Sampling-Based Categorization of Employee Turnover in IBM HR Analytics

Amaresh Bose[1], Naghma Khatoon[2]
[1,2] Faculty of Computing and Information Technology
Usha Martin University
Ranchi-835103, India
[1]Corresponding Author

ABSTRACT

In this paper, we analyse the reasons for employee turnover using data from IBM's attrition survey. To begin, we used the correlation matrix to eliminate characteristics that were not strongly linked to others. Second, we found that a worker's monthly income, age, and the number of companies they've worked for all had significant effects on turnover, and we discovered this using Random Forest. Then, we used K-means Clustering to divide the population into two groups. The final quantitative analysis we performed used binary logistic regression, and it revealed that frequent travellers were 2.4 times more likely to abandon the group than infrequent ones. In addition, we discovered that human resources employees are more likely to resign than employees in any other division.

Keyword: Attrition, Random Forest, K-Means, Support Vector Machines

1. Introduction:

Employee attrition refers to the gradual loss of workers over time due to normal causes like retirement or voluntary separation from the workforce[1], and it has become a top concern for all businesses today due to the negative impact it has on workplace productivity and on meeting organizational objectives on time.[2] Businesses should make it a priority to reduce staff turnover in order to maintain a competitive edge over rival businesses. As a result, it is essential for the growth of a business that its management identify the leading causes of employee turnover and work to improve the company's productivity, workflow, and overall performance.

Contributions:

The following are the paper's key findings.

- We classify potential deserters using the K-means clustering algorithm after using Random Forest to identify the most influential factors in employee turnover. Clustering

- Through the use of quantitative analysis, we were able to compare the rates of employee turnover between different demographics.

This section will explain the remaining sections of the paper.

In Section 2, we present the methods we'll be using and conduct an analysis of the data.

In Section 3, we process the raw dataset by removing irrelevant variables and cleaning up the data.

In Section 4 Our machine learning model is implemented.

In the final section, we briefly review the key points.

2. Analysis of the Data

Since data on employee turnover is not generally considered to be public knowledge, we utilized the data set made available by Kaggle in this article. The data set contains 1471 records, and its 34 feature variables can be grouped into three categories: demographics, employment, and presence. In this research, we analyzed the impact of personality and length of service on employee turnover. The extent to which particular characteristics affect employee turnover. We also used machine learning algorithms to determine and foresee causal elements in employee turnover. In this paper, we applied the Logistic Regression, Decision Tree, and k-means clustering machine learning techniques.

2.1 Random Forest

Random forest is a type of machine learning that can be used for classification, regression, and other tasks by combining the results of many decision trees. By breaking the pattern of decision trees' over-reliance on the training data, random forest improves the model's accuracy. To get started, we can randomly replace pieces of the original data.[3] While each of the sub-datasets has information that is unique from the others, there may be some redundancies. Each decision tree has an outcome, so we can use the sub-dataset to build a sub-decision tree, which we can then use the voting mechanism of to arrive at a conclusion. Figure 1 depicts the results of the four datasets, three of which favored Alternative A and one favoring Alternative B, with Alternative A emerging as the winner.

