Model Optimisation and Feature Selection of IBM HR Employee Attrition

Ziyue(Hannah) Zhang

A paper presented for Data Analytics Major

DA 401: Seminar in Data Analytics

Data Analytics Major

Denison Unversity

United States

November 15th 2022

Abstract

Applying computational techniques into business strategy analysis has become an increasing trend for many firms. In this research, we analyzed human resource attrition. A series of supervised machine learning techniques had been performed to predict employee attrition based on personal characteristics and job-related profiles. In the data preprocessing process, we dug into the linearity between the independent and predicted variables, then we classified all the factors into two classes: linear and nonlinear. Logistic regression, K-nearest clustering (KNN), and Support Vector Machine (SVM) are the three models used in this paper. In terms of examining the model performance, we used precision, recall, and F-1 score in this case, as it is an imbalanced dataset. Computation time is another factor that we want to take into account. The results showed that the proper use of the combination of logistic and KNN algorithms provided accurate results and was able to answer all the questions. We expect this method can be widely used by other firms in their dataset in the future.

Introduction

Employee is a valuable asset for many firms. The Human Resource department monitors the employee inventory for the firms and wants to reduce business risks. They need to collect and analyze the employee movement data on a regular basis to track the company's recruitment situation. However, recently as the unemployment rate goes higher after the COVID crisis, there is a significant increase in employee turnover in many industries and that led to a high extra cost for the firms. In order to make a long-term business plan to ensure the company's health, employee attrition has been highlighted in HR's work. It is an important topic for a company's development: a low turnover rate is 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 for this paper is to find out what are the influential factors that make employees leave their current job, which we also called employee attrition. In this

paper, we used IBM HR attrition dataset to conduct a feature selection research. This dataset is public data that has been used for research since 2018. A lot of research has been done based on other sectors of employee attrition, like the hotel industry, banking, and technology firms. Also, different algorithms had been tested using this dataset, from basic statistical analysis to deep learning. However, some may contain flaws, like performance comparison and linearity between variables, while others did not realize the importance of imbalanced data affecting the model results. In this paper, I highly emphasized the property of this imbalanced dataset and made adjustments to my models. I explored two methods to approach this dataset in this paper: one self-designed logistic regression (Cox, 1958) with K-nearest clustering (KNN) (Mucherino, 2009) and another commonly used Support Vector Machine (SVM) (Cortes, 1995) method. Then in this paper, I plan to compare the results of both and find out the most influential factors. The overall performance of my self-designed Logistic regression with the KNN model performs very well. And that model satisfied all the statistical assumptions, which can be applied in the future. The other goal of the paper is to identify a reliable model that can be duplicated and reused widely for other firms in the employee-oriented industry using their company data.

Literature Review

Many labor-intensive service sectors: Finance, Tourism, Information Technology (IT), etc, put employees as their priority. The purpose of the Human Resource department is to manage and assist employees' wellness and maintain a stabilized development of the company. However, because of the uncertainty of the current job market, the time for an employee to stay in a firm has become shorter gradually. Employee attrition has become an increasingly important topic in long-term company strategies. Employee attrition had been defined as an employee leaving the firm both voluntarily and involuntarily. There are two numerical measurements to compare companies' turnover status: monthly turnover and annualized turnover (Bright Hub, 2011). HR departments usually track both rates as they can show the overall situation of a firm's labor stability.

$$Monthly\ Turnover = \frac{Total\ Number\ Of\ Termination}{Headcount\ Number} * 100\%$$

$$Annualized\ Turnover = \frac{Total\ Termination\ Across\ the\ Period}{Average\ Headcount\ in\ that\ Period} * 100\%$$

A lot of service sector firms started to experience a rise in turnover rates, which became a chronic problem for firms. The hotel industry, for example, had an annualized employee turnover rate of 73.8% in 2016 (Haussman, 2016). In other words, if we started a hotel that had 100 employees, at the end of the period, 73 contracts had been terminated. Low-skilled jobs, requiring less training time and no bachelor's degree, tend to have a high attrition rate (Toporek, 2020). However, by observing the current industry market, this trend has started to impact those middle to high-skilled jobs. And the IT industry is one of them. Different from traditional sectors, the IT industry had been established and grown in recent decades. Moreover, IT firms are an essential part of the US economy. In 2021, it generated 1.8 trillion US dollars in revenue and contributed 9.3% of the US GDP (Sava, 2022). Having a high attrition rate leads to high extra costs for the firms, including recruitment costs, costs related to training and development of new employees, productivity lost, and morale loss (Khera et al, 2019). A case study developed a model to mitigate the monetary and time-related loss of employee turnover (Sikaroudi et al., 2015).

