Large Scale Data Analytics
ENGI 5631

Project: Human Resource Analysis

Submitted by Sahil Dhankhad

Student ID: 0674692

Submitted to Dr. Emad Mohammed

Table of content

Abstract

Visualization of data method to understand the data perfectly has been done in this project. A perfect model to understand the reasons why a person leaves a job at company has been developed in this project with the help of data. The model developed is then applied on the company's current employees to predict the reasons for people leaving the company and preventing them from doing so. Different methods were used to find out the risk factors responsible for employees leaving the job at the company. The model was then used for the current employees and comparison was done. From the model the company can predict which employees are more susceptible to leave the job soon and things can be done to prevent the skillful employees to leave the company. The company can then work on the risk factors that are making the experienced employees leave the job.

Keywords: Decision tree, Random forest, Prediction model.

Chapter 1

Introduction

According to a lot of people and experts, an employee leaving a job at a company is very unpredictable. But there are some ways and methods to predict the factors responsible for this problem. Statistical procedures can be used for this prediction. This project shows how employee leaving a company can be predicted using different models such as logistic regression, decision tree, random forest, Gaussian naïve Bayes.

Employees are the ones who shape the culture of a company, they are the ones who do most of the work and act as the pillars of the company on which the whole company depends. If a company starts losing its experienced employees at a high rate, then it shows that there is something wrong within the company. For a company to be successful they need to have a safe and healthy environment for their employees and if a lot of employees leave the job that sends a very negative message to the rest of the workers working in the company which in turn affects the success of the company and hinders its growth.

In a company's balance sheet, staffing is the highest expense. A lot of money is invested in employee salaries, incentives, benefits, training, etc. That is why when a company loses its employees a lot of its investments are drained away and the environment in the company also degrades as well. A healthy environment plays a very important role in the successful achievement of the goals that the company aims to achieve.

Understanding the factors responsible for employee leaving the company or having knowledge about them allows the company to limit this problem from happening again and again. This increases the employee productivity and company environment. There are several factors that influence the employee leaving a company such as relationship with supervisor and co-workers, job satisfaction, salary, incentives, work load, etc. The individual departments can also be analysed to know more about the risk factors.

The predictivity model gives the higher authorities an opportunity to take necessary steps to prevent the employees from leaving the company. Having an employee prediction model gives a better approach for retention strategies of employees.

Chapter 2

Literature Review

Large scale data is a detailed version of a very large amount of data that can be accessed easily for prediction of things that might happen in the future depending upon the data that is with someone in the present.

In this project I am going to predict whether am employee is going to leave or stay at a company on the basis of data that is been provided by Kaggle.com (Anon n.d.). The first step after a data is collected is data preparation. It is one of the most fundamental stages of data analysis. The main aim is to get cleaned data so that it is used for many valuable purpose (Zhang et al. 2003).

Correlation matrix tests the correlation between two different variables. It can tell us how one variable is dependent on the other and vice versa (Dziuban & Shirkey 1974). With the help of different models the data collected is analysed and predictions are done (Silahtaroglu & Donertasli 2015).

Logistic regression was done to organize data analysis (Dayton 1992). There are two possible outcomes of logistic regression (Silahtaroglu & Donertasli 2015). Decision tree uses a tree like structure of decision model which includes the chances of event outcomes (Inese et al. 2010). In this project decision tree has been used.

Chapter 3

Material and Methods

Data Collection

The data used in this report is collected from Human Resource Analytics dataset provided from the Kaggle website. The whole data is provided by a company so, the data is related to real world scenario. The data has information of 14,999 employees.

```
: mydataset = pd.read_csv('HR_comma_sep.csv')

: mydataset.shape

: (14999, 10)
```

Figure 1: Loading the data.

Cleaning and preparation

The data is loaded into the system before its preparation. Data preparation is an initial step and is very important before the analysis of the data is done. Data preparation is a lot of work and can be a difficult job. The data has to be easily understandable before proceeding to the next step. Kaggle is clean and contains no missing values. Having no missing values saves us from calculating and replacing the missing data. It should be made sure that the data is readable and all the values match with each other.

1. Conversion of numerical data into categorical data, like satisfaction, department and focus of work
2. Renaming some features if you want for better readability.
3. Data cleaning is involved to check for any missing values and invalid points.
4. There is one dataset named “left”. This data set contain the option of “o” and “1” depending on whether the employee stays or leaves. So, check that dataset. For us it sa dependent variable.
5. Convert if we have any string value into binary value. So, our model can run smoothly without having any trouble.

```
# now i'm converting string value into Binary form; so we can use them in our proejct
mydataset['salary'].replace({'low':1,'medium':2,'high':3},inplace=True)
mydataset['sales'].replace({'IT':11,'RandD':12,'accounting':13,'hr':14,'management':15,'marketing':16,'product_mng':17
```

Figure 2: Converting string to binary.

In the data there are some independent variables present that will be used in the model. The following factors will be considered in the data:

1. **Satisfaction:** Employee satisfaction with the job.
2. **Evaluation:** Time to time performance evaluation of employees.
3. **Project Count:** Number of projects that has been undertaken by an employee and successfully completed
4. **Average monthly hours:** Average number of hours worked in a month.
5. **Number of years worked in the company:** How long a person has been doing the job for.

6. **Work accidents:** Whether an employee had any work-related accidents or not.
7. **Promotions:** any promotions given to the employee since his/her joining the company.
8. **Department:** The department to which employee belongs to, for example HR, sales, marketing etc.
9. **Salary:** The type of salary being received by the employee.

	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_last_5years	sales	salary
0	0.38	0.53	2	157	3	0	1	0	sales	low
1	0.80	0.86	5	262	6	0	1	0	sales	medium
2	0.11	0.88	7	272	4	0	1	0	sales	medium
3	0.72	0.87	5	223	5	0	1	0	sales	low
4	0.37	0.52	2	159	3	0	1	0	sales	low

Figure 3: Data Features.

