Ankit Kumar Pradhan

ML FOR PREDICTIVE ANALYSIS

9 October 2021

**Decision Tree Based Employee Churn Prediction**

**Abstract**

Employee churn is a major issue in companies as it affects the company cost and to replace the left employee with a new employee. So, predictive models are useful in this type of scenarios. Employee churn and employee attrition are closely related but they are not the same. Employee churn affects its overall revenue and brand image of the organization. Employee churn leads to customer dissatisfaction. New hiring will consume money as well as time and freshly hired employees take time to make the respective organization profitable. Correlation matrix plot is also done on our dataset to see which features are more related to our target variable. Some more visualization works have been done like comparing how many employees left the company, left vs. satisfaction level, left vs. promotion since last 5 years. Well- known classification algorithms including, Decision Tree, SVM, KNN have been applied. Different performance measures like precision, recall, F-measure and overall accuracy are taken into consideration. From the simulation results it can easily be concluded that decision tree performs better as compared to other models with the accuracy of 98.0677%.

**Keywords**

Employee churn, employee attrition, correlation matrix, classification, Decision Tree, SVM, KNN

1. **Introduction**

In the IT industry the employees churn rate is approximately 12-15%. This churn rate is quite high and assuming even a lower churn rate of 5%, the cost involved in an employee leaving a firm is approximately 1.5 times the annual salary of an employee. Assuming that organizations strength (in terms of number of employees) is 140,000 (organization under study) and employee’s average salary is $12,000.00 annually, then organization has to lose $126000000 (7000 \* 1.5 \* 12000). This amount is certainly not good news for organizations with high employee churn rate (attrition) [1].

An employee would decide to join or leave an organization based on several reasons, for instance, work environment, work place, gender equity, pay equity etc. Others may consider personal reasons such as relocation due to family, maternity, health, conflict with the managers or colleagues in a team. Employee churn is a big issue for the organizations specially when trained, technical and key employees leave for a better opportunity in a competitor organization. It requires time, effort and results in financial loss to replace a trained employee. Therefore, we use the current and past employee data to analyze the common grounds for employee attrition. The employee churn prediction helps in identifying and solving the issues that results in attrition. We can use this information for possible retention of the current employees. In this study, we implement some of the well-known techniques of data classification namely, Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbour (KNN), and Artificial Neural Network (ANN) on the data set. The dataset includes 1499 records with 9 features including categorical and numeric features. Before implementing method, we calculated the correlation of the features in order to avoid features with high correlation. The results of this methods have been analyzed then by their accuracy, precision, recall, and F-measure values. Then, the method with best performance has been conducted. Finally, we implement a feature selection method to select the most important features of the dataset [2].

1. **Problem Definition**

Churn analysis and prediction is widely studied for customer churn, since it is a notorious issue and results in revenue loss. Employee churn is a similar problem for organization, however to predict employee churn is rather more complex than customer churn prediction. Employee churn leads to issues such as efforts and time to get the replacement and retraining, financial loss, customers dissatisfaction and many more. Therefore, for smooth running of an organization, the key is to retain its trained workforce. Employee churn can be categorized in two types; (1) voluntary, those who leave for their own reasons, and (2) involuntary, those who are released from their services by the organization. Usually companies focus on voluntary churn, where an employee would either leave for a better opportunity in terms of pay, benefits, work environment etc, or due to negative reasons at the present organization such as conflict with the supervisors, lack of opportunities for promotion, lack of interesting work and many more. In this study, we also focus on predict voluntary churn employees [2].

