The Analytics and

Prediction about IMB Employee Attrition

yingrui Lyu
16/12/2020

Abstract

Under the conditions of the new economy, employees have become the most important strategic resource of modern enterprises. However, almost every company has employee attrition to varying degrees. This article uses the 1470 observations and 35 variables of IBM employee attrition provided by the Kaggle website to do visual analysis and univariate analysis between attrition and other variables. Further use GBM algorithm to model the dataset. And use weighting, up-sampling, and down-sampling to optimize the model. The analysis found that the GBM model optimized by the weighting method has a higher AUC value, which is 0.7689, and the model is more accurate. It is further found that the top five factors affecting employee turnover are OverTimeYes, Age, MonthlyIncome, JobSatisfaction, StockOptionLevel. And the probability of employees leaving the sales department is significantly higher than that of other departments

Key word: Employee Attrition Univariate Analysis GBM Weighting up-sampling down-sampling AUC Value

Introduction

Statistical analysis is everywhere in life. For example, in medical clinical trials, it analyzes the efficacy of drugs and provides better services to patients; in economics, it assists economists to analyze market trends. The correct and efficient use of statistical knowledge to discover the cause of a certain phenomenon and predict its trend will bring great convenience to life.

Employee turnover is one of the key factors that plague enterprises. This article selects job resignation data that is closely related to people's lives from the Kaggle website [1], and analyzes IBM's resignation data. On the basis of observing the influencing factors of turnover rate, establish a model and predict which employees are more likely to leave. This helps the company's HR to prepare for personnel turnover.

In this report, this dataset consists of 1470 records of observations and 35 variables of IMB Employee Attrition. Facing these data, my curiosity trend me to ask the following questions: Why do employees leave? What are the reasons for leaving? Which types of employees are easier to leave? Which model should be selected to predict the employee turnover rate in order to achieve the desired prediction accuracy?

Generally speaking, data analysis is divided into three steps: data collection and cleaning, exploratory analysis and modeling prediction. The data set in this article is simulated data used by IBM to study employee forecasts. The data is very complete and does not need to be cleaned. Therefore, this article is mainly divided into three parts: Exploratory analysis of some important variables; Use the univariate analysis [2] to analyze the factors that lead to employee turnover, and dig out the impact of related factors; Build an effective model which is GBM [3], through algorithms to predict whether employees will leave. ROC value[4][5][6] was used to verify the validity of the model.

By employees of IBM data practices, we hope to discover factors that affect employee turnover, and were reviewed using the R language for data analysis process, deepen the understanding of the

meaning of data analysis.

Due to time and article space constraints, it is impossible to further explore which employees are most willing to stay in the company. Analyzing the reasons why these employees are willing to stay is more helpful for headhunters to retain elite talents and tap corporate advantages. This will be the future research direction of this article.

Data

The dataset is consisted by 1470 records of observations and 35 variables of IMB Employee Attrition The specific field descriptions are shown in Table 1. This article uses the variable Attrition as the explanatory variable to be predicted, and the other 34 fields as potential influencing variables to analyze the factors that lead to employee attrition.

Table 1: The Variable Explanation of IBM Employee Attrition

Variable	Type and Explanation
Age	The age of employee
Attrition	Employee leaving the company, 0=no, 1=yes
BusinessTravel	1=No Travel, 2=Travel Frequently, 3=Travel Rarely
DailyRate	Salary Level
Department	1=HR, 2=R&D, 3=Sales
DistanceFromHome	the distance from work to home
Education	the education level of the employee, from 1 to 5, 5 means the highest level of education
EducationField	1 'low' 2 'medium' 3 'high' 4 'very high'
EmployeeCount	The count of employee
EmployeeNumber	employee id
EnvironmentSatisfaction	satisfaction with the environment ;1 'low' 2 'medium' 3 'high' 4 'very high'
Gender	1=Female, 2=Male
HourlyRate	hourly salary
JobInvolvement	1 'low' 2 'medium' 3 'high' 4 'very high'
JobLevel	level of job
JobRole	1=HR Rep, 2=HR, 3=Lab technician, 4=Manager, 5= Managing director, 6= Research director, 7= Research scientist, 8=Sales executive, 9= Sales representative
JobSatisfaction	satisfaction with the job ;1 ='Low' ,2 ='Medium', 3 ='High' ,4 ='Very High'
MaritalStatus	1=divorced, 2=married, 3=single
MonthlyIncome	monthly salary
MonthlyRate	monthly rate
NumCompaniesWorked	Number of companies where employees have worked
Over18	whether age over 18 years old;1=YES, 2=NO
OverTime	1=NO, 2=YES
PercentSalaryHike	Percentage increase in salary
PerformanceRating	Performance rating
RelationshipSatisfaction	Relations satisfaction
StandardHours	Standard hours
StockOptionLevel	Stock options, Higher the number, the more stock option an employee has

TotalWorkingYears	Total years worked
TrainingTimesLastYear	Hours spent training
WorkLifeBalance	Time spent between work and outside
YearsAtCompany	Total number of years at the company
YearsInCurrentRole	Years in current role
YearsSinceLastPromotion	Last promotion
YearsWithCurrManager	Years spent with current manager

Due to the data is very complete, so it does not need to be cleaned. Based on the question raised in the introduction part, we need to make a descriptive statistic and plot the graphs about the data.

