STAT 642-674: Data Mining for Business Analytics

Why Do Employees Leave?

A comprehensive analysis of critical factors impacting employee turnover rate & a potential approach to predicting attrition

By-
Nam Dang
Tram Phan
Linh Ha
Prat Patodkar

DREXEL UNIVERSITY
LeBow College of Business

Date of submission: 06/08/2022

Executive Summary

This report aims to elaborate the problem of employee attrition, present the findings of research conducted on factors influencing it, and offer potential solutions to curb attrition. It also provides a deeper insight into the process Group 7 followed while conducting the analyses.

Introduction

Attrition is one issue that is agnostic of industries. It continues to cost businesses in the United States over 1 trillion USD in revenue every year & this estimate does not include the loss of potential profit. Apart from monetary loss, voluntary turnover also causes loss of resources, good employees, and talent. The absence of these factors adversely impacts organizations. Therefore, it is critical to determine the causes of attrition and deploy stable mechanisms to restrain it. The unprecedented situations triggered by the global pandemic have further solidified the importance of retaining good employees.

A technology company wants to help human resources develop strategies to keep employees and automate the process of predicting attrition. This will not only result in curbing the loss of revenue, it'll also help increase their profit margins, and overall satisfaction of the employees in the company.

To predict which employees are more likely to leave the company, it is pertinent to understand the relationships between a range of factors, whether some factors affect the process more than others, which factors play a major part in employees deciding to leave the company, and who the at-risk employees are. For this project, Group 7 identified some variables of interest, conducted multiple analyses, and tested whether these variables were impacting employee

attrition. We did this using supervised & unsupervised methods of building prediction models in R.

Dataset & Exploratory Data Analysis

The dataset used in this project is titled **employeeretention.csv**. It consists of **1470 rows, 30 columns** and predicts whether or not an employee will stay in the company. Therefore, the **target variable** in this dataset is **Attrition**.

Following is a classification of the variables:

Numerical Variables	Categorical Variables	
▪ Age ▪ Distance From Home ▪ Hourly Rate ▪ Monthly income ▪ No. Companies Worked ▪ Percent salary hike ▪ Total working years ▪ Training time last year ▪ Years at company ▪ Years in current role ▪ Years Since Last Promotion ▪ Years With Current Manager	**Nominal Variables** ▪ Gender ▪ Department ▪ Education field ▪ Job Role ▪ Marital Status ▪ Over 18 ▪ Attrition ▪ Overtime	**Ordinal Variables** ▪ Performance rating ▪ Environment Satisfaction ▪ Business travel ▪ Education ▪ Job involvement ▪ Job level ▪ Job satisfaction ▪ Relationship Satisfaction ▪ Stock option level ▪ Work life balance

We included below summary statistics of the target variable as well as of other predictors included in our model:

Numerical Variables

Numerical Variables	Age	Distance From Home	Hourly Rate	Monthly Income	No. Companies Worked	Percent Salary Hike
Min.	18.00	1.00	30.00	1009.00	0.00	11.00
1st Q.	30.00	2.00	48.00	2911.00	1.00	12.00
Median	36.00	7.00	66.00	4919.00	2.00	14.00
Mean	36.92	9.19	65.89	6503.00	2.69	15.21
3rd Q.	43.00	14.00	83.75	8379.00	4.00	18.00
Max.	60.00	29.00	100.00	19999.00	9.00	25.00

Numerical Variables	Total Working Years	Training Times Last Year	Years At Company	Years At Current Role	Years Since Last Promotion	Years With Current Manager
Min.	0.00	0.00	0.00	0.00	0.00	0.00
1st Q.	6.00	2.00	3.00	2.00	0.00	2.00
Median	10.00	3.00	5.00	3.00	1.00	3.00
Mean	11.28	2.80	7.00	4.22	2.19	4.12
3rd Q.	15.00	3.00	9.00	7.00	3.00	7.00
Max.	40.00	6.00	40.00	18.00	15.00	17.00

Ordinal Variables (*)

