Student Name	**Suohuijia Wang**
Student Number	**17200170**
Assignment Title	**Hackathon Report**
Module Code	**MIS40970**
Module Title	**Data Mining for Business Analytics**
Lecturer	**David Fagan**
Date Submitted	**28/04/2018**
Submission Zip	**17200170_Hackathon**

STUDENTS SHOULD KEEP A COPY OF ALL WORK SUBMITTED.

Procedures for Submission and Late Submission
Ensure that you have checked the Adult Education Centre's procedures for the submission of assessments.
Note: There are penalties for the late submission of assessments. For further information please see the Adult Education *Assessment Guidelines* publication or the website at www.ucd.ie/adulted.

Plagiarism is the unacknowledged inclusion of another person's writings or ideas or works, in any formally presented work (including essays, examinations, projects, laboratory reports or presentations). The penalties associated with plagiarism are designed to impose sanctions that reflect the seriousness of the University's commitment to academic integrity. Ensure that you have read the University's ***Briefing for Students on Academic Integrity and Plagiarism*** and the UCD ***Plagiarism Statement, Plagiarism Policy and Procedures***, (http://www.ucd.ie/registrar/)

Declaration of Authorship
I declare that all material in this assessment is my own work except where there is clear acknowledgement and appropriate reference to the work of others.

Signed: Suohuijia Wang **Date**: 28/04/2018

Table of Contents

Executive Summary

Currently, our company is struggling to keep our employees. Our data science department has been asked to investigate the recent high turnover to other companies based on the dataset provided by HR department.

- Why we need to write this report

  Employee turnover plays a very important role in companies in the business world. Since if there are so many employees choose to leave, it will get in a series of troubles. The first and most important issue is the decrease of profit. As a result, for every company, the HR team need to pay more attention to their employees who are likely to leave. So our task is to help HR team to analyze the patterns of our employees dataset. Find the relationships between different factors.

- The questions

  We have four issues to solve.
  (1) What staff members are likely to leave next?
  (2) Which department is most at risk to lose employees?
  (3) Is there anything else about the potential ex-employees.
  (4) What strategies can we employ to try and keep our current employees?

- Our solution

  Our team is going to do data mining for this dataset, and use different analyzing tools to process our data. The whole process includes data pre-processing, data visualization, building models and evaluation, prediction. Our goal is to find general patterns and relationships between different features of our employees in order to see what kinds of employees are more likely to leave, and how to retain our current employees. In addition, we build different kinds of models to do future prediction.
  Based on our results, HR team can make specific plans for different employees, help them perform better and solve their current problems.

- Our analyzing tools

  We use a variety of tools, Excel, Python, R. During the process of data mining, we have lots of methods. In the process of data pre-processing, we do basic description, use one hot encoding to convert all features to numeric data in order to put them into our model.
  In the process of data visualization, we plot histograms, scatter plots, and box plots.
  In the process of building models, we use random forest and decision tree to train our dataset, choose different parameters to get a better result. We use classification report to analyze the accuracy of our prediction.

- Our conclusion

  In conclusion, employee retainment and employee turnover prediction are very complex problems, and there are several aspects that can affect our employees. we need to analyze and take actions according to different situations. Overtime, monthly income, age, gender, marital status, stock option level, education, department are the main features for employees to stay or leave. Sales department is most at risk to lose employees.

- Our recommendation

  Based on our analysis, we give a list of recommendations, and HR team can go through the last section of our report, to see more details.
  (1) Provide more training courses for employees
  (2) Increase salary for outstanding staff
  (3) Adjust stock option level
  (4) Treat differently to males and females
  (5) Provide good working environment
  (6) Build a corporate core values
  (7) Concern new employees about their work and experience
  (8) Provide benefit for overtime

1. Introduction

Currently, many companies are struggling with one problem, that is, how to keep their employees. For a company, if there are so many employees choose to leave, it will lead to a series of issues. Firstly, the company's reputation will be affected. Secondly, employee departure will affect other staff's enthusiasm, which may reduce their work efficiency. Thirdly, it can also reduce business productivity and then cause high turnover. According to incomplete statistics, employees' work efficiency will reduce to 40%~50% for one to three months before they leave. However, those new employees' work efficiency can only up to 60% in the first 3 months after they join in the company. As a result, business managers often pay more attention to employee's turnover rate.

Based on data mining of a dataset from HR team, we are going to answer two main questions. First is about the prediction. Predicting what kind of staff members are more likely to leave next, and which department is most at risk to lose employees. Secondly, give some strategic recommendations to HR team about how to retain current employees. The ultimate goal is to minimize the turnover of a company and maximise the profit for a company by make good use of every employee's strength.

Generally, in employee analysis, the focus is on structured historical data. For this study we have been provided with a database containing several variables regarding to the current employees in the company and some confidential information. In this report, we conduct a data mining process for a dataset which is from our HR team.

Section 2 is about basic analysis for our dataset. Section 3 provides data visualization, discovers some useful and interesting patterns from dataset by plotting different kinds of figures. Section 4 introduces the random forests and decision trees, exploit the most relevant variables into our trees to train models and give predictions to answer our questions. The last part section 5 is our conclusion and recommendation.

2. Data Exploration

2.1 Preliminary Exploration

This dataset has 1470 rows and 35 columns. We can check first 10 lines to see the basic information.

	Age	CurrentEmployee	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber	...	Relat
0	41	No	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	1	...	
1	49	Yes	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1	2	...	
2	37	No	Travel_Rarely	1373	Research & Development	2	2	Other	1	4	...	
3	33	Yes	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1	5	...	
4	27	Yes	Travel_Rarely	591	Research & Development	2	1	Medical	1	7	...	
5	32	Yes	Travel_Frequently	1005	Research & Development	2	2	Life Sciences	1	8	...	
6	59	Yes	Travel_Rarely	1324	Research & Development	3	3	Medical	1	10	...	
7	30	Yes	Travel_Rarely	1358	Research & Development	24	1	Life Sciences	1	11	...	
8	38	Yes	Travel_Frequently	216	Research & Development	23	3	Life Sciences	1	12	...	
9	36	Yes	Travel_Rarely	1299	Research & Development	27	3	Medical	1	13	...	

10 rows × 35 columns

Figure 2.1.1 First 10 lines of the dataset

Obviously, our target is the current employee column, others are features of different employees. Next, we observe the basic statistics for this dataset.

	count	mean	std	min	25%	50%	75%	max
Age	1470	36.92380952	9.135373489	18	30	36	43	60
DailyRate	1470	802.4857143	403.5090999	102	465	802	1157	1499
DistanceFromHome	1470	9.192517007	8.106864436	1	2	7	14	29
Education	1470	2.91292517	1.024164945	1	2	3	4	5
EmployeeCount	1470	1	0	1	1	1	1	1

