ISM6405: Business Intelligence

PROJECT REPORT

Exploratory Data Analysis of Human Resource Data

Project Group 9
Aditi Jayaprakash
Ektaa Rochiramani
Harsh Vora
Sudnyana Kadagadkai

Table of contents

Executive Summary

A company wishes to analyze why their best and most experienced employees are leaving prematurely. Employee attrition is bad for a company as it leads to poor morale, which in turn impacts effectiveness and efficiency. The dataset was found publicly for analysis from Kaggle.com. The purpose of this study is to investigate individual employee characteristics and organizational variables that may lead to employee attrition.

Introduction

Data Mining is the process of examining a large amount of data to derive meaningful data from it and generate new information. Our team is an HR Analytics team which analyzes and predicts employee behavior in a company and the company's performance.

Our analysis will help the company understand how the employees feel about the company and how their performance is impacting the company's progress. It can also help the organization understand where they are lacking in their services provided to employees by analyzing factors such as employee satisfaction, work hours, number of years worked, whether the employee has left the organization, promotions in the department, and so on. By analyzing these metrics, the organization can take steps towards greater employee satisfaction. The overall employee satisfaction currently ranges between 58% and 62%. A good employee satisfaction rating typically starts at 75% and above.

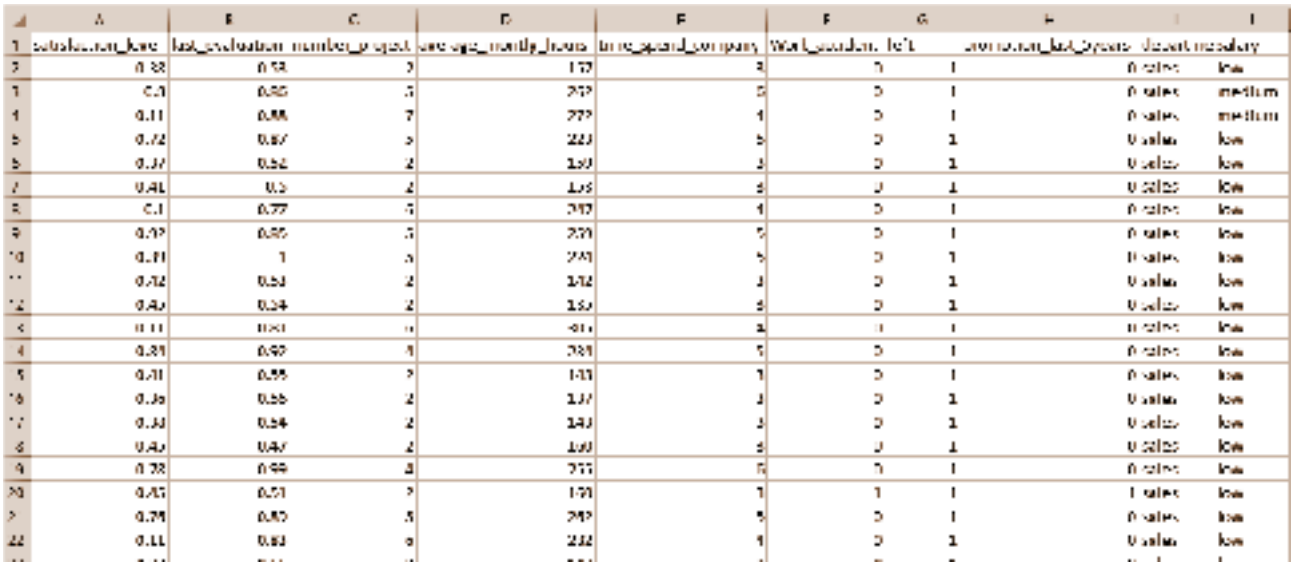
Predicting trends/behaviors will help the organization prepare themselves for a time when an employee might leave, using methods such as data exploration, principal component analysis, and neural networks.

About the data

The data set contains the following fields:

1. *satisfaction_level:* How satisfied the employee is with the company, and his/her work with the organization. The satisfaction level depends on a lot of factors such as promotions, number of years worked, work accidents, salary level, etc. The values fall in the range of 0 and 1.
2. *last_evaluation:* The performance evaluation an employee received for his/her work in the past year. The values fall in the range of 0 and 1.
3. *number_projects:* The total number of projects that are currently being handled by an employee.
4. *average_monthly_hours:* The average number of hours spent at work by an employee every month.
5. *time_spend_company:* The number of years an employee has worked in the organization.
6. *left:* Binary values indicating if the company left the company (1) or not (0).
7. *promotion_last_5years:* Binary values indicating whether an employee received a promotion in the past 5 years (1) or not (0).
8. *department:* This field describes which department an employee belongs to.
9. *salary:* The salary an employee receives. It has been classified into high, medium, and low.

