Trainee Churn Prediction using Machine Learning: A Case Study of Data Scientist Job*

Oanh Thi Tran** and Ly Phuong Nguyen

International School, Vietnam National University, Hanoi, Vietnam
{oanhtt,lynp}@isvnu.vn

Abstract. The number of positions for data scientists is increasing. The companies working on big data and data science usually receive many registrations for the training programs of the companies before officially giving them a permanent role. Among those trainees, the companies want to know which candidates are really want to work for them or will look for a new employment after training time. This will help to reduce the training cost, and bring higher levels of satisfaction and retention. This work is performed to interpret the main factors impacting to candidate decision and then build a prediction model to predict the probability of a candidate will look for a new job or will work for the company using the current credentials, demographics, experience data, etc. To this goal, different robust machine learning methods are carefully investigated which are single classifiers such as decision trees, naive bayes, KNNs, SVMs and ensemble classifiers such as random forest, voting strategies, Xgboost and LGBM on a public dataset. The experimental results show that the ensemble classifiers have achieved relatively higher performance in comparison to the single classifiers. The LGBM classifier was the best one which yielded up to 80% in the F1 score using the selected feature sets. This research shows promising results and provides a strong preliminary result on this interesting yet unexplored problem.

Keywords: Churn prediction · machine learning method · data science.

1 Introduction

Churn prediction [1,9,13] is very common for any company or organization to know when and why the employees are likely to leave the company. This research direction is attracting the attention of many researchers over the world. Recently,

** Corresponding Author.

the application of machine learning in this field is blooming thanks to data for churn prediction is now available in considerable quantity. For examples, there existed a lot of research using different robust machine learning methods such as SVM [2], logistic regression [14], Xgboost [15], or tree-based classifiers like decision trees [17], random forest [17] , etc. on many public datasets.

While many researches have been done for employee churn prediction, to our knowledge, there is no published research on trainee or candidate churn prediction. Nowadays, companies which are active in Big Data and Data Science want to hire data scientists among people who successfully pass some courses which conduct by the company. Learning and developing at the training time is win-win for both the companies and the trainees. Typically, these companies receive multiple candidate signups for their training programs. Hence, they wants to know which of these candidates really want to work for the company after training time or looking for a new employment at other companies. This prediction would be extremely useful because it helps to reduce the cost and time as well as the quality of training or planning the courses and categorization of candidates.

This prediction problem is considered to be quite close to the problem of employee churn prediction. In fact, the data about the current and past candidates can be used to analyze to figure out the common characteristics of the candidates targeted to making prediction about the possible retention of the potential candidates in the future. In this paper, we aim at systematically studying about the trainee churn prediction. We exploited the public data available[1] at Kaggle to conduct the research.

This dataset designed to understand the factors that lead a person to leave the company after training programs. By model(s) that uses these data, we can predict the probability of a candidate to look for a new job or will work for the company, as well as interpreting affected factors on employee decision. Specifically, we conducted a systematic study on different robust machine learning techniques as follows:

- *Single classifiers*: decision tree [8], logistic regression [16], multilayer perceptron, k-nearest neighbors, and support vector machine [6].
- *Ensemble classifiers*: random forest [4], voting strategies, XGBoost [5] and LightGBM [3].

Before implementing the different models, we also performed explanatory data analysis to get more insights from this dataset. We also performed preprocessing to make the data in a good quality before feeding into the models. Finally, we also conducted feature selection method to select the most important features for building the best model. Experimental results on the public dataset are quite promising. The SVM method was proved to be the best model among single classifiers, while the LGBM classifier was the best one among ensemble classifiers. LGBM even outperformed SVM for all evaluation metrics and yielded 80% in the F_1 score. This result was slightly improved with the selected 26 feature

[1] https://www.kaggle.com/arashnic/hr-analytics-job-change-of-data-scientists

set. Specifically, using LGBM on these feature sets, we achieved nearly 80% in the F1 score.

The rest of this paper is organized as follows: Section 2 presents the related work on employee churn prediction. Section 3 shows some preliminary scan on the data using exploratory data analysis before developing the prediction models using the proposed methods mentioned in Section 4. Section 5 describes experiments setups, experimental results and some discussion on the results. Finally, we conclude the paper and show some lines of future work in Section 6.