1. **Literature Survey**

Middle level officers are more likely to leave, may be due to some disagreement with their senior officer as proposed by [3]. They observed major factors that influenced employee abandonment from the firm. The two rules are moderately derived by him. Some sets of questions are asked with the both parties and depending upon their answers he concluded some facts based on workload, objectives, carrier opportunity and firm management. Human resource management [4] endeavours on basically termination rates and dismissal rates but actual content of them is enormously different. The previous model shows that, there are several distinct levels of attrition and turnover. Some research dictates that the consequences of dismissal and termination rates are at organizational level. Allen & Meyer (1990) [5] described the three-basic entity for the negative side of the turnover. Regulating officer will more probably leave from the organization because of a contention with the higher administration than a representative who is in struggle with his prompt director. He recognized the determinant figures that influence employee acceptance [5] without protest from the organization. Two arrangements of information social occasion techniques were directed. An equivalent number of representative and officer respondents were solicited to answer a set from polls that were ordered by workload, objectives, identity, professional success, and hierarchical administration. The after-effects of the two information gathering methods demonstrated that the most noteworthy component that adds to employee rejection [6] is money related compensation

1. **Theoretical Background**

In this chapter, we present the theoretical background of the methodology that used in this study.

1. *Decision Tree*

Decision trees become famous because of the ease of interpretation of the discovered rules. This

learning algorithm constructs a tree with a training data set in which each node is an attribute and branches of the nodes are corresponding attribute values. Nodes in the decision tree are arrived on the basis of the explanatory power of the attributes [7]. One problem with decision tree learning algorithm is that it is the small changes in the training data lead to large variations in the classification performance so it would be an unstable algorithm. To solve this problem, Breiman proposed random forests [8]. His approach is to construct multiple decision trees on sampled data (obtained through bootstrap re-sampling technique) using only a subset of attributes. The final class label for a new data point is obtained by combining (using a voting-based scheme) the decisions of all the trees thus built.

1. *Support Vector Machine (SVM)*

Support vector machine (SVM) is a supervised learning algorithm used for binary classification.

Considering that the input training data points (each with a class label 1 or +1) are ‘‘scattered” in an n-dimensional space, SVM identifies a suitable linear function (hyper plane) which optimally separate the training data points belonging to the two classes. Thus, all points with class label 1 are on one side of the discovered hyper plane and all points with class label +1 are on the other side of the discovered hyper plane. SVMs use what is widely known as the kernel trick to obtain a nonlinear decision function for data points that are not linearly separable. In the kernel trick, data points in the input space are mapped using a non-linear mapping φ(.) into a high dimensional space (called the feature space). The inner product of two points from the feature space is replaced with a kernel function:

K(X*i*, X*j*) = { φ (X*i*), φ(X*j*) }

The class of algorithms that employ this kernel trick for learning is known as kernel methods. The optimal separating hyper plane is found by maximizing the margin i.e., the distance between the closest data points lying on either side of the hyper plane. This maximization is formulated as a quadratic programming problem with the constraints that all points belonging to one class (say + 1) should lie on one side of the hyper plane and points belonging to different class (say \_1) should lie on other side of the hyper plane [9].

1. *K-Nearest Neighbour*

K-NN classifier [10] is known as lazy learner in machine learning community. It never learns

from the data and do not build any models. Rather, it finds out the examples from the train dataset which are closest to the unknown example. Based on the neighbours’ examples it will predict the new example. The value of ‘k’ determines the no. of closest data points or examples to be selected from the training example.

1. *Artificial Neural Network*

Artificial Neural Network (ANN) is an information processing mechanism which is which is based on a large collection of simple neural units (artificial neurons). ANN comprises some interconnected elements, or neurons, working together as one system to overcome problems. The structure of an ANN is determined by both the inter-neuron connections’ arrangement and the nature of these connections. The way of training or adjusting the strengths of these connections so as to achieve a desirable overall behaviour is known as learning process [11]. ANNs are well known for their well-established empirical modelling power. Their most outstanding feature is their ability to learn automatically from available data in order to provide a means for predictions. They also are able to impose blind (hidden) insights into the hidden relationships [12]. ANNs can be categorized into two different groups considering their structures: Feed-Forwards and Recurrent. A Feed-Forward network’s neurons are grouped into input, hidden and output layers. Feed-Forwards networks can be recognized by unidirectional arrows which flow from input layer through the output, not connecting neurons in a same layer but connecting them from previous layer just to the next one [13].

1. *Evaluation Measurement*

Classification model produces four scores which generate by confusion matrix. We evaluate the classification models using confusion matrix as shown in table I. The thorough explanations are:

1) TP (True Positives): the number of not transferring correctly labelled as not transferring.

