401 Methods Section

Ziyue(Hannah) Zhang

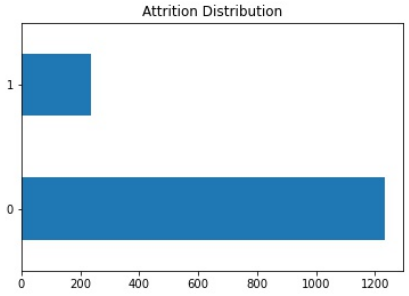
Due 14th Oct.

This research project is feature selection data analytics research, basically comparing several machine learning models and conducting various statistical tests to choose the most significant variables that influence people's willingness of attrition in the near future. IBM HR Attrition data has been used here. IBM is an international Information Technology firm, which has been defined as a part of the labor intensive industries. There are 35 attributes in the raw datasets, including one responsive variable employee attrition and some independent variables. Attrition status dataset is a binary list that shows whether this employee is still actively working for IBM or not. From the plot(a) Attrition distribution, we noticed it is an imbalance data. 87.87% of observations in the dataset shown a 0 means they are still active in the firms. Only 16.13 % people are inactive, with a indicator of 1. Through a series of model testing and validations, we should be able to get conclusions about some noticeable factors for the company's HR department in terms of maintaining a stable employee situation. Also, which model works the best for those kind of data. The results can be used for other firms in the future.

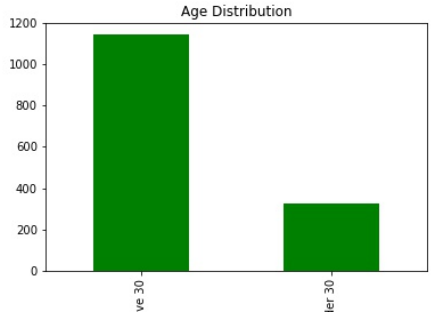
Data Exploration

As a feature selection analysis, it is reasonable to dig deeper into the independent variables first. As a high-dimensional data, we want to remove the irrelevant attributes and nonsense data first. After scan through the dataset, "employee count", "standard hours" and "over 18" attributes can be moved. All the values within all three columns are 1, which we suspect are missing data and not helpful for our analysis (Mansor et al.). Other than that, "EmployeeNumber is found not useful because they are just some random generated integers, which can be removed from the dataset too. After deleting all these, there are 30 independent variables in the datasets. There are 24 numerical variables and 6 categorical

variables.



(a) Attrition Distribution



(b) Age Distribution

In order to make variables more interpretable, we classified them into two groups based on their contents: job related reasons and personal reasons. There are 12 variables full into the category of job related reasons. Then, we divided them into three sub-genres: Job descriptions, financial reasons, and non-financial reasons. Jobs descriptions refers to people's departments, job level, job role, etc. Most of them are categorical data. Financial reasons mainly focus on people's compensation. In common, one of the goals that people change jobs and work is to earn money (Good, 2015). In many public technology firms like IBM, they also provide company stocks to their employees as a part of the benefit. This stock option level is included in this category too. The last subcategory non-financial reason is about job satisfaction: employee involvement, work life balance, job satisfaction and so on. Those data usually had been self-reported by the employees and that reflect whether they are satisfied with the current work. A potential indicator of whether they may decide to leave their current job in the near future. On the other side, some factors related to the employees, self identity and family status, is an essential part of our research too. In the self identity part, we included people's personal backgrounds: gender, education, age, etc. We want to know whether a particular personal background may lead to a more likely attrition. Family is very important for some people, distance from home, business travel frequency, marital status, that may affect people's job stability. In this case, we explore deeper about whether people in different age groups react differently to their family concerns. We splitted

the data based on the employee's age into two groups: below 30 and above 30. We created interactive variables based on variable "Below 30" and other independent variables and put them into our models, so we can find out whether the age changing the factors' impacts for the employee attrition.