Data science technology has been involved in trying to mitigate this growing issue. After the yearly development of machine learning and deep learning, the idea of using models to help to make decisions started to be part of the company's goals. This research used the IBM Watson Human Resources Attrition data to identify the important features affecting employee attrition, as well as compare the models to find the best way to identify problems. It is an employee attrition analysis, and I conducted a case study focused on IBM. Some literature review papers have been written, which summarized the previous research in a detailed way. It is a publicly available dataset, so various machine learning algorithms had been applied previously for this dataset before: 1. Decision tree method; 2. A random

forest method 3. Gradient boosting methods; 4. Logistic regression methods: 5. Support vector machine (SVM); 6. Neural networks; 7. Linear discriminant analysis; 8. A naive Bayes method; 9. K-nearest neighbor method (Zhao et al, 2019). In Zhao's paper, the author contrasted the pros and cons of those models in a broad way. The paper mentioned the model performances, in terms of the accuracy rate. By making the comparison between the size of the data and how that mattered to the attrition predictions, the author got the conclusion that for a small HR dataset, trying different algorithms is a good strategy to find the optimal model. However, for a larger dataset, extreme gradient boosting was suggested (Zhao et al, 2019). It is a meaningful consideration for people when they try to pick the proper models for their own dataset. In Mansor et al's paper, they provided a table that concluded all previous research using this IBM HR dataset until 2021. The author included the paper title, data, and results. In the third column of their table, they pointed out the recommended ML method for each paper (Mansor et al, 2021). Another literature review paper summarized the results from six different papers each of which uses different algorithms. The author made the table about accuracy rate, precision, and recalls. A table visualization is a good approach to capture all the results altogether.

Adapting Mansor's methodology to my paper, I created Figure 1, which classified all the literature into three categories based on the result formats: classification, prediction possibility, and feature selection (Zhao, 2021). Moreover, I included the latest research literature as well as the previous lost ones. The formats of the results of each are different which leads to a different final goal for the analysis. Classification is a good method for this IMB HR attrition dataset. The goal of the classification method will return class: the result is binary 0 or 1 or in one of several categories (Bruce, 2020). The drawbacks of some of those methods are the reasons that classification usually can not be observed. Another paper utilized K-means Clustering (KNN) to classify people into two clusters: prone to leave and less likely to quit. The results showed that people who worked in 3-4 companies are less likely to quit. On the other hand, employees who have more than four companies working experience indicate that they are unstable and often change jobs (Yang et al, 2021). Binder

Figure 1: Literature Review paper for IBM Attrition Data

Category	Objective of Study	Techniques studied	Recommendation of Techniques or results
Classification	To predict whether an employee is likely to quit. (Mansor et al, 2021)	DT, SVM, ANN	SVM
	To classified people into likely to quit, and less likely to quit. (Yang et al, 2021)	K-means Clustering	K-means Clustering
	To predict whether a employee will leave their current company (Mohbey, 2020)	Naïve Bayes, SVM, Decision Tree, random forest, logistics regression	Decision Tree
	To predict whether an employee will leave in immediate future (Tharani et al), 2020.	Logistic Regression and XG boost	XG boost
	To present a model to predict employee attrition (Frye et al, 2018.),	Logistic Regression, KNN, Random Forest	Logistic Regression
Prediction	To find out how business travel influence attrition. (Yang et al, 2021)	Binary Logistic Regression	Binary Logistic Regression
	Size of dataset affect the method choice (Zhao et al., 2019)	Ten ML methods	Gradient Boosting trees for large dataset
	Prediction of employee attrition (Poornappriya, 2021)	SVR, ANN, Neutral Network based Regressor	NNR is the best methods
	Employee attrition prediction using data mining techniques (Sukhadiya et al, 2018)	Random Forest, Support vector machine, Gradient Boosted Classifier, and LR.	Extreme Gradient Boosting provides the best results.
Feature Selection	To find out the top seven factors for IBM attrition dataset. (Khera et al, 2019)	NumPy and Matplotlib	Age, Gender, Marital status, etc.
	To find out top three factors for IBM Attrition dataset. (Yang et al, 2021)	Random Forest	Monthly Income, age, and the number of companies worked before.
	To find out the top five factors for IBM attrition dataset (Bindra et al, 2019).	C5.0 Decision Tree	Gender, Distance from Home, Environment Satisfaction, work-life balance, education.
	To find out the charactertics of employee turnover (Zhang et al., 2018)	GBDT algorithm and LR methods	Travel frequency, sales department, males, single is the strong indictors.

et al. used a decision tree algorithm and C 5.0 methods to make the binary prediction. Out of 1151 instances, 90 percent of the results were correctly predicted. The prediction time is 0.02 ms, which is much more quicker than the traditional C5.0 model. On the other hand, the RAM consumption is less compared with traditional methods (Bindra et al, 2019). This paper took the insights from this method about model improvement and designed a new combined KNN and logistic regression model that showed in the methods and results section later.

Prediction Possibility is to find out the possibilities that a person will quit this job, the results are shown in a numerical format. Logistic regression is one of the techniques. When the probability is close to 1, it means people are likely to leave and vice versa. The best prediction had been defined by the highest accuracy rate with the shortest time needed to run the machine learning model (Mansor, 2021). Sexton et al. developed a neural network-based predictive model, also called NNSOA. This model modified the algorithm that applied the search techniques and let us train the model (Sexton et al., 2005). Heiat (2016) created two ML models based on IBM HR attrition data to determine the employee turnover rate: an ANN and DT (Decision Tree). ANN model had a higher accuracy rate which was about 85.33%. However, for the decision tree method, the rate is only 80.89 percent per time (Heiat, 2016).