2) TN (True Negatives): the number of records that correctly labelled as transfer.

3) FP (False Positives): the number of who are transfer but the algorithm incorrectly categorized them as not transferring.

4) FN (False Negatives): the number of employees who are not transferring but the algorithm incorrectly categorized them as transfer.

Table I. Confusion Matrix

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Predicted | |
| Not Transferring | Transfer |
| Actual | Not Transferring | True Positive (TP) | True Negative (FN) |
| Transfer | False Positive (FP) | True Negative (TN) |

Beneficial to find the best classification model, it is crucial to evaluate the performance of each model. There are prominent parameters for evaluating the classification models which are classification accuracy, precision, recall, and F1 or F-measure [14]. Those parameters successfully generate the performance of each model. The parameters which are being used in this study for the evaluation measurement of each model is shown in Table II.

Table II. Parameter Measurement

|  |  |
| --- | --- |
| **Parameter** | **Description** |
| CA | Accuracy classification score |
| F1 | weighted average of the precision and recall |
| Precision | Ratio of measurement on how appropriate the model in predicting the class |
| Recall | the proportion of positives that are correctly identified |

The parameters are calculated from the number that produce by the confusion matrix. The number including TP, TN, FP, and FN. The mathematical equation and explanation of the parameters are explained below [14]:

1. Accuracy is the number of all the correct predictions and calculated using the Equation:

Accuracy =

1. F1 or F-measure is the harmonic average of Precision and Recall and calculated using the Equation:

F-measure = 2 ×

1. Precision is the ratio of predicted churners which are correct and calculated using the Equation:

Precision =

1. Recall is the ratio of real churners which are correctly identified and calculated using the Equation:

Recall =

1. **Methodology**
2. *Data Set Analysis*

There are two datasets, obtained from Kaggle website, is used in this paper for the experimental verification. One dataset name is hr\_data in which there are 14,999 rows and 9 features and other dataset name is employee\_satisfaction\_evaluation in which there are 14,999 rows and 3 features.

Table III shows the features of data and their type and definition in the hr\_data dataset

Table III. HR dataset features

|  |  |  |
| --- | --- | --- |
| **No.** | **Feature** | **Data Type** |
| 1 | employee\_id | Numerical |
| 2 | number\_project | Numerical |
| 3 | average\_monthly\_hours | Numerical |
| 4 | time\_spend\_company | Numerical |
| 5 | Work\_accident | Numerical |
| 6 | Left | Numerical |
| 7 | promotion\_last\_5\_years | Numerical |
| 8 | department | Categorical |
| 9 | salary | Categorical |

Table IV shows the features of data and their type and definition in the employee\_satisfaction\_evaluation dataset.

|  |  |  |
| --- | --- | --- |
| **No.** | **Feature** | **Data Type** |
| 1 | EMPLOYEE # | Numerical |
| 2 | satisfaction\_level | Numerical |
| 3 | last\_evaluation | Numerical |

The two datasets were merged using the employee\_id feature in hr\_data and EMPLOYEE# in employee\_satisfaction\_evaluation and the main data frame have been created using these two primary keys.