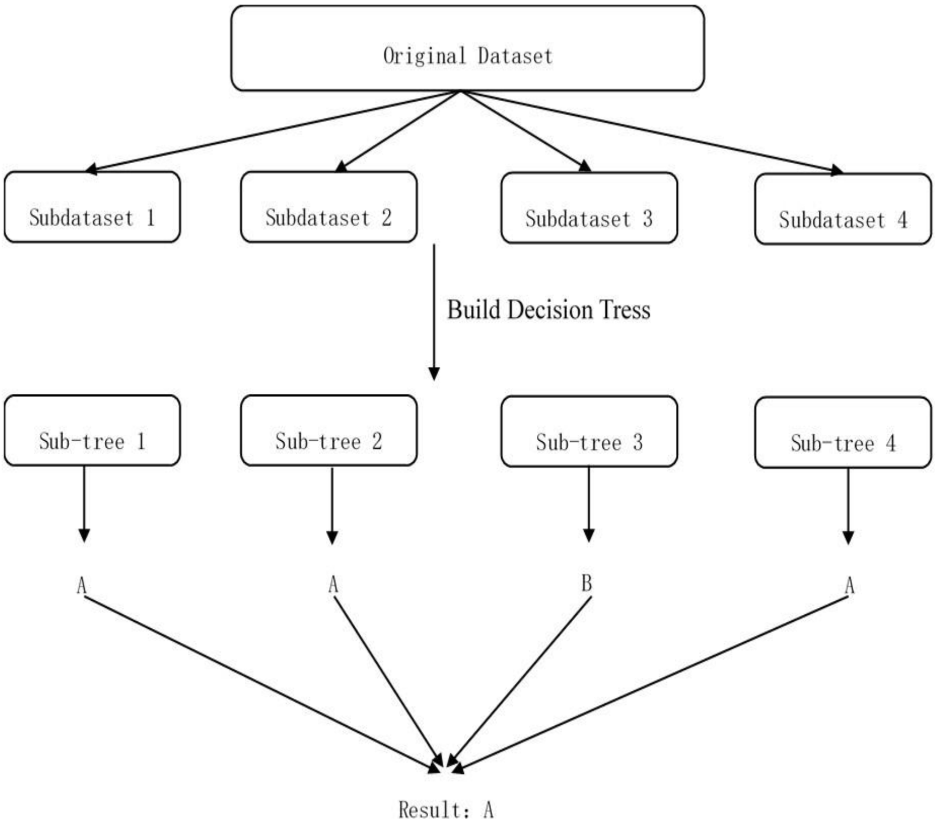


Figure 1: Decision-making process of random forest

Random forest is a popular machine learning technique. The random forest technique is widely used in the realm of machine learning. It combines precise forecasting with a straightforward explanation of the underlying model. Predictions and generalizations in RF can be improved with the help of random sampling and ensemble methods. Because of this, it's also a great explanation. It can instantly calculate the relative importance of each factor. That is to say, it is possible to easily ascertain the relative significance of each factor in the decision-making procedure. To analyse the factors that lead to employee turnover, we built a Random Forest model for this study. We generate 100 replacement-based random samples from our dataset, and from those we select k(k34) employee characteristics at random to use in our decision tree. In total, there are 34 such characteristics. Each node in the decision tree will be subdivided into more "pure" (i.e., gender-specific) child nodes for the outcome variable based on the prediction factors. As a result, there will be a rise in the birth rate of "pure" (gendered) offspring. The next child node will not split off from its parent if the trait it will inherit is already present in the parent. Each decision tree resolves to a conclusion, and the one with the most votes is selected.[4]

2.2 Binary Logistic Regression

The two main goals of logistic regression analysis are (1) estimating the likelihood of an event and (2) investigating the factors that contribute to the problem. In medicine, logistic regression is frequently used to investigate potential disease origins5. Potential risk factors for diabetes are being investigated, including age, smoking, alcohol consumption, and diet. Non-scale questions can be analyzed with logistic regression in questionnaire research, for example, to see how factors like age, gender, and household income affect consumers' propensity to make a purchase decision. Binary logistic regression is the most common statistical method.

$$logit_{(p)} = \ln \frac{p(y=1)}{1-p(y=1)} = \beta_0 + \beta_1 x_1 + ..$$

Employee turnover is 50% if y is between 0 and 1. Coefficients (values from 0 to n) stand in for the influence of independent variables on the dependent one, such as gender, education, extra hours worked, employment status, etc. (n=34, x1, x2,..., x34). Whether the regression coefficient is positive or negative, the value must have a justification. A positive regression coefficient for the dependent variable indicates a positive effect, while a negative value indicates a negative effect.

2.3 K-means Cluster Analysis:

K-means Clustering is used by the vast majority of existing apps. Given a set of K numbers and an equally large set of K starting cluster centers, move each point as close as possible to the center of its cluster. After compiling all the information, the new cluster center is calculated as an average. Only at the very end, in fact, does the cluster's center undergo any significant motion at all. This section's intent is to:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i^{(j)} - c_j||^2$$

If there are k cluster centers, then the jth cluster will have n total data points, and the Euclidean distance between cj and xi can be calculated as ||xij cj ||. The k-means technique is used for data classification, where the set of data points is denoted by X = x1, x2, x3,..., xn, and the set of centers is denoted by c = c1, c2,..., ck.

To get started, pick 'k' centers at random from your data set. Second, determine how far each data point is from the origin of each cluster.

Third, have the information be located at the center of the cluster that is closest to the target location.

After additional cluster centers have been identified, it is necessary to revise the formula n 1 cj = xi (3) n i=1 (step 4) to account for the new distances between individual data points and their respective cluster centers.

Figure 2 depicts the steps that make up K-means Clustering; if you get to step (6) without making any reassignments, you'll need to start over at step (3).

$$c_j = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{3}$$

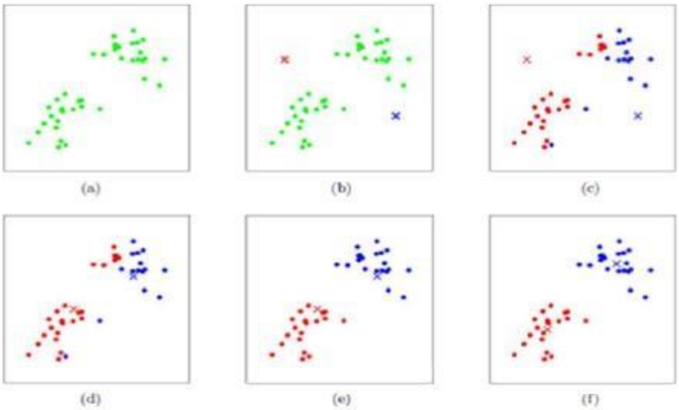


Figure 2: Clustering process of K-means Clustering

3. Data Processing:

Our studies include 34 variables in total, but we did not include Employee Count, Over18, or

Standard Hours because they only have one level of data. Getting rid of features that aren't strongly correlated with other properties can help speed up computations. To see how strongly the factors are related, we created a "correlation matrix," a sort of table. The table below displays the levels of correlation between each factor.

