Title

Optimisation of logistic regression with KNN to analyze IBM Employee Turnover

Problem Statement or Introduction

Human Resources is an essential department for almost all the firms. Nowadays, there is a significant increase in employee turnover in many industries and that led to a high extra cost for the firms. Employee attrition is an important topic for a company's development: a low turnover rate usually linked to high productivity, expected profit, and stability of the firm. Therefore, it is important to understand what leads to this situation and what can we do to mitigate it. My research question is to find out what is the most influential factor that determines an employee's willingness of attrition using IBM HR attrition dataset (Pavansubhash, 2016), a feature selection research. I built two models in this paper: one self-designed logistic regression with KNN and a commonly used SVM method. Then compare the results of both and find out the most influential factors.

Findings and Results

0.1 Statistical Analysis

The age in the dataset is from 18 to 60 and we divided them into two parts: Under 30 and Above 30. According to the figure, most of the people showed an age above 30. Also, there are five different education levels that have been indicated clearly in the model: 1 means under college, 2 stands for some college, 3 is for bachelor's degree, 4 is for master's degree, and 5 means doctoral degree. We made paired heat maps: one is about the education level to the attrition status, and the other is about how age affected the attrition. From the education one, we observed the percentages of employee attrition for under college, some colleges, bachelors, and masters are very similar. However, for the highest doctoral degree, it showed a low rate of about 10% instead. For the age variable, we grouped them into 6 equal length groups and conducted the percentages of attrition of each group. It showed

a linear trend, which means as people get older, they are less likely to leave their current firms. For employees in their 18-25, the percentage of leave is 28%. However, for the age group from 53 to 60, the percent fall to 12%. We suspect that age may be a factor that influences employee attrition in this dataset.

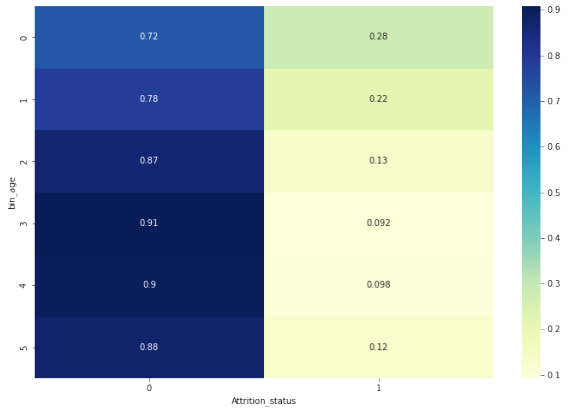


Figure 1: Age-Attrition Heatmap

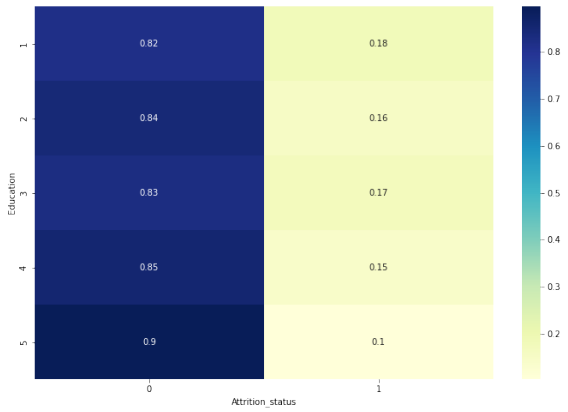


Figure 2: Education-Attrition Heatmap

We proposed the hypothesis that for the group of people before 30 and people after 30, they may have different extent of the correlation that how education level contributed to