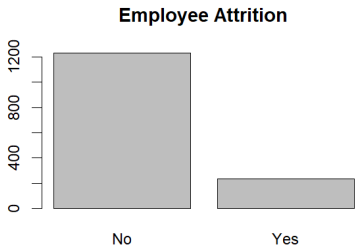
Values	Environment Satisfaction	Job Involvement	Job Level	Job Satisfaction	Performance Rating	Relationship Satisfaction	Stock Option Level	Work Life Balance
0	-	-	-	-	-	-	631	-
1	284	83	543	289	-	276	596	80
2	287	375	534	280	-	303	158	344
3	453	868	218	442	1244	459	85	893
4	446	144	106	459	226	432	-	153
5	-	-	69	-	-	-	-	-
Avg.	2.72	2.73	2.06	2.73	3.15	2.71	0.79	2.76

Factor Variables (*)

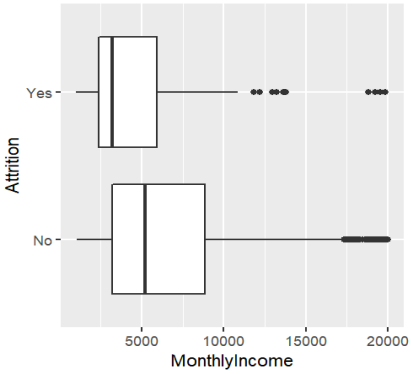
Categorical Variables	Values	Count
Gender	Male	882
	Female	588
Business Travel	Non-Travel	150
	Travel Frequently	277
	Travel Rarely	1043
Department	Human Resources	63
	Research and Development	961
	Sales	446
Education Field	Human Resources	27
	Life Sciences	606
	Marketing	159
	Medical	464
	Technical Degree	132
	Other	82
Job Role	Sales Executive	326
	Research Scientist	292
	Laboratory Technician	259
	Manufacturing Director	145
	Healthcare Representative	131
	Manager	102
	Other	215
Marital Status	Divorced	327
	Married	673
	Single	470
Over Time	Yes	416
	No	1054
Over 18	Yes	1470
	No	0

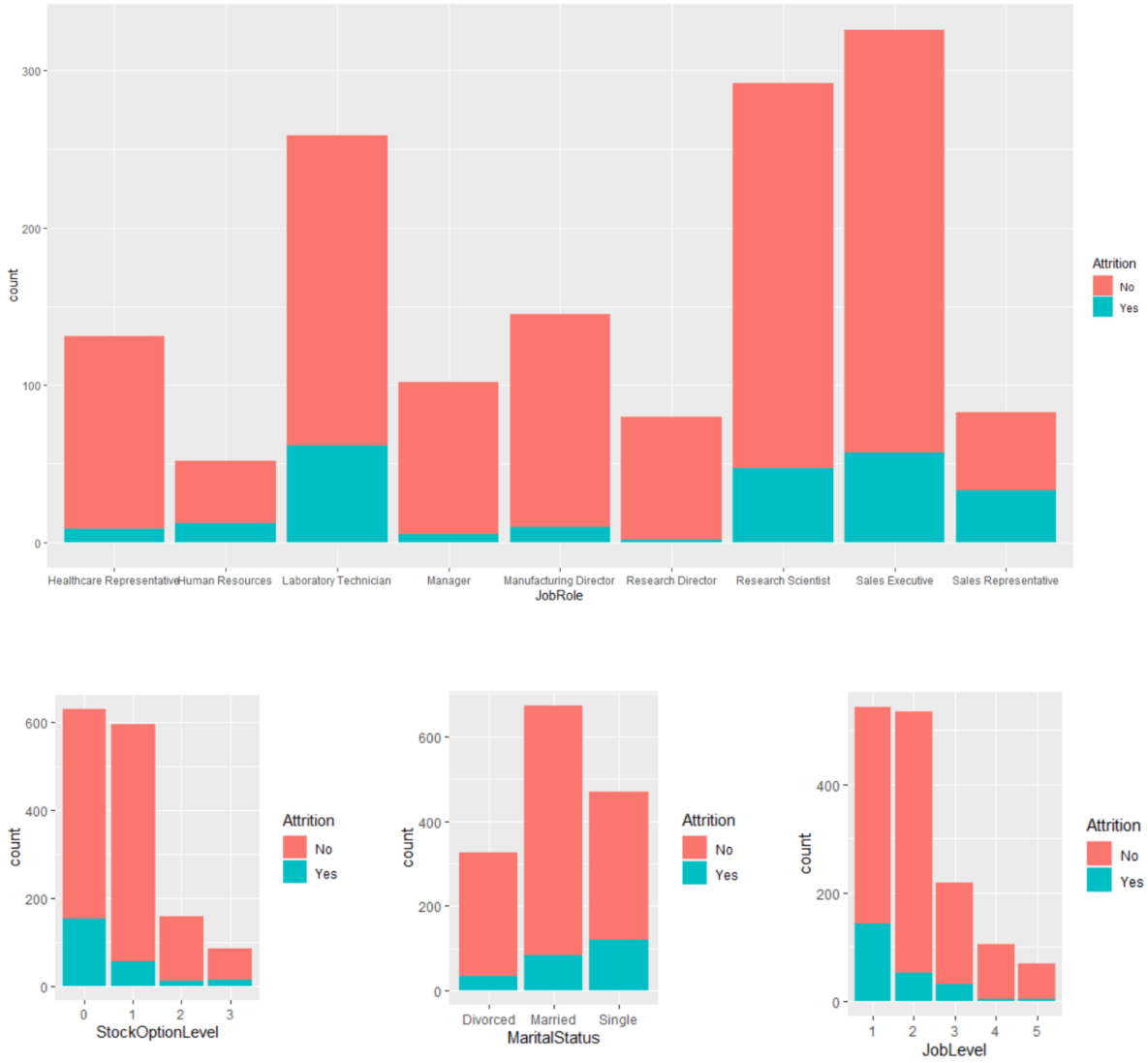
(*) Bar charts are integrated in the table to visualize the distribution of the values within each variable.

