Employee Turnover Prediction – Nidhi Bodar , Foram Shah

bodar.n@northeastern.edu
shah.fo@northeastern.edu

Abstract

Employee Turnover Prediction shows up the ratio of number of employees leaving the firm in a particular year. Employee Turnover means whether the employee is going to leave the organization in the upcoming period of time. Employee Turnover Prediction allows us to predict which employee or how many employees might leave the organization in near future based on a few features. Machine learning Classification algorithms will be used to predict the number of employees leaving the organization. Such classifier would help an organization predict employee turnover and be pro-active in helping to solve such costly matter. We can determine what are the responsible features that leads to Employee Turnover. We used two models i.e Random Forest Classification and Logistic Regression to get the prediction whether the employee might leave the organization in the upcoming period.

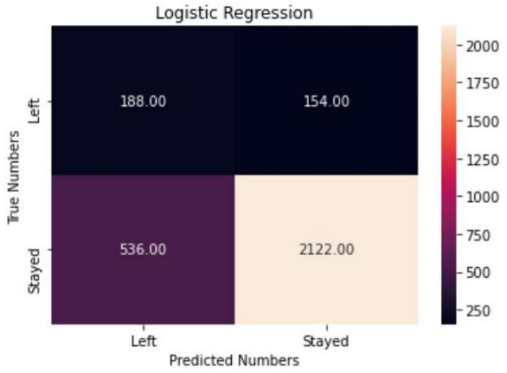
Introduction

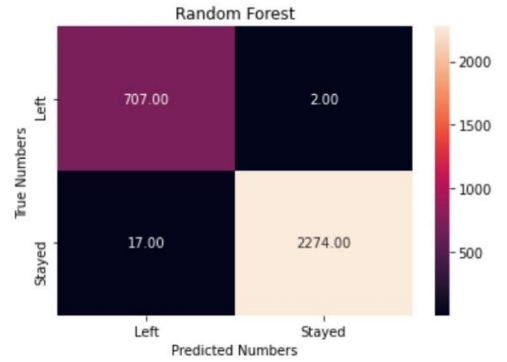
Prevention of employee turnover is a challenging task a human resources management field can encounter for organizations. As it brings the low productivity, high hiring costs, overtime costs. It is very costly for the organization where costs include but is not limited to: separation, vacancy, recruitment, training and replacement. On an average, any firm/organization usually invests three to four weeks for training any employee and it will turn into organizational loss if the employee leaves within a year . Service firms recognize that the timely delivery of their services can become compromised, overall firm productivity can decrease significantly and, consequently, customer loyalty can decline when employees leave unexpectedly. As a result, it is imperative that organizations formulate proper recruitment, acquisition and retention strategies and implement effective mechanisms to prevent and diminish employee turnover, while understanding its underlying, root causes. Most recently, the prevalence of intelligent machine learning algorithms in the field of computer science has led to the development of robust quantitative methods to derive insights from industry data. To reduce highly probable turnovers, the information of employee turnover prediction plays an essential role which can be exploited by the companies to take precautions against such situation as well as to derive valuable insights, which all together can result in ultimate enhancement of the efficiency of the organization. In this project, we would like to work on a problem which contributes towards predicting whether an employee will leave a company or not based on such number of projects, average monthly hours, promotions, salary and satisfaction level using the data available publicly. Hence, our project is inclined towards using supervised machine learning model, in particular logistic regression model which can be used for binary classification problem. With the help of random forest classification model, our project will be able to calculate accuracy of data. Graph analysis, scatter plot, roc curve will assist our model to interpret several experiments which will be computed on our model.

Data Analysis

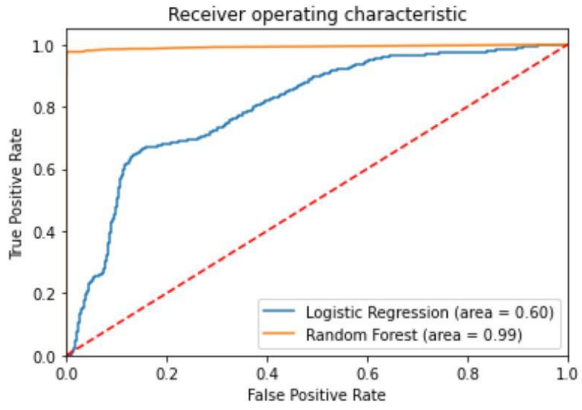
We analyzed our data to preprocess them to perform model and found categorical value which needed to be converted into numerical values for which we used Label Encoder. We implemented two models, logistic regression, and Random Forest to train our dataset. Confusion matrix and ROC curve were used to visualize prediction, for evaluating the accuracy of the machine learning

classification model and to compare these two model's implementations with each other.