Correlation Matrix

Correlation matrix is used to perform the investigation between multiple variables at the same time. A results table is obtained after that which contains the correlation coefficients between each variable and others.

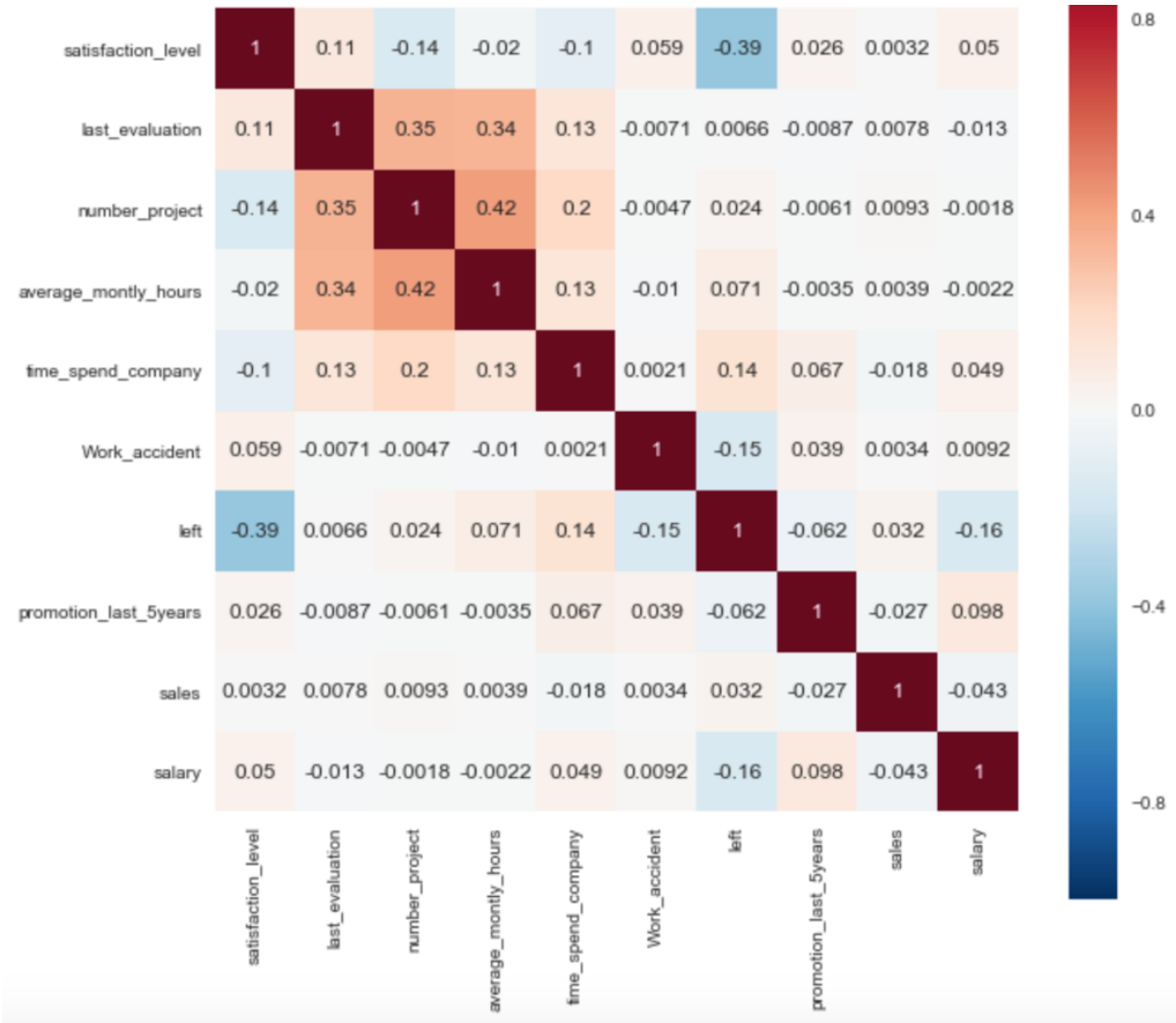


Figure 4: Correlation Matrix.

In other words, it measures the linear relationship between the two variables. After looking at the head map both the negative and positive correlation can be identified.
In case of a positive correlation, r > 0, which means both the variables move in the same direction. In case of negative correlation, r<0, which means the variables are moving in different directions.
No relationship between the variables is indicated by zero correlation and a good relationship between the variables is indicated by one correlation.

1. Evaluation and satisfaction level scores are high which shows that less number of employees left the company because of this reason.
2. Getting promotions makes people happy and they are less likely to leave the job.
3. The higher the number of projects in hand the more will be the average monthly hours spend in office and it results in good evaluation but on the other hand it makes employees unhappy as they are spending more time at work place and have less time for themselves.

Correlation matrix provides a little bit of the idea about the factors that can be considered by a person in his/her project. The mean of all the independent variables of employee database gives a rough idea about the number of each variable. For example, satisfaction variable has 61% of satisfied employees.