Data visualization And Analysis

Next, this part is mainly done through the visualization of univariate analysis. The data visualization will be divided into seven sections to analyze the relationship between employee attrition and 34 variables.

Section1: Table2 shows the basic information of the employee variable of attrition, employee age, and monthly income.

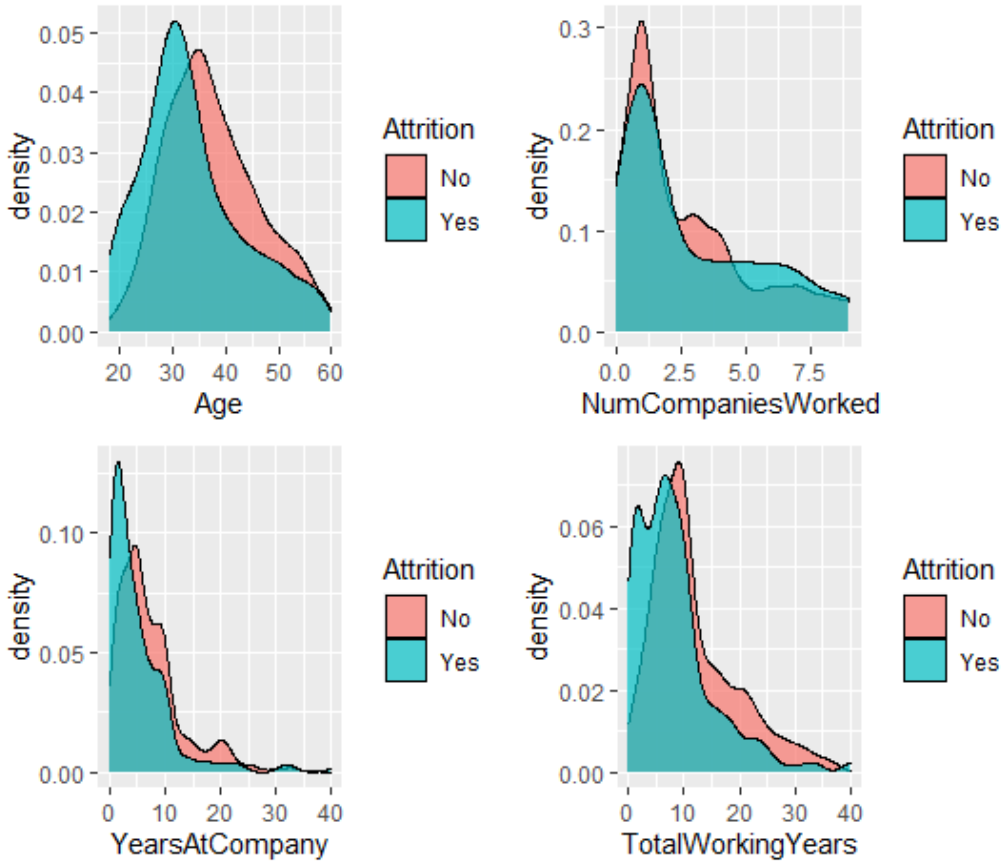
Table 2: Basic Information about Employee

	Age	**Attrition**	**Monthly-Income**
Min	18	No=1233	1009
Mean	36.93		6503
Max	60	Yes=237	19999
Median	36		4919

It can be seen from theTable2 about the dataset, the company's employee turnover rate is about 1:5, and the employee attrition rate is 16%; the average age of the company's employees is 36-37 years old and the oldest is 60 years old, the youngest is 18 years old; the monthly salary is about 4900 US dollars (because the distribution is not a normal distribution, the median here is more representative than the mean).

Section 2: Conduct a univariate analysis to explore the relationship among the IBM employee Attrition with the variable of Age, NumcompanierWorked, YearsAtComany and TotalWorkingYears (which uses the R tool to plot this graph)

Figure 1: The Distribution of Attrition with the Variable of Age, NumcompanierWorked, YearsAtComany And TotalWorkingYears