Given below is the screenshot of a part of the data set:



Project Motivation

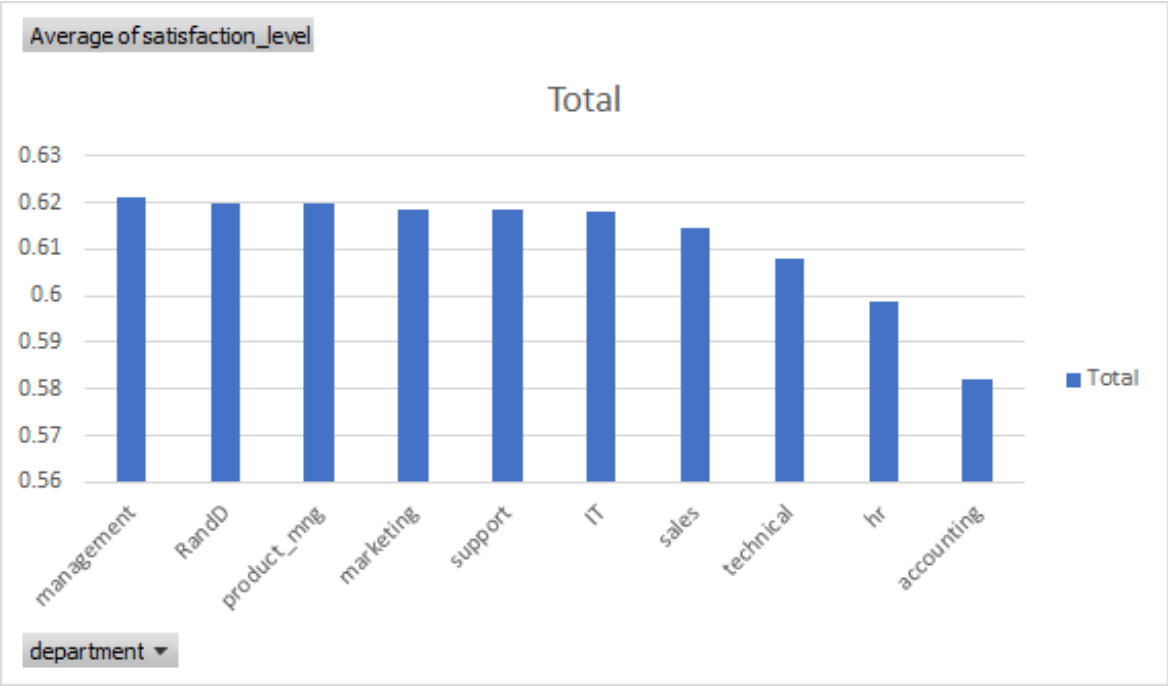
Human Resources Data should be analyzed to manage the employees effectively for reaching business goals quickly and efficiently. The performance of a company depends on the performance and satisfaction of its employees, hence it is important to leverage that data. Insights can be gained and capabilities can be predicted by modeling the data. Employee satisfaction is key to retaining employees and better performance of the company. Our motivation for this project is to help the company identify the factors affecting attrition so that they can take steps to reduce it.