Probability of Employee leaving a department:

Information about the different departments in the company has also been given in the data indicating the number of employees working in each department. Data related to these departments was taken and analysed.

		satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	promotion_last_5years
sales	left							
IT	0	0.677170	0.713050	3.756813	198.886792	3.356394	0.162474	0.000000
	1	0.411868	0.730037	4.025641	213.849817	3.860806	0.032967	0.010989
RandD	0	0.653799	0.706081	3.822823	198.951952	3.252252	0.186186	0.040541
	1	0.432810	0.745372	4.024793	210.975207	4.000000	0.082645	0.000000
accounting	0	0.647211	0.726128	3.808171	199.037300	3.424512	0.154529	0.024867
	1	0.402598	0.694510	3.872549	207.029412	3.794118	0.044118	0.000000
hr	0	0.666679	0.720802	3.702290	199.250000	3.192748	0.156489	0.028626
	1	0.433395	0.679721	3.539535	197.306977	3.753488	0.032558	0.000000
management	0	0.654861	0.723451	3.812616	200.233766	4.395176	0.181818	0.122449
	1	0.422857	0.727253	4.142857	207.263736	3.758242	0.054945	0.032967
marketing	0	0.669878	0.723282	3.720611	198.888550	3.480916	0.195420	0.065649
	1	0.453153	0.692020	3.581281	200.990148	3.857143	0.049261	0.000000
product_mng	0	0.658466	0.711435	3.795455	197.765625	3.330966	0.174716	0.000000
	1	0.481566	0.726566	3.848485	207.787879	3.989899	0.045455	0.000000
sales	0	0.668548	0.709223	3.789187	199.571657	3.443698	0.173065	0.029750
	1	0.447663	0.711243	3.736686	205.041420	3.812623	0.045365	0.006903
support	0	0.673799	0.721714	3.783751	199.140980	3.213859	0.187575	0.010155
	1	0.450901	0.727315	3.864865	205.636036	3.933333	0.055856	0.005405
technical	0	0.668319	0.716609	3.814632	198.471083	3.222442	0.172022	0.012358
	1	0.432525	0.734132	4.061693	214.183644	3.959828	0.047346	0.004304

Figure 5: Percentage Probability of an employee leaving the company.

From the picture, satisfaction levels of employees who belong to the accounts department can be seen. They are either staying or leaving depending upon the satisfaction level they have with their job.

- **Staying (0):** If the satisfaction level of an employee in the accounts department is 0.64 (64%) then they are going to stay in the company.
- **Leaving (1):** If the satisfaction level of an employee in the accounts department is 0.43 (43%) then they are going to leave the company.

Other variables for the accounts department can also be taken under consideration to help in the decision-making process of whether an employee will stay back in the company or leave it soon.

Department Vs Salary

For each department there are three different levels of salary that has been taken into consideration; low, medium and high. A graph showing various employees in each department having different salary levels have been plotted.

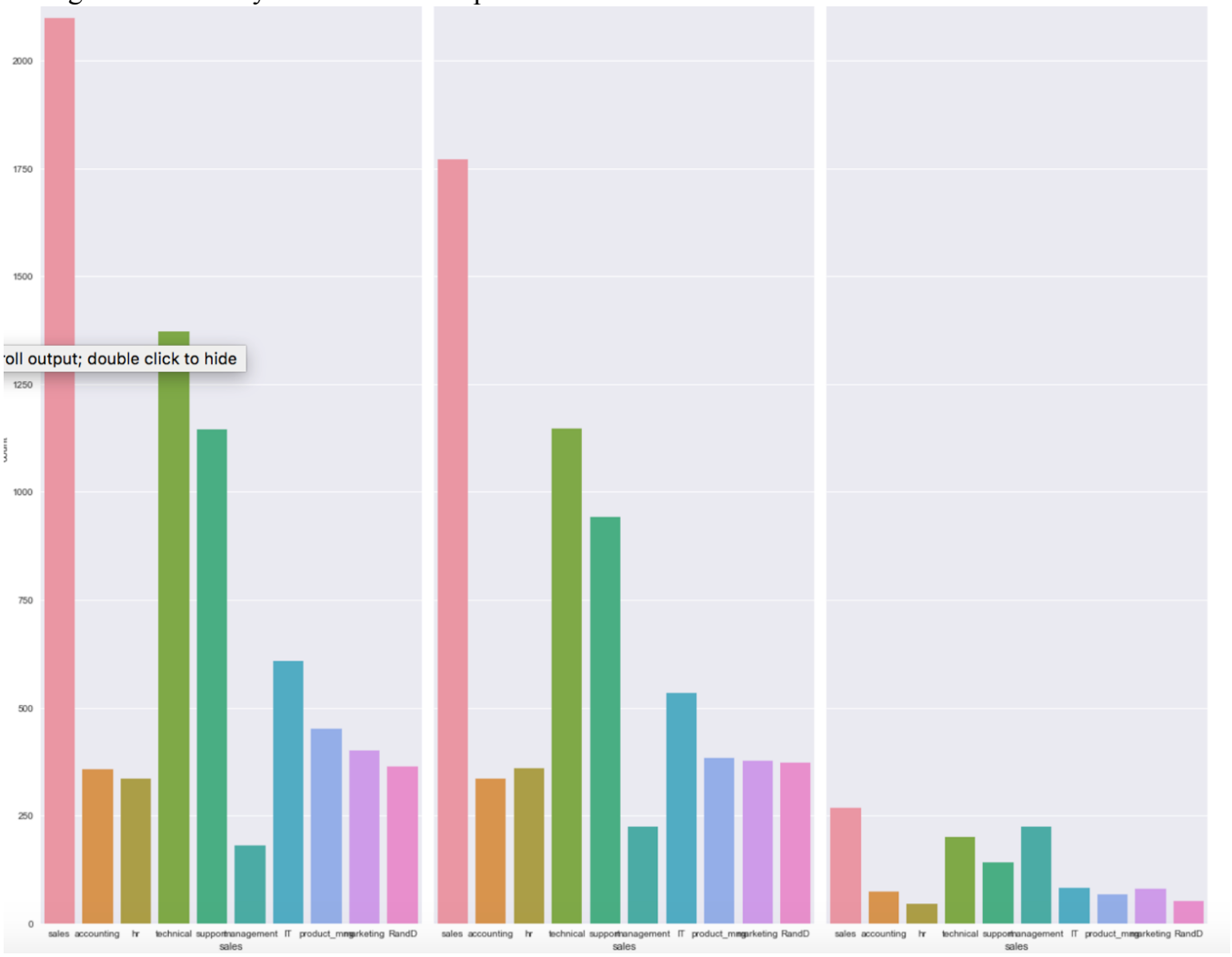
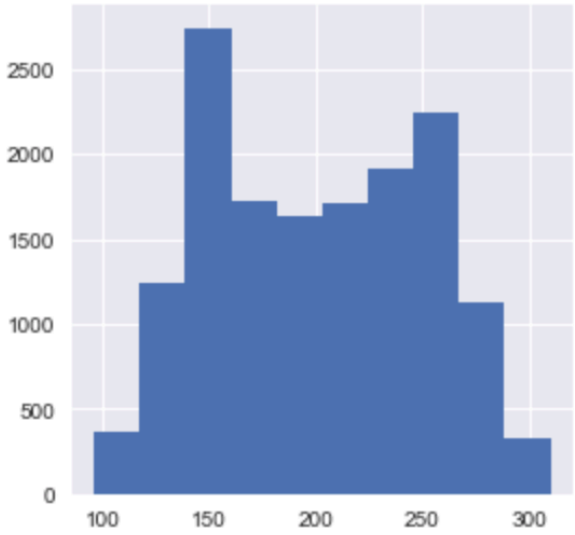