EmployeeNumber	1470	1024.865306	602.0243348	1	491.25	1020.5	1555.75	2068
EnvironmentSatisfaction	1470	2.721768707	1.093082215	1	2	3	4	4
HourlyRate	1470	65.89115646	20.32942759	30	48	66	83.75	100
JobInvolvement	1470	2.729931973	0.711561143	1	2	3	3	4
JobLevel	1470	2.063945578	1.106939899	1	1	2	3	5
JobSatisfaction	1470	2.728571429	1.102846123	1	2	3	4	4
MonthlyIncome	1470	6502.931293	4707.956783	1009	2911	4919	8379	19999
MonthlyRate	1470	14313.1034	7117.786044	2094	8047	14235.5	20461.5	26999
NumCompaniesWorked	1470	2.693197279	2.498009006	0	1	2	4	9
PercentSalaryHike	1470	15.20952381	3.659937717	11	12	14	18	25
PerformanceRating	1470	3.153741497	0.360823525	3	3	3	3	4
RelationshipSatisfaction	1470	2.712244898	1.081208886	1	2	3	4	4
StandardHours	1470	80	0	80	80	80	80	80
StockOptionLevel	1470	0.793877551	0.852076668	0	0	1	1	3
TotalWorkingYears	1470	11.27959184	7.780781676	0	6	10	15	40
TrainingTimesLastYear	1470	2.799319728	1.289270621	0	2	3	3	6
WorkLifeBalance	1470	2.76122449	0.70647583	1	2	3	3	4
YearsAtCompany	1470	7.008163265	6.126525152	0	3	5	9	40
YearsInCurrentRole	1470	4.229251701	3.623137035	0	2	3	7	18
YearsSinceLastPromotion	1470	2.187755102	3.222430279	0	0	1	3	15
YearsWithCurrManager	1470	4.123129252	3.568136121	0	2	3	7	17

Figure 2.1.2 basic statistics for the dataset

From figure 2.1.2, we can see that there is no NAN value, which means our dataset is completed. Then we can check the mean columns for several features. See highlight rows. Most of them are rated as 1,2,3,4,5 format. So we can check the mean value to see the overall performance.

As for environment satisfaction, job satisfaction and relationship satisfaction, we can see that basically, employees are satisfied in this company, which is a good thing.

As for job level, 2.06, which is a little lower than 2.5. This is reasonable, since the number of managers and directors are surely smaller than employees.

In conclusion, in this step, we can have a whole recognition for our dataset, and then we will do more detailed analysis to find patterns.

2.2 Further Details

Here, we analyse some specific columns to see if there are some interesting patterns.

2.2.1 Department

There are three departments in this company, Sales, Research & Development and Human Resources. They have 446, 961 and 63 employees respectively. This means Research & Development is the main department, also sales department is very important, since according to common sense, sales is a high turnover job, we will confirm this assumption later.

Next, observe the mean salary in different departments, since salary is often a very important factor for turnover. Following are the results.

1) Sales

Here we can see according the average salaries, Sales department has the most significant difference, from 2626 to 16986. Lower income may lead to more turnover.

Sales Executive Average Salary	6924.28
Manager Average Salary	16986.97
Sales Representative Average Salary	2626

2) Research & Development

In this department, employees are divided into three classes, for each class, the mean salary is similar.

Research Scientist Average Salary	3239.97
Laboratory Technician Average Salary	3237.17
Manufacturing Director Average Salary	7295.14
Healthcare Representative Average Salary	7528.76
Research Director Average Salary	16033.55
Manager Average Salary	17130.33

3) Human Resources

Human Resources department only contains two job roles, however, the lower class's mean salary is not the smallest in the company, we can assume that the turnover in this department is not very high. Later we will confirm this assumption.

Human Resources Average Salary	4235.75
Manager Average Salary	18088.64

2.2.2 Job Level

There are five levels for all jobs, 1,2,3,4,5. Following is the average salary for every level. We can see that when your job level is higher, your salary is higher, which is reasonable.

level 1 average monthly salary	2786.92
level 2 average monthly salary	5502.28
level 3 average monthly salary	9817.25
level 4 average monthly salary	15503.78
level 5 average monthly salary	19191.83

2.2.3 Education Field

There are 6 categories for education field, 'Life Sciences', 'Medical', 'Marketing', 'Technical Degree', 'Human Resources' and 'Other'. Figure 2.2.3.1 shows the overall distribution of employees based on different Education field.

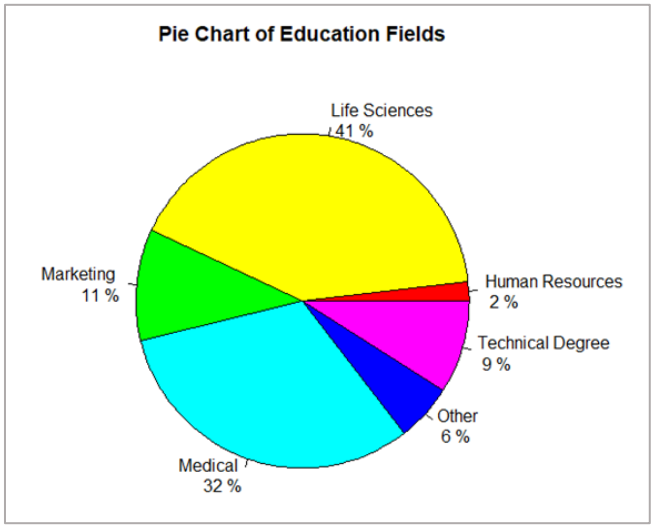

Figure 2.2.3.1 Overall distribution of the Education field (all employees)

In this column, we want to analyse whether one employee's education field matches his/her job, since this is a possible factor which will let employee leave. And I define Match factor by myself, which are match, not match, not clear.

Life sciences, Technical Degree, Medical **matches** Research & Development department

Marketing **matches** Sales

Human Resources **matches** Human Resources

Following is the result for every employee.

	Sales	Research & Development	Human Resources	EducationField	job_match	job_not_match	job_unclear
0	1	0	0	Life Sciences	0	1	0
1	0	1	0	Life Sciences	1	0	0
2	0	1	0	Other	0	0	1
3	0	1	0	Life Sciences	1	0	0
4	0	1	0	Medical	1	0	0
5	0	1	0	Life Sciences	1	0	0
6	0	1	0	Medical	1	0	0
7	0	1	0	Life Sciences	1	0	0
8	0	1	0	Life Sciences	1	0	0

Figure 2.2.3.2 one_hot encoding for education field with current jobs

2.2.4 Gender

We use one hot encoding method to convert gender as 0 and 1. And rename the column as Male, so it is very clear now. In Male column, 1 means male, 0 means female.

2.2.5 Marital Status

As for marital status, we have three categories, single, married, divorced. Still, we use one hot encoding to redefine them. For instance, if one employee is single, then his marital status is [1,0,0] corresponding to three new columns [singe, married, divorced]

2.2.6 Over Time

Similar method to overtime column, we set overtime equals to 1, otherwise, 0.

3. Data Visualization

Until now, we have already transform all columns to numeric columns. So that we can use this updated dataset to do data visualization.

3.1 Correlation Matrix

we plot the correlation matrix of every pair of data and then rank them. Then we choose the most important pairs and analyse their patterns. From the correlation pair ranking, we define high correlation with >0.3, or <-0.3.