Job Related		**Personal**	
• Department	• Relationship Satisfaction	• Age	• Distance From Home
• Job Level	• Job Role	• Gender	• Business Travel
• Environment Satisfaction	• Performance Rating	• Total Working Years	• Work-Life-Balance
• Hourly Rate	• Years In company	• Education	• Number of Company worked
• Monthly Rate	• Years Since Promotion	• Training Time	
• Stock Option Level	• Years In Current Role	• Martial Status	
• Job Involvements	• Years With Current Mgr		
• Job Satisfaction			

Before making a regression models, we conducted the correlation matrix for our variables. Job level and monthly income are highly correlated with a correlation coefficient 0.95. Also, the job level variable and total working years variable have a correlation coefficient 0.78. We want to make sure our model follows all assumptions and there is no multicollinearity problem. The threshold for correlation score is 0.8. After we decided to remove the job level variable, all other variables do not have multicollinearity problems. Also, according to the summary statistics, there are no outliers in the data. There is no missing value in the dataset too. Our finalized IBM HR attrition data for this research is 1470 rows with 29 valid columns.

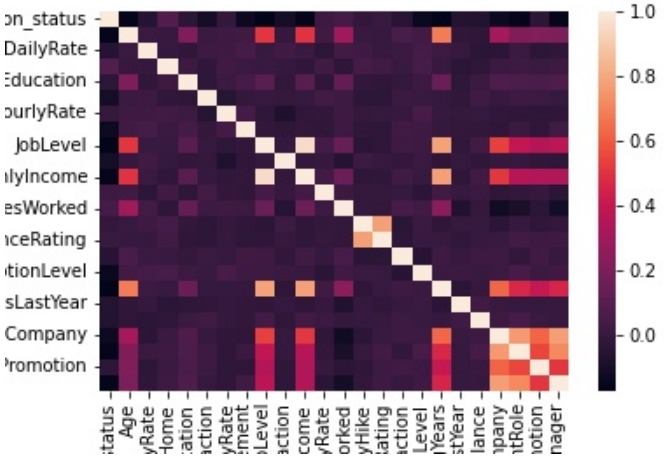


Figure 1: Correlation Matrix for Selected Variables (Yang et al. 2021)

Tools and Platforms

All the models had been built using Python, within the Jupyter Notebook environment. In this research, I mainly used the Scikit-learn package (Pedregosa et al., 2011) and pandas package to load the dataset and perform machine learning models. Seaborn and matplotlib libraries are used for visualization. Numerical computation had been conducted by the NumPy package (Van Der Walt, Colbert, Varoquaux, 2011). Stata had been used rarely for the data exploration part. All the codes can be found in the supplement material in github (https://github.com/HannahZZY/Senior_Project_Hannah).

Feature Selection

After understanding the variables in the data and classifying them by contexts, we want to explore deeper about the correlation between those features to the response variable attrition. As we already separated our variables into two categories: job related reasons and personal reasons. We want to find the most influential variables from both perspectives using machine learning algorithms. Based on the current literature, SVM (Support Vector Machine) is a generally recommended approach for this dataset. There are 6 out of

10 research paper recommend this algorithm. At the same time, I made a approach that is a combination of logistic regression and K-Nearest Neighbors (KNN). And we want to compare pros and cons of both approach to get our final conclusion.

I conducted a logistic regression and K-Nearest Neighbors (KNN) in this research. Logit is for the linear variables and KNN is for non-linear variables. Logistic regression is a statistical method which we use to predict the possibility of a binary response variable using one or more independent variables (Sukhadiya et al, 2018). In order to make logistic regression to be valid, the data in columns that we put into it should be linear. Logistic regression can help us to conclude which factors enhance the probability of a given case: for example willingness of attrition in our case. KNN algorithm is a supervised machine learning method to predict the label for the responsive variables, based on some of the close observations. Here in this case, KNN method will give us the concrete answer whether the people are predicted to be terminated or continue to work for the firms in the near future. Also, there is no limitation about data itself. The ways we did is to split the column into equal length bins based on the range. Then, we drew the stack bar plots to count the frequency of that variable appeared in each bin. If the relationship is linear, we should observe a clear linear trend in our plot, and vice versa. All the linear variables, we did logistic regression to them. For the remaining variables, we did KNN. KNN is less efficient compared with logistic regression, and we also want to take it into considerations.