Feature selection is dedicated to finding out some top-rated variables among many attributes that influence the dependent variable. For the HR data, some features come from people's profiles (age, education, work experience, etc), and others are from work-related information (compensation, travel, distance from workplace to home, etc). All the attributes had different importance, so it is necessary to gain a better understanding of those factors. In a past paper that used the GBDT algorithm and logistic regression, they concluded that frequent travel will lead to a high turnover rate. HR and technical people are more likely to change jobs. For marital status, a single group is prone to change jobs compared with a married group. Overtime is a risk factor that pushes people to leave their current work-

places (Usha. P.M et al, 2020). Previous researchers did a lot of work to find out their choices using different algorithms and the results are not exactly the same. Starting with another service sector, banking, Y. Zhao used bank attrition data and got the results that the last pay raise, job tenure, and age are the three factors for banking employee attrition using XGB methods (Zhao et al, 2019). Khera and Divya identified three irrelevant factors for employee attrition in the Indian IT industry: business travel, gender, and the number of companies worked before (Khera et al, 2019). At the same time, Khera et al believed age, gender, marital status, job level, job profile, job role, traveling, and attrition are the top seven features that were important to this dataset. We used the some of the packages mentioned in this paper. In this research, it used real HR data from several top-rated global IT firms. A research used a random forest model to select the top three important factors for this dataset: monthly income, age, and the number of companies worked, with 0.84561 average accuracies (Yang et al. 2021). Another research using IBM HR attrition data said gender, education, environment satisfaction, distance at home and work-life balance are the top 5 factors instead (Bindra et al, 2019). The number of companies that worked before may be a special variable that played a different role using different methods that we can dig into further in this research.

In conclusion, HR attrition data has been defined as social science data that personal decisions had been highly involved. The decision to quit the job may not be predictable all the time: some unexpected reasons include family problems, disease, unexpected emergence, etc. Among all the models and algorithms mentioned above in various literature using this dataset, accuracy rates are around 85 percent. The precision is around 30 percent, while the recalls are from 20-35 percent (Usha.P.M et al., 2020). And the author also made tables about computation time and memory utilization comparison. We admitted that the background noise of the data is inevitable. That may be one of the reasons why the results are varied. Thus, in this paper, we tried to identify the most important factor and reduce the noise after applying different models.

Table 1: Variables in IBM Attrition Data

Category	Variables	
Job Related Reasons	• *Department*	• *Job Level*
	• *Environment Satisfaction*	• *Hourly Rate*
	• *Monthly Rate*	• *Stock Option Level*
	• *Job Involvements*	• *Job Satisfaction*
	• *Relationship Satisfaction*	• *Job Role*
	• *Performance Rating*	• *Years In company*
	• *Years Since Promotion*	
	• *Years In Current Role*	
	• *Years With Current Mgr*	
Personal Reasons	• *Age*	• *Gender*
	• *Total Working Years*	• *Education*
	• *Training Time*	• *Martial Status*
	• *Distance From Home*	• *Business Travel*
	• *Work Life Balance*	
	• *Companies Worked*	
Response Variables	Attrition Status	

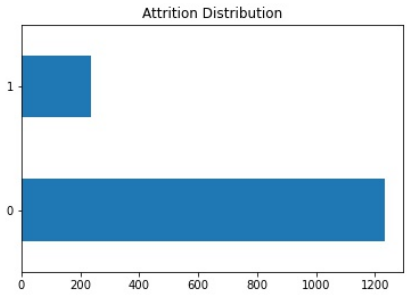
Methods

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 to leave the firm 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

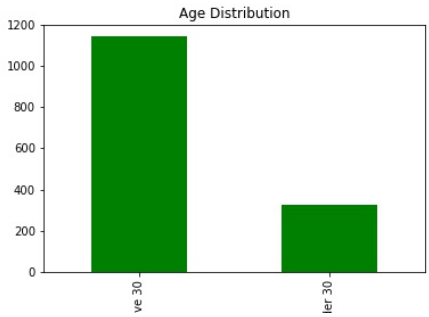
some independent variables. The attrition status dataset is a binary list that shows whether this employee is still actively working for IBM or not. From plot(a) Attrition distribution, we noticed it is an imbalance of data. 87.87% of observations in the dataset show a 0 means they are still active in the firms. Only 16.13% of people are inactive, with an 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 based on their specific needs. 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 high-dimensional data, we want to remove the irrelevant attributes and nonsense data first. After scanning 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 is 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. We will use this for the statistical result analysis. 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. Job descriptions refer to people's departments, job levels, job roles, 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). Many public technology firms like IBM, also provide company stocks to their employees as a part of the benefit. This stock options 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 reflects 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, are 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 more likely attrition. Family is very important for some people, distance from home, business travel frequency, and marital status may affect people's job stability. In this case, we explore deeper whether people in different age groups react differently to their family concerns. We split the data based on the employee's age into two groups: below 30 and above 30. We explore further in our statistical analysis part in results section. That we want to know how age played an role in 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 regression models, we need to ensure that all the assumptions had been satisified. We conducted the correlation matrix for our variables. Job level and monthly income are highly correlated with a correlation coefficient of 0.95. Also, the job level variable and total working years variable have a correlation coefficient of 0.78. We want to make sure our model follows all assumptions and that 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 table (Appendix 1), 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.