3.1.1 Strong positive correlation

	f1	f2	corr_value
302	PerformanceRating	PercentSalaryHike	0.773549996
477	TotalWorkingYears	MonthlyIncome	0.772893246
687	YearsWithCurrManager	YearsAtCompany	0.769212425
636	YearsInCurrentRole	YearsAtCompany	0.758753737
688	YearsWithCurrManager	YearsInCurrentRole	0.714364762
465	TotalWorkingYears	Age	0.680380536
931	job_not_match	Sales	0.654878849
609	YearsAtCompany	TotalWorkingYears	0.628133155
661	YearsSinceLastPromotion	YearsAtCompany	0.618408865
891	job_match	Research & Development	0.613598308
845	job_level_5	MonthlyIncome	0.598335232
662	YearsSinceLastPromotion	YearsInCurrentRole	0.548056248
808	job_level_4	MonthlyIncome	0.533144161
600	YearsAtCompany	MonthlyIncome	0.514284826
689	YearsWithCurrManager	YearsSinceLastPromotion	0.510223636
820	job_level_4	TotalWorkingYears	0.508647767
186	MonthlyIncome	Age	0.497854567
633	YearsInCurrentRole	TotalWorkingYears	0.460364638
684	YearsWithCurrManager	TotalWorkingYears	0.459188397
32	Divorced	StockOptionLevel	0.446284745
857	job_level_5	TotalWorkingYears	0.430751694
658	YearsSinceLastPromotion	TotalWorkingYears	0.404857759
624	YearsInCurrentRole	MonthlyIncome	0.363817667
649	YearsSinceLastPromotion	MonthlyIncome	0.344977638
675	YearsWithCurrManager	MonthlyIncome	0.344078883
795	job_level_4	Age	0.323998708
825	job_level_4	YearsAtCompany	0.313098649
588	YearsAtCompany	Age	0.31130877

Figure 3.1.1.1 strong positive correlation (>0.3)

Analysis:

1) Performance rating versus Percent salary hike, which is true. If an employee performs outstanding, he/she will get more salary in the future.

2) Total working years versus Monthly income, which means that basically, if an employee work in this company for a long time, he/she will get more income.

3) Job not match versus Sales department. This indicates that in Sales department, most employees' education field doesn't match their jobs. This is reasonable in the business world, especially for those sales representatives. Since

they are not required for high degrees or technical skills. However, this will be a potential problem, employees are more likely to change their jobs since their salary is very low and they can easily find other new jobs. I will show this later.

4) Years since last promotion versus Years at company. This indicates that when employees work in this company for a long time, they will need more time to get a promotion. This is a flag that we need to pay attention. We all know it is harder for a director to become a manager, but if it will take many years, some directors may choose to leave, since they have other things to consider. So our company need to think about this situation, that is, not only take care of those basic employees, but also those directors and managers, which are our very important staff.

5) Job match versus Research & Development. this is reasonable.

3.1.2 Strong negative correlation

	f1	f2	corr_value
341	Research & Development	Sales	-0.906818254
948	job_not_match	job_match	-0.855944691
531	Travel_Frequently	Travel_Rarely	-0.753091732
930	job_not_match	Research & Development	-0.703055742
408	Single	StockOptionLevel	-0.638956874
166	Married	Single	-0.629980781
703	job_level_1	MonthlyIncome	-0.604300461
724	job_level_1	job_level_2	-0.578086399
892	job_match	Sales	-0.569789255
715	job_level_1	TotalWorkingYears	-0.530190718
235	Non-Travel	Travel_Rarely	-0.526849874
19	Divorced	Married	-0.491506349
988	job_unclear	job_match	-0.406603337
720	job_level_1	YearsAtCompany	-0.384434838
690	job_level_1	Age	-0.368663637
30	Divorced	Single	-0.366690527
721	job_level_1	YearsInCurrentRole	-0.36249204
723	job_level_1	YearsWithCurrManager	-0.353212187
793	job_level_3	job_level_1	-0.319364126
794	job_level_3	job_level_2	-0.315180098

Figure 3.1.2.1 strong negative correlation (<-0.3)

Analysis:

1) Single versus Stock option level. This is an interesting result. However, it is reasonable. Since for most single employees, they are young, so their working years are short, so their stock option level is low.

2) We can see the several rows of job level 1, their monthly incomes are low, their working years are short, they are very young, etc. These are reasonable.

3.2 Data distribution

We plot every column's data distribution to see overall pattern. Basically, age obeys normal distribution, which is reasonable for this company. Daily rate and monthly rate are nearly evenly. See distance from home, most people live near company. The numbers of employees who get married or not are similar.

3.3 Department distribution

Different departments have different structures and jobs. We have already seen above.

Now we can plot the distribution of employees based on our department. Following is the figure. We can see that, in current employee bar, Research & Development department occupied a large part, which means that this department is very important in this company, we should pay more attention to this department. In no current employee bar, the percentage of Sales department rises. This means that Sales department is more likely to lose employees than other departments.

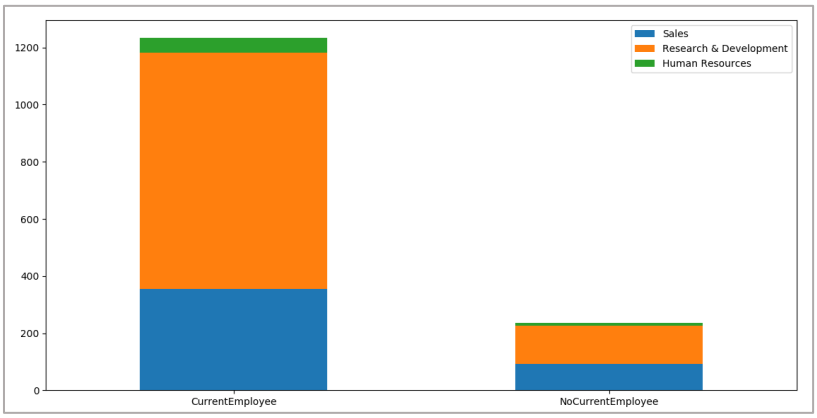

Figure 3.3.1 data distribution based on different departments

3.4 Job Level distribution

Before we have analysed that job levels are corresponding to the salaries. In addition, different job roles have different job levels. So we can know the patterns of job roles by analysing the job levels distribution. Following is the figure about employees for different job levels.

From this figure, we can see that in current employee bar, most employees belong to level 1 and 2, which is reasonable, since in a normal company, most staff are basic employees. From no current employees, we can see that job level 1 occupies the most percentage, this means those employees are more likely to leave. Also, as their job level getting larger, they are more unlikely to leave.

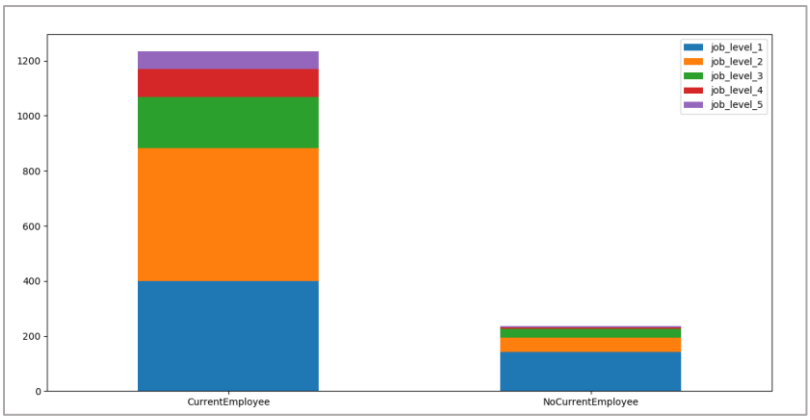

Figure 3.4.1 data distribution based on different job levels

3.5 Overtime distribution

Working overtime is very common in many companies, especially in those technical companies. However, Excessive overtime may affect employees' daily life, so here we want to analyse whether overtime is an important factor for employees.