Figure 6: Salary distribution in departments.

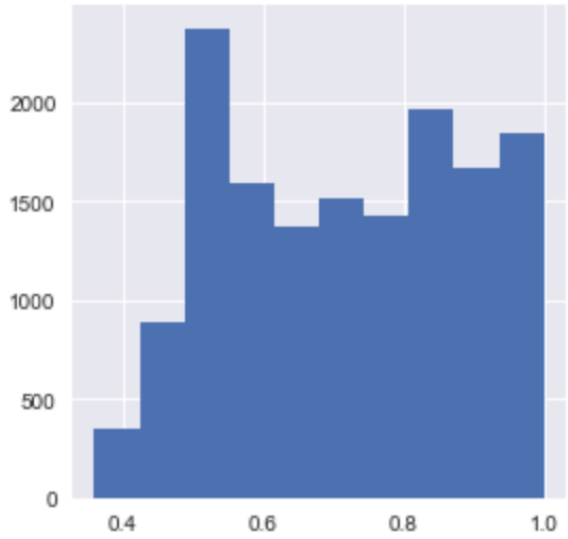
- **Low Salary:** Sales department has the most number of employees with low salary followed by the Management department.
- **Medium Salary:** Again, sales department has the highest number of employees getting medium salary as compared to other departments. Management department has the second highest number of employees getting medium salary.
- **High Salary:** Management department has more number of employees which receive higher salary than the employees in the other departments.

Histogram of all variables

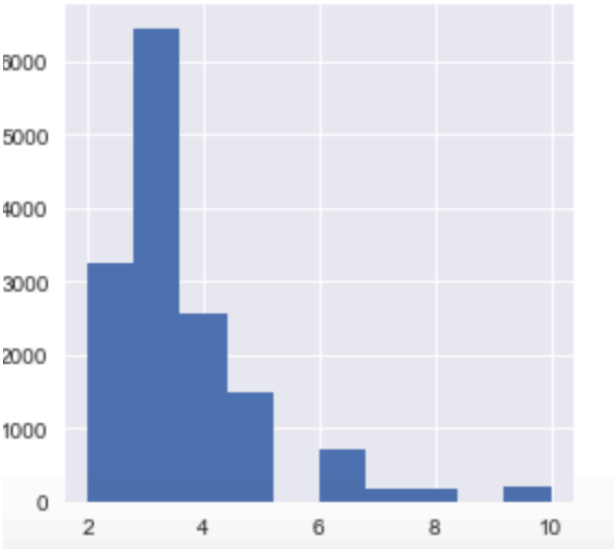
To know the distribution of data amongst the different levels histograms are plotted.



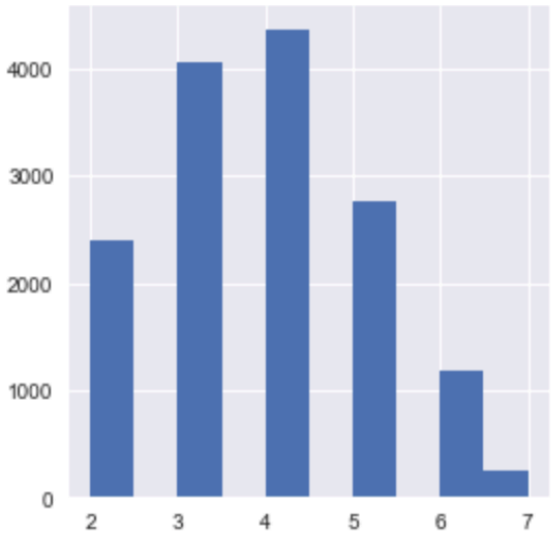
(A)



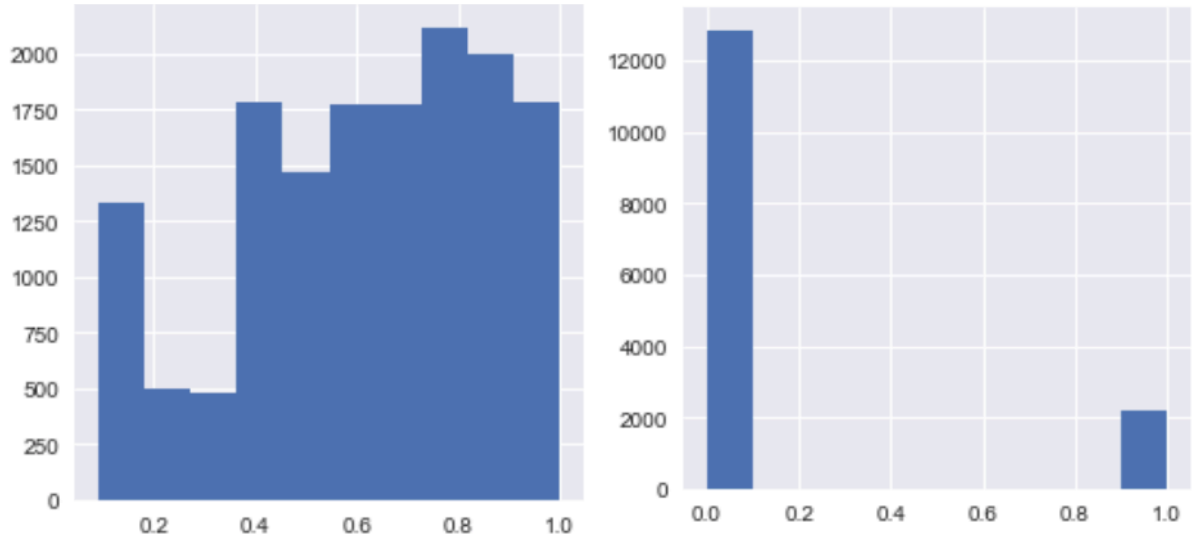
(B)