SVM is a generalized machine learning method that is used to solve classification problems. The objective of SVM is to find the most accurate hyperplane in the N dimensional dataset. The way of doing this is to find the maximum margin between the different classes (Gandhi, 2018). In the present study, we have 29 independent variables, which transfer to a hyperplane with 29 dimensions using linear algebra.

We used the Scikit-learn learn package in python to perform both models. The first step of making machine learning models is to split the train and test data. After considering the

size of the dataset, we decided on a 2:8 distribution with 20% of data used for training and 80% of the data used as the testing data. Then we build models with different algorithms and alter the parameters, like the K value for KNN to find the most accurate model. A smaller K may lead to a less accurate model. However, a large K will lead to overfitting problems and need a long time to run. After fitting the test data with the model, it is very important to conduct validation techniques to make sure the model is valid. We chose the cross-validation techniques to split data into k groups (Mansor er al., 2021). We calculated the absolute sum of the residuals for both training and testing data. If the sum of residuals for testing is close to the value that we got in training part. That means we do not have overfitting/underfitting issue in our model. After all the models had been completed. We summarized the results from each model and compared the results.

From the distribution of attrition status plot, we know it is an unbalanced dataset, which means the number of 0s and 1s are not equal. We can not simply report and compare the accuracy rate, as the smaller 1s are the target that we want to focus and we want to know the factors influence people's attrition. In order to make the comparison in this case, we want to calculate the precision and recalls for the model. Both of them are ways to compute the difference between the predicted results and the actual results. Precision and recall both close to 1 means the model is a good classifier. We use F1-score as a matrix to take into account both precision and recall and report F1- score as part of the result (Harikrishnan, 2019).

$$Precision = \frac{TP}{FP + TP} \qquad (1) \qquad\qquad Recall = \frac{TP}{FN + TP} \qquad (2)$$

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

*TN: True Negative; TP: True Positive; FN: False Negative FP: False Positive

Based on our central question. We have two goals for the paper. On one side, we want to compare the performance of the models: F1-score and time needed to run, etc. That can

be utilized by firms in the future. On the other side, we want to conclude the important features for each category as our key results. In the following sections, we perform all the models and present the results.

Appendix

Table 1. Summary of Numerical Variables

Variable	Obs	Mean	Std. dev.	Min	Max
Daily Rate	1,470	802.4857	403.5091	102	1499
Distance From Home	1,470	9.192517	8.106864	1	29
education	1,470	2.912925	1.024165	1	5
Environmental Satisfaction	1,470	2.721769	1.093082	1	4
Hourly rate	1,470	65.89116	20.32943	30	100
Job involvement	1,470	2.729932	.7115611	1	4
Job Satisfaction	1,470	2.728571	1.102846	1	4
Monthly income	1,470	6502.931	4707.957	1009	19999
Monthly rate	1,470	14313.1	7117.786	2094	26999
# of companies worked	1,470	2.693197	2.498009	0	9
Percent salary hike	1,470	15.20952	3.659938	11	25
Performance	1,470	3.153741	.3608235	3	4
Relationship Satisfaction	1,470	2.712245	1.081209	1	4
Stock option Level	1,470	.7938776	.8520767	0	3
Total working years	1,470	11.27959	7.780782	0	40
Training time	1,470	2.79932	1.289271	0	6
Work life balance	1,470	2.761224	.7064758	1	4
Years at company	1,470	7.008163	6.126525	0	40
Years In Current role	1,470	4.229252	3.623137	0	18
Years since Promotion	1,470	2.187755	3.22243	0	15
Years with current Mgr	1,470	4.123129	3.568136	0	17

Reference

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.

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.