Following figure shows the overtime distribution, we can see that in current employee bar, most employees don't work overtime, but in no current employee bar, there is no significant difference between work overtime or not.

We can consider in another way, most employees who don't work overtime choose to stay, so it does mean that

work overtime is a pretty important feature for our goal.

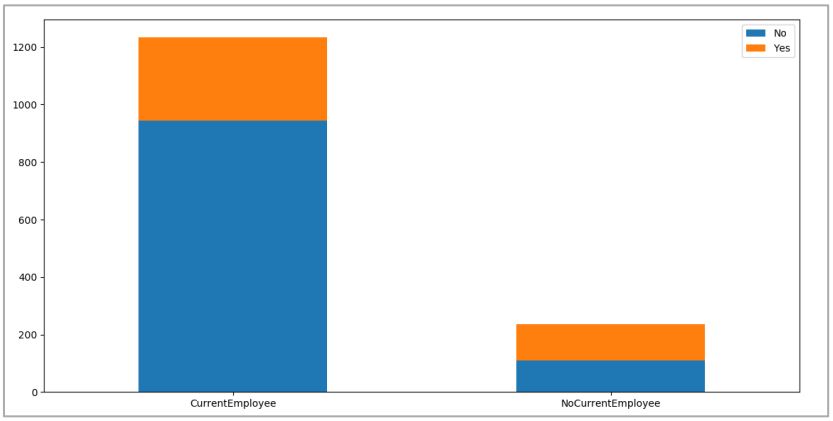

Figure 3.5.1 data distribution based on whether work overtime or not

3.6 Gender distribution

Gender is also a very important feature for a company. It varies in different departments. For example, in technical department, males are main employees, in design department, females are main employees. In addition, males and females have different responsibilities in the world.

From the following figure, we can see, firstly, the number of males is greater than females. Moreover, in no current employee bar, there are more males, which indicates that males are more likely to leave. Females are more likely to stay in a company for a long time.

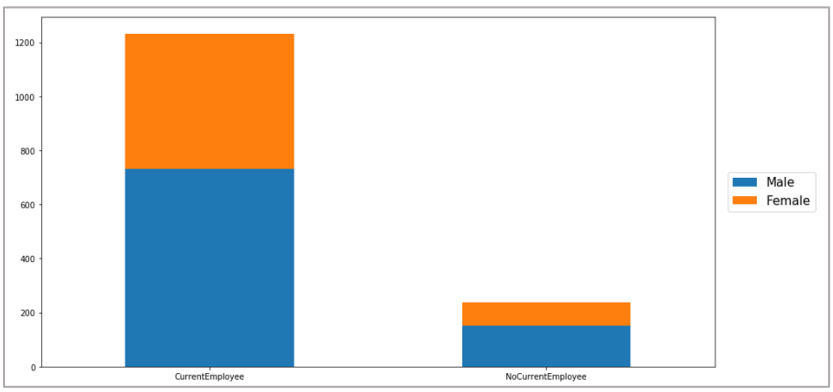

Figure 3.6.1 data distribution based on gender

3.7 Job Match distribution

Generally, whether your education field matches your job is not a very important thing for your job. What matters is your interest and your potential. However, here we consider a common case. We assume that your education field is what you love, so that you are more interested in your work if you choose these corresponding jobs. Then you will be unlikely to leave and would love to pay more attention to your work since you love it. Based on these assumption, we plot this following figure.

From this figure, we can see that, in our company, most employees choose jobs which is corresponding to their education. If you remember the correlation matrix, job match has a high positive correlation with Research & Development, and there are the most employees in this department. So that we can get a result that most employees especially those who are in Research & Development department get their job corresponding to their education field. So this is not a problem in our company.

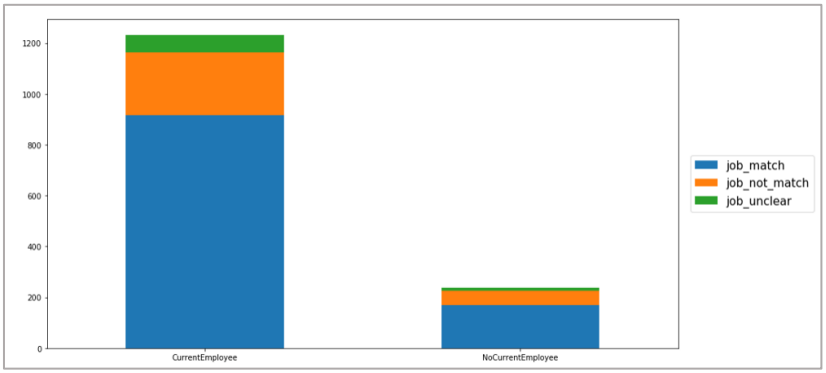

Figure 3.7.1 data distribution based on if education field match jobs or not

3.8 Marital Status distribution

Marital status is also a factor that we want to think about. Since people in different stages have different thought and consideration. If we can find the patterns between them, we will know how to treat them.

Following figure is the marital status distribution. It can be seen that in current employee bar, the married employees occupy the most percentage. This might indicate that for employees who get married, they are more willing to have a stable life, and they don't want to change jobs since it will lead so some risks, and they need to support their family. In no current employee bar, we can see single employees occupy the most percentage, which is reasonable. They are single, so they don't need to consider other things. All they think are themselves. If they can get more satisfactory job, they are more likely to leave.

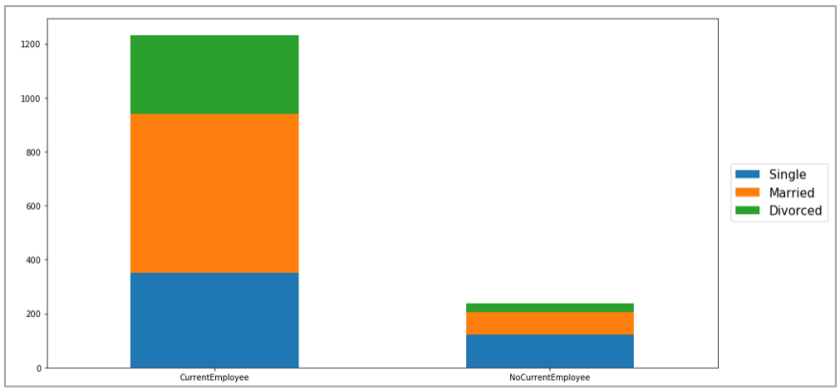

Figure 3.8.1 data distribution based on different marital status

3.9 Education distribution

Plotting education distribution can not only help us to analyze the Education composition of our company, but also observe which group will be more likely to leave, especially for those with high degree.

In our company, we have five education degrees, Below college, College, Bachelor, Master and Doctor. We can see in both current employee bar and no current employee bar, the education distribution is similar. Bachelor is the most, next is Master, then College. Doctor has the smallest percentage.