2 Related Work

Alamsyah et al., 2018 [1] used three popular models for prediction which are Naïve bayes, decision tree, and random forest using a Human Resource Information System (HRIS) from a well-known telecommunications company in Indonesia. Punnoose et al.2016 [13] also used data from the HRIS of a global retailer to compare XGBoost against six historically used supervised classifiers and demonstrate its significantly higher accuracy for predicting employee turnover. The same, Jain et al. 2021 [9] used dataset from the HRIS and showed that the system using the CatBoost algorithm outperforms other ML algorithms. Alduayj et al. 2018 [2] conducted experiments using a synthetic data created by IBM Watson and using the following machine learning models: SVM, random forest and KNN. Aseel Qutub et al. 2021 [14] used IBM attrition dataset for training and evaluating machine learning models. Their result suggestion that Logistic Regressor had the highest values and Decision tree had the lowest scores. Khera et at., 2019 [11] used support vector machine (SVM) for prediction based on archival employee data collected from Human Resource databases of three IT companies in India, including their employment status at the time of collection. The same dataset, however, Yue Zhao et al. 2019 [17] used tree-based classifiers (XGB, GBT, RF, DT) and showed that they worked well in general. Srivastava et al.2021 [15] established the predictive power of Deep Learning for employee churn prediction over ensemble machine learning techniques on real-time employee data from a mid-sized Fast-Moving Consumer Goods (FMCG) company. Nguyen et al. 2020 [12] applied a case study of an organization with 1470 employee positions to demonstrate the whole integrating churn predict, EVM and machine learning process.

These researches mostly focused on the target of employees who are permanently working for the companies using a wide range of machine learning techniques. In this paper, we target to candidates or trainees of the company to see whether or not they are likely to leave the company after training time. We performed a systematic research on this task using a public dataset on Kaggle.

3 Explanatory Data Analysis

Here are some analysis on this dataset using explanatory data analysis techniques such as histogram, box plot, corelation analysis, etc.:

- Number of candidates '*leaving*' only accounted for 25%, while number of candidates '*not leaving*' made up 75%. Hence, this is an imbalanced class problem.
- It is noted that the majority of '*leaving*' are Male (89%). This is not surprising given that the dataset features a higher relative number of male than female and other.
- People who work in Data Science for the first eight years are more likely to look for a new job, and more than half of those who have been in the field for more than 20 years are not looking for a new job.
- Candidates work in small company are more likely to look for a new job, while medium and large company has a smaller number of seeking new opportunities.
- Candidates with graduate education are more likely than others to look for a new job.
- The majority of the candidates who do not leave the company are from cities with city indexes ranging from 0.8 to 0.9, whereas the candidates who do leave the company are from cities with city indexes ranging from 0.6 to 0.9.

4 Proposed ML classifiers

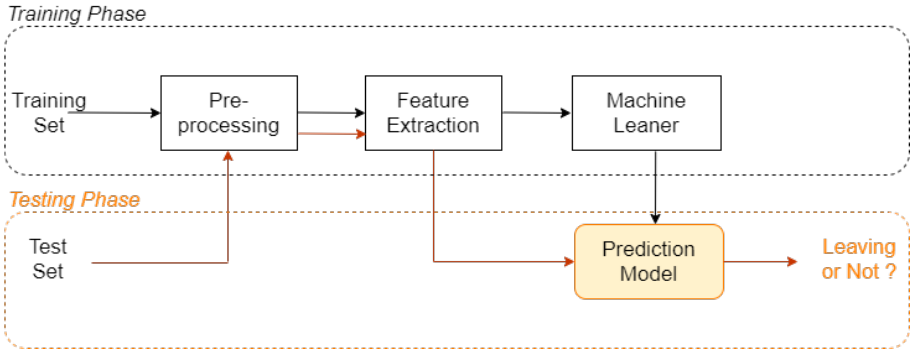

Fig. 1. The general framework for predicting trainee churn using ML methods.

Figure 1 shows the architecture for predicting hypertension risks using the machine learning approach. It consists of two main phases: training and testing phases. First, the data (both training and testing sets) will be pre-processed to remove noises and make data in a good quality.

After pre-processing, we performed extracting features for the machine learning methods used. That is, each sample will be represented by a vector $F = \{f_1, f_2, ..., f_n\}$. Labels are encoded into values of 0 and 1. The first phase uses training data including of D={train_X, train_Y} to help computers learn the pattern of hypertension or not hypertension. The prediction model will be later used to make prediction on unseen data set.

In this work, we exploited both single classifiers and ensemble classifiers to train the prediction models.

5 Experiments

5.1 Data Pre-processing

Dealing with missing data There are 8 features containing missing values including experience, enrolled university, last new job, education level, major discipline, gender, company type and company size. To handle this problem, we used the method *fillna()* to replace *NaN* values with '*unknown*' for these eight columns.

Converting categorical features Because all predictor variables in many models must be numeric. Therefore, these categorical variables must be properly transformed into numeric representations using dummy encoding methods.

Feature Scaling Feature scaling is to transform the values of different numerical features into the similar range of [0,1] using the *StandardScaler* function.