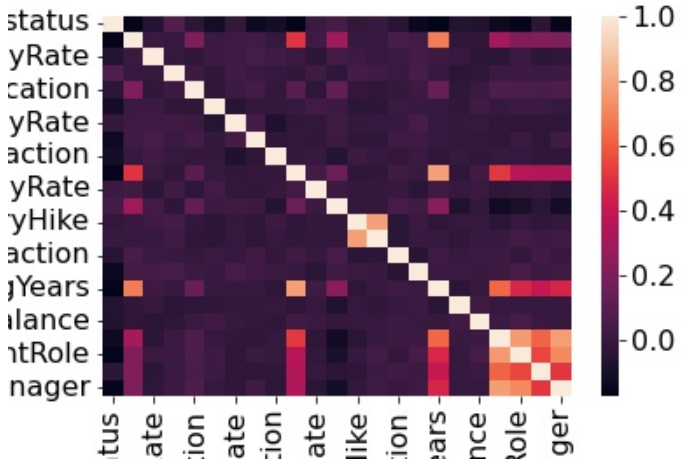


Figure 2: Correlation Matrix for Selected Variables (Yang et al. 2021)

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 (McKinney et al, 2010) package to load the dataset and perform machine learning models. We used Matplotlib (Hunter, 2007) library for data visualization. Numerical computation had been conducted by the NumPy package (Van Der Walt et al. ,2011). Stata had been used rarely for the data exploration part. All the codes and supplementary material can be found in the Ziyue(Hannah) Zhang's Github repository. (https://github.com/HannahZZY/Senior_Project_Hannah)

Model Selection and Assumptions

After understanding the variables in the data and classifying them by contexts, we want to explore deeper 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, according to Figure 1, literature review table. At the same time, I made an approach that is a combination of logistic regression and

K-Nearest Neighbors (KNN). And we want to compare the pros and cons of both approaches and analyze which one will perform the best.

Logistic regression and K-Nearest Neighbors (KNN) are the main algorithms that had been used in this research. Logit is for the linear variables and KNN is for non-linear variables. Logistic regression is a statistical method that 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 the 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, the KNN method will give us a concrete predicted labels: 0 for working or 1 for termination. Also, there is no limitation to the data itself. The way we did this 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 that appeared in each bin. If the relationship is linear, we should observe a clear linear trend in our plot and vice versa. For 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 consideration.

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). During the data cleaning process, we change the business travel to the numerical data ranging from 0 to 2. 2 means the most frequent travel, while 0 means no travel. Also, we changed the categorical variables department into dummy variables, which is the accepted numerical format for SVM. In this study, we applied SVM to all 30 variables, which transfer to a hyperplane with 29 dimensions using linear algebra. The next step is to report the precision, recall, F-1 score, and computation time generated by the SVM model

to measure the model performance.

We used the Scikit-learn package in python to build all models. The first step of making machine learning models is to split data into the training set and the testing set. After considering the size of the dataset, we decided on a 7:3 distribution with 70% of the data used for training and 30% of the data used as testing data. Because of the randomness of machine learning in general, we specified the random state so all the results can be duplicated easily to get the same results. 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. We tested the different times to use a build-in function in python and record the time needed to perform the model. 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 et 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 the training part. That means we do not have overfitting /underfitting issues in our model. After all the models had been completed. We summarized the results from each model and compared the results.

From Figure (a) Attrition Distribution, we know it is an unbalanced dataset, which means the number of 0s and 1s are not similar to each other. We can not simply report and compare the accuracy rate, as it will cause inaccurate predictions and biased results. In order to make the comparison, in this case, we want to calculate the precision, recall, and F-1 score for the model, in order to measure the performance of the models. Both of them are ways to compute the difference between the predicted results and the actual results. Precision and recall both a percentage number smaller than 1, if the number is close to 1 means the model is a good classifier. F1-score is a matrix that takes into account both precision and recall and reports the F1-score as part of the result (Harikrishnan, 2019). A smaller difference between the two classes' F-1 scores means the overall performance of the model is good.

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

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

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

As mentioned in our introduction section, 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. The best model can be utilized by firms in the future. On the other side, we want to conclude the important features of each category as our key results. In the following sections, we perform all the models and present the results.

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, five different education levels have been indicated clearly in the model: 1 means under college, 2 stands for some college, 3 is for a bachelor's degree, 4 is for a 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

rate fall to 12%. We suspect that age may be a factor that influences employee attrition in this dataset.