As we can from the above confusion matrix that the first block is True positive value, and second block in the first row represents False Positive value which is 188 and 154 respectively for Logistic Regression. While for the Random Forest model, the numbers are 707 and 2 which are strangely accurate as Random Forest gives only 2 False Positive numbers, and the rest are true values that it has predicted.



We also compared two models in terms of ROC curve in which the red dotted line represents the ROC curve of a purely random classifier. It is plotted with TPR against the FPR where TPR is on the y-axis and FPR is on the x-axis. if the curve of any model is closer to the 45-degree diagonal of the ROC space, the test is less accurate which we can see that in our case is logistic regression curve. While the curve is near the 90-degreee and far from the ROC space such as Random Forest curve herr then it is expected to perform better.

Furthermore, the findings that we obtained from the feature importance parameter were quite surprising as we expected that salary and promotion obtained in the last five years will play essential role for employees to make the decision for leaving the company though the features that we acquired are satisfaction level of the employee, the number of projects they have done also amount of time spent at the company which can be seen from the following horizontal bar graph.

Models

In Modelling phase, we chose the best algorithm used for predicting employee turnover. As every supervised machine learning task, in order to train the predictive model, dataset is split into a Training Dataset, where the model is trained and where its parameters are fine-tuned to best fit the target variable, and into a Test Dataset, where the performance of the trained model is tested. We can use many different models like Decision trees, Random Forest, XGBoosted tree, KNN, SVM, Neural networks, etc. and then decide which model performs better than others. Looking at the data we have for prediction, we chose Random Forest Classification and Logistic Regression.

Random Forest Classification:

Random forests take an ensemble approach that provides an improvement over the basic decision tree structure by combining a group of weak learners to form a stronger learner. Ensemble methods utilize a divide-and-conquer approach to improve algorithm performance. In random forests, a number of decision trees, i.e., weak learners, are built on bootstrapped training sets, and a random sample of m predictors are chosen as split candidates from the full set P predictors for each decision tree. As $m << P$, the majority of the predictors are not considered. In this case, all of the individual trees are unlikely to be dominated by a few influential predictors. By taking the average of these uncorrelated trees, a reduction in variance can be attained, making the final result less variable and more reliable.

Logistic Regression:

Logistic Regression is a traditional classification algorithm involving linear discriminants, as originally proposed in 1958 by Cox . The primary output is a probability that the given input point belongs to a certain class. Based on the value of the probability, the model creates a linear boundary separating the input space into two regions. Logistic regression is easy to implement and work well on linearly separable classes, which makes it one of the most widely used classifiers.

What the models achieved

Random Forest Classification's feature selection method helps us in obtaining the best features that are useful for predicting the employee turnover with good accuracy. We had 18 features that were brought down to 10 features by the feature selection method. Fitting the training data on Random Forest Classification we got 99.1% accuracy through accuracy score attribute from metrics in sklearn. On the other hand we just got 77.1% accuracy while training the same data in Logistic Regression which can clearly tell us that Random Forest works better on our data set than Logistic Regression. Through Random Forest's Feature importance method we determined that Satisfaction level and Time spent at the company are the most important features that contributes the most for an employee leaving the organization.

Conclusion

Employee leaving the organization has been the greatest problem for any firm as it results in lots of loss to the firm in many ways. To solve this problem we made Employee Turnover prediction model using Random Forest Classification and Logistic Regression . The accuracy for Random Forest turned out to be better than Logistic Regression. The features affecting employees to leave the organization the most were satisfaction level, number of projects and the time spent at the organization. We can try different models to check what model gives better accuracy. Apart from that we can find out the aspects of an employee being happy at the organization and compare it with the ones leaving which may help the organization create a better environment.

References

1. 1. Alao, D., Adeyemo, A.B.: Analyzing employee attrition using decision tree algorithms. Comput. Inf. Syst. Dev. Inform. Allied Res. J. 4 (2013)
2. Sexton, R.S., McMurtrey, S., Michalopoulos, J.O., Smith, A.M.: Employee turnover: a neural network solution. Comput. Oper. Res. 32, 2635–2651 (2005)
3. https://www.displayr.com/what-is-a-roc-curve-how-to-interpret-it/
4. https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62
5. https://searchbusinessanalytics.techtarget.com/definition/logistic-regression
6. https://www.andrew.cmu.edu/user/yuezhao2/papers/18-intellisys-employee.pdf