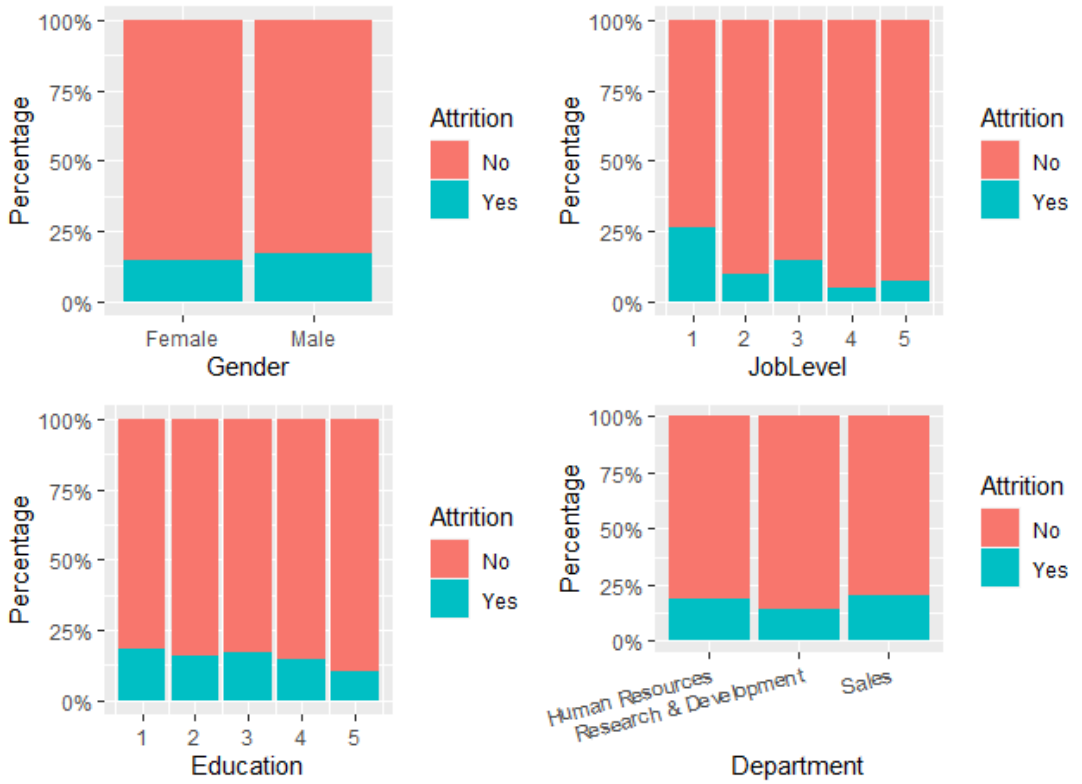


The following conclusions can be drawn from Figure 2:

Lower age employees have a higher turnover rate. Mainly concentrated in employees under the age of 30. The more companies that have worked, the easier it is to leave. The longer they work in the company, the less likely they are to leave. Employees with low working experience are more likely to leave.

Section 3: Do the analysis and observe the responses of gender, job level, education background, department with the employee attrition.

Figure 2: The Distribution of Attrition with Gender, Joblevel, Education and Department



These following conclusions can be draw from figure2:

The employee attrition rate of men is slightly higher than that of women. The higher the level, the less likely to leave. However, the main concentration of newcomers in the job level is level=1. 3. There is not much correlation between academic qualifications and attrition rate. The turnover rate of the sales department is relatively high compared to the other two departments.

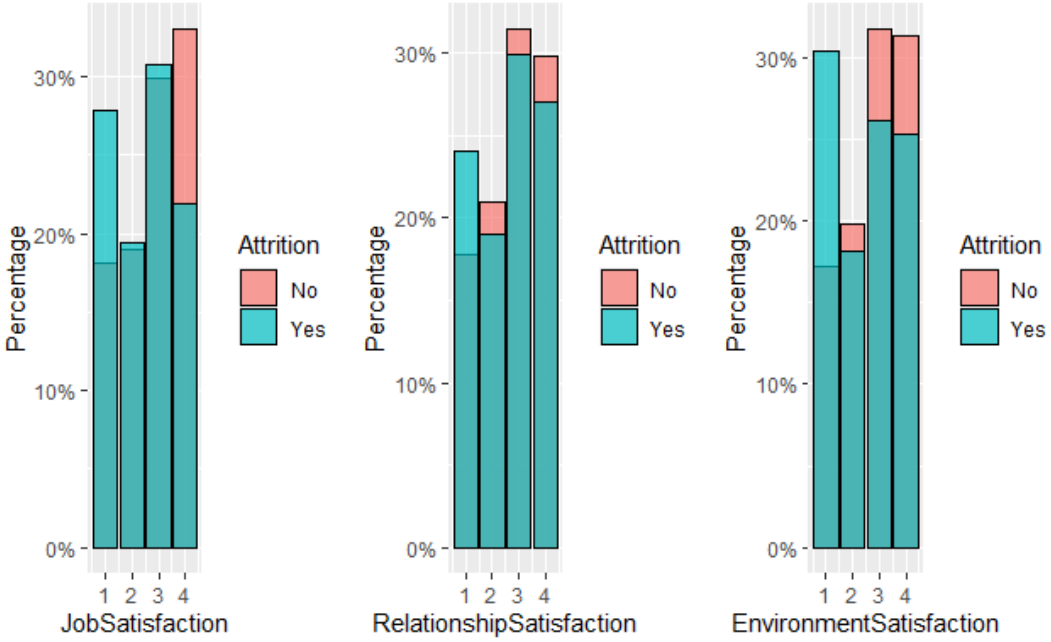
Section 4: Analyze the responses of PercentSalaryHike, TrainingTimesLastYear, YearsSinceLastPromotion, StockOptionLevel with the employee attrition.