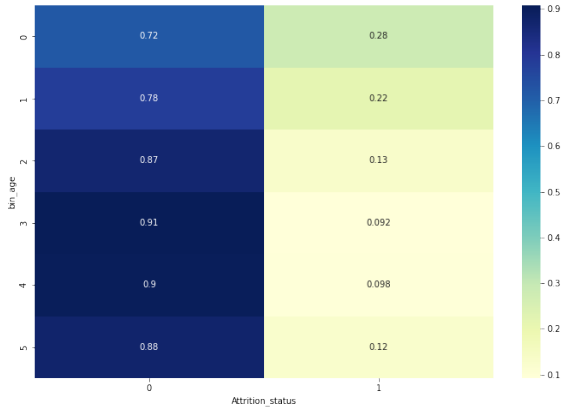


Figure 3: Age-Attrition Heatmap

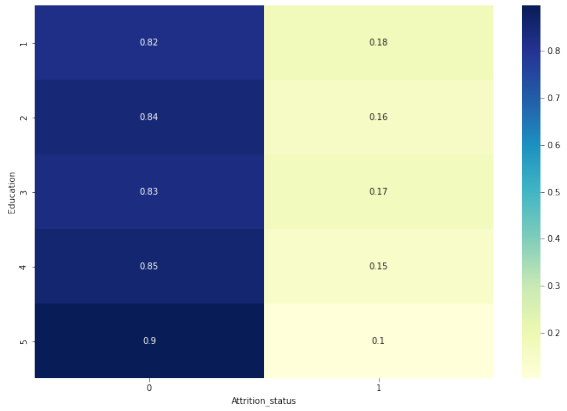


Figure 4: Education-Attrition Heatmap

We proposed the hypothesis that age varied the attrition: below 30 and greater than 30 years old. They may have a 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 statistic

of -10.1841 with a p-value much smaller than 0.05. In terms of the 95% of the confidence interval, the lower bound is -0.757, and the upper bound is -0.509. We can conclude that it is meaningful to explore deeper. The degree of freedom of the two sample t-tests is 514.14. 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 to 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 the 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 function in pandas and then made bar plots to observe the frequency of 1 appearing 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 (M. Lavin, Personal Communication, 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 datasets based on the linear

Figure 5: 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 using the KNN method. Therefore, we can conclude the results of each and get the conclusion. Then, we apply SVM to the whole dataset that contains both linear and non-linear variables.

Table 2: Linear and Non-linear Variables Result

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

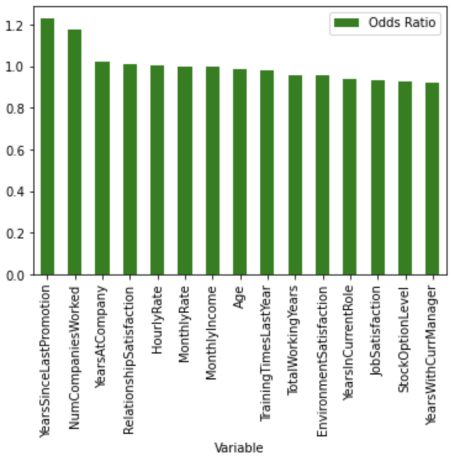
Logistic regression is a good approach to finding out the probability of an event happening in our dependent variable in binary. In this case, predict dependent variables from the logit should be the likelihood of whether a person with those parameters will leave the current company or not. The various parameters for each independent variable should reflect the extent of influence on 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 30% of the data for training purposes and 70% of the data for testing purposes. For the logistic model, we set parameter class weights as balanced, which is commonly used in the imbalance data by changing the possibility of 0 and 1 as the initial number assumption. Random state parameter equals 11 in order for the replication purpose. We built the model

using training data and then applied the model to the testing data. The model running time was 331 ms. For one more variable added, the computation time increases by 22 ms. The coefficient for each variable and its odds ratio tables are shown in Table 3. From the table, Years since promotion, the number of companies worked, and years at the company are the top three most significant variables in this model that affect employee attrition.

Figure 6: Logistic Regression Results



In order to make the results more comparable, we made a bar plot using the odds ratio for each variable (Figure 6). The accuracy rate we got for the linear variables is 74.83%. Then, we generated the confusion matrix for this model (See Figure 7). We set the x-axis as the predicted values and the true value as the y-axis. In terms of achieving the best result, we want a very high rate of true positive and true negative, which indicates the correct predictions. However, even though false positive means incorrect prediction, in this case, it is also acceptable because their results are good, which is to stay in the firm. And we want to reduce the percentage of false negatives, which means people left the company that we did not expect. There are 302 correct predictions for people who did not leave and the model gave the same result, which is the correct prediction. There were 28 correct predictions for the leaved class. The false positive value is 71, which means people did not leave the company but the model predicted he/she left. We alter the parameter about weight in the model set and find the balance between classes. To make our FP to be smaller, we have to sacrifice some of the FN results. And the false positive is very low here: 40 wrong case predictions for people who left the company but not be correctly identified by this logistic model.

As it is an imbalanced dataset, we also computed the precision, recall, and F-1 score as