Business Intelligence Techniques Used

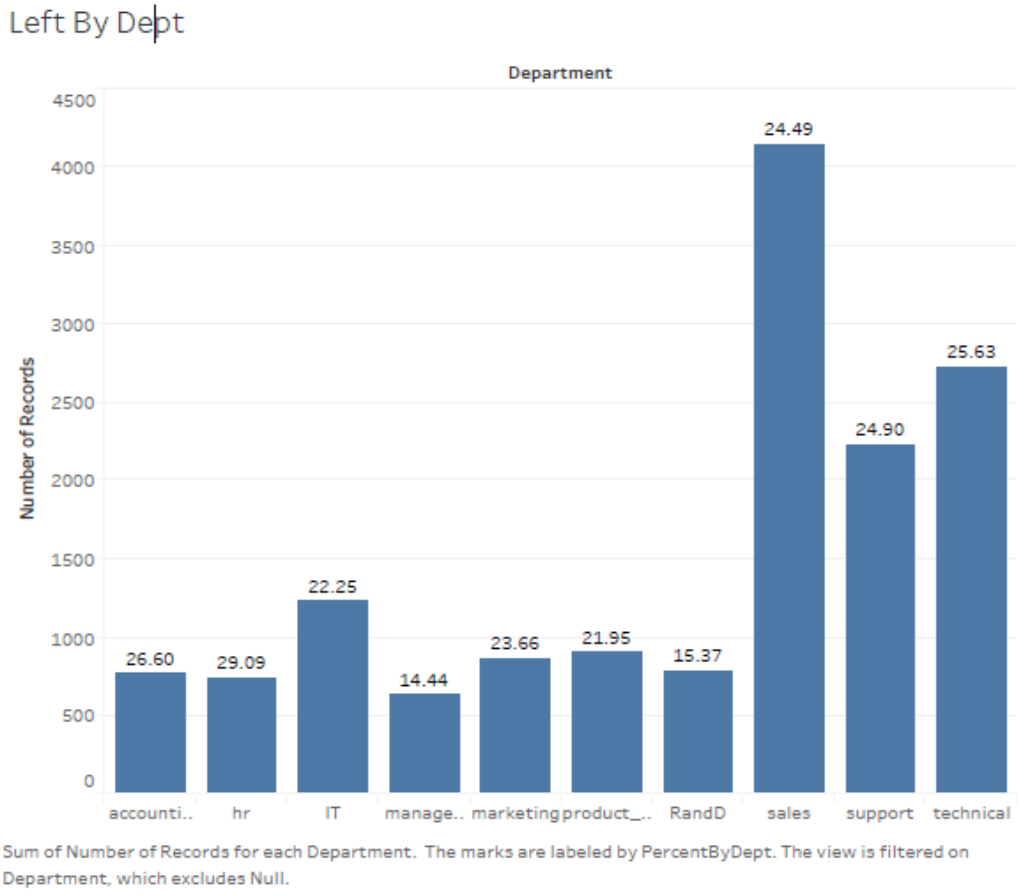
I. First we used Pivot Tables in Excel to find out the average satisfaction level of employees by department and sorted them in descending order of satisfaction level.

Row Labels	Average of satisfaction_level
management	0.621349206
RandD	0.619822109
product_mng	0.619634146
marketing	0.618601399
support	0.618299686
IT	0.618141809
sales	0.61444686
technical	0.607897059
hr	0.598809202
accounting	0.582151239
Grand Total	**0.612833522**

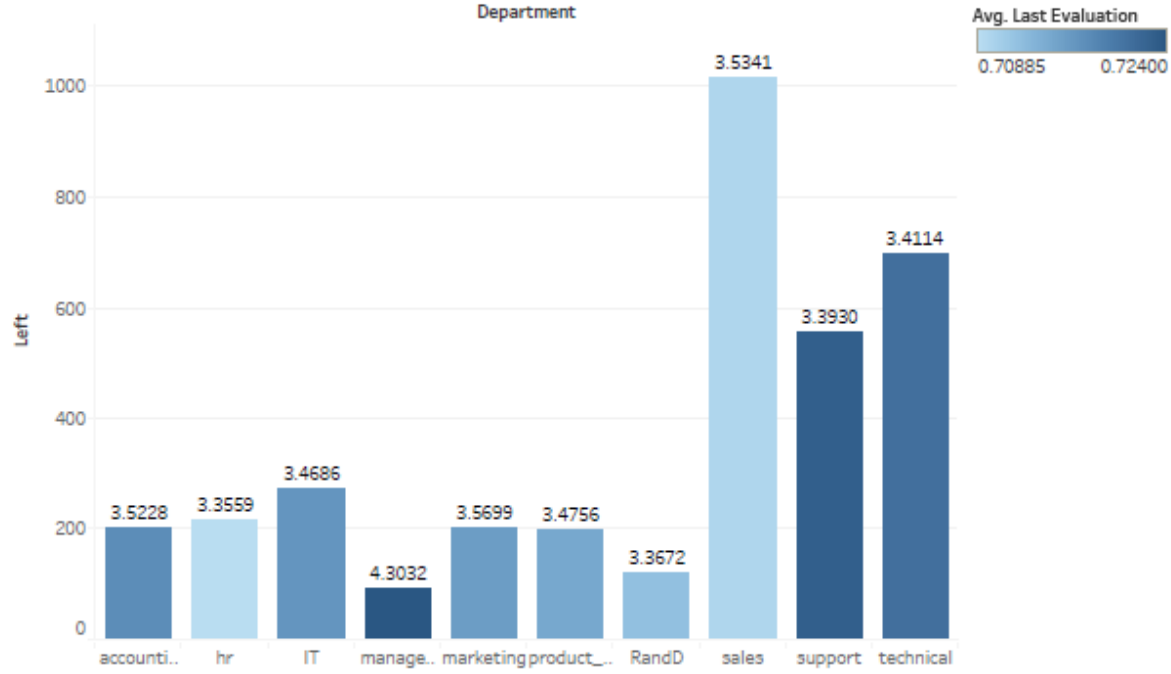
II. Then, we plotted the above values on a bar chart and noticed that the satisfaction levels did not vary a lot across departments. The management department is the highest in the satisfaction level while the accounting department is the least satisfied.