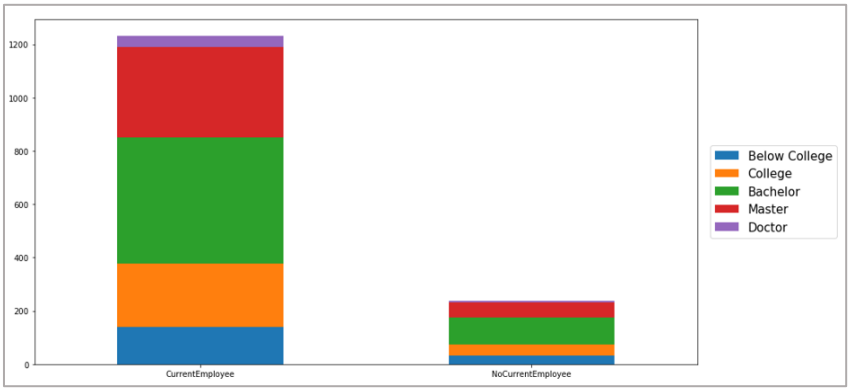

Figure 3.9.1 data distribution based on different education degrees

3.10 Work Life Balance distribution

Currently, people pay more attention to their life and work balance. They prefer those jobs which allow them to spend time with their families and friends. So we choose the work life balance column and plot this following figure.

In this figure, we can see that most employees are able to balance their job and life. And since the distribution of different categories are similar in these two bars, this factor is not important for our goal. Since there are most betters in current employees, of course the most betters in no current employees is reasonable.

As a result, we could say, in our company, most employees are satisfied according to this factor, which is a good news.

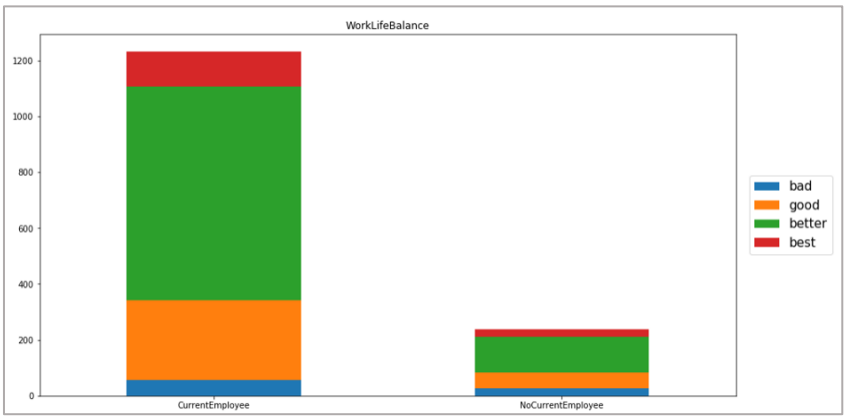

Figure 3.10.1 data distribution based on work life balance (current employee column)

For further research, we also plot this factor with gender distribution, to see whether males and females are different.

In following figure, there is still similar patterns in two bars. So in our company, both males and females are able to balance their life and work.

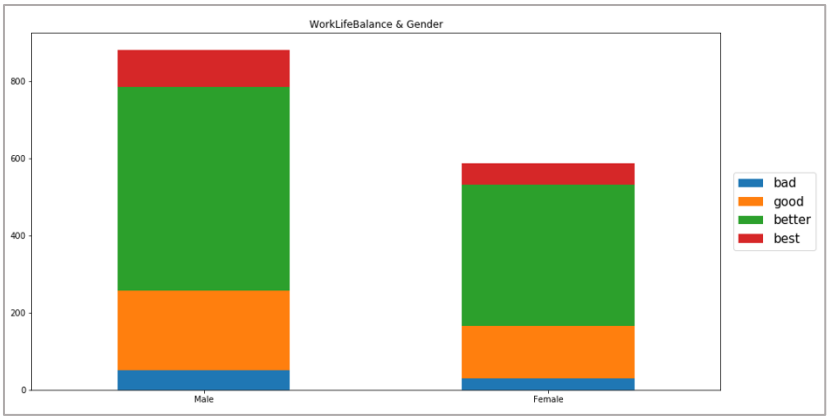

Figure 3.10.2 data distribution based on work life balance (gender column)

3.11 Monthly Income with Age

Next, we want to do analysis about two features and plot scatter plots to see patterns.

From this figure, we can see that those employees who are young and get low monthly income are more likely to leave, see red points in left down corner.

As employees getting older, for instance, see the monthly income 2500 with different ages, although they still get low monthly income, they are not willing to leave. Which indicates that when people are young, they are more likely to work harder and choose higher salary jobs. They consider more about their career.

So we can pay more attention to those young employees, provide more chances for them, training, promotion, etc.

When monthly income is getting larger, all employees are not likely to leave, which means that high monthly income is an important factor.

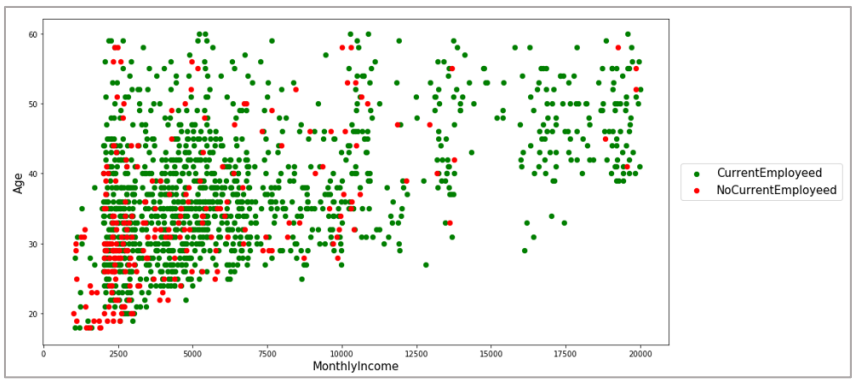

Figure 3.11.1 scatter plot for Monthly Income and Age

3.12 Years Since Last Promotion with Age

Promotion is also a very important factor for our target. We have already analyzed before. Employees will be likely to leave if they cannot get promotion for a long time. Here we plot Years Since Last Promotion with Age scatter plot to see the relationship between them.

Firstly, we can see the left down corner, it contains the most red points, which means no current employees, this result is interesting, it means that even we give those young employees promotion in a short time, they still want to leave. So in this case, we could say, age is a very important factor compared with promotion.

We should pay more attention to those young employees. Talk to them, understand what they think and what they really want.

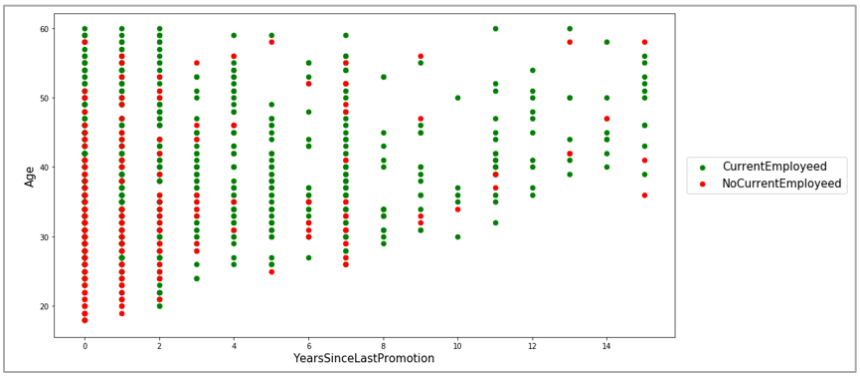

Figure 3.12.1 scatter plot for Years Since Last Promotion and Age

3.13 Monthly Income with Years at Company

According the correlation matrix, we know that monthly income has a strong positive correlation with years at company, so we plot this figure to see what the pattern looks like.