There seems to be a disproportion in our target variable, as only about 16% of the employees in our dataset have decided to leave the company.



Monthly Income of the employees in our data set has an average of ~$6,500. People who stayed at the company, on average, have a higher salary compared to the salary of people who left. Based on the boxplot, Monthly Income has a lot of outliers that can potentially affect our prediction accuracy.





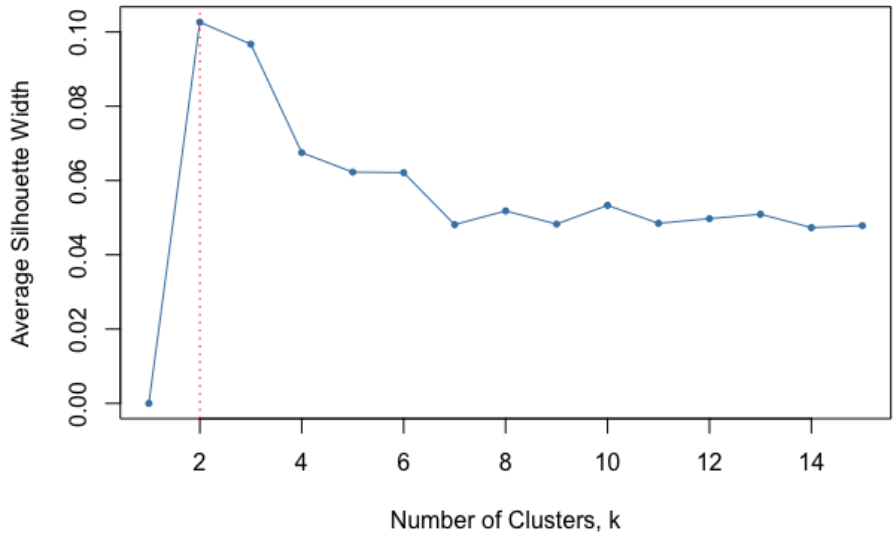
Further exploratory analysis of data suggests that highest attrition levels were observed among lab technicians, research scientists, and sales executives. Position of Research Director had the lowest attrition level. More employees with no stock options & lower job levels were leaving the company. This could be because there aren't many tangible bonds (stocks, equity) keeping them from leaving the company. Lastly, most attrition occurred in people who were single. This could be because single people possibly have lesser financial obligations compared to those that are married or divorced. **During this process, the group also recognized the need**

to conduct deeper analysis to validate or nullify these assumptions. Based on the preliminary exploration, the data set has good quality, since there are no missing values. From the list of variables, we decided to omit Over 18, because it only takes one value, which is "Yes", so it would not be a good predictor for our models. There are outliers in the Monthly Income variables, which should be considered during modeling. For all models, we use 80/20 as our training/testing split ratio.