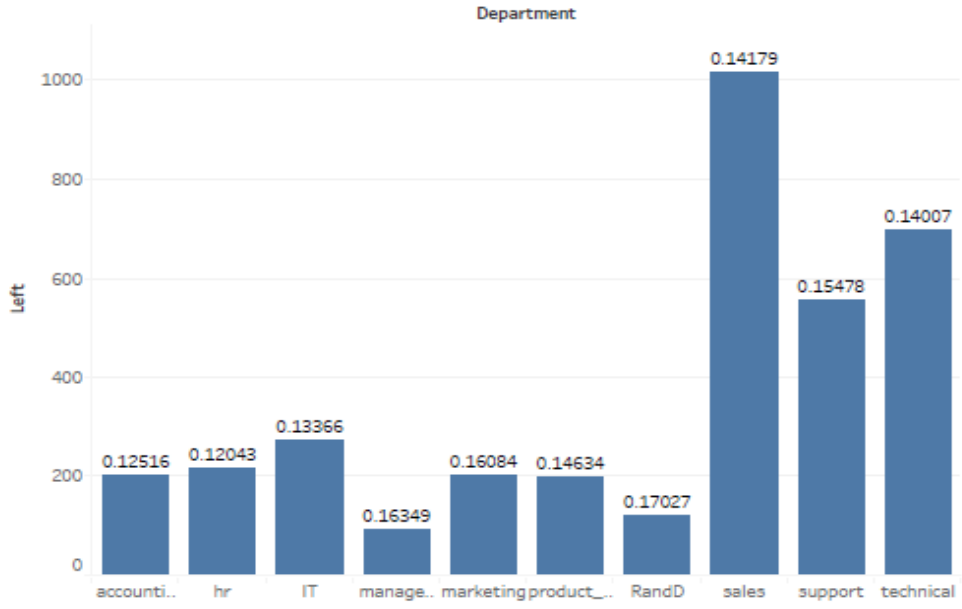
The graph above shows that the overall satisfaction level in the company is not very high and ranges between 58% and 62% across all departments.



This graph shows the employees who have left by their specific departments. Even though more number of people from Sales have quit the company, it is only because the Sales dept has more employees. But from the graph above we can conclude that the HR dept has a higher rate of attrition.



Sum of Left for each Department. Color shows average of Last Evaluation. The marks are labeled by average of Time Spend Company.



Sum of Left for each Department. The marks are labeled by average of Work accident.

Principal Component Analysis

It is a procedure that converts a set of possible correlated variables into a set of linearly uncorrelated variables. It helps to reduce the number of variables and to remove the overlap of information. In this project, we performed Principal Component Analysis.

We ran PCA through XL Miner and included all the variables in the analysis except the output variable and categorical variables. We fixed the number of components at 3 and ran the analysis. From the analysis results, we focused on the PCA Component output page and were able to determine the variables which had high correlation with all other variables in each of the three components.