In this figure, we can see that monthly income still plays a very important role in this situation.

When monthly income is lower and years at company is shorter (see left down corner), employees are more likely to leave. When income gets larger, most employees choose to stay. Also we can see that there is no data in the left up corner, which means that if you work in this company for a long time, you are unlikely to get few income.

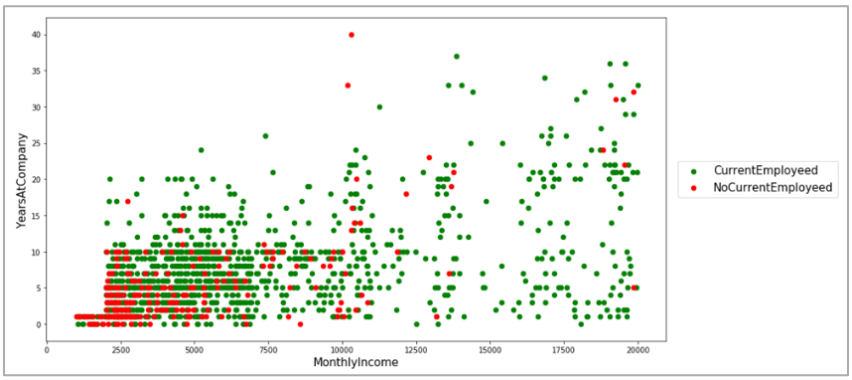

Figure 3.13.1 scatter plot for Monthly Income and Years at Company

3.14 Monthly Income with Stock Option Level

To some extent, stock option level shows the importance and status of an employee. Since we know monthly income is a very important factor, so we compare it with stock option level to see if there are some changes.

Firstly, we can see that low monthly income with low stock option level (see left down corner) has the most no current employees (see red points), which is obvious.

Next, surprisingly, we get a new conclusion. See the left-up corner, and we can choose the 2500 income, you can see the red points get fewer from down to up. This indicates that for our employees, even they have very few income, if they have a very high stock option level, they are more likely to stay and continue to work in our company.

In conclusion, stock option level is also a very important factor for our target. Our company can pay more attention to the influence that this factor gives to our employees.

Figure 3.14.1 scatter plot for Monthly Income and Stock Option Level

3.15 Monthly Income with Training Times Last Year

Currently, most employees especially those young employees are more willing to those companies which can provide them with more training courses. So we plot Monthly Income with Training Times Last Year scatter plot to see the

pattern.

From the following figure, we can see that for employees with different income, they all have chances to attend training courses. In addition, we have most no current employees with 2 or 3 training times. However, we can see that when they have more training times, they are more unlikely to leave.

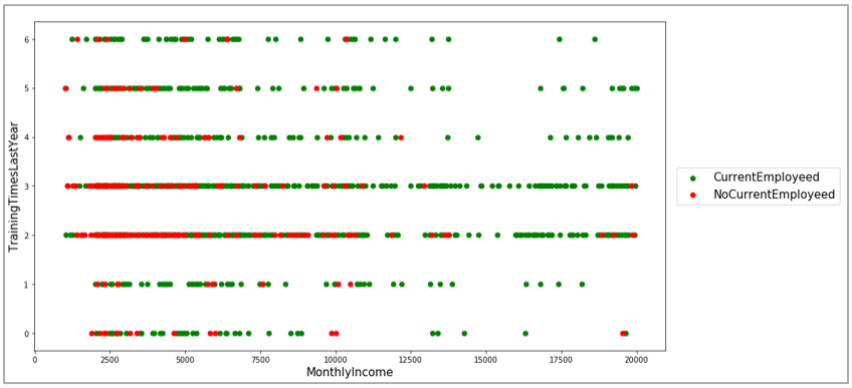

Figure 3.15.1 scatter plot for Monthly Income and Training Times Last Year

3.16 Percent Salary Hike

Percent salary hike indicates the improvement in salary, that is, the potential of your job in the future in this company. So almost every employee cares about this factor. We plot a box plot to see the distribution of percent salary hike among current employees and no current employees.

The result in following figure is obvious, in no current employee box, the percent salary hike is lower. Apart from those low salary job, we need to pay more attention to those employees who are in important stage but have low percent salary hike.

Figure 3.16.1 box plot for Percent Salary Hike (current employee column)

4. Models and Prediction

In this section, we build different models for our dataset. Then we analyze the accuracy of our test data in order to do future prediction.

Our goal is to build a model which can get the pretty high accuracy when we use it to predict new data. In other words, when we have a new employee, and we have all the features for him/her, we can use our model to predict whether he/she is likely to leave or not. Or how to retain our current employees. This is very important for our company.

We use two classifiers, random forest classifier and decision tree classifier. We pay more attention to random forest

classifier.

4.1 Observe dataset

4.1.1 Partition dataset

Firstly, we divide our data into two parts, train set and test set. Train set is used to build models, test set is used to test our models. Since we already known the actual value for the test set, so we can compare the predictive results with actual results to see if our models perform well.

For instance, there is an employee in the test set, which is a current employee, that is 1 (YES). If we put this employee in our model, and we get the prediction yes or 1, it means that our model predicts correctly, otherwise, wrong.

4.1.2 Observe dataset

Table 4.1.2.1 is our result for the current employee column, we can see that this is a biased dataset, so we need to use a parameter in random forest classifier called class_weight to help us deal with this kind of dataset. In addition, we use cross validation to avoid being overfitting.

train set	current employee sample num is 986	no current employee sample num is 190
test set	current employee sample num is 247	no current employee sample num is 47

Table 4.1.2.1 The number of employees in train set and test set

4.2 Importance for every feature

Technically, we cannot use the whole dataset to train our model, since it can be overfitting. Overfitting means our model learns too much from our dataset, and it performs very well in the train set, since it has already known everything. But this can cause a problem, when we use our model in the test set, or in the business world, the performance is not good or even bad. Since test set has many uncertainties, and all these uncertainties may mislead our model to get a wrong conclusion.

As a result, we need to choose the most important features from our dataset, then get a more general model which can fit a general dataset.

Following is the importance ranking figure. From this figure, we define 'importance>0.02' as pretty good features for our target. Now we have a list of important features, they are:

MonthlyIncome, TotalWorkingYears, Age, OverTime, DailyRate, DistanceFromHome, MonthlyRate, HourlyRate, YearsAtCompany, StockOptionLevel, JobSatisfaction, PercentSalaryHike, YearsWithCurrManager, NumCompaniesWorked, YearsInCurrentRole, EnvironmentSatisfaction, YearsSinceLastPromotion, TrainingTimesLastYear, RelationshipSatisfaction, WorkLifeBalance, Education, JobInvolvement. Then we use these features to build models.