Analysis Results:

1. **Unsupervised methods:**

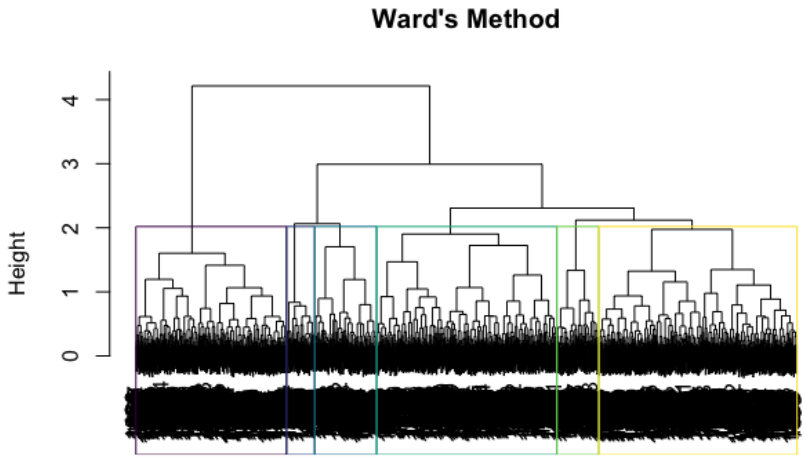
From the exploratory step, we were interested in seeing if employees can be divided into certain groups and we can identify any groups of employees that have similar traits resulting in a higher chance of attrition. **Moreover, in the data exploration section, it appears that attrition is more likely to happen at lower monthly income positions.** Hence, the cluster analysis also serves as a confirmative statement to our assumptions drawn from data observations which is the reason behind attrition is because of lower compensation. Observing that the dataset has varying scales affecting the distribution, we normalized and scaled the original dataset so we can obtain a new normalized dataset. We apply the YeoJohnson method to process the newly obtained dataset and take care of any outliers. To obtain the distance metric, we used the Gower method as it can handle all mixed types of data in our dataset. We used the pre-built function given in class to plot a chart of possible k. From the below chart, it looks like there are 3 possible values of k: 2, 4 and 6.



To be able to decide the optimal k, we choose to perform 2 unsupervised methods in this research including:

- **Hierarchical Cluster Analysis (HCA):**

In HCA, our selected Ward's method provides a dendrogram that aids in the process of choosing a k. According to the chart, there seems to be 6 clusters.



The next step in our process is to obtain centroid information for Ward's HCA clusters. This information is conducted separately for numerical and categorical variables. Based on our result, employees in cluster 3 have the highest chance of quitting their jobs. The majority of employees in cluster 3 are single, have no stock option package, and have a performance rating of 3. Also, they have average work environment satisfaction, job involvement, and job satisfaction. Those employee groups rarely travel so they are probably working in departments such as Research & Development or Human Resources.

We got a result of 0.0572 for the Adjusted Rand Index value, which is decent because it's closer to 1 than -1, but still relatively low. Therefore, this clustering solution is not good. However, the information revealed from HCA is consistent with what we observed from the data exploration step and our assumption.

- **K-Medoids Analysis (PAM):**

We perform PAM analysis on the processed dataset but it does not give us much helpful information. In fact, PAM shows that all 6 clusters are employees who decide to stay with the company.

In the last step of assessing the unsupervised method, we compare the two clustering solutions using (HCA and PAM) based on the Dunn Index, the average distance between clusters, and the average distance within the clusters.

	HCA	PAM
average.between	0.3762943	0.3716984
average.within	0.3221324	0.3257058
dunn	0.1789674	0.1179906

For average distance between clusters, we want the solution with a higher result, so HCA is preferred. For average distance within clusters, we want the solution with lower result, so HCA is still preferred. For the Dunn Index, we want the solution with higher results, so HCA

performs better than K-Medoids. Thus, overall, HCA performs better than PAM based on the above measures.