Figure 2: The Distribution of Attrition with PercentSalaryHike, TrainingTimesLastYear, YearsSinceLastPromotion, StockOptionLevel

It can be seen from Figure 3: The attrition rate of employees without salary increase plan is higher. The number of training sessions and turnover rate did not have much impact. Employees who have not been promoted since last year have a higher turnover rate. The turnover rate of employees without stock option is higher.

Section 5: Explore the responding changes of the variable between job satisfaction, relationship satisfaction, environment satisfaction with the employee attrition.

Figure 4: The Distribution of Attrition with JobSatisfaction, RelationshipSatisfaction, EnviromentSastictifaction

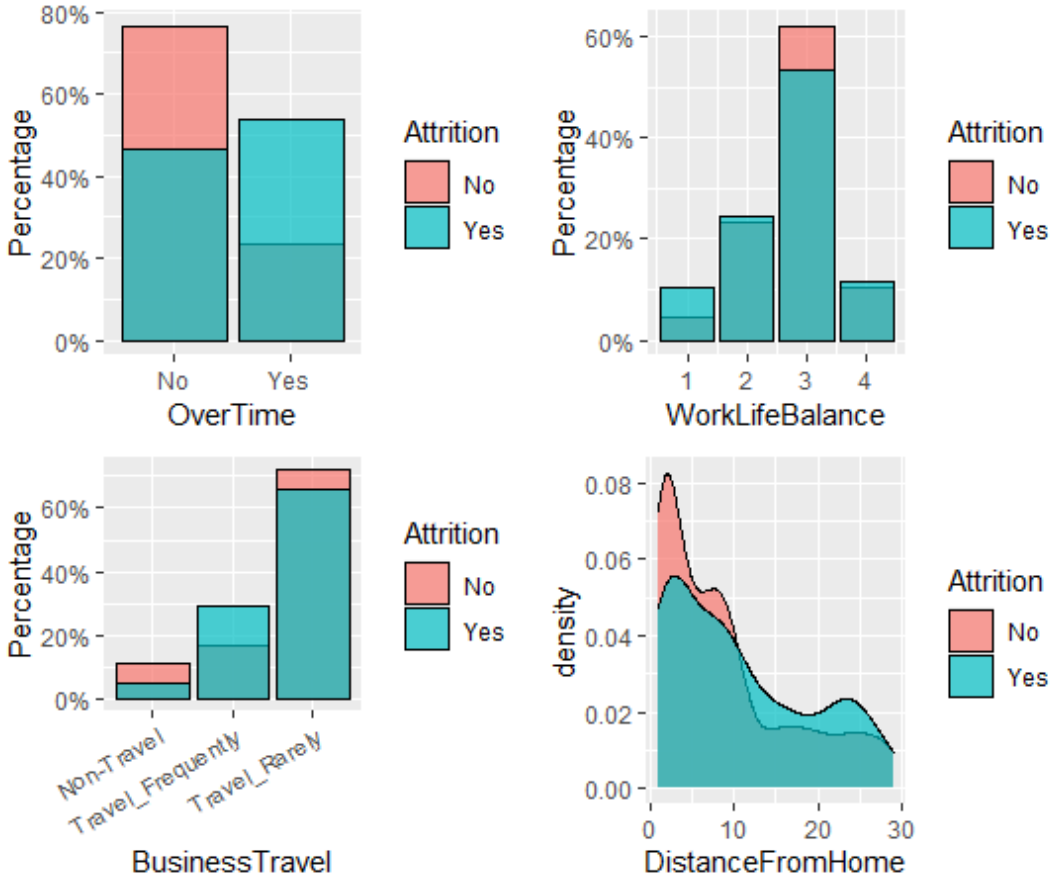


By observing the Figure 4, it can find the conclusion: The higher satisfaction on the job, relationship

and job's environment, the less likely to leave

Section 6: Find the relationship between the factors of Overtime, WorkLifeBalance, BusinessTravel and DistanceFromHome and the employee attrition respectively.

Figure 5: The Distribution of Attrition with Overtime, WorkLifeBalance, BusinessTravel and DistanceFromHome



The following conclusions can be observed from Figure 5:

The more overtime, the higher the turnover rate. Employees who believe that work-life balance level equals 1 have a higher turnover rate. Employees who travel frequently have a higher turnover rate. The farther away from work, the higher the turnover rate of employees.