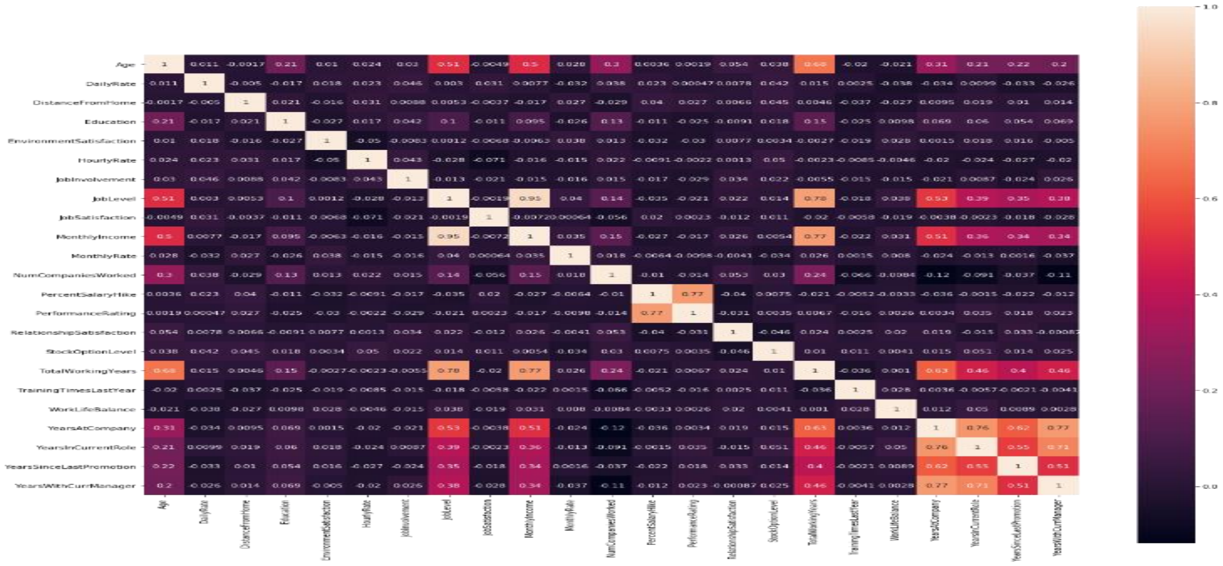


Figure 3: Correlation Matrix

The correlations between job level and monthly pay are 0.95, job level and total years worked are 0.78, and monthly pay and total years worked are 0.77 (Figure 3), while the correlations between daily, hourly, and monthly rates are very low or nonexistent. Therefore, we are clearing the books by erasing the daily, hourly, and monthly rates.

4. Model Construction:

Using the data provided, we will construct a machine learning model to predict employee turnover by selecting influential variables and classifying them by cause of departure.

4.1 Feature Selection

We've all had to figure out how to pick out the important features from a dataset and get rid of the irrelevant ones in order to make our model more accurate. As a result, we need to zero in on the causes of employee turnover that matter the most and ignore the rest.

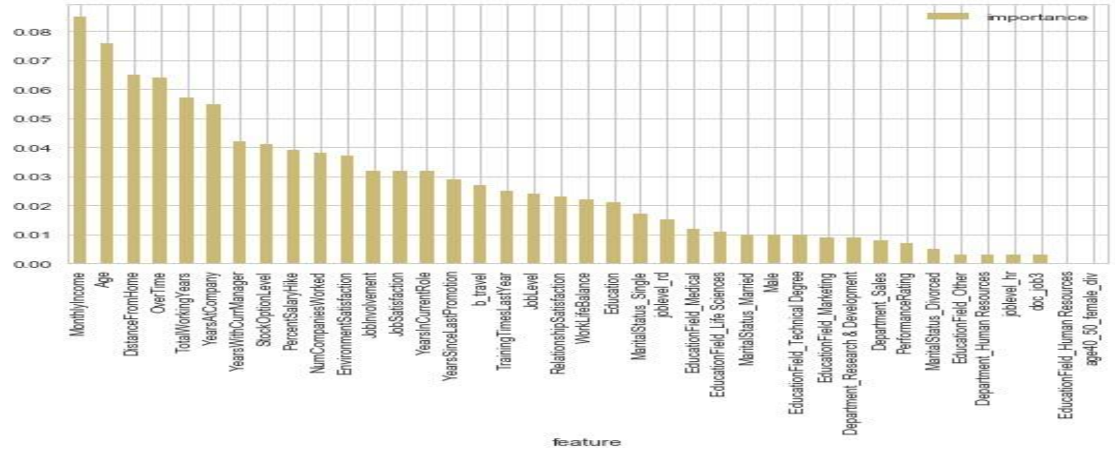


Figure 4: Important Features

Figure 4 shows that a worker's likelihood of quitting increases with their monthly income, age, and distance from home, but is mitigated by marital status and being a female worker aged