2. **Supervise methods:**

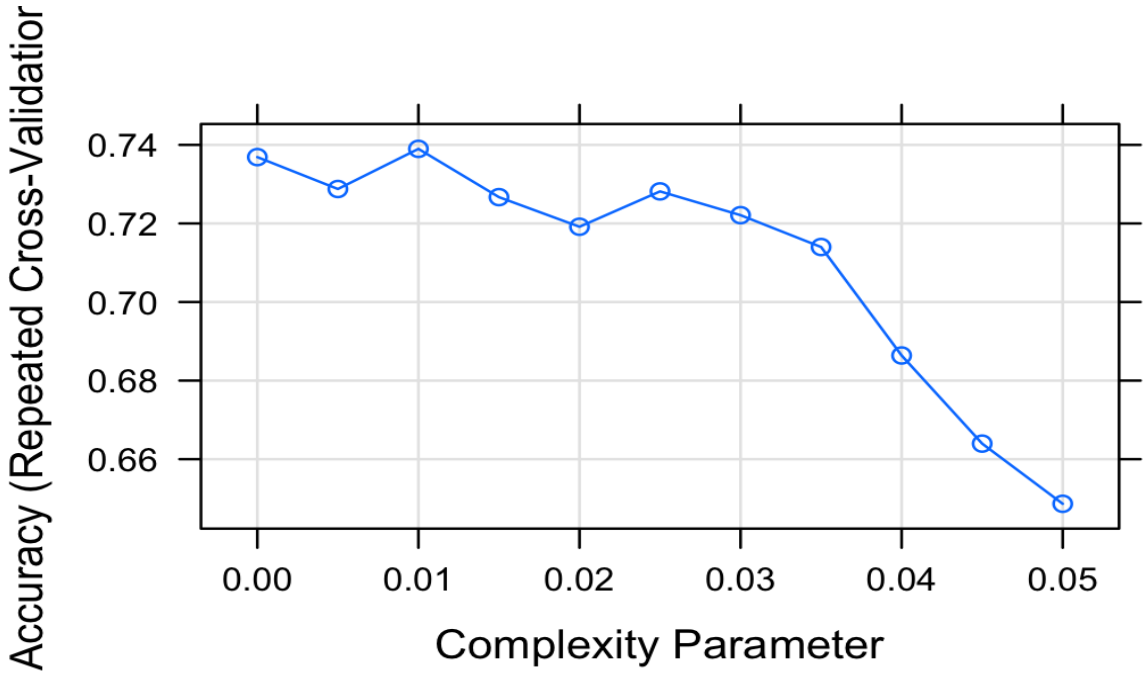
- **Decision Tree**

Our goal is to explore the main reasons that contribute to employee attrition and predict the attrition rate based on their features. The reason we choose the decision tree is because it is inexpensive to construct and has fast speed to do classification, which can avoid decision error and all kinds of deviation. Moreover, the model does not require a lot of data pre-processing steps such as normalization, transformation, and scaling. Overall, the decision tree model is also intuitive and easy to present to managers and can be implemented across the company's department.

The decision tree model can handle the missing value, however, we decided to check the missing value since it will impact our implementation if they exist. Our model is tuned to the cost complexity parameter by performing a grid search for the optimal complexity parameter (cp) value. We searched over a grid of values from 0 to 0.05 and set up a train control object with 10-fold cross validation, repeated 3 times and specified search = "grid" for a grid search. We decided to perform random oversampling since we experienced class imbalance in our target variable "Attrition".

The optimal complexity parameter is 0.01 with an average accuracy of 73.9%. The cp value helps determine the size of the tree by penalizing it for having too many splits. Since the size of the tree is inversely proportional to the value of the complexity parameter, because cp is low (0.01), the tree is larger. Based on the result of our decision tree model, the top three

variables which affect the company employee attrition are total working years, over time and monthly income.



Overall, the result from the table below for both training and testing are not optimal. The specificity of our decision tree model is high with 80% for training and 77% for testing. This means the model is good at predicting who will choose to stay at the company. However, our model is overfitting with resampling, since there is a 5% gap between training and testing accuracy. Other measurements such as Precision, Recall and F1 for the testing model are not high as well. Therefore, we would like to try other models to find the optimal prediction for attrition rate.

	Training	Testing
Accuracy	80%	75%
Sensitivity	79%	62%
Specificity	80%	77%
Precision	44%	35%
Recall	79%	62%
F1	57%	45%

- **Support Vector Machine**

For this study, we decided to develop a Support Vector Machine (SVM) model. A problem with our data set that we mentioned in the exploration section is that Monthly Income, a variable of interest in our data set, is plagued with outliers. SVM can handle this problem because the model itself is robust to redundant variables and outliers. Also, since the result of the Decision Tree model indicates some overfitting, SVM could potentially provide a more balanced model since it is not very prone to overfitting. For the above reason, we believe that SVM could be a good candidate to be our prediction model for this data set.