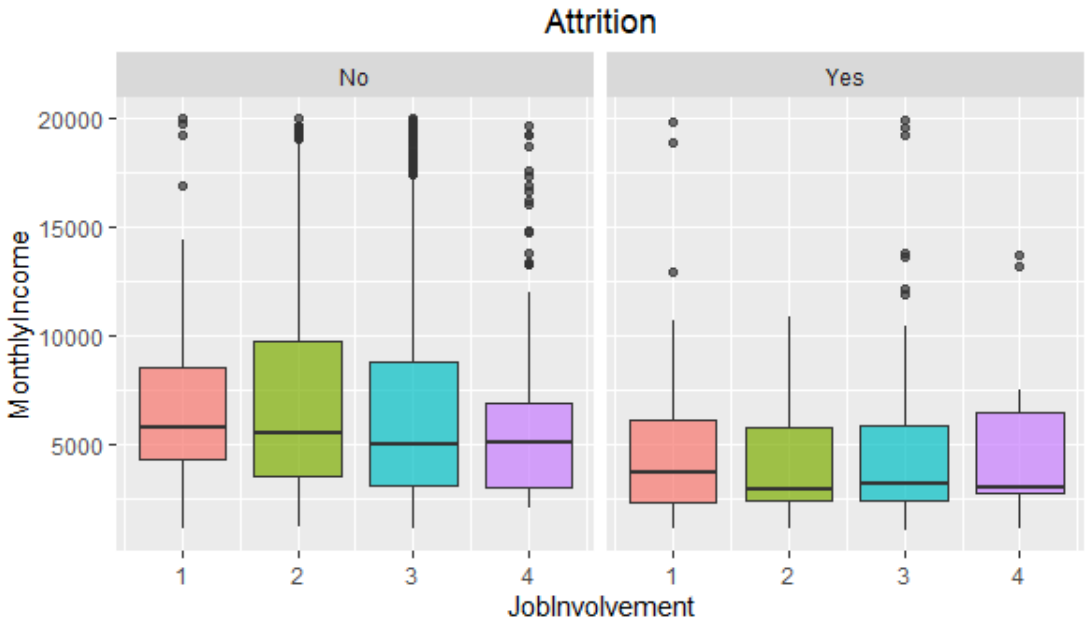
Section 7: Investigate the relationship between monthly income and job involvement and attrition.

Figure 6: The Distribution of Attrition with MonthlyIncome and JobInvolvement

It can be clearly concluded from Figure 6: Employees with low monthly salaries are easy to leave. The turnover rate of low-level positions is higher, but it is not obvious.

Further analyze the relationship between monthly income and job level

Figure 7: The Distribution Between MonthlyIncome and JobInvolvemen



From figure 7 can clearly get the results that the level of income is not the most important factor affecting employee attrition; If the pay and the return are not proportional, there will be a great turnover of employees.

Model and Results

Before modeling, some redundant variables that have little effect on the model need to be deleted to reduce the computational complexity of the model algorithm. Based on the exploratory analysis above, the variable of EmployeeCount, EmployeeNumber, Over18, StandardHours has little effect on employee attrition, and need to delete. Next, we select the GBM model to build and optimize the model, and finally predict employee attrition.

First, briefly describe the basic principles of the GBM model. The full name of GBM is gradient boosting machine, which is generally called gradient boosting tree algorithm. It is a commonly used integrated learning algorithm in machine learning. It is usually used to solve large sample data classification, regression and prediction. The learning model can be defined as a problem of minimizing the loss function. Due to the complexity of the specific algorithm principle and the space limited, the article will not explain it in detail. The specific algorithm[7] can be introduced and found in github. Secondly, the AUC standard is adopted for the evaluation of the model. AUC (Area Under Curve) is defined as the area under the ROC curve and the coordinate axis. The closer the AUC is to 1.0, the higher the authenticity of the model. Among them, the ROC curve is based on the confusion matrix [8] which is explanted in table 3.

Table 3: Confusion Matrix

		Reference	
		Positive	Negative
Prediction	Positive	TPR	FPR
	Negative	FNR	TNR

The abscissa of the ROC curve is False Positive Rate, and the ordinate is the True Positive Rate. Correspondingly, there are True Negative Rate and False Negative Rate. The calculation methods of these four types of indicators are as follows:

- False Positive Rate (FPR): the probability of being judged as a positive but not a true example, that is, the probability of being judged as a positive in a true negative example.
- True Positive Rate (TPR): The probability that it is judged to be a positive case is also a real case, that is, the probability that a real case is judged to be a positive case (also named Sensitive).
- False Negative Rate (FNR): The probability that it is judged to be a negative case but not a true negative case, that is, a real case is judged to be negative the probability of a case.
- True Negative Rate (TNR): The probability that a negative example is also a true negative example, that is, the probability of a negative example in a true negative case.