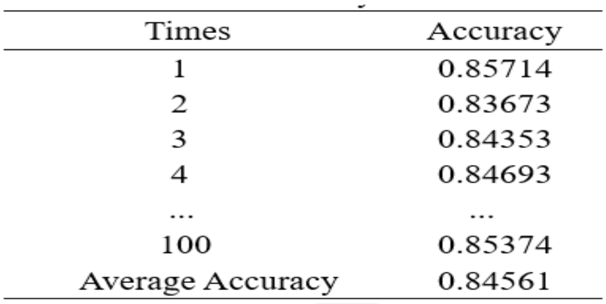
40–50. Salary is a major motivator for employees to leave their jobs, and it also has a direct effect on people's standard of living. High-priced services (such as medical care) and the benefits of a healthy way of life are more accessible to the wealthy. People are willing to risk leaving their current jobs in search of those that pay more because they are so concerned about their financial stability. It's also common for recent graduates to bounce around from job to job before finding something that truly fulfills them. Leaving a job is easier for younger individuals since they are less committed, since they have their own families and children, seniors want to make sure they are financially stable. Too much of a commute to and from work is a common reason for people to quit their jobs. Research in the social sciences and the field of public health has found that commutes are bad for both commuters and society as a whole. Longer commute times have been linked to an increased risk of health issues like obesity, hypertension, high cholesterol, back and neck pain, marital strife, and early death.

To accomplish this, we need both a training set and a validation set of information regarding employee turnover rates. The dataset was split in two, with 80% going to the training set and 20% to the test set. We made 100 predictions using RandomForest Classifier about employee turnover and recorded the average accuracy of these forecasts. To improve the clarity of the picture and to prevent any potential harm from using too many numbers, we intentionally omitted some data. Table 1 shows that overall, our model fits the data well, with an average accuracy of 0.84561.

4.2 Classification of People

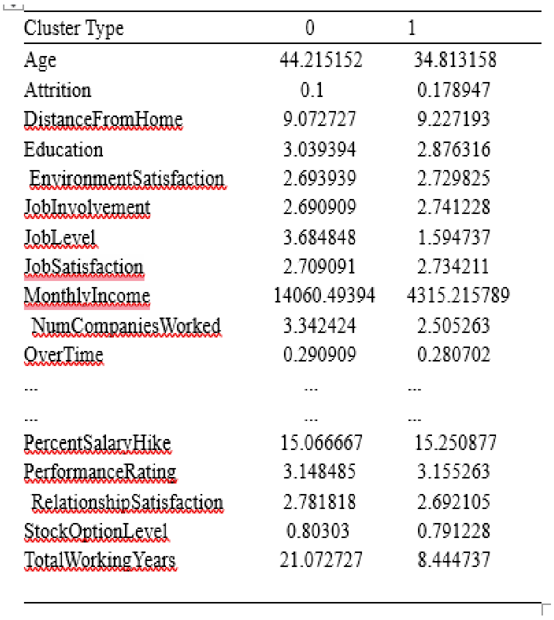
K-means clustering is used to sort the information and find out what makes people more likely to give up. The former tend to lose interest, while the latter are more dedicated. Table 2 (expanded version in appendix table 4.4) shows that workers are less likely to quit their jobs when they are older, have more education, are more satisfied with their jobs, make more money each month, and have worked for more companies. Attrition rates range from low (cluster set 0) to high (1 cluster set). The results are consistent with observations and prior studies of Random Forest feature selection. We showed how a worker's likelihood of quitting changes with age, monthly salary, and commute time in the previous section. The number of employers an individual has worked for turns out to be a significant factor as well. People with three to four years' experience in the workforce are less likely to resign because they have a solid idea of where they want to work, while those with more than four years' experience in the workforce show signs of job-hopping instability. People at higher levels of an organization are less likely to leave because they are paid more and are held in higher esteem. The majority of entry-level workers are dissatisfied with their jobs and eager to move up the corporate ladder. The turnover rate is largely determined by employees' levels of contentment with their current positions. There is a direct correlation between employees' levels of job satisfaction and their likelihood to stay with an organization[7].

Table 1: Prediction Accuracy of Random Forest



Times	Accuracy
1	0.85714
2	0.83673
3	0.84353
4	0.84693
...	...
100	0.85374
Average Accuracy	0.84561

Table 2: Clustering



Cluster Type	0	1
Age	44.215152	34.813158
Attrition	0.1	0.178947
DistanceFromHome	9.072727	9.227193
Education	3.039394	2.876316
EnvironmentSatisfaction	2.693939	2.729825
JobInvolvement	2.690909	2.741228
JobLevel	3.684848	1.594737
JobSatisfaction	2.709091	2.734211
MonthlyIncome	14060.49394	4315.215789
NumCompaniesWorked	3.342424	2.505263
OverTime	0.290909	0.280702
...	...	...
...	...	...
PercentSalaryHike	15.066667	15.250877
PerformanceRating	3.148485	3.155263
RelationshipSatisfaction	2.781818	2.692105
StockOptionLevel	0.80303	0.791228
TotalWorkingYears	21.072727	8.444737