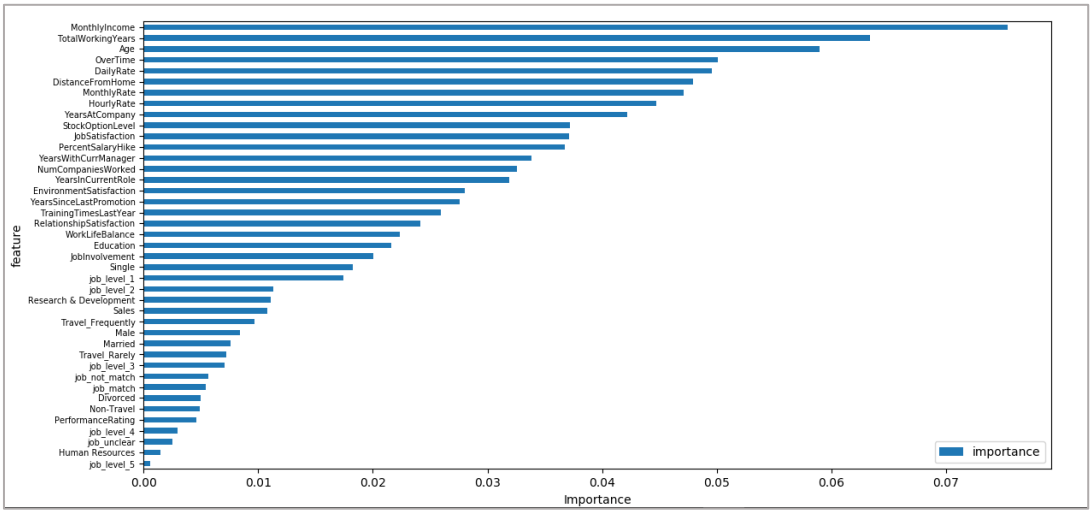

Figure 4.2.1 importance for every feature in the dataset

4.3 Random Forest Classifier

Random forest classifier uses random forest algorithm to do classification. Then we use grid search method to automatically generate a best set of parameters for random forest algorithm.

We will change a parameter called scoring in grid search algorithm to get the best parameters. We have two options for scoring, accuracy and roc_auc. And then we use these parameters to train our model. Then we will get confusion matrix as well as classification report to analyze our models' performance.

Our confusion matrix's format is like the table 4.3.1

	Prediction	
Outcome	C_{00}: True Negative	C_{01}: False Positive
	C_{10}: False Negative	C_{11}: True Positive

Table 4.3.1 standard confusion matrix

4.3.1 Accuracy

We use 'accuracy' as our scoring parameter.

1) Confusion Matrix

	Prediction	
Outcome	TN:11	FP: 36
	FN: 8	TP: 239

Table 4.3.1.1 confusion matrix for the model with accuracy parameter

TN and TP cells mean predicting correctly. So we can calculate the overall accuracy. (8+242) / (8+39+5+242) = 0.85. So this model will get a 85% accuracy.

For our dataset, the current employees are defined as 1. Since we want to predict which employees will leave, so we pay more attention to those no current employees. We now see the accuracy of our prediction to the no current employees, which means that if the employees will leave, can we identify that? 11 / (11+36) = 0.234. This result means that if the employees will leave, we have 23.4% accuracy to identify them.

To some extent, this result is not good, since if we cannot identify those employees who will leave, then we won't take actions to retain them, then we are likely to lose them, which can increase the turnover. So we need to improve

our model.

2) Classification Report

In the classification report, we need to understand three indicators.

① Precision for each category = the number of predicting correctly / the number of predicting that category

That is, precision for 0: TN / (TN + FN), precision for 1: TP / (TP + FP)

② Recall for each category = the number of predicting correctly / the number of prediction

That is, recall for 0: TN / (TN + FP), recall for 1: TP / (TP + FN)

③ F1 = 2 * precision * recall / (precision + recall)

test set				
	precision	recall	f1-score	support
0	0.58	0.23	0.33	47
1	0.87	0.97	0.92	247
avg / total	0.82	0.85	0.82	294
train set				
	precision	recall	f1-score	support
0	0.97	0.95	0.96	190
1	0.99	0.99	0.99	986
avg / total	0.99	0.99	0.99	1176

Table 4.3.1.2 classification report for the model with accuracy parameter

We can see from the table above, in train set, every indicator is very high, but in our test set, the recall for 0 is very low, 0.23. This means our model may be overfitting, or we still cannot perfectly deal with our unbalanced dataset.

4.3.2 ROC_AUC

We use another criteria as our scoring parameter, roc_auc, and get the result.

1) Confusion Matrix

In Table 4.3.2.1, we can see the whole accuracy is (8+242) / (8+242+39+5) = 85.03%.

	Prediction	
Outcome	TN:8	FP: 39
	FN: 5	TP: 242

Table 4.3.2.1 confusion matrix for the model with roc_auc parameter

2) Classification Report

In this classification report, we get a similar result. In addition, roc_auc parameter performs worse than accuracy, since the recall of 0 in test set is very low, only 17%.

test set				
	precision	recall	f1-score	support
0	0.62	0.17	0.27	47
1	0.86	0.98	0.92	247
avg / total	0.82	0.85	0.81	294

train set				
	precision	recall	f1-score	support
0	1	0.99	1	190
1	1	1	1	986
avg / total	1	1	1	1176

Table 4.3.2.2 classification report for the model with roc_auc parameter

3) ROC Curve

See the following figure. ROC curve is used to illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. AUC value is equivalent to the probability that a randomly chosen positive example is ranked higher than a randomly chosen negative example.

See following figure, this model does work, since ROC is above the diagonal (blue dashed). And AUC area is 0.74. The greater the AUC is, the better our model performs.

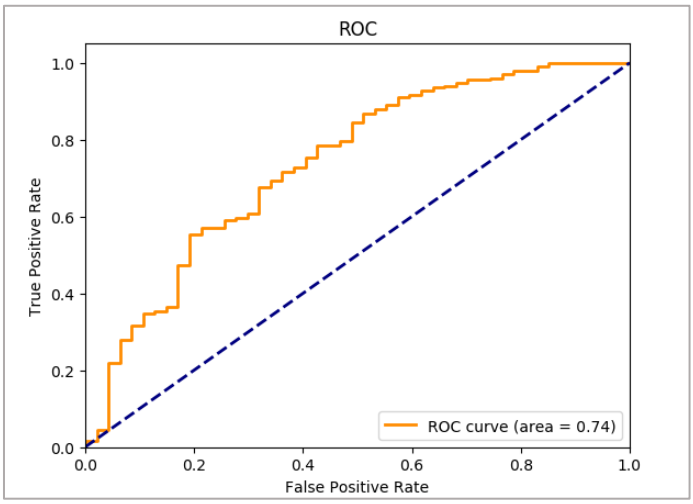

Figure 4.3.2.3 ROC curve

4.4 Decision Tree Classifier

Decision tree is a type of supervised learning algorithm (having a pre-defined target variable) that is mostly used in classification problems.

We can use decision tree to classify our dataset. A good decision tree can help us to correctly identify what kind of data belong to our target.

During our data mining process, we use the most important features in our decision tree and train our model. We have tried different parameters in the Decision Tree Classifier, and then we choose one of the decision tree in our report to interpret it meaning. In this model, we pre-process our unbalanced dataset, and limit the max leaf in the decision tree to avoid being overfitting.