XLMiner: Principal Component Analysis

Output Navigator			
Principal Componen	Variances	Scores	Summary

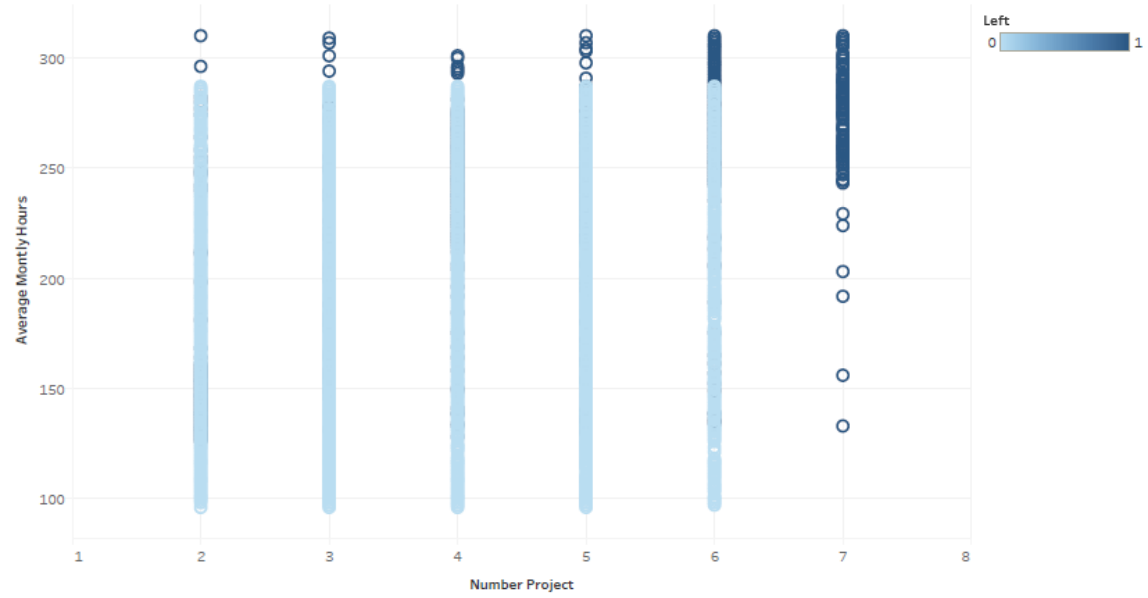
Principal Components			
Feature\Co	1	2	3
satisfaction	-0.14699	0.7269	0.265852
last_evalua	0.461401	0.382365	0.171426
number_pr	0.567629	0.013052	-0.07747
average_m	0.519033	0.16652	0.062026
time_spen	0.412745	-0.33446	-0.167
Work_accid	-0.04421	0.281247	-0.0864
promotion_	-0.0292	0.211417	-0.67777
pay	-0.02716	0.248699	-0.62878

From this analysis, we can conclude that the variable actually affecting attrition rates are as follows:

1. **Satisfaction Level**
2. **Last Evaluation**
3. **Average Monthly Hours**
4. **Number of Projects**
5. **Work Accidents**

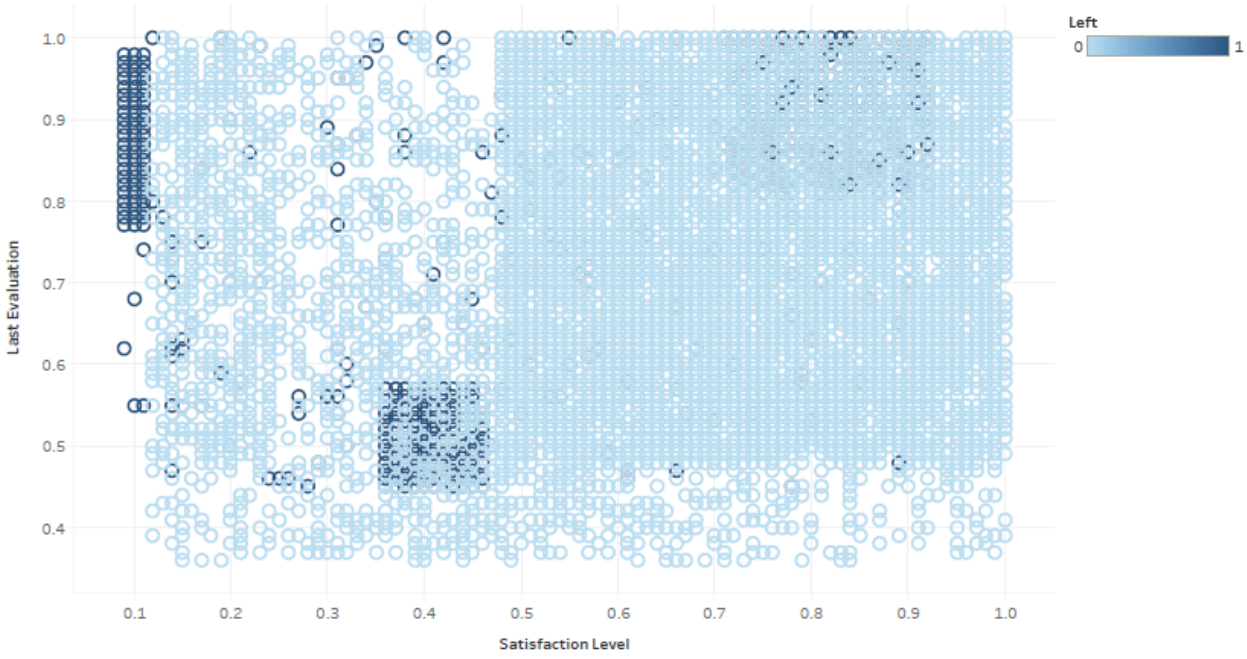
Tableau Visualization of Above Variables:

Once we knew what variables were affecting attrition rates at the company, we plotted them against each other and color coded them with employees who left. We obtained the following graphs.