4.3 The Logistic Regression

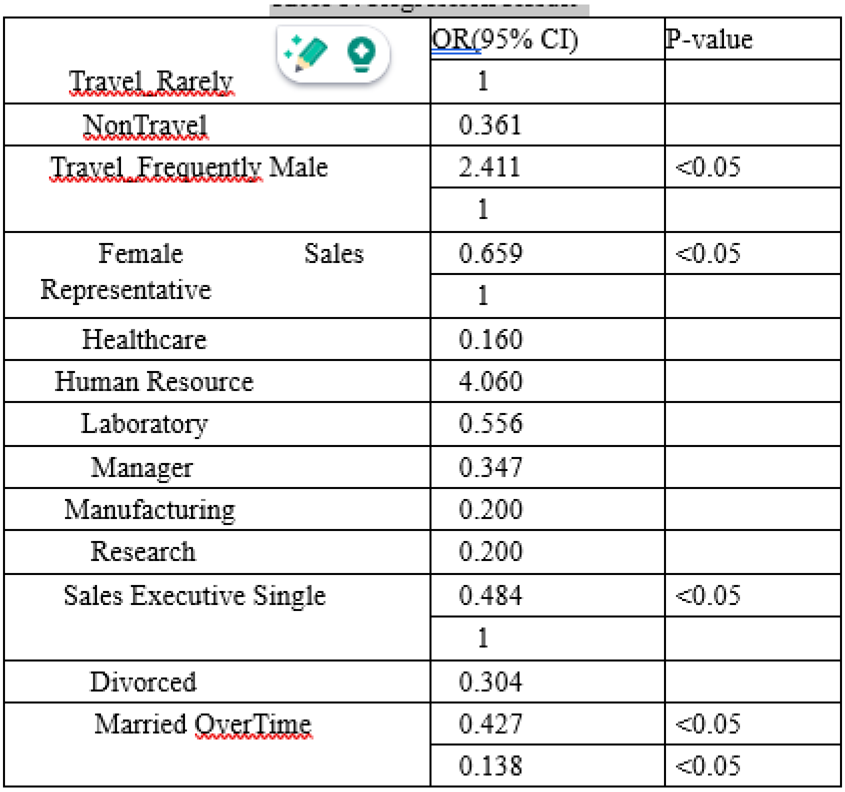
Here, we used binary logistic regression to forecast the connection between some explanatory variables (employee traits) and some latent ones (turnover rates). This model attempts to predict the relationship between indicators (employee traits) and an expected variable (employee turnover), where the dependent variable is a binary (NO:0, YES:1). We'll also look at the subgroups to see if there are any differences in what helps people quit smoking (see Table 4.5 of the annex for the full regression table).
According to Table 4.3, frequent travelers have a turnover rate that is 0.361% higher than the overall employee turnover rate. Employees who never leave the office have a turnover rate nearly twice as high as frequent flyers. Assuming people don't take many vacations is a necessary condition here. And there is a 0.659% gender gap in attendance. The employee turnover rate in Human Resources is significantly higher than in other departments, whereas it is lower in Research Science, Laboratories, Manufacturing, and Healthcare. Consequently, the likelihood of job turnover is higher among those who have never been married or divorced than among those who are married. Finally, and this shouldn't come as a surprise, people who put in a lot of overtime are eager to find a new position. The accuracy of the Python logistics regression model needs to be verified before any conclusions can be drawn from it. The value of 0.8843 indicates that our model provides a close approximation to the data.

The aforementioned illustration demonstrates that our findings are consistent with the opinions of other specialists and with empirical evidence. Through the use of Random Forest and K-means Clustering, we identified the primary causes of employee turnover. To begin with, a person's monthly salary, the number of employers they have worked for in the past, and their age are the primary factors that influence whether or not they will quit their current job. According to the findings of Random Forest. According to the K-means Clustering analysis, long-term employees tend to have characteristics such as age, level of education, job satisfaction, monthly income, and number of employers. However, since everyone's motivations are different, more in-depth studies using qualitative analysis are necessary. The binary logistics regression was utilized for this purpose. Compared to the overall population, we found that the attrition rate was 0.659 times higher for females than for males, 0.427 times higher for married people, and 0.304 times higher for divorced people. In addition, the rate of attrition was 2.4 times higher for frequent flyers compared to those who only flew occasionally. We also learned that there is a high rate of turnover in the HR division. Finally, other interesting findings from our study include the lower attrition rate among those who have worked for only two to four companies compared to those who have worked for more, the lower attrition rate among women compared to men after six years in the workforce, and the lower attrition rate among those with a doctoral degree.
The dataset used to make turnover predictions was split in half (80% for training and 20% for testing), with the test set's accuracy being recorded. The results show that Logistics Regression is better suited for prediction in our dataset (accuracy of 0.8843 vs. 0.8456 for Random Forest).

We hope that the company will take this research and the suggestions we make to heart and show more concern for their employees and work to increase their job satisfaction. Concurrently, they should pay more attention to HR because many of its employees are unhappy in their roles. The company also has an obligation to provide its employees with reasonable time off to rest and spend with their loved ones. It's generally accepted that workers perform better when they are allowed more frequent breaks.