4.4.1 Interpretation

Class 1 means current employees, class 0 means no current employees.

left hand side means true, right hand side means false.

Gini index says, if we select two items from a population at random then they must be of same class and probability for this is 1 if population is pure. So Gini index means the probability of an item which is misclassified.

Sample means the number of our current data.

4.4.2 Analysis

Following is our chosen decision tree. In this decision tree, we analyse the condition of class=0 to predict which employees will leave, and class=1 to consider how to retain our good employees.

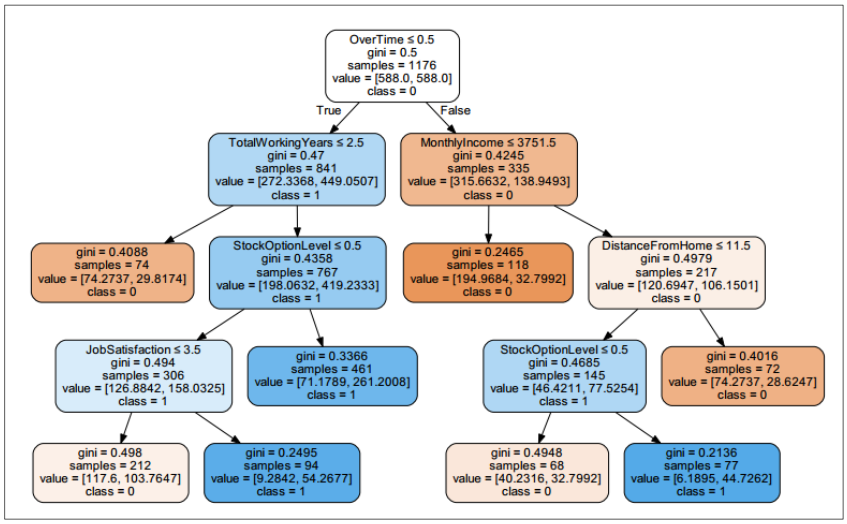

Figure 4.4.2.1 decision tree with chosen features and proper parameters

1) Class = 0 (No current employees)

Employees who work overtime may leave, no matter how many monthly income they will get. So work overtime is a very important factor.

Employees who work overtime and their monthly income is less than 3751.5 may leave.

Employees who work overtime and their home is far from our company, although their monthly income is greater than 3751.5, they may leave.

Employees who work overtime, monthly income is greater than 3751.5, their live not far from their company, and their stock option level is low, they may leave.

Employees who don't work overtime but work in our company for a short time may leave.

Employees who don't work overtime, they work in our company more than 2.5 years, their stock option level is less than 0.5, their job satisfaction is less than 3.5, they may leave.

2) Class = 1 (Current employees)

Employees who don't work overtime and work in our company more than 2.5 years may stay.

Continue to the situation above, if they are provided for a high stock option level, most of them may stay. In addition, if they are satisfied with their job, they are more likely to stay.

Employees who work overtime, but have high monthly income and live not far from our company, they still are likely to leave, if they are provided with a high stock option level, they are more likely to stay.

5. Conclusions and Recommendations

5.1 Conclusions

In conclusion, we conduct a data mining based on the employee dataset. We summarize some reasons for the employee departure. And then, we answer the questions that HR team have asked before.

5.1.1 What staff members are likely to leave next

Based on our analysis, employees who are likely to leave may have these features.

1) Young employees

   Young employees are emotional and ambitious, they are willing to learning new things. However, some of them are single, and are not able to settle in a city for a short time.

   As a result, if their salary is very low and cannot get many training courses in company, they are likely to leave. In addition, single status means that they are not bound with their family, so they have many choices, in this case, some employees may leave if they have a better chance for their jobs.

2) Insufficient career development space

   As the increase of employees' working years and working skills, a clear career development path plays a very important role in retaining outstanding employees. According to our analysis, those employees who cannot get a promotion for a long time are likely to leave.

3) Low-level employees

   Apart from the salary, stock option level and job level gradually become very important in a company.

   Based on our analysis, employees whose job level is low are likely to leave. Based on our decision tree, although some employees' monthly income is greater than 3751.5, and they live not far from their company, if their stock option level is low, they may leave. In the contrary, some employees with low salary, but they have high stock option level, they may continue to stay in their current company.

4) Low salary

   Apparently, salary is everyone's mainly focus in their work. Employees with low salary may leave, however, the number of young employees is smaller than old employees, which means age affect this result as well. In addition, employees with low percent salary hike, especially for young people, may leave.

5) Work overtime

   From our decision tree, we can see that employees who work overtime may leave, no matter how many monthly incomes they will get. So work overtime is a very important factor, Especially for those who have low salary.

6) Male employees

   Males are more likely to leave compared with females.

7) Live far from home to company

   Based on our analysis, most employees live not far from our company, however, from our decision tree, employees who work overtime and their home is far from our company, although their monthly income is greater than 3751.5, they may leave.

8) Work for a short time

   Those employees who work in a company for a short time can hardly build loyalty for their current company. according to our result, employees who don't work overtime but work in our company for a short time may leave.

5.1.2 Which department is most at risk to lose employees

Sales department is more likely to lose their employees. For some following reasons:

1) Employees' salary in this department varies from 2626 to 16986, sales representatives have the lowest salary, in order to have a better life, they are more likely to leave and search for a job with higher salary.
2) Employees in Sales department are not required to have specific technical skills. Even those who are below college still can get this job. Some of them don't consider their job as a career. Based on this situation, they are more likely to leave when they are not satisfied with their current jobs.
3) Most employees in Sales department do the work which is not related to their education. Based on our analysis, most current employees have a common feature, their work matches their education.

5.2 Recommendations

What strategies can we employ to try and keep our current employees?

1) Provide more training courses for employees
   Employees especially those young employees are enthusiastic about their work, they are more willing to rich themselves. So our company can provide some training courses for them, help them to improve their working skills.
2) Increase salary for outstanding staff
   According to our analysis, salary has positive correlation with working years. However, we should pay more attention to those young but outstanding employees. Although they work for a short time, they can make a huge contribution for our company. If we give them more chances, they are more willing to work here.
3) Adjust stock option level
   We know stock option level plays a very important role for employees' stay or leave. So we can adjust the structure of stock option level. For those employees who create a huge profit for our company, we can properly increase their stock option level to encourage them.
4) Treat differently to males and females
   Males focus more on their career, so they are more interested in salary, job level, percent salary hike, etc. Females focus more on their work life balance, so they are more interested in overtime, distance from home, etc. Our HR team identify what they really want, give them what they want.
5) Provide good working environment
   We can provide multiple benefit for our employees. For instance, different kinds of insurance, welfare, gifts for different festivals, etc. If employees feel our company as a whole family, they are unlikely to leave. Since, money is not all.
6) Build a corporate core values
   A company should have its own core value. It is the key of a company. If employees agree with you, they may work for you even they are not provided adequate salary.
7) Concern new employees
   New employees are not familiar with our company. And their work efficiency may be very low in the first 3 to 6 months. We need to pay more attention to them, especially to those young employees, help them to join in their team as soon as possible.
8) Benefit for overtime
   Based on our analysis, overtime is a very important factor. We shouldn't encourage overtime, however, if employees really need to work overtime, we can give them some benefit, for example, food, allowance, or relaxation.