Divide the data into training set and test set, 70% as training set, and the rest as test set. Then use the original data to build the model, and the RUC value obtained at this time is 0.6359. It can be seen from the value of RUC that the model built is not ideal. This is mainly due to the imbalance of the sample. It can be seen from Table 1 that the employee attrition ratio of yes and no is 1:5, and the data distribution is seriously unbalanced, So the model needs to be further optimized. This paper selects three methods of Weighted Random Sampling[9], up-sampling, and down-sampling [10][11]to optimize the GBM model.

Weighting aims to reduce errors in minority groups. In this article, it refers to people whose employee attrition value is NO; up-sampling refers to randomly deleting instances from the majority class. Down-sampling refers to copying instances from a few classes.

Table 4: The Value of AUC

	Raw data	Weighted-Fit	Up-Fit	Down-Fit
AUC	0.6359	0.7689	0.7183	0.6821
TPR(Sensitive)	0.2985	0.7463	0.5970	0.6716

It can be concluded from Table 4 that the Weighted-Fit model has the best fit, with the AUC value rising from 0.6359 to 0.7689. The sensitivity increased from 29.85% to 74.63%. Moreover, as can be seen from the table, the optimization effect of up-sampling on the model is not particularly good. This is because the method reduces the sample size, so the fitting effect is not particularly good. The

following analysis will use the best fit Weighted-Fit model to predict.

Next, use the established model for further analysis. The model gives a series of influencing factors, if these are included in the model, it will undoubtedly increase the complexity of the model. Therefore, it is necessary to check the importance of variables and see which factors contribute to determining the outcome of employee departures in general. This is helpful for determining where human resources or management should carry out their work. Variable importance is shown in Table 5.

Table 5: GBM-Weighted-Fit variable importance

Variable Name	Overall（%）	Variable Name	Overall（%）
OverTimeYes	100	NumCompaniesWorked	27.61
Age	73.91	EnvironmentSatisfaction	26.23
MonthlyIncome	52.99	YearsWithCurrManager	26.01
JobSatisfaction	49.53	YearsAtCompany	19.65
StockOptionLevel	44.75	JobLevel	18.94
JobInvolvement	44.11	PercentSalaryHike	18.81
TotalWorkingYears	34.12	DepartmentResearch&Development	17.2
MonthlyRate	31.87	YearsSinceLastPromotion	16.08
DistanceFromHome	31.12	WorkLifeBalance	15.41
DailyRate	30.09	BusinessTravelTravel_Frequently	14.93

It can be seen from the table 5 there is only 20 most important variables shown out of 35 variables. And the five main factors that affect employee turnover are: OverTimeYes, Age, MonthlyIncome, JobSatisfaction, StockOptionLevel.

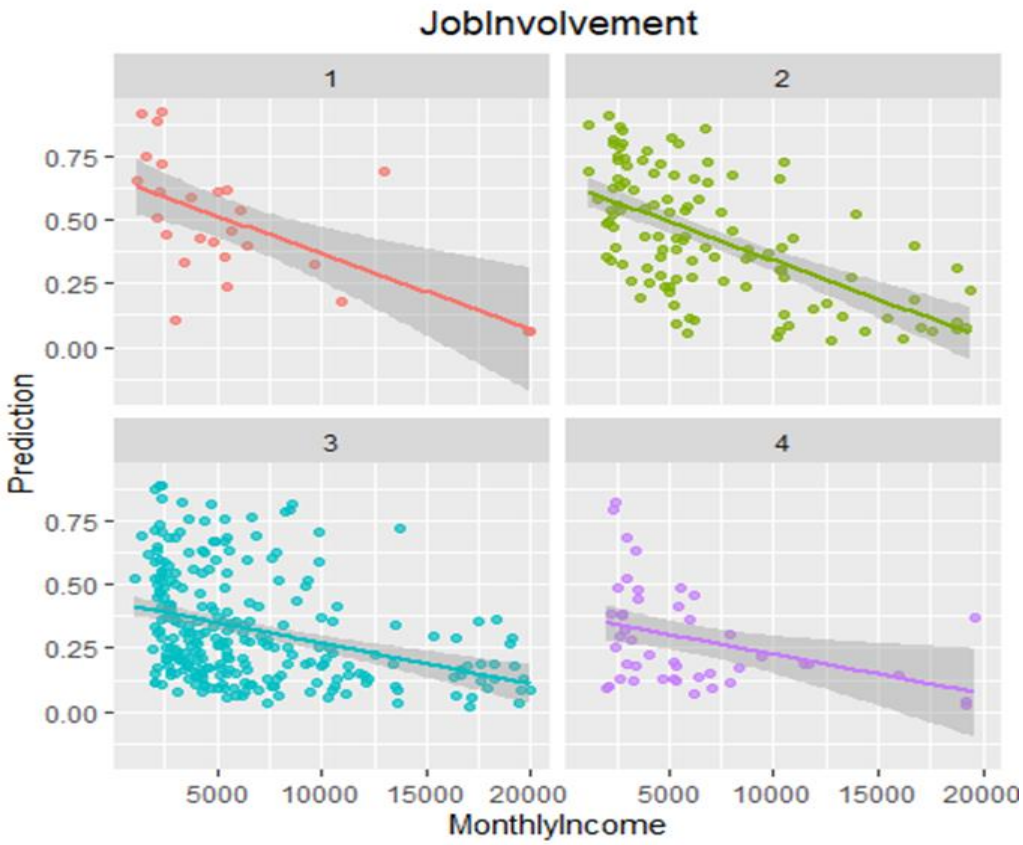
Overtime and monthly income from the previous exploratory analysis results, we can clearly realize the important impact of these two factors. From the results, it seems that the company should take some measures for those who work overtime and then leave and those with low monthly income. Measures. Finally, based on the previous data visualization, we have paid attention to the variables related to work-life balance. The four related variables WorkLifeBalance, DistanceFromHome, OverTime, and BusinessTravel are all in the importance list. It can be seen that the relationship should be paid attention to the fact of employee attrition.