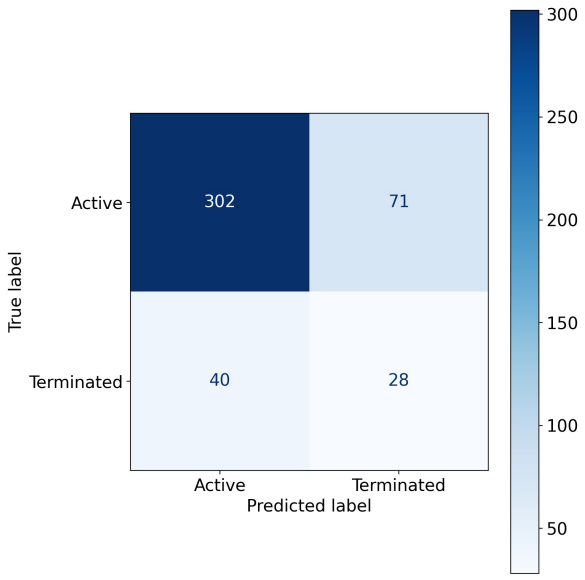


Figure 7: Logistic Regression Confusion Matrix for Linear variables

extra measurements of the model. Precision and Recall are use different formula to show the quality of the model. We analyzed the precision and recall from two classes' perspectives: active and terminated. The precision for the active employee class is 88.30%, and for the termination, class is 28.28%. The recall showed more similar scores between the two classes: 80.96% for the first class and 41.17% for the second class. Then, we used these two to compute the F-1 scores for the model, which are 84.47% and 33.53% respectively. The difference between the two classes F-1 scores is 50.94%.

0.4 K-Nearest Neighbors (KNN)

For the next step, we built a KNN model for the non-linear datasets. KNN captures the closest observations and makes a prediction about the data point that we want it to predict. For KNN, there are not a lot of assumptions needed to fit, but only all data should be in numerical format. We made the dummy variables for the education field variable. Also, we changed the travel frequency from strings to the numerical format: 0 means no travel, 1 means rarely travel, and 2 means travel frequently. The drawback of KNN is that no

Table 3: Logistic Regression Coefficient Table

Variable	Coefficient	Odds Ratio
YearsSinceLastPromotion	0.206713	1.229630
NumCompaniesWorked	0.160702	1.174335
YearsAtCompany	0.020929	1.021150
RelationshipSatisfaction	0.010660	1.010717
HourlyRate	0.005337	1.005351
MonthlyRate	0.000029	1.000029
MonthlyIncome	-0.000068	0.999932
Age	-0.015104	0.985010
TrainingTimesLastYear	-0.022325	0.977922
TotalWorkingYears	-0.044691	0.956293
EnvironmentSatisfaction	-0.045862	0.955173
YearsInCurrentRole	-0.060386	0.941401
JobSatisfaction	-0.067663	0.934575
StockOptionLevel	-0.074973	0.927768
YearsWithCurrManager	-0.084213	0.919235

coefficients had been produced by the model, but the labels only. And the training time is longer than the logistic model. The choice of K needs to fit with the size of the dataset and the number of variables. A too-large K easily causes the overfitting problem which means the model may predict the training data very well, but not be useful for the testing data. Also, a large K will lead to a long computing time and make the model less useful in large datasets. In this case, the CPU running time for KNN is 585 ms. For one more variable added, there is an average increase of 31 ms. KNN took longer compared to the logistic model we did above. For this KNN model, we choose K equals 5 and lets weights be the distance. We decided to use the 5 closest data points to predict our unknown label. Then, we created the confusion matrix for this KNN model. Using a similar interpretation format for the logistic regression, we have a higher overall correct prediction rate, with a significant fall in false positive results. It means this KNN model works better to predict people who did not leave the company. According to the confusion matrix, the number of true positives is very high, which is 356 correct predictions. It can predict a lot of correct cases of actively employed people. However, there are only 10 true negative cases, which means the correct prediction for people who leave the firm. In other words, if we want to predict the people who left the firm, the predictability is not as good as the previous logistic regression. In the

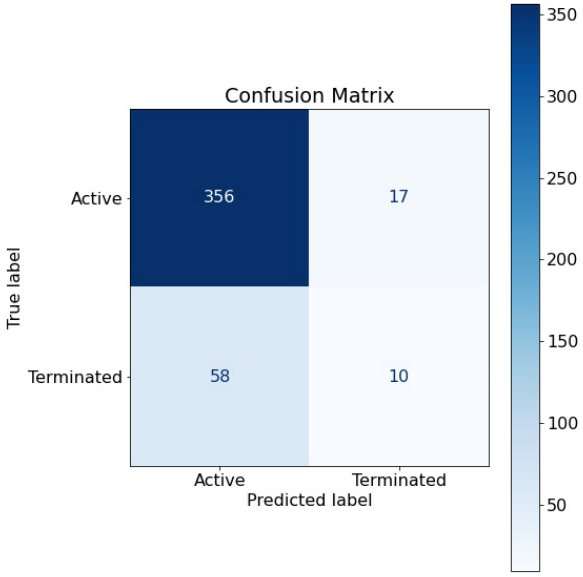


Figure 8: KNN Confusion Matrix for Non-Linear variables

KNN model, we did not know which variables contributed to employee attrition to most as we did not have coefficients. But we are still able to analyze the performance by comparing the model scores. We reported the model performance scores for both classes. The accuracy rate is 88.99% for this KNN model. In terms of the first class, the precision for the model is 85.99% and the recall is 95.46%. The F-1 score is 90.47%. However, for the second class, values are around 40% and 20% respectively. Table 5 in the discussion section provides all concrete numbers with explanations.