Class imbalance we used the SMOTE method for the tuned LGBM Classifier that is the best model. What it does is, it creates synthetic (not duplicate) samples of the minority class. Hence making the minority class equal to the majority class. SMOTE does this by selecting similar records and altering that record one column at a time by a random amount within the difference to the neighboring records [10].

5.2 Experimental Setups

We conducted 5-fold cross validation test. All experiments were performed using Google colab and evaluated using precision, recall, F1 and accuracy scores.

5.3 Experimental Results

Experimental results of different ML methods Table 1 shows the experimental results of models with Precision, Recall, F_1 score and the accuracy score.

Among single classifiers, the worst performance is the performance of the decision tree method, followed by the MLP method. The SVM method significantly outperformed other methods and yielded the highest performance on all four evaluation metrics. In comparison to the second and third best methods of KNN and logistic regression, it boosted the F1 and accuracy scores by approximately 3%. Using SVM, we achieved 78.81% in the F1 score and 79.22% in the accuracy score.

Table 1. Experimental results of different single classifiers and ensemble classifiers.

	Precision	**Recall**	**F1-score**	**Accuracy**
Single classifiers				
Decision tree	72.19	72.09	72.14	72.09
MLP classifier	74.57	75.20	74.86	75.20
Logistic Regression	75.71	77.85	75.85	77.85
KNN	75.49	76.77	75.96	76.77
SVM	**78.53**	**79.22**	**78.81**	**79.22**
Ensemble classifiers				
Soft Voting classifier	76.14	77.86	76.53	77.86
Hard Voting classifier	77.41	78.71	77.79	78.71
Random Forest	78.52	79.35	78.83	79.35
XGBoost	78.88	79.43	79.12	79.43
LGBM Classifier	**79.78**	**79.64**	**79.71**	**79.64**

As shown in Table 1, the ensemble classifiers have achieved relatively higher performance in comparison to single classifiers. The simple voting techniques could not enhance the performance even using strong single classifier like SVM. For the random forest technique, its performance was competitive with the best single SVM classifier. Two variants of gradient boosting architectures which are Xgboost and LGBM proved to be quite effective in predicting the likelihood of candidate churn on this dataset. Among two classifiers, LGBM was slightly better than Xgboost. It boosted the F1 score by nearly 1% in comparison to the single SVM classifier. This best classifier yielded quite good performance with 79.71% in the F1 score and 79.64% in the accuracy score.

Experimental results using SMOTE to handle imbalanced data Table 2 illustrates the model evaluation without SMOTE and with SMOTE using the best ensemble classifier of the LGBM method. The SMOTE technique can slightly improve the performan in all evaluation metrics. For the F1 score, using it enhanced the F1 score by 0.24% in comparison to not using it.

Table 2. Experimental results of the best LGBM methods with or without SMOTE.

	Precision	**Recall**	**F1-score**	**Accuracy**
Without SMOTE	79.78	79.64	79.71	79.64
With SMOTE	80.65	79.49	79.95	79.72

We also measured the performance of each class using SMOTE and realized that the prediction of the *class 1* is more difficult than the prediction of *class 0*. In more details, we gained 86% and 61.64% in the F1 score for *class 0* and *class 1*, respectively.

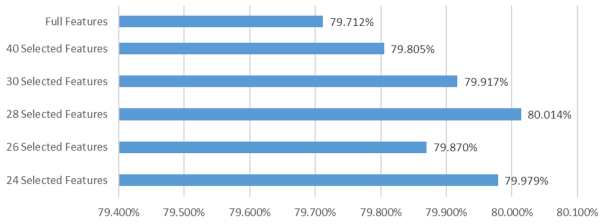

Fig. 2. Weighted F1 score of the best LGBM models using the selected features sets and full features.

Comparing experimental results between selected features and full features. This greatly impacts the performance of the models. In this study, we investigated three popular feature selection methods including univariate Selection with chi-squared statistical test, feature importance used the tuned LGBM classifier, and heatmap. Among selected top 50 best features for each technique, we found that all 3 methods shared the same 28 features. Based on these feature sets, we built the best models using the best tuned LGBM classifier. To have a better picture about the best feature sets, we also tried with other options around these 28 features. Figure 2 depicts that using only shared features of common 28 features yielded a sightly better performance than using their feature subsets. Using the best set of 28 features yielded the best performance with 0.3% improvement in the F1 score in comparison to using the full feature set.

6 Conclusion

This paper presented a work on predicting the likelihood of the candidates with the intention to leave or do not leave the company after training periods. This work was performed to interpret the main factors impacting to candidate decision and then build a prediction model to predict the probability of a candidate will look for a new job or will work for the company using the current credentials, demographics, experience data, etc. We conducted extensive experiments using different machine learning methods in order to look for the best prediction model. Experimental results on a public dataset showed that in general the ensemble classifiers gave the relatively higher performance in comparison to the single classifiers. The LGBM classifier was the best one which yielded up to 80% in the F1 score using selected feature sets. Among two classes, the experimental results showed that predicting the class 1 – the candidate leaving the company is more difficult than predicting the class 0 – the candidate doesn't not leave the company. We don't expect a perfect model but the promising results suggested that the best model could be used in the companies today.