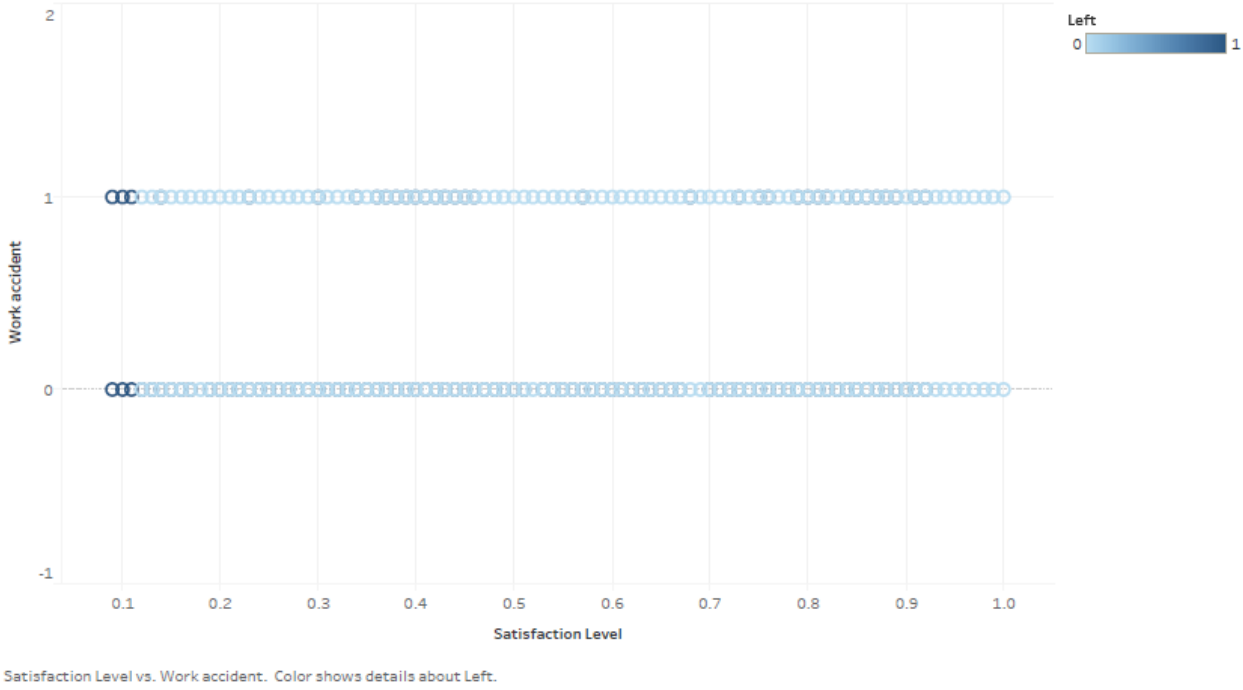
Number Project vs. Average Montly Hours. Color shows details about Left.

This graph gives a scatter plot of Number of Projects vs Average Monthly Hours and is color coded by employees who have left. From this graph we can conclude that, employees who had more than 6 projects and had to work at least 250 hours each month were likely to quit the company.