0.5 Support Vector Machine (SVM)

SVM is a supervised algorithm that works well for small datasets, compared with traditional regressors like ANN (Artificial neural network) (Poornappriya et al). SVM works with many hyperplanes that have been created by independent variable classes. It will find the optimal separation plane using the methods we specified in the kernel parameter, based on the maximum margin, means the distance measured by the data points from each class (Gandhi, 2018). In this dataset, we want to find the linear combination of features for both

linear and nonlinear variables. The training and predicting time for this model is much longer than the previous model, which took about 1 min 26s. In other words, it may lead to a high time cost if we want to work with a large dataset. We use the coefficient function to find out the coefficients created by this SVM. The coefficients of SVM shown as an array are the coordinates of a vector that is orthogonal to the found hyperplane. The input data for this model contains 30 columns, then the hyperplane is also 30 dimensions.

Table 4: SVM Coefficients Table

Variables_0	Coefficient_0	Variables_1	Coefficient_1
JobInvolvement	-25.073940	NumCompaniesWorked	11.597471
StockOptionLevel	-20.212368	YearsSinceLastPromotion	12.786225
EnvironmentSatisfaction	-15.349191	BusinessTravel	27.087020

- Notes: 0 represent the active employee class, 1 means termination class

Then we can take the dot products to the point of the vector. If that is a positive number, it belongs to category 1 which means people who left the company. In contrast, if the dot product is negative, that means the predicted result is 0, which indicates that the employee will still work for the company. From the Table 4, we can conclude that the number of companies that have worked before, business travel, Years since promotion are the three top choices that contribute to employee attrition. On the other hand, people with a high job involvement, a high stock option level, and a high environmental satisfaction are more likely to have a low turnover rate, which is to said they are more likely to stay in the firm. The confusion matrix shown in the Figure 9. There are 381 correct predictions in total. The model had a false positive 55 and a True negative 13.

It means the model predictability for the people who left the company class is not solid. The overall accuracy of the model is 86.39%. The average precision of this model is 79.51% and the average recall is 58.88%. Then, we calculated the average F-1 score for this model is 61.34%, with the classes difference 62.23%, which is a smaller difference than both logistic regression and KNN.

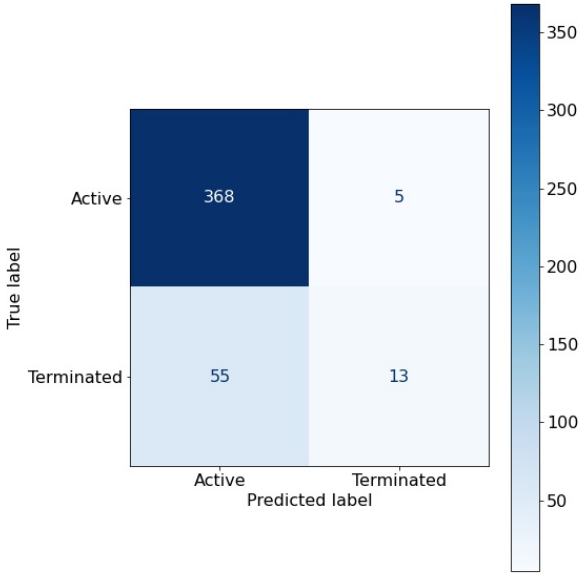


Figure 9: SVM Confusion Matrix for All variables

Discussion

This IBM Human Resource Attrition dataset is a publicly available dataset that has been used and investigated for years. A lot of angles and directions had been explored in the past. There are two main reasons that this dataset has been populated and still used for research today. Because of the latest algorithm developments, people are passionate about using the most updated methods and want to develop a better way to interpret it. Another reason is that for machine learning it has uncertainty and randomness in its operation process. Because of the inconsistency, people tried different parameters and model settings and found the best solutions.

The feature selection results is one of the central questions that we want to analyze in this paper. From the literature review section, we found that researchers did get various conclusions regarding the most important factors. The different choice of algorithms and the parameters will be one of the possibilities of what happened. Another reason may be because people have different ways of cleaning and preprocessing the data. For example,

some people change categorical variables as the dummy variables, while others use different ways to handle this non-numerical data problem.

In my linear data set, the logistic regression coefficient found that years since promotion, number of companies worked, and years at company are the top factors that contribute to employee attrition. It is inconsistent with Bindra's model that the authors concluded that gender, education, and environmental satisfaction contributed the most (Bindra, 2017). However, KNN can not give us the results because of its limitation in parameters. It can only provide the label prediction instead. SVM, on the other hand, came to the conclusion that the number of employees at the company worked, business travel, and years since promotion are the most important factors that push the employee to leave the current company. It is very reasonable that if a person had a lot of experience changing jobs, the possibility that the people would leave this job is absolutely higher. It has been a habit for them to live in a changing environment. In a company like IBM, the average age for the employees is lower than the national average (Cascade, 2017). People tend to seek promotion opportunities. If the employee remains in a position longer than they expected, the model showed that they are very likely to change a job. It may be a good way to get promotion in a different organization within the same business line. From the feature selection perspective, if we prefer to get the predicted results rather than understanding the machism, this logistic and KNN combined method is a good one to use. The results from here are very straightforward and the computation is easy. However, if we want to understand how each variable contributed to the overall results. SVM models can provide more comprehensive and detailed instructions. However, the predictability of the SVM model depends on the data quality as well as it may take much longer time compared to the Logistic and KNN methods.