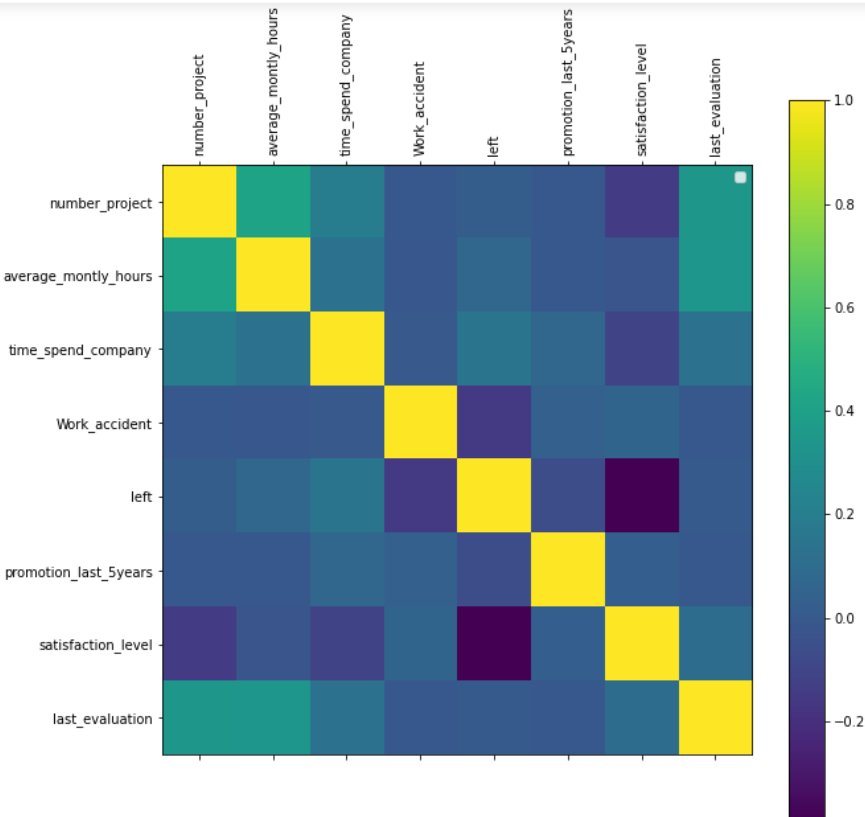


Fig I. Correlation matrix

Figure I represent the correlation matrix which helps to identify attributes with the strong or weak correlation

1. *Data pre-processing*

Data pre-processing is one of the key steps in our approach, since a clean data gives us very good results even by using simple algorithms. We may need some more attributes that are not directly observed in the employee data but can be derived from inside the data. Moreover, we may encounter some missing data and can use several imputation techniques [2]. The categorical values are converted to numeric values in order to make the classification algorithm more efficient. For example, categorical attribute ‘salary’ contains three values such as low, medium and high. Hence it is converted to 0, 1 and 2 respectively [15].

If some attributes have not been recorded, they are filled up using appropriate domain knowledge (e.g., if the satisfaction\_level is not available, then it is filled up with the mean of all satisfaction\_level).

1. *Model Construction*

In this phase we generate the classification model using decision tree, KNN, SVM and ANN. In constructing the prediction model, the dataset is divided into two sets of data, training dataset and testing dataset. In this study, we use the ratio of 80:20, 80% for training data and 20% for testing data [14].

1. *Model Comparison*

To ensure the prediction model is able to predict the employee churn, we compare three models [14]. This phase aim to know the best classification model to predict employee churn. The measurement parameters are listed in table II. The best score of each parameter is 1 and the worst score is 0.

1. **Results and Analysis**

In this experiment, we divide the employee dataset into training and testing dataset is 80:20 [14]. The 80% percent of the data is adequate in order to construct the model or else known as the model learning while the latter 20% is needed in order to validate the model. Therefore, table V shows the number of samples that we use in this study. The training dataset contains of data and testing dataset contains the remaining dataset or unseen data.