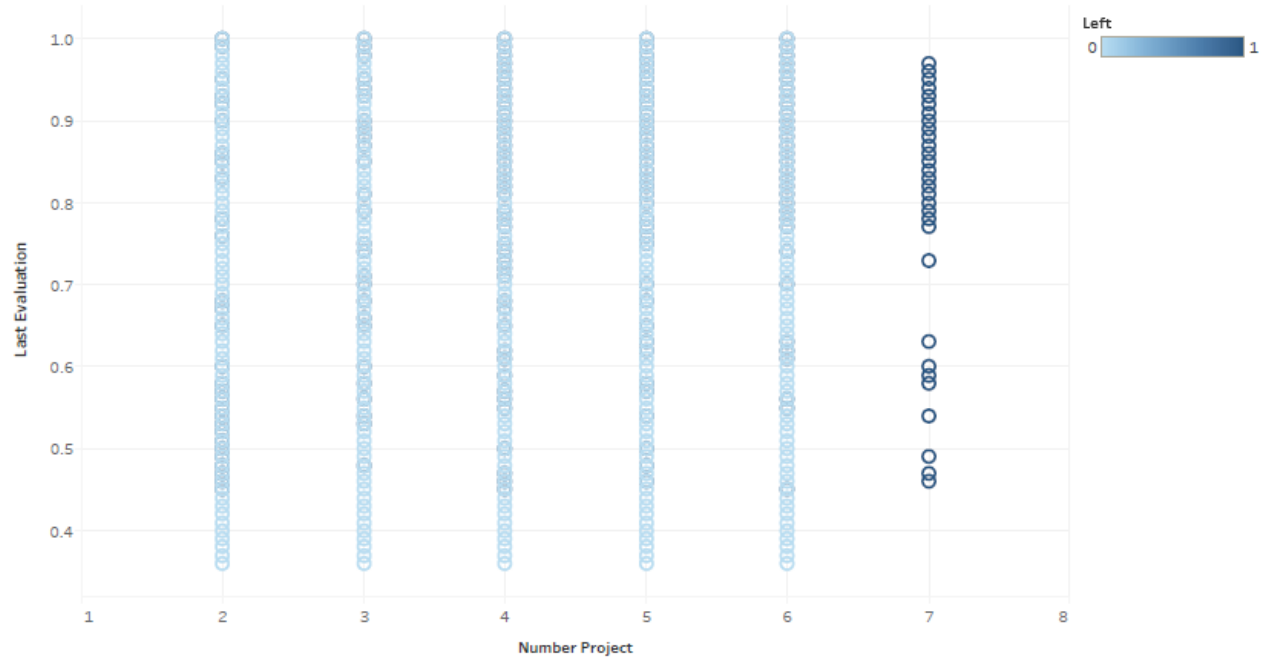


Satisfaction Level vs. Last Evaluation. Color shows details about Left.

This plot compares Satisfaction levels and last evaluation which is color coded by employees who have left. From the above graph we can conclude that, even if employees are given high evaluations, they tend to leave because they are not satisfied.

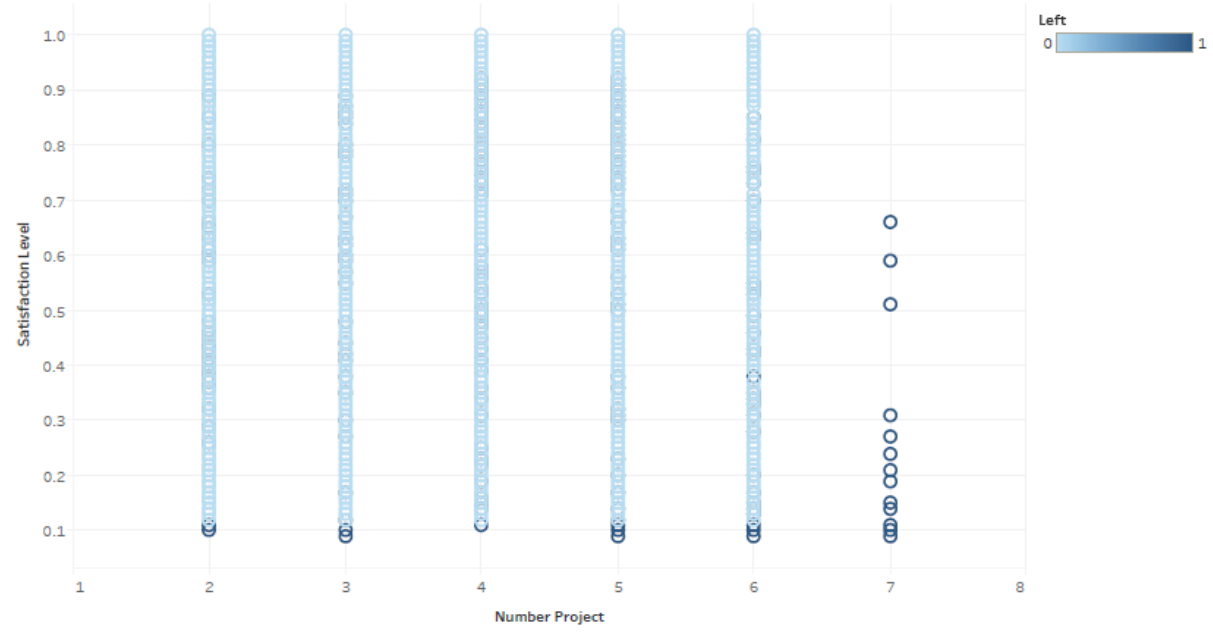


This graph plots satisfaction levels against number of accidents at work, and employees who left are color coded. It is evident from the graph that employees who are not satisfied and have between 0-1 work accident tend to quit the company.