Discussion

Further Analysis

Based on the importance list of the above variables, further use the model to predict whether people with high investment in exploration work and low income are more likely to leave.

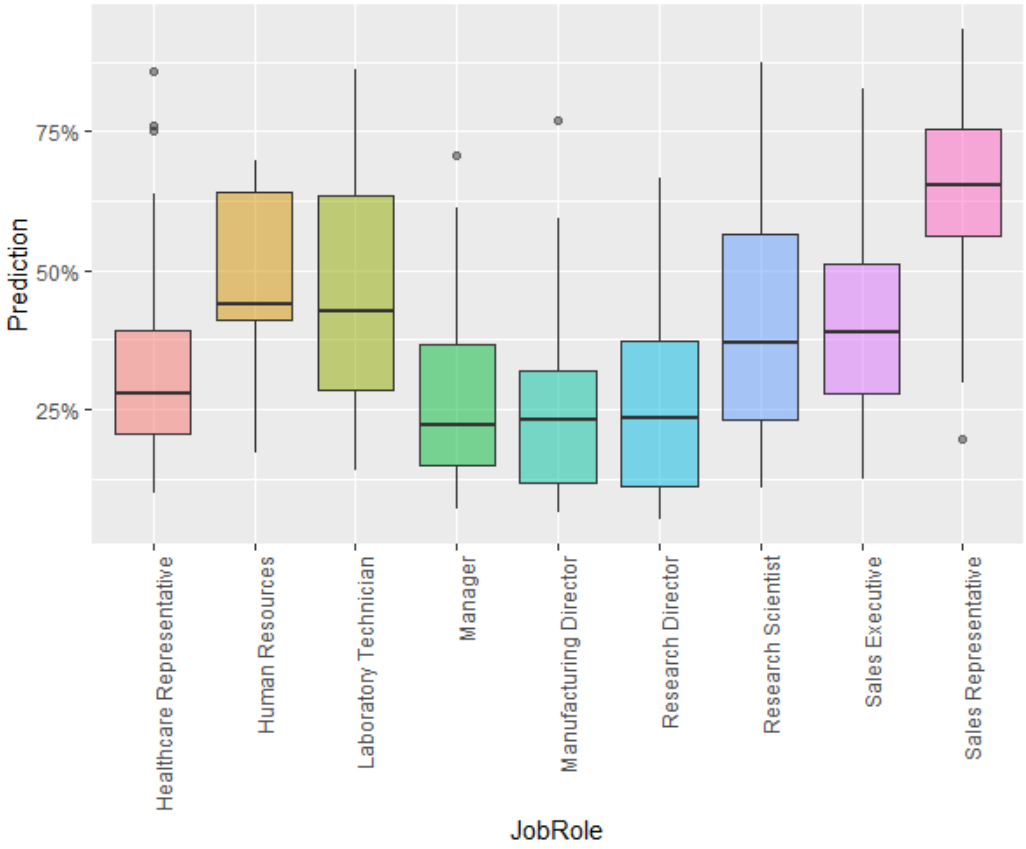
Figure 7: The Relationship Between JobInvolvement and MonthlyIncome



The predicted result is a bit unexpected. It can be seen from the figure 7 that people with high investment and low income are less likely to leave, and the slope of the regression line is lower and tends to be flat. The reason may be that they have a sense of belonging to or belong to the company. Other benefits are better. If can explore the further reason, maybe HR can gradually explore the factors that can keep employees firmly.

Finally, use the model to predict which departments and job roles are more likely to leave.

Figure 7: The Predicting Results of Departments and Job Roles



It can be drawn from the results of the forecast that the probability of sales representatives leaving the company is significantly higher than that of other departments, exceeding 60% on average.