Table V. Dataset Portioning

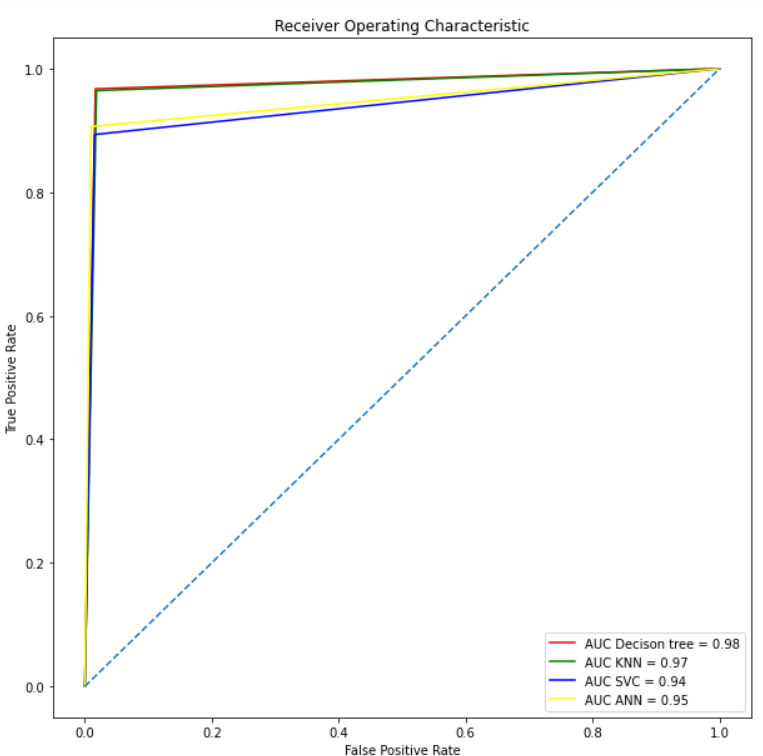
|  |  |  |
| --- | --- | --- |
| **Training/Testing** | **Proportions** | **No. of Samples** |
| Training | 80% | 11,999 |
| Testing | 20% | 3000 |

The first experiment is to produce a confusion matrix. Table VI shows summary of confusion matrix score that generated by decision tree, SVM, KNN. ANN. False Negative (FN) and False Positive (FP) represent a number of instances that misclassified by those three models. The highest number of misclassified generated by SVC, with the number of FN 2.50% and 1.17%% of TP.

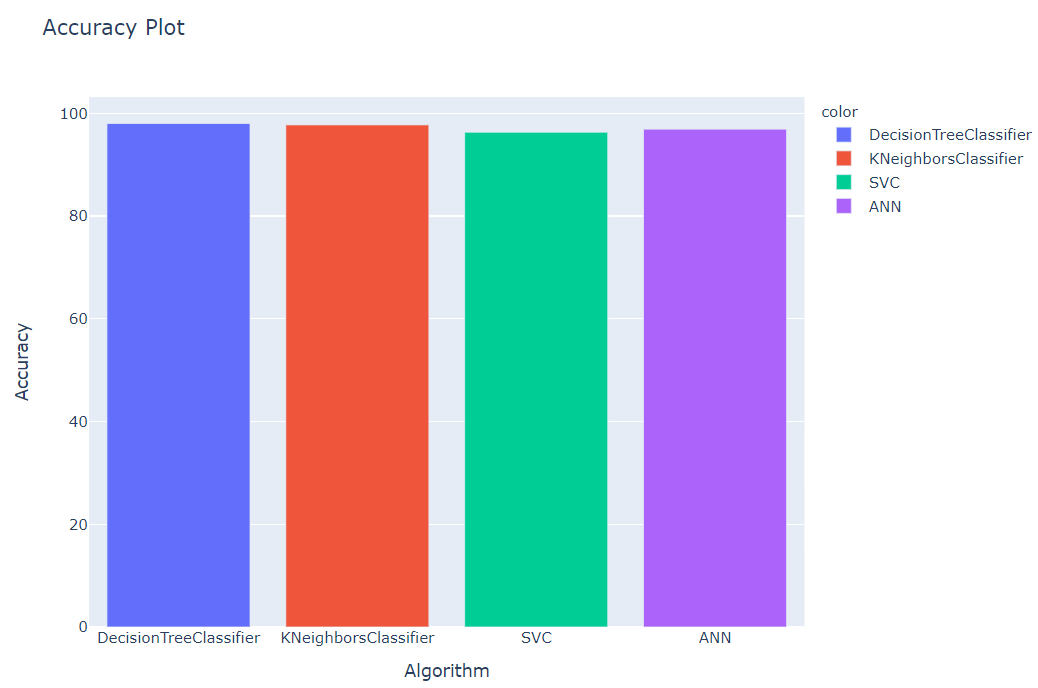
Table VI. Confusion Matrix Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Decision Tree** | **SVC** | **KNN** | **ANN** |
| **True Positive (TP)** | 22.63% | 20.93% | 22.60% | 22.70% |
| **False Negative (FN)** | 0.80% | 2.50% | 0.83% | 1.97% |
| **False Positive (FP)** | 1.17% | 1.17% | 1.40% | 1.10% |
| **True Negative (TN)** | 75.40% | 75.40% | 75.17% | 74.73% |