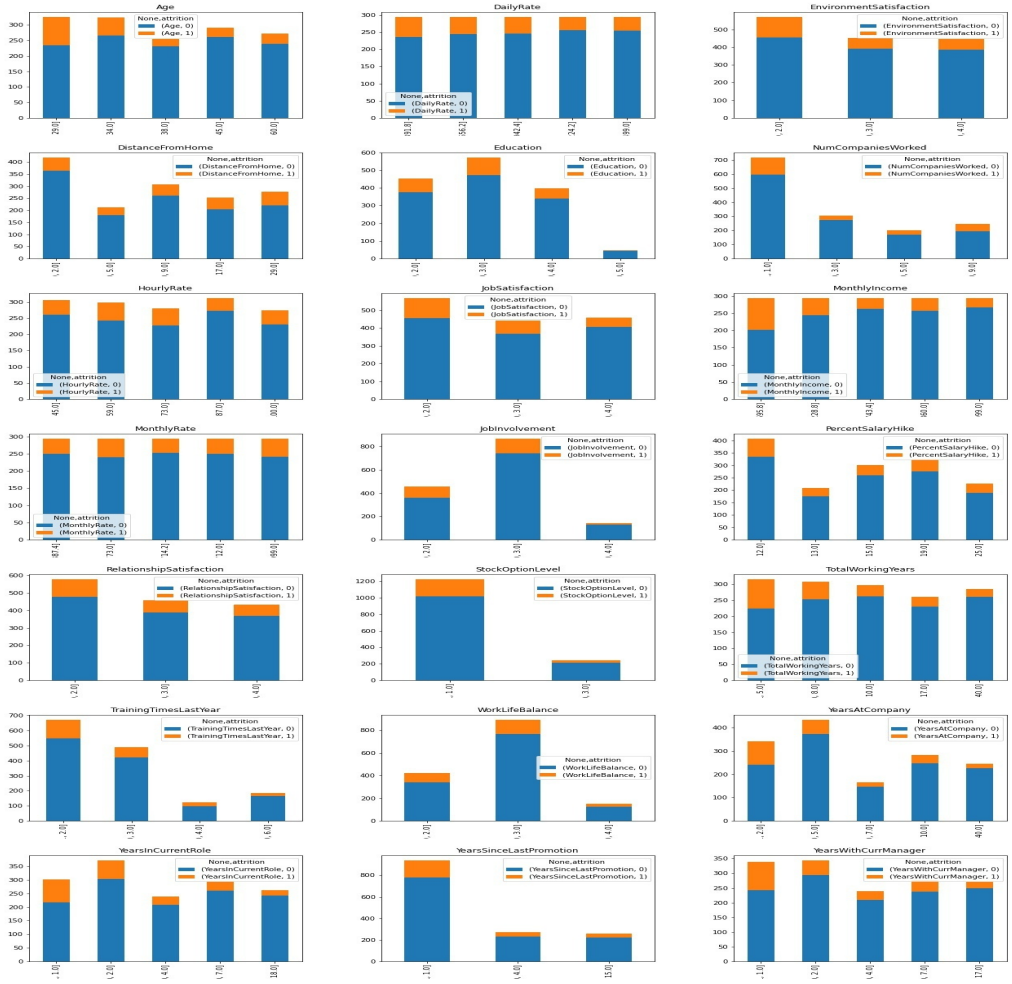
the attrition. We connected a two sample t-test: our first group is the education under 30 and another group for the education above 30. The result showed a t-test statistics -10.1841 with a p-value much smaller than 0.05. We can conclude that it is meaningful to explore deeper. In order to figure it out, we made the education variable into five dummy variables and then made the interaction variable between the Under 30 column and the education dummy variables. Then, we plan to add those variables into the logistics regression that had been explained in the following section.

0.2 Linear Testing

Linear Testing of all the independent variables is an important stage of the research. We used a binary response variable attrition, so it will be inaccurate if we conduct simple linear regression tests and get p-value. There is no missing data in the dataset and 22 independent numerical variables are continuous. We clustered all data points for all 22 variables into 5 equal length bins using qcut and then made bar plots to observe the frequency of 1 appeared in each range. From the stacked bar plot, we observed the trend of split lines between 0 and 1 for attrition. If the trending line for each plot showed a straight line, either upwards or downwards, we can conclude that the relationship between the testing independent variable and the attrition is linear. The reason behind this is we want to investigate the distribution of the number of 1s in each bin. If the percentage always goes up or down, we can conclude that changing in that independent variable can lead to a rise or fall in dependent variable attrition (Lavin, 2022).

Because of the repetition of data in each column, some of the columns can not be equally distributed into 5 bins, but less. It is also reasonable to observe the trending line if we have fewer bins. For the performance rating variable, there is one bin left, which does not make sense. We classified it into nonlinear categories. We can conclude that there are 14 variables showing a linear relationship to the response variable attrition, while 8 variables do not show a linear correlation. We divided the whole dataset into two sub dataset based on the linear

Figure 3: Linear Testing for 22 Numerical Independent Variables



classification. Then, we will use logistic regression to train the linear variable dataset. At the same time, we train the non-linear dataset use KNN method. Combining the results, we get the performance of model 1. Then, we apply SVM to the two datasets but specify the different kernels.

Linear-Variables		**Nonlinear-Variables**	
• Age	• Years In company	• Education	• Performance Rate
• Relationship Satisfaction	• Years Since Promotion	• Number of Company worked	• Percent Salary Hike
• Environment Satisfaction	• Years In Current Role	• Monthly Rate	• Job Involvement
• Job Satisfaction	• Years With Current Mgr	• Hourly Rate	
• Hourly Rate	• Total Working Years	• Work-Life-Balance	
• Monthly Income	• Training Time		
• Stock Option Level			
• Number of Company worked			

0.3 Logistic Regression Model

Logistic regression is a good approach to find out the probability of an event happening if our dependent variable is in binary. In this case, predict dependent variables from the logit should be the likelihood about whether a person with those parameters will leave the current company or not. The various parameters for each independent variable should reflect the extent of influence to our predicted variable. We take the coefficient β generated by the model and do the logarithms operation to it. It will give us the odd ratio which can be used in order to compare various independent variables.

$$e^{\beta} = Odds\,Ratio$$

In order to train our model, we conducted the train-test-split for our linear variable dataset. We set 20% of the data for training purpose and 80% of data for testing purpose. We built

the model using training data and then applied the model to the testing data. The accuracy rate we got for the linear variables is 85.034% which is close to the figure reported by the literature. And then , we built a table that showed the coefficient of each variable, and did the logarithm operation to get the odds ratio. I will discuss more in the next paper.