(C)



(D)



(E) (F)

Figure 7: Histograms. (A) Average monthly hours; (B) Last evaluation; (C) Time spend in company; (D) Number of projects; (E) Satisfaction level; (F) Work accidents.

- **Satisfaction:** There are three different levels on which the satisfaction of an employee with his/her job can be seen. They are:
    a) Level 1: 0-0.3
    b) Level 2: 0.3-0.5
    c) Level 3: 0.5-1.0
- **Evaluation:** In evaluation there are three different levels on which the employees will be judged. They are as follows:
    a) Level 1: 0-0.55
    b) Level 2: 0.55-0.7
    c) Level 3: 0.7-1.0
- **Average monthly hours:** Three groups can be seen in this. They are:
    a) Group 1: 100-150 hours
    b) Group 2: 150-250 hours
    c) Group 3: 250-300 hours
- **Work accidents:** This variable has been divided into two categories.
    a) Zero (0): Employees who does not have any job accidents.
    b) One (1): Employees who have work accidents.
- **Promotions:** This variable has been divided into two categories:
    a) First: people who doesn't have any promotions.
    b) Second: people who got promoted in last five years.

Department Vs Left (employees leaving)

Even though the top three departments which have the highest number of employees are sales, technical and support then also accounts department is the department where the number od employees leaving the company are the most.

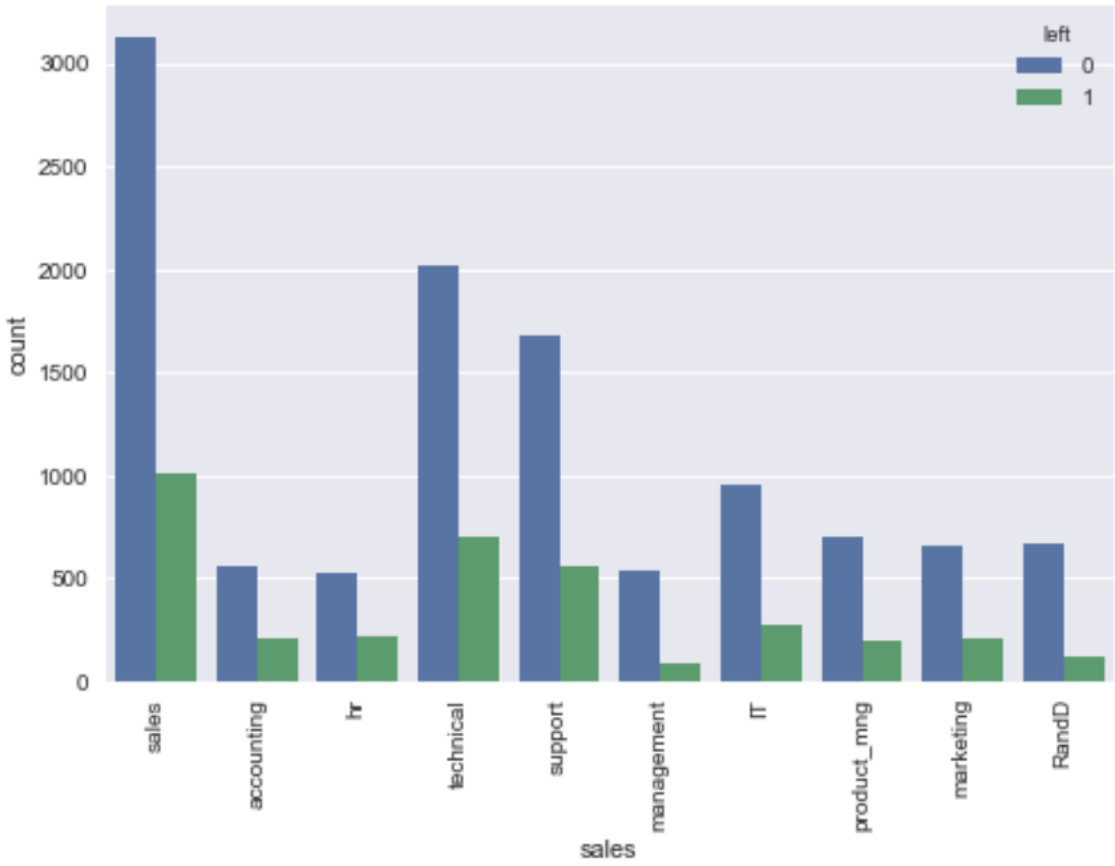


Figure 8: Number of people left from different departments.