Afterwards, a model evaluation is performs using parameter as shown in table II for testing dataset in order to find the best model. In the table VI its shows that decision tree is the best model, it gains 98.03% of accuracy. SVC has 2.50%% of FN means that SVC is not a strong model in term of correctly classified employee that not churn as churner. For FP KNN also has 1.40% which means KNN still left behind in term of correctly classified churner as not churn. In Figure 7 for different classifiers. The higher the area under the curve, greater is the accuracy of the classifiers.



|  |  |  |
| --- | --- | --- |
| **S.NO.** | **Algorithm** | **Accuracy** |
| 1 | Decision Tree | 98.033 |
| 2 | SVC | 97.7667 |
| 3 | KNN | 96.3333 |
| 4 | ANN | 96.9333 |

****

1. **Conclusion**

Employee churn leads to customer dissatisfaction. In this project, hr\_data and employee\_satisfaction\_evaluation datasets have been considered. Different machine learning algorithms, Decision tree, KNN, SVM, and ANN have been used to find out which employee is going to quit the respective organization based on their working details and environments. Different performance measures like confusion matrix and overall accuracy have been used. From the experimental results, Decision tree is found to be better as compared to other models considered in this study. Thus, this model can be used for the real-life applications.

References

[1] Linoff, Gordon S., and Michael JA Berry. Data mining techniques: for marketing, sales, and customer relationship management. John Wiley & Sons, 2011.

[2] Yiğit, İbrahim Onuralp, and Hamed Shourabizadeh. "An approach for predicting employee churn by using data mining." *2017 International Artificial Intelligence and Data Processing Symposium (IDAP)*. IEEE, 2017.

[3] O. Ali and N. Z. Munauwarah, “Factors affecting employee turnover in organization/Nur Zuhan Munauwarah Omar Ali,” 2017.

[4] A. Frederiksen, “Job Satisfaction and Employee Turnover: A firmlevel perspective,” Ger. J. Hum. Resour. Manag., vol. 31, no. 2, pp. 132–161, 2017.

[5] H. Ongori, “A review of the literature on employee turnover,” 2007.

[6] V. V. Saradhi and G. K. Palshikar, “Employee churn prediction,” Expert Syst. Appl., vol. 38, no. 3, pp. 1999–2006, 2011.

[7] Duda, Richard O., Peter E. Hart, and David G. Stork. Pattern classification. John Wiley & Sons, 2012.

[8] Breiman, Leo. "Random forests." Machine learning 45.1 ,5-32 , 2001.

[9] Saradhi, V. Vijaya, and Girish Keshav Palshikar. "Employee churn prediction." Expert Systems with Applications 38.3,1999-2006, 2011.

[10] L. E. Peterson, “K-nearest neighbor,” Scholarpedia, vol. 4, no. 2, p. 1883, 2009

[11] Polak, Sebastian, et al. "Prediction of the hERG potassium channel inhibition potential with use of artificial neural networks." Applied Soft Computing 11.2 ,2611-2617 , 2011.

[12] Burges, Christopher JC. "A tutorial on support vector machines for pattern recognition." Data mining and knowledge discovery 2.2, 121- 167, 1998.

[13] Özbakır, Lale, Adil Baykasoğlu, and Sinem Kulluk. "A soft computing-based approach for integrated training and rule extraction from artificial neural networks: DIFACONN-miner." Applied Soft Computing 10.1 ,304-317, 2010.

[14] S. Khodabandehlou and M. Rahman, "Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior," Journal of Systems and Information Technology, vol. 19, no. 1/2, pp. 65-93, 2017.

[15] Sisodia, Dilip Singh, Somdutta Vishwakarma, and Abinash Pujahari. "Evaluation of machine learning models for employee churn prediction." *2017 International Conference on Inventive Computing and Informatics (ICICI)*. IEEE, 2017