According to the preprocessing checklist for the SVM model, we need to handle missing values, binarize categorical variables, and rescale numerical variables. Since our data does not have any missing values, our preprocessing only consists of binarizing and rescaling. For binarizing, the factor and ordinal variables in our data set are binarized using the class2ind() and dummyVars() function from the caret package. For rescaling, we standardize our numerical

variables during the modeling. We also treat class imbalance in our target variable using the downSample() function to randomly undersample the training data.

We perform parameter tuning for the SVM model using the radical kernel approach using the kernlab package. We specify our approach of 5-Fold cross validation, repeated 3 times. Instead of a grid search, for this model, we use the random search of 10 different combinations of gamma (sigma) and cost value (C). The optimal values for sigma and C according to our parameter tuning are 0.0067 and 0.9131, respectively.

Using the optimal parameters, we train our SVM model and achieve the results below. Compared to the result from the Decision Tree model, the SVM model is more balanced, with accuracy metrics of Training and Testing are 78% and 77%. The SVM model also does better in terms of Sensitivity and Recall, indicating a high true positive rate - ability to accurately predict employees that will leave the company. Notably that high Sensitivity and Recall also means that the false negative rate is low, meaning the model is less likely to wrongly predict an employee that will remain at the company. This is one of the important factors, which contributes to our decision to choose SVM over Decision Tree (detailed explanation included in section F).

	Training	**Testing**
Accuracy	78%	77%
Sensitivity	77%	76%
Specificity	79%	77%
Precision	41%	40%
Recall	77%	76%
F1	54%	52%

F. Discussion & Conclusion

Based on the results of the applied unsupervised methods (HCA and K-Medoids), there is observed that attrition tends to happen among employees who have characteristics as below:

- They are single with no family
- Employees that do not have any stock options granted (level = 0)
- Their jobs are entry-level positions in the R&D and Sales Department
- The majority of those employees seem to be Lab Technician and Sale Representatives whose jobs are not required much travel
- They have a good work-life balance and above-average satisfaction in their job, work environment, and relationship with colleagues
- Some of them usually work overtime

The main reason for a turnover seems to be because of our compensation structure which does not provide a competitive package for employees in the R&D and Sales Department. Our Lab Technician and Sales Representative are having an average of $2k6 - $3k2/monthly. **Those employees quitting their jobs have been paid 15.5% - 16.46%. The national average income for lab technicians and sales representatives is higher than our employee average income. According to ZipRecruiter, a lab technician's monthly income ranges from $1,750 to $4,958 with an average of $3,309 per month (ZipRecruiter).**

Based on the result of our prediction model, we would recommend the SVM to predict attrition rate for the company. First, our SVM model is more balanced between the training and testing compared to the decision tree model. This shows how well a set of actual observed values

match those predicted by the model. Second, the recall/sensitivity of SVM is higher than our Decision Tree model. This means our model can accurately predict employees who will leave the company. More importantly, high recall/sensitivity also indicates a low false negative rate. **In other words, the chance that we predict an employee will stay at the company, but eventually end up leaving, is low**. This is very important in the Human Resources context, because such a situation would cause an immediate personnel crisis in the office, leading to unexpected cost in hiring and delayed working schedule. Third, the sensitivity of the Decision Tree is low with 62%, compared to 76% of the SVM model. This indicates that the SVM model is better at predicting employees who will leave the company. For these reasons, we believe our SVM model is better at predicting the rate of attrition in the company than the Decision Tree model.

To summarize, from our analysis, we would recommend the following approach:

1. The company should restructure its compensation model & align it with the national averages across the country
2. Conduct bi-weekly/Monthly meetings to understand what the employees feel about the organization, what their demands are, the organization's areas of improvement etc.
3. Pace over-time working hours. The employees shouldn't work too many hours. That may lead to feelings of discontent and further attrition within the organization
4. Work with lower ranking employees to set up recreational projects, invest in their growth, set up clubs for them and find ways of engaging this strata of employees
5. Integrate the SVM model to predict employees that are likely to leave the company to support HR management and strategies.

References

How much do lab technician jobs pay per month? - ziprecruiter. (n.d.). Retrieved June 8, 2022, from https://www.ziprecruiter.com/Salaries/Lab-Technician-Salary-per-Month