Conclusion

By using the IBM HR Attrition dataset, I had some understanding about all those features at this stage. I explored the linearity between independent variables and my responsive variable and classify them into two categories: linear and nonlinear. I built two datasets to represent each and plan to work on them separately. Also, I am starting to work on the different model training and regressions. For the next steps, I plan to run all the regressions: build the models for training data and apply the model to the testing data. Then, I will do the model validation and compare the results with other scholars' literature.

Reference

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python*. O'Reilly Media.

B, H. N. (2020, June 1). Confusion matrix, accuracy, precision, recall, F1 score. Medium. Retrieved October 13, 2022, from https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd

Bruce, P. C., Gedeck, P., Sawka, K., & Danch-Wierzchowska, M. (2021). Statystyka praktyczna W Data Science: 50 kluczowych zagadnień W językach R I python. Helion.

Gandhi, R. (2018, July 5). Support Vector Machine - introduction to machine learning algorithms. Medium. Retrieved October 13, 2022, from https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Good , B. (2018, November 6). 11 factors to consider when making a career change. GQR. Retrieved October 13, 2022, from https://www.gqrgm.com/11-factors-to-consider-when-making-a-career-change/

Haussman, G. (2016, October 25). 4 truths about hotel employee retention. Hotel Management. Retrieved September 21, 2022, from https://www.hotelmanagement.net/4-truths-about-hotel-worker-employee-retention#:~:text=It's%20so%20serious%2C%20there's%20an,must%20be%20hired%20and%20trained.

Immaneni, Kiran & Vedala, Naga & Vedala, Sailaja. (2019). Article ID: IJM_10_06_017 Cite this Article: Kiran Mayi Immaneni and Dr. Vedala Naga Sailaja, A Study on Factors

Effecting the Employees Attrition in Hotel Industry with Reference Hyderabad. 170-176.

Khera, S. N., & Divya. (2019). Predictive modelling of employee turnover in Indian IT industry using Machine Learning Techniques. Vision: The Journal of Business Perspective, 23(1), 12–21. https://doi.org/10.1177/0972262918821221

McKinney, W., & others. (2010). Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51–56).

Mansor, N., Sani, N. S., & Aliff, M. (2021). Machine learning for predicting employee attrition. International Journal of Advanced Computer Science and Applications, 12(11). https://doi.org/10.14569/ijacsa.2021.0121149

Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830.

Poornappriya, T. S., & Gopinath, R. (2021). Employee Attrition In Human Resource Using Machine Learning Techniques. In Webology (6th ed., Vol. 18, Ser. 2021, pp. 2844–2856). essay.

Sava, J. A. (2022, April 7). Tech GDP as a percent of total U.S. GDP 2021. Statista. Retrieved September 21, 2022, from https://www.statista.com/statistics/1239480/united-states-leading-states-by-tech-contribution-to-gross-product/

Sky Ariella. "27 US Employee Turnover Statistics [2022]: Average Employee Turnover Rate, Industry Comparisons, And Trends" Zippia.com. Aug. 30, 2022, https://www.zippia.com/advice/employee-turnover-statistics/

Toporek, A. (2020, October 27). Employee retention: Low-skilled, hourly jobs. Medium. Retrieved September 23, 2022, from

https://medium.com/a-level-capital/employee-retention-low-skilled-hourly-jobs-9b5a6a269853

V, L. G. (2022, August 2). Cross-validation techniques in machine learning for better model. Analytics Vidhya. Retrieved October 13, 2022, from https://www.analyticsvidhya.com/blog/2021/05/4-ways-to-evaluate-your-machine-learning-model-cross-validation-techniques-with-python-code/

Yang, S., & Islam, M. T. (2020). IBM Employee Attrition Analysis. arXiv preprint arXiv:2012.01286.

Zhang, Z. (n.d.). *Hannahzzy/senior_project_hannah*. GitHub. Retrieved October 27, 2022, from https://github.com/HannahZZY/Senior_Project_Hannah

Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., & Zhu, X. (2018, September). Employee turnover prediction with machine learning: A reliable approach. In Proceedings of SAI intelligent systems conference (pp. 737-758). Springer, Cham.