Projects Vs Left (employees leaving)

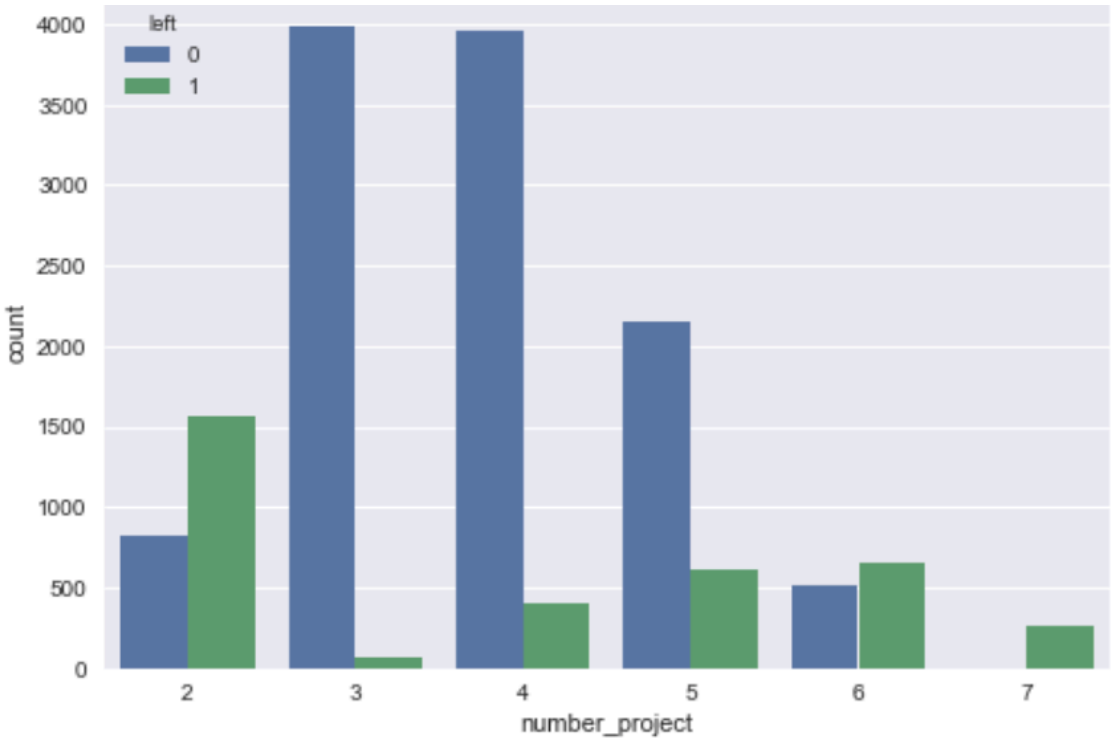


Figure 9: Number of people doing project and still they left the company.

From the it can be concluded that more than half of the employees who were working on 2,6 or 7 number of projects were the ones who left the company. Employees working on 3 projects are the least number of employees who left the company. It is followed by the employees working on 5 and 6 number of projects respectively. Employees who had 7 projects left the company in the end.
So, employees who had 2 projects they left the company because they felt they had less work to do ad hence had less number of hours. On the other hand, in case of employees having more than 5 projects left the company because they were over working and their schedule was too hectic for them to manage.

Projects Vs Salary

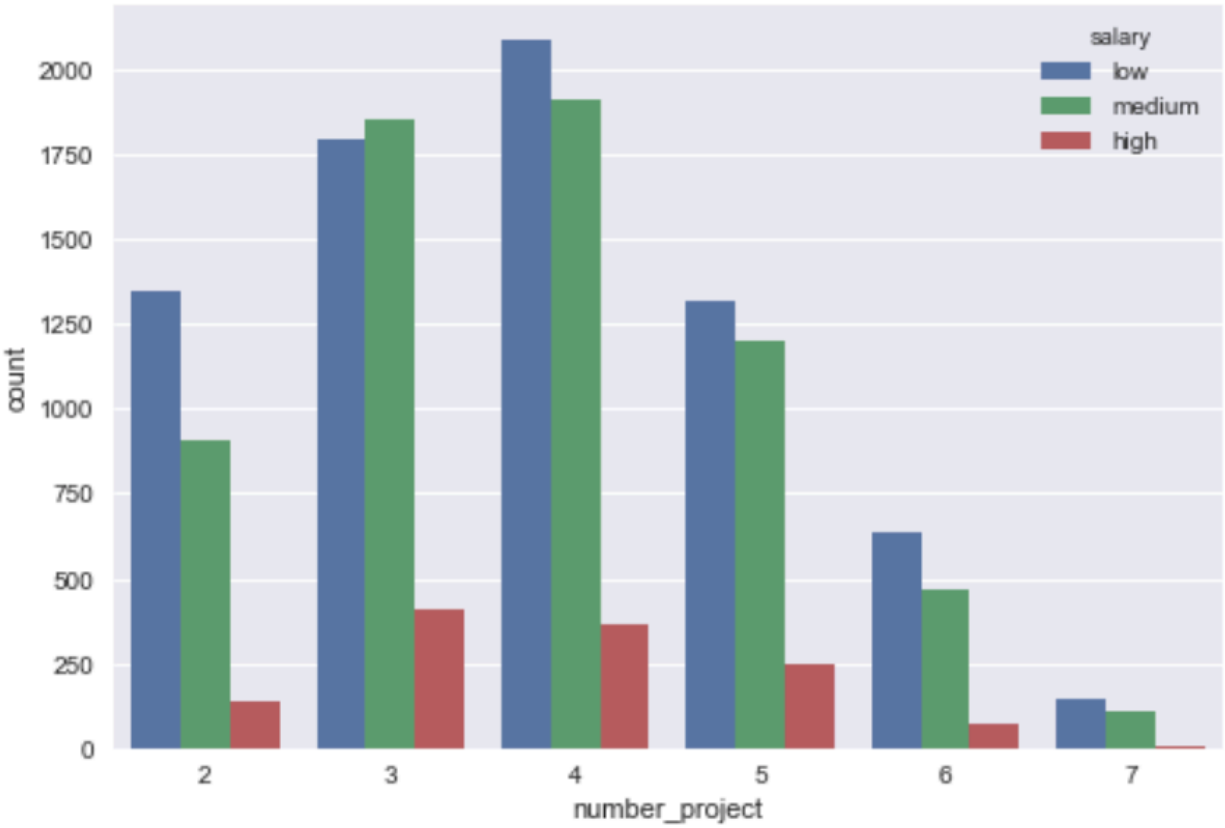


Figure 10: Number of project Vs Salary.