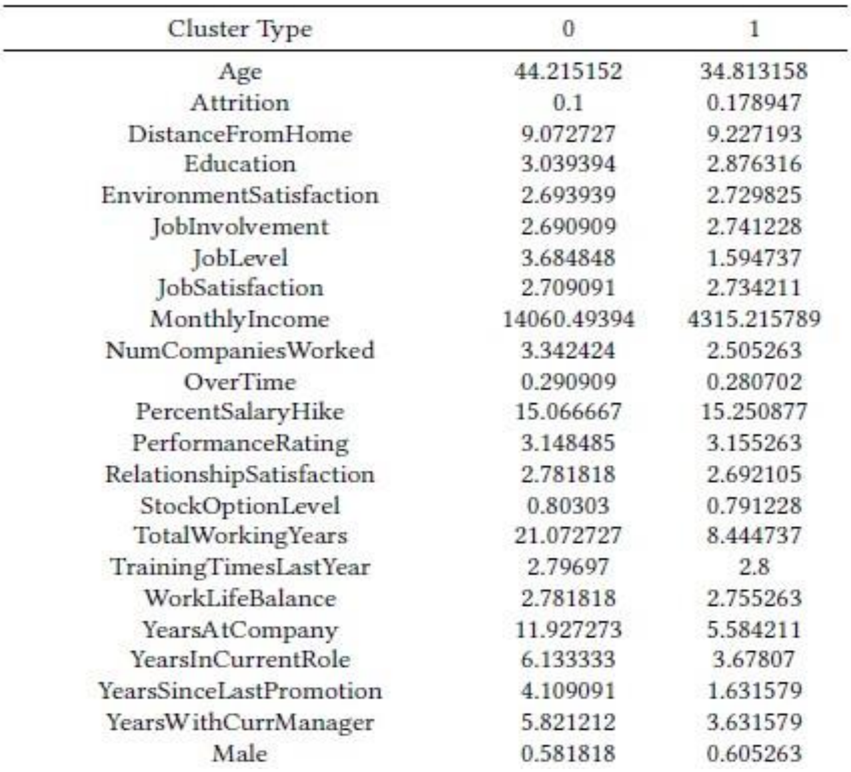
Using a combination of statistical and visual methods, we can examine how the values in each column of satisfaction differ in relation to one another.

Table 3: Regression Result



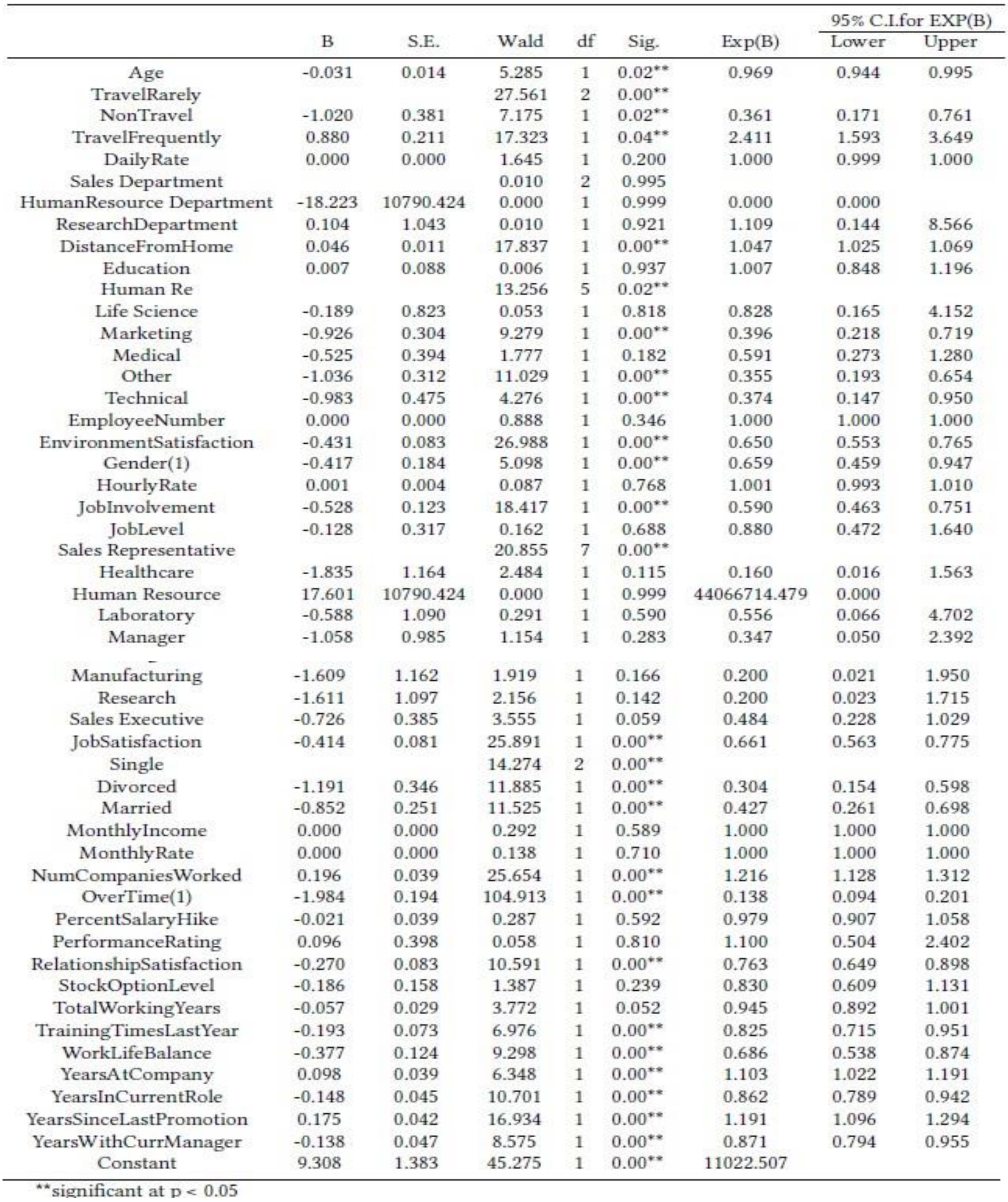
	OR(95% CI)	P-value
Travel_Rarely	1	
NonTravel	0.361	
Travel_Frequently Male	2.411	<0.05
	1	
Female Sales Representative	0.659	<0.05
	1	
Healthcare	0.160	
Human Resource	4.060	
Laboratory	0.556	
Manager	0.347	
Manufacturing	0.200	
Research	0.200	
Sales Executive Single	0.484	<0.05
	1	
Divorced	0.304	
Married OverTime	0.427	<0.05
	0.138	<0.05