Among three models that we built in this paper, we want to compare the precision and recall for both classes: stay in the firm (0), and leave the firm (1). Terminated group is the class that we want to highly emphasize in this research, because that is how the high turnover problem occurred. And we are seeking for a low difference of F-1 score among two

classes, which means it gave a good prediction overall. Logistic regression showed the best performance for both classes. For KNN, the prediction results for the stay in the firm group are really robust.

Table 5: Summary Table

Models	Accuracy	Precision_0	Precision_1	Recall_0	Recall_1	F-1 Score_0	$\text{F}_1 Score_1$
Logit	0.7483	0.883	0.2828	0.9096	0.4118	0.8448	0.3353
KNN	0.8299	0.8599	0.3703	0.9544	0.147	0.9047	0.2105
SVM	0.8639	0.868	0.7222	0.9865	0.1912	0.9246	0.3023

Notes: 0 represent the active employee class, 1 means terminated class

However, for the other group, the recall value is relatively low, compared with the average across the other literatures. SVM showed a similar trend as KNN: a strong precision, recall, and F-1 score for the first class, but not very solid result for the second class. We suspect the reason is because of the limited data that we used to train. This IBM HR attrition dataset only contains about 1400 rows, which is relatively weak for a deep learning algorithm. Another consideration is about the properties of imbalance. The imbalance rate for this data is about 1:6. That will significantly influence the predictions for the attrition class. The choice of parameter always depends on the researcher's insights. There are not a right or wrong answers. It is also meaningful to attempt various parameters in order to get a better performance using the limited data. In the future analysis, we suggested trying a large size of data with a greater balance in terms of the predictive binary variable. Therefore, we may create a model with a better performance and give the most accurate results.

The model selection depends on the properties of the dataset they plan to use: balance or imbalance binary data, the size of the dataset, and the expected computation capacities. Also, the type of targeted answer is also another consideration that alters the choice of algorism. The model results can be varied from many coefficients representing the probabilities of features to just simple predicted labels results. We need to take into account all those factors before we start to perform the best model.

Conclusion

Human capital is one of the essential resources to enhance the economy. This research will boost people's understanding about what factors that employee cares about and what approaches a company can do to assist their employees in a better way. According to our model results, we found that the conclusions may differ based on the model setting and the input dataset. With some similarities to the results that other scholars did, we found that the SVM method can be very strong in some situations as it provides an elegant accuracy, precision, and recall scores in general. However, the results from logistic regression were solid too. It can give a lot of information about how each variable contributes to the overall results. On the other hand, KNN puts the attrition result labels in a very straightforward way that people with a little statistical knowledge also can interpret and understand. This paper provides a lot of possible directions that people who used the results can dig further. We are looking for a model with increasingly better predictions in the future.

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

Bindra, H., Sehgal, K., Jain, R. (2019). Optimisation of C5.0 Using Association Rules and Prediction of Employee Attrition. International Conference on Innovative Computing and Communications. Lecture Notes in Networks and Systems, vol 56. Springer, Singapore. https://doi.org/10.1007/978-981-13-2354-6_3

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

Harikrishnan, N.B (2019, Dec 10). 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.

Campbell, S. (2019, October 30). Ageism in tech: The silent career killer. Cascade Insights. Retrieved November 14, 2022, from https://www.cascadeinsights.com/ageism-in-tech-the-silent-career-killer/

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.

Cox, D. R. (1958). The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological), 20(2), 215–232.

Gandhi, R. (2018, July 5). Support Vector Machine - introduction to machine learning algorithms. Medium. 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/

Harris, C.R., Millman, K.J., van der Walt, S.J. et al. Array programming with NumPy. Nature 585, 357–362 (2020). DOI: 10.1038/s41586-020-2649-2.

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.

J. D. Hunter, "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.

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

Lavin. M (2022). personal communication.

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

Mucherino, A., Papajorgji, P.J., Pardalos, P.M. (2009). k-Nearest Neighbor Classification. In: Data Mining in Agriculture. Springer Optimization and Its Applications, vol 34. Springer, New York, NY. https://doi.org/10.1007/978-0-387-88615-2_4

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/

Sexton, Randall & McMurtrey, Shannon & Michalopoulos, Joanna & Smith, Angela. (2005). Employee turnover: A neural network solution. Computers & Operations Research. 32. 2635-2651. 10.1016/j.cor.2004.06.022.

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/

Sukhadiya, J. (2018). Employee Attrition Prediction using Data Mining Techniques. International Journal of Management, Technology And Engineering, 8(X), 2882–2888.

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

Usha, P. M., & Balaji, N. V. (2020). An Analysis of the Use of Machine Learning for Employee Attrition Prediction – A Literature Review. Journal of Information and Computational Science, 10(3), 1429–1438. https://doi.org/10.12733.JICS.2020.V10I3.535569.12053

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, Heng & Xu, Lexi & Cheng, Xinzhou & Chao, Kun & Zhao, Xueqing. (2018). Analysis and Prediction of Employee Turnover Characteristics based on Machine Learning. 371-376. 10.1109/ISCIT.2018.8587962.

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.