From the data provided in figure 10 it can be seen that the employees doing 3 projects have more salary than employees having 6 projects. Most number of employees in the company work on 4 projects and they receive medium to low salary. It can be concluded from the data that the employees doing more number of projects are considered less valuable by the company and hence they receive low salaries.

Projects and average monthly hours Vs left (employees leaving)

Employees who stayed at the company are doing 2 to 6 projects and are investing 180-210 hours in a month.

From the graph which shows the employee who left the company, as the number of project increases so does the average monthly hours increases which in turn promotes the increase in employees leaving the company.
People who have less projects get less number of hours which causes them to leave the company.
People who have 3 projects they put in approximately 195 hours stayed back but if their average monthly hours increases to 220 hours then they leave the company.

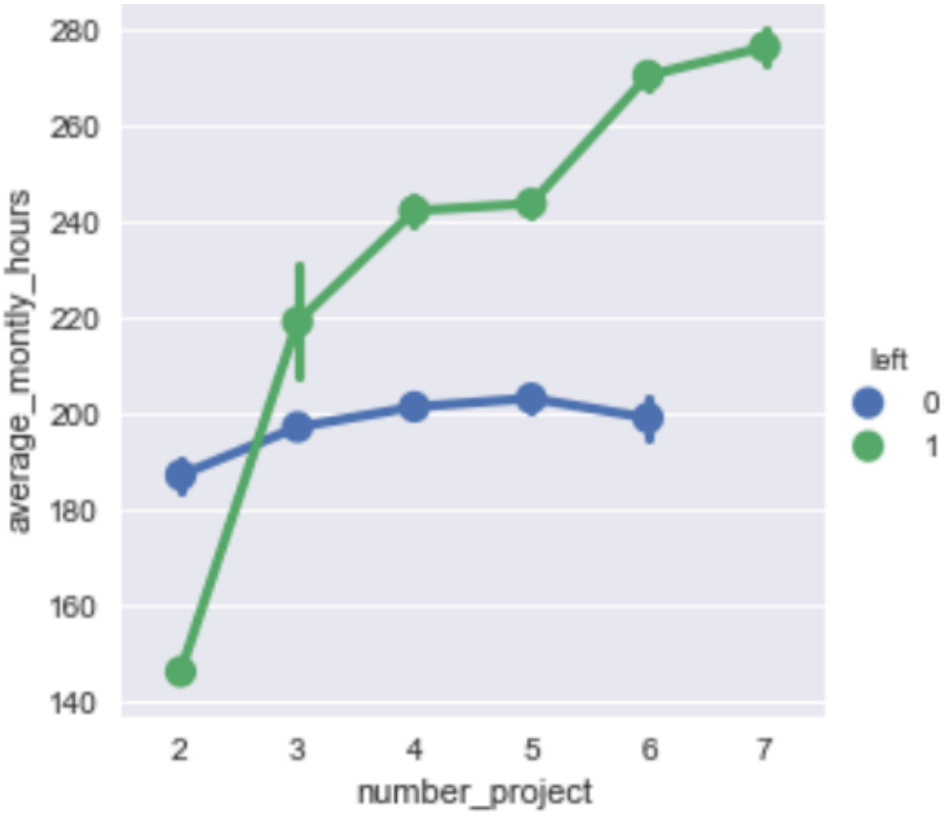


Figure 11: Employees left on considering the number of projects and average monthly hours.

People having more than 3 projects put in approximately 210 hours which in turns becomes the reason for them to leave the company.
People working on more projects have more working hours which results on over time and causes them to leave the company.

Projects and Satisfaction vs left

Employees having number of projects between 2 to 5 and have satisfaction between 0.6 to 0.7 won't leave the company. Employees who are working on 6 to 7 projects have the lowest satisfaction level. People who work on 4 to 5 projects have the highest satisfaction level but it does not stop them from leaving the company.

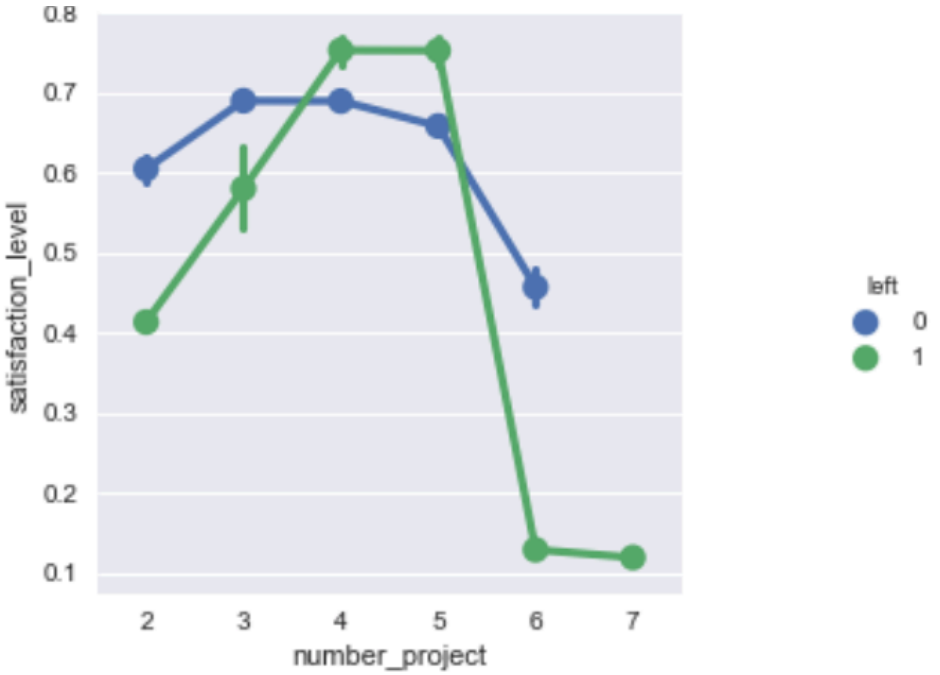


Figure 12: Number of projects Vs satisfaction level based on that people leaving and staying.