Table 4: Results of Clustering



Cluster Type	0	1
Age	44.215152	34.813158
Attrition	0.1	0.178947
DistanceFromHome	9.072727	9.227193
Education	3.039394	2.876316
EnvironmentSatisfaction	2.693939	2.729825
JobInvolvement	2.690909	2.741228
JobLevel	3.684848	1.594737
JobSatisfaction	2.709091	2.734211
MonthlyIncome	14060.49394	4315.215789
NumCompaniesWorked	3.342424	2.505263
OverTime	0.290909	0.280702
PercentSalaryHike	15.066667	15.250877
PerformanceRating	3.148485	3.155263
RelationshipSatisfaction	2.781818	2.692105
StockOptionLevel	0.80303	0.791228
TotalWorkingYears	21.072727	8.444737
TrainingTimesLastYear	2.79697	2.8
WorkLifeBalance	2.781818	2.755263
YearsAtCompany	11.927273	5.584211
YearsInCurrentRole	6.133333	3.67807
YearsSinceLastPromotion	4.109091	1.631579
YearsWithCurrManager	5.821212	3.631579
Male	0.581818	0.605263

Table 5: Results of Logistics Regression



	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
							Lower	Upper
Age	-0.031	0.014	5.285	1	0.02**	0.969	0.944	0.995
TravelRarely			27.561	2	0.00**			
NonTravel	-1.020	0.381	7.175	1	0.02**	0.361	0.171	0.761
TravelFrequently	0.880	0.211	17.323	1	0.04**	2.411	1.593	3.649
DailyRate	0.000	0.000	1.645	1	0.200	1.000	0.999	1.000
Sales Department			0.010	2	0.995			
HumanResource Department	-18.223	10790.424	0.000	1	0.999	0.000	0.000	
ResearchDepartment	0.104	1.043	0.010	1	0.921	1.109	0.144	8.566
DistanceFromHome	0.046	0.011	17.837	1	0.00**	1.047	1.025	1.069
Education	0.007	0.088	0.006	1	0.937	1.007	0.848	1.196
Human Re			13.256	5	0.02**			
Life Science	-0.189	0.823	0.053	1	0.818	0.828	0.165	4.152
Marketing	-0.926	0.304	9.279	1	0.00**	0.396	0.218	0.719
Medical	-0.525	0.394	1.777	1	0.182	0.591	0.273	1.280
Other	-1.036	0.312	11.029	1	0.00**	0.355	0.193	0.654
Technical	-0.983	0.475	4.276	1	0.00**	0.374	0.147	0.950
EmployeeNumber	0.000	0.000	0.888	1	0.346	1.000	1.000	1.000
EnvironmentSatisfaction	-0.431	0.083	26.988	1	0.00**	0.650	0.553	0.765
Gender(1)	-0.417	0.184	5.098	1	0.00**	0.659	0.459	0.947
HourlyRate	0.001	0.004	0.087	1	0.768	1.001	0.993	1.010
JobInvolvement	-0.528	0.123	18.417	1	0.00**	0.590	0.463	0.751
JobLevel	-0.128	0.317	0.162	1	0.688	0.880	0.472	1.640
Sales Representative			20.855	7	0.00**			
Healthcare	-1.835	1.164	2.484	1	0.115	0.160	0.016	1.563
Human Resource	17.601	10790.424	0.000	1	0.999	44066714.479	0.000	
Laboratory	-0.588	1.090	0.291	1	0.590	0.556	0.066	4.702
Manager	-1.058	0.985	1.154	1	0.283	0.347	0.050	2.392
Manufacturing	-1.609	1.162	1.919	1	0.166	0.200	0.021	1.950
Research	-1.611	1.097	2.156	1	0.142	0.200	0.023	1.715
Sales Executive	-0.726	0.385	3.555	1	0.059	0.484	0.228	1.029
JobSatisfaction	-0.414	0.081	25.891	1	0.00**	0.661	0.563	0.775
Single			14.274	2	0.00**			
Divorced	-1.191	0.346	11.885	1	0.00**	0.304	0.154	0.598
Married	-0.852	0.251	11.525	1	0.00**	0.427	0.261	0.698
MonthlyIncome	0.000	0.000	0.292	1	0.589	1.000	1.000	1.000
MonthlyRate	0.000	0.000	0.138	1	0.710	1.000	1.000	1.000
NumCompaniesWorked	0.196	0.039	25.654	1	0.00**	1.216	1.128	1.312
OverTime(1)	-1.984	0.194	104.913	1	0.00**	0.138	0.094	0.201
PercentSalaryHike	-0.021	0.039	0.287	1	0.592	0.979	0.907	1.058
PerformanceRating	0.096	0.398	0.058	1	0.810	1.100	0.504	2.402
RelationshipSatisfaction	-0.270	0.083	10.591	1	0.00**	0.763	0.649	0.898
StockOptionLevel	-0.186	0.158	1.387	1	0.239	0.830	0.609	1.131
TotalWorkingYears	-0.057	0.029	3.772	1	0.052	0.945	0.892	1.001
TrainingTimesLastYear	-0.193	0.073	6.976	1	0.00**	0.825	0.715	0.951
WorkLifeBalance	-0.377	0.124	9.298	1	0.00**	0.686	0.538	0.874
YearsAtCompany	0.098	0.039	6.348	1	0.00**	1.103	1.022	1.191
YearsInCurrentRole	-0.148	0.045	10.701	1	0.00**	0.862	0.789	0.942
YearsSinceLastPromotion	0.175	0.042	16.934	1	0.00**	1.191	1.096	1.294
YearsWithCurrManager	-0.138	0.047	8.575	1	0.00**	0.871	0.794	0.955
Constant	9.308	1.383	45.275	1	0.00**	11022.507		