Conclusion

- A major reason for employee resignation is due to overtime, or the disproportionate contribution and return.
- Compared with the attractiveness of high salaries, employees are more aware of the enjoyment of equity, and employees who enjoy equity dividends are less likely to leave.
- Age, years in the company and working experience are also some important indicators that affect employee turnover
- In some aspects of life, such as frequent business trips and long distances to work, it is also a secondary reason for employees to leave.
- The probability of sales representatives leaving the company is significantly high.

Limitations and Future Work

This article does not explain the GBM model in detail due to the length of the article, which may make the reading for readers of non-statistics students a little bit boring. Therefore, deepening the understanding of the GBM model after class and considering how to describe it in a popular and vivid manner will be the place where this article will continue to work hard.

It is necessary to further study whether the age of employees and the number of companies they have served is a matter of recruitment strategy or corporate culture. If companies often hire freelancers and other factors, this will also cause some misleading analysis results. If not, then from the result, the younger the person is, the higher the instability. This point needs further study.

Explore the turnover rate of the elite talents that companies are most concerned about, and predict which elite talents will leave. After all, companies can't put all their energy on everyone. The rule of

28 can be used to focus on 20% of the core elite employees. Here Scope for in-depth thematic analysis.

Further explore who is more willing to stay in the company, and focus on the factors that make employees care about and why they don't leave. Assist human resources to do a good job in the employment of the company, especially in the construction of corporate culture and newcomers in terms of training, effectively reduce the overall management cost of personnel.

Appendicies

Citations

[1]. Data source. Employee Attrition: https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset.

[2]. Univariate Analysis. https://www.statisticshowto.com/univariate/.

[3]. LightGBM's documentation. https://lightgbm.readthedocs.io/en/latest/.

[4]. ANIRUDDHA BHANDARI (2020). AUC-ROC Curve in Machine Learning Clearly [3]. Explained. https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/.

[5]. Terry M Therneau [aut, cre], Thomas Lumley [ctb, trl](2020). survival: Survival Analysis. Version: 3.2-7. https://cran.r-project.org/web/packages/survival/(AUC).

[6]. Xavier Robin ORCID iD [cre, aut], Natacha Turck [aut].(2020).Version:1.16.2. pROC: Display and Analyze ROC Curves. https://cran.r-project.org/web/packages/pROC/.

[7]. Developers [aut] (2020). gbm: Generalized Boosted Regression Models. Version:2.1.8. https://cran.r-project.org/web/packages/gbm/.

[8]. Pablo Diez,(2018)Confusion Matrix in Machine Learning.https://www.sciencedirect.com/topics/engineering/confusion-matrix.

[9]. Pavlos EfraimidisPaul Spirakis(2005). Weighted Random Sampling.https://link.springer.com/referenceworkentry/10.1007/978-0-387-30162-4_478.

[10]. Up sampling and Down sampling. (2013). https://www.projectrhea.org/rhea/index.php/Upsampling_and_downsampling_lab.

[11]. Dragoș Dumitrescu, Costin-Anton Boiangiu(2019);A Study of Image Upsampling and Downsampling Filters; DOI https://doi.org/10.3390/computers8020030.

[12]. Hadley Wickham (2020).ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics. Version:3.3.2. https://CRAN.R-project.org/package=ggplot2.

[13]. Carson Sievert.plotly: Create Interactive Web Graphics via 'plotly.js'. Version:4.9.2.1 https://CRAN.R-project.org/package=plotly.

[14]. Baptiste Auguie [aut, cre], Anton Antonov [ctb] (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. Version:2.3. https://cran.r-project.org/web/packages/gridExtra/.

[15]. Hadley Wickham [aut, cre](2020). plyr: Tools for Splitting, Applying and Combining Data. Version:1.8.6. https://cran.r-project.org/web/packages/plyr/.

[16]. Terry Therneau [aut], Beth Atkinson [aut, cre], Brian Ripley [trl](2019). rpart: Recursive Partitioning and Regression Trees. Version:4.1-15. https://cran.r-project.org/web/packages/rpart/.

[17]. Stephen Milborrow(2020). rpart.plot: Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'. Version:3.0.9. https://cran.r-project.org/web/packages/rpart.plot/.

[18]. Max Kuhn [aut, cre], Jed Wing [ctb], Steve Weston [ctb](2020). caret: Classification and Regression Training. Version:6.0-86. https://cran.r-project.org/web/packages/caret/.

[19]. Brandon Greenwell ORCID iD [aut, cre], Bradley Boehmke ORCID iD [aut], Jay Cunningham [aut], Luis Torgo(2013). DMwR: Functions and data for "Data Mining with R".Version:0.4.1. https://cran.r-project.org/web/packages/DMwR/

[20]. Hadley Wickham [aut, cre], Dana Seidel [aut], RStudio [cph](2020). scales: Scale Functions for Visualization. Version: 1.1.1. https://cran.r-project.org/web/packages/scales/

Code

https://github.com/yingrui-lyu/Attritioncode.