Number Project vs. Last Evaluation. Color shows details about Left.

This graph compares number of project with last evaluation and employees who left are color coded. From this graph we can conclude that employees who get more than 6 projects tend to quit even if they are given high evaluations.



Number Project vs. Satisfaction Level. Color shows details about Left.

Lastly, in this plot we compare number of projects with the level of satisfaction and employees who leave the company are color coded. It is evident from the graph that, employees who have very low level of satisfaction tend to leave the company regardless of how many projects they get to work on.

Neural Networks

Now, we can use Neural Networks to help us predict employees who will quit the company. We use classification technique of data mining in XL Miner and run the manual neural network analysis. With probability of success cut off at 0.55 and three layers in the network, we obtain the following results.

Training Data Scoring - Summary Report

Cutoff probability value for success (UPDATABLE)	0.55	Updating the value here will NOT update value in detailed report

Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	901	107
0	119	3373

Error Report			
Class	# Cases	# Errors	% Error
1	1008	107	10.61508
0	3492	119	3.407789
Overall	4500	226	5.022222

Performance	
Success Class	1
Precision	0.883333
Recall (Sensitivity)	0.893849
Specificity	0.965922
F1-Score	0.88856

Validation Data Scoring - Summary Report

Cutoff probability value for success (UPDATABLE)	0.55	Updating the value here will NOT update value in detailed report

Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	556	51
0	72	2021

Error Report			
Class	# Cases	# Errors	% Error
1	607	51	8.401977
0	2093	72	3.440038
Overall	2700	123	4.555556

Performance	
Success Class	1
Precision	0.88535
Recall (Sensitivity)	0.91598
Specificity	0.9656
F1-Score	0.900405

Test Data Scoring - Summary Report

Cutoff probability value for success (UPDATABLE)	0.55	Updating the value here will NOT update value in detailed report

Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	340	45
0	58	1356

Error Report			
Class	# Cases	# Errors	% Error
1	385	45	11.68831
0	1414	58	4.101839
Overall	1799	103	5.725403

Performance	
Success Class	1
Precision	0.854271
Recall (Sensitivity)	0.883117
Specificity	0.958982
F1-Score	0.868455

The Test set and Validation set both provide satisfactory results based on the value of F1- score which is derived from Precision and Recall values. These values are in no way concrete and there can many other interpretations of the model by altering the layers in the network or the partition of the dataset.

Conclusion

To summarize, we first determined what variables actually affect rate of attrition with the help of Principal Component Analysis and were able determine the following five variables/factors:

These are the 5 factors which affect employee attrition rate:

- Satisfaction Level
- Number of projects
- Average number of monthly hours worked
- Last Evaluation
- Number of work accidents

Then, multiple combinations of these variables were used to plot visualizations using Tableau and confirmed the analysis of PCA about the cause of attrition in the company.

Lastly, we also developed a prediction model using manual neural networks to help us predict employees who would quit the company.

Recommendations

Some ways in which employee satisfaction can be improved:

- Reducing work hours: As observed from the data, employees are working close to 50 hours/week on an average. This is not healthy as it prevents them from having a good work/life balance.
- Increasing the workforce: As more number of projects allocated per employee increases the attrition rate, it would be a good idea to hire more employees in order to distribute the work equally and not over allocate some resources.
- Recognizing/Rewarding good performances: It is important to recognize and reward employees from time to time for good performance to maintain high morale in the workplace.
- Organize trainings: If employees are not adequately skilled, arrange for trainings and certifications that can help them perform better and get higher evaluation scores.

Tools Used:

XL Miner
Tableau

References

Data Source:

https://www.kaggle.com/ludobenistant/hr-analytics

Other Resources:

- https://deedata.wordpress.com/2016/01/11/employee-attrition-exploratory-data-analysis-and-predictive-modeling-using-r-part-1/
- http://abbottanalytics.blogspot.com/2010/02/prinicpal-components-for-modeling.html