**significant at $p < 0.05$

CONCLUSION

Our findings are consistent with both empirical data and those of other researchers, as this model shows. Through the use of Random Forest and K-means Clustering, we were able to determine the factors that had the greatest impact on employee turnover. First, according to Random Forest, the main factors influencing employees' decisions to leave their jobs are their monthly wage, their age, and the number of companies they've worked for. According to the K-means Clustering study, long-term employees tend to have characteristics such as age, level of education, job satisfaction, monthly income, and number of employers. Nonetheless, given the diversity of human motivation, additional research employing qualitative methods is required. For this, we used the binary logistics regression. Females had an attrition rate that was 0.659 times higher than males', married people had an attrition rate that was 0.427 times higher than the general population, and divorcees had an attrition rate that was 0.304 times higher than the general population. In addition, the attrition rate was 2.4% higher for frequent fliers than it was for those who took fewer trips. We also learned that the human resources department has a high turnover rate. Finally, other interesting findings from our study include a lower attrition rate among those who have only worked for two to four companies compared to those who have worked for more, a lower attrition rate among women compared to men after six years in the workforce, and a lower attrition rate among those with a doctoral degree.

The accuracy of the turnover predictions was reported based on the performance of the test set, which was comprised of 20% of the original dataset. Our data analysis indicates that Logistics Regression performs better than Random Forest at making predictions (accuracy of 0.8843 versus 0.8456).

We're crossing our fingers that the company will take our findings and recommendations seriously and do more to make their employees happy at work. At the same time, they need to focus more on human resources because many of its workers are dissatisfied. The company must also ensure that workers are given adequate time off to relax and spend with their families. It's common knowledge that allowing employees more frequent breaks boosts productivity.

REFERENCES

1. According to Bhatnagar (J. (2007). Employee engagement in Indian IT service providers: a key retention strategy for talent management. Relations with staff.

2. R. Jain and A. Nayyar. (The month of November in the year 2018). Using xgboost, a machine learning technique, we can foresee employee turnover. System modeling and progress in research trends (smart): Proceedings of the 2018 International Conference (pp. 113-120). IEEE.

3. Pal, M. (2005). Classification in remote sensing using a random forest classifier. The international remote sensing journal, volume 26 issue 1, pages 217–222.

4. Qi, Y. (2012). In bioinformatics, a random forest is used. Methods and applications for ensemble machine learning. Pages 307-323. Published by Springer US in Boston.

5. Meng X H, Huang Y X, Rao D P, etc. Predicting diabetes and prediabetes using risk factors: a comparison of three data mining models[J]. 2013;29(2):93-99 The Kaohsiung Journal of Medical Sciences.

6. The choice of K in K-means clustering: D. T. Pham, S. S. Dimov, and C. D. Nguyen, J. Cluster Analysis. Part C: Journal of Mechanical Engineering Science, Proceedings of the Institution of Mechanical Engineers, 2005, 219(1), pages 103-119.

7. Chen Weiqi. Chinese Education& Society 40, no. 5 (2007): 17-31 "The structure of secondary school teacher job satisfaction and its relationship with attrition and work enthusiasm."