References

1. Alamsyah, A., Salma, N.: A Comparative Study of Employee Churn Prediction Model. In Proceedings of the 4th International Conference on Science and Technology (ICST), pp. 1-4 (2018), doi: 10.1109/ICSTC.2018.8528586.
2. Alduayj, S.S., Rajpoot, K.: Predicting Employee Attrition using Machine Learning. In: Proceedings of the 2018 International Conference on Innovations in Information Technology (IIT), pp. 93-98 (2018), doi: 10.1109/INNOVATIONS.2018.8605976.
3. Amin, A., Rahim, F., Ali, I., Khan, C., Anwar, S.: A Comparison of Two Oversampling Techniques (SMOTE vs MTDF) for Handling Class Imbalance Problem: A Case Study of Customer Churn Prediction. In: Rocha A., Correia A., Costanzo S., Reis L. (eds) New Contributions in Information Systems and Technologies. Advances in Intelligent Systems and Computing, vol 353. Springer, Cham. https://doi.org/10.1007/978-3-319-16486-1_22.
4. Breiman, L.: Random forests. Machine Learning, 45(1), pp. 5–32 (2001).
5. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y.: Xgboost: extreme gradient boosting. R package version 04-2, pp. 1–4 (2015).
6. Cortes, C., Vapnik, V.: Support vector machine. Machine Learning, 20(3), pp. 273–297 (1995).
7. Gao, X., Wen, J., Zhang, C.: An improved random forest algorithm for predicting employee turnover. Mathematical Problems in Engineering, pp. 1–12 (2019).
8. Jin, C., De-lin, L., Fen-xiang, M.: An improved ID3 decision tree algorithm. In Proceedings of the 4th International Conference on Computer Science and Education: IEEE (2009). https://doi.org/10.1109/iccse.2009.5228509.
9. Jain, N., Tomar, A., Jana, P.K. A novel scheme for employee churn problem using multi-attribute decision making approach and machine learning. J Intell Inf Syst 56, pp. 279–302 (2021). https://doi.org/10.1007/s10844-020-00614-9.
10. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y.: LightGBM: a highly efficient gradient boosting decision tree. In Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 3149–3157 (2017).
11. Khera, S., Divya.: Predictive Modelling of Employee Turnover in Indian IT Industry Using Machine Learning Techniques. Vision: The Journal of Business Perspective, 23(1), pp.12-21 (2019).
12. Nguyen, T.N.A., Nguyen, D.T., Vijender, K.S., Nguyen, L.G., Vu, H.T., Luong, N.S., Nguyen, D.L., Vu, T.N.: Integrating Employee Value Model with Churn Prediction. International Journal of Sensors, Wireless Communications and Control; 10(4) (2020), https://doi.org/10.2174/2210327910666200213123728.
13. Punnoose, R., Ajit, P.: Prediction of Employee Turnover in Organizations using Machine Learning Algorithms. International Journal of Advanced Research in Artificial Intelligence, 5(9) (2016).
14. Qutub, A., Al-Mehmadi, A., Al-Hssan, M., Aljohani, R., Alghamdi, H.: Prediction of Employee Attrition Using Machine Learning and Ensemble Methods. International Journal of Machine Learning and Computing, 11(2), pp. 110-114 (2021).
15. Srivastava, P., Eachempati, P.: Intelligent Employee Retention System for Attrition Rate Analysis and Churn Prediction. Journal of Global Information Management, 29(6), pp.1-29 (2021).
16. Webb, G.I., Sammut, C., Perlich, C., Horvath, T., Wrobel, S., Korb, K.B., Noble, W.S., Leslie, C., Lagoudakis, M.G., Quadrianto, N., Buntine, W.L., Quadrianto, N., Buntine, W.L., Getoor, L., Namata, G., Getoor, L., Xin Jin, J.H., Ting, J.A.,

Vijayakumar, S., Schaal, S., Raedt, L.D.: Logistic regression. In Encyclopedia of Machine Learning (pp. 631–631) (2011).
17. Zhao, Y., Hryniewicki, M.K., Cheng, F., Fu, B., Zhu, X.: Employee Turnover Prediction with Machine Learning: A Reliable Approach. In: Arai K., Kapoor S., Bhatia R. (eds) Intelligent Systems and Applications. IntelliSys 2018. Advances in Intelligent Systems and Computing, vol 869. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-01057-7_56.