced variables)    | 98.64 | 97.93  | 95.27 | 93.67 | 79.33 | 95.93 |

### 4.4. Applied ensemble method named bagging and boosting

In this study I used the bagging and boosting of the ensemble algorithm. After applying the Bagging, it is randomized that the random forest classifier and decision tree offers the best accuracy in cases only and it is 98.0%. Table 6 shows the results of the accuracy of the different classifiers after the Bagging process.

After applying the Boosting, it is randomized that the random forest classifier and decision tree offers the best accuracy in cases only and it is 99.4%. Table 6 shows the results of the accuracy of the different classifiers after the Boosting process.

| Path            | Random forest | Decision tree | KNN   |
|-----------------|---------------|---------------|-------|
| Training accuracy (Bagging) | 99.20 %       | 99.30 %       | 95.30 %       |
| Test accuracy (Bagging)    | 98.00 %       | 98.00 %       | 94.10 %       |
| Training accuracy (Boosting) | 100.00 %      | 100.00 %      | 100.00 %      |
| Test accuracy (Boosting)   | 98.50 %       | 98.30 %       | 98.30 %      |
### Table 7. Comparison with reference work [11]

| Model                                      | RF   | DT   | SVM  | MLP  | GNB  | KNN  |
|--------------------------------------------|------|------|------|------|------|------|
| Proposed Model - Reduced features with     | 99.40% | 98.32% | 95.9% | 80.21% | 97.26% |
| 10-Fold Cross validation                   |      |      |      |      |      |      |
| Reference Model - Reduced                  | 98.03% | 97.00% | 95.8% | 92.8% | 77.4% | 96.00% |

### 4.5. Applied classifier with 10-fold Cross-validation

From the ROC curve, we can clearly see that the random forest still offers better model functionality as the region under the value holds more value. After applying the cross validation, it is randomized that the random forest classifier offers the best accuracy in both cases only and it is 99.40%. Table 7 shows the results of the accuracy of the different classifiers after the Cross Validation process.

### 4.6. Comparison with reference work

In Table 7, we have shown that our model perform better than reference model [11]. We have achieved the highest accuracy 99.4% using the reduced feature with 10-Fold Cross-validation by applied the Random Forest classifier and achieved 98.64% accuracy with reduced feature by applied the Random Forest classifier whereas the reference work best model achieved accuracy 99.3% with reduced feature based model. We have also achieved higher accuracy using two different approach named the reduced feature with 10-Fold Cross-validation based model and reduced feature based model using DF, DT, MLP and GNB classifier algorithms which was mentioned in the Table 7.

### 5. Conclusion

Employee turnover rate is the measurement of the number of employees within typically consider the one-year time frame who leave the institution. An organization considers the sum of employees who left, turnover can also apply to subdivisions within an industry like independent departments. But most of the time, it is very difficult to control employee turnover rates within the organization. The purpose of this study was to determine the significant causes of employee turnover and these are satisfaction level, final evaluation, how many projects the worker has done, average monthly hours, time spent. These factors must be taken into account when an employee is in the company. Newly hired employees must be well-behaved so that they can adapt to the company environment. We have executed five different models to get better accuracy named as with all features, with reduced features with SBS, Chi-square and Random Forest Importance, with 10-fold cross-validation and with bagging, booting.

We have achieved the highest accuracy 99.4% using the reduced feature with 10-Fold Cross-validation by applied the Random Forest classifier and which is higher than the mentioned reference work best model accuracy 99.3%.

### References

[1] Sikaroudi, E., Mohammad, A., Ghousi, R., & Sikaroudi, A. (2015). A data mining approach to employee turnover prediction (case study: Arak automotive parts manufacturing). Journal of Industrial and Systems Engineering, 8(4), 106-121.

[2] Gao, Y. (2017). Using decision tree to analyze the turnover of employees.
[3] Ajit, P. (2016). Prediction of employee turnover in organizations using machine learning algorithms. Algorithms, 4(5), C5.

[4] Liu, H., & Setiono, R. (1995, November). Chi2: Feature selection and discretization of numeric attributes. In Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence (pp. 388-391). IEEE.

[5] Hall, M. A., & Holmes, G. (2003). Benchmarking attribute selection techniques for discrete class data mining. IEEE Transactions on Knowledge and Data engineering, 15(6), 1437-1447.

[6] Alam, M. M., Mohiuddin, K., Islam, M. K., Hassan, M., Hoque, M. A. U., & Allayear, S. M. (2018, July). A machine learning approach to analyze and reduce features to a significant number for employee’s turn over prediction model. In Science and Information Conference (pp. 142-159).

[7] Robbins, S. D. D.(2001). Fundamentals of Management. Translated by: Aarabi S et al.(1th ed). Tehran: Cultural Research Bureau.

[8] Moureen, M. (2004). Human Resource Planning.

[9] Krueger, J., Bernini, M., & Wilkinson, S. (2014). Introspection, isolation, and construction: mentality as activity. Commentary on Hurlburt, Heavey & Kelsey (2013).” Toward a phenomenology of inner speaking”.

[10] Huang, J., Gu, M., Lai, Z., Fan, B., Shi, K., Zhou, Y. H., & Chen, Z. (2010). Functional analysis of the Arabidopsis PAL gene family in plant growth, development, and response to environmental stress. Plant Physiology, 153(4), 1526-1538.

[11] Howley, T., Madden, M. G., O’Connell, M. L., & Ryder, A. G. (2005, December). The effect of principal component analysis on machine learning accuracy with high dimensional spectral data. In International Conference on Innovative Techniques and Applications of Artificial Intelligence (pp. 209-222). Springer, London.

[12] Fan, C. Y., Fan, P. S., Chan, T. Y., & Chang, S. H. (2012). Using hybrid data mining and machine learning clustering analysis to predict the turnover rate for technology professionals. Expert Systems with Applications, 39(10), 8844-8851.

[13] Kuldeep, L.: Human Resources Analytics (2016, Fall). https://www.kaggle.com/ludobenistant/hr-analytics/data. (Last accessed at 01-11-2019).

[14] Sklearn.preprocessing.StandardScaler. (n.d.). http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html. (Last accessed at 01-11-2019).

[15] Raschka, S.M.: Python Machine Learning -. S.l.: Packt Publishing Limited (2017)

[16] Pandas.get_dummies(n.d.). pandas.pydata.org/pandas-docs/stable/generated/pandas.get_dummies.html (last accessed at 01-11-2019).

[17] Liu, H., & Setiono, R. (1995, November). Chi2: Feature selection and discretization of numeric attributes. In Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence (pp. 388-391). IEEE.

[18] Breiman, L. (1996). Bagging predictors. Machine learning, 24(2), 123-140.

[19] Freund, Y., & Shapire, R. (1996). 13th International Conference on Machine Learning. Experiments with a new boosting algorithm, 148-156.