Work accidents Vs Left (employees leaving)

From the graph, it can be seen that very few people had any work-related accidents and a lesser number of people left the company. The number of employees leaving the company without having any work-related accidents is more.

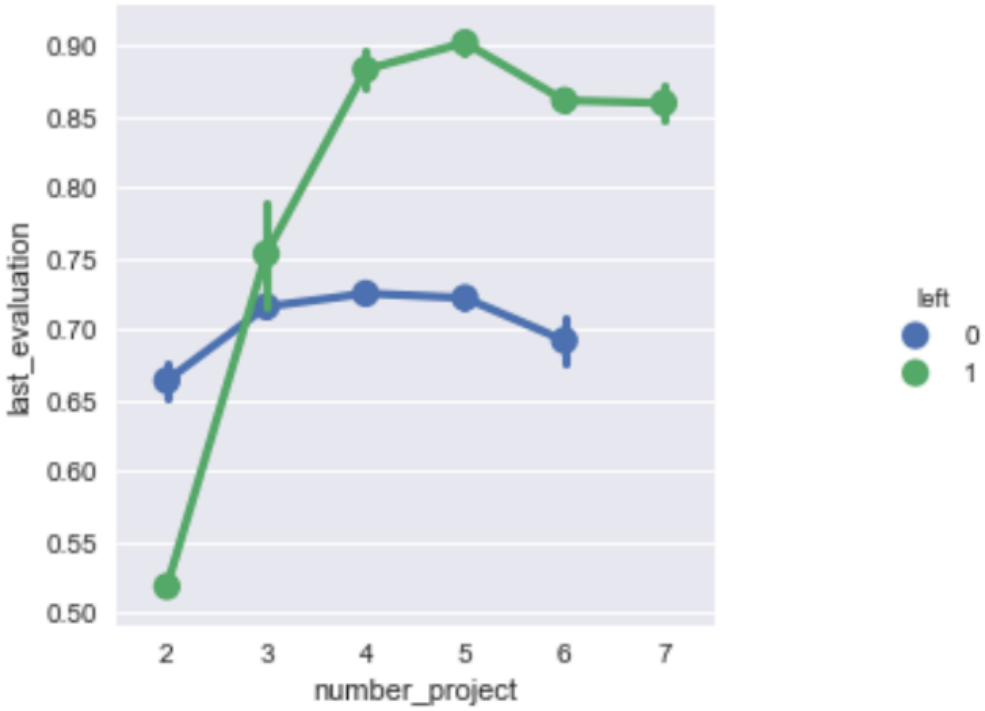


Figure 13: Number of project Vs satisfaction level based on that people leaving or staying in the company.

Time spent at company Vs Salary

From the analysis of graph, if an employee has spent 3,4 and 5 years at a company they tend to leave the company more often. The chance of an employee leaving the company increases as they spend more than 3 years at a company because they end up getting bored at the same job and look out for new and challenging opportunity outside the company.

```
]: sns.countplot(x="time_spend_company",hue='salary', data=mydataset)
   plt.title('Salary based on the time spend in company')

]: <matplotlib.text.Text at 0x115afc850>
```

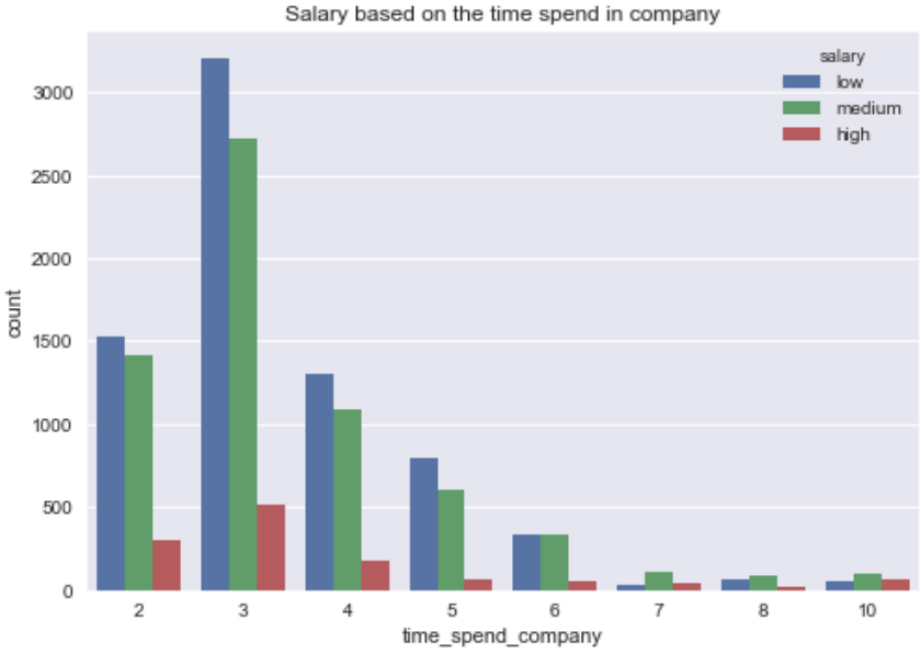


Figure 14:Time spend company vs Salary

Time spent at company and average monthly hours Vs Left (employees leaving the company)

People who have spent 2 to 10 years in the company and have 190 to 200 average working hours per month then they tend to spend more time at the company.
After looking at the figure it can be said that if a person is working 150-200 hours in a month then they are more likely to stay.
Employees who have spent 4,5 or 6 years in the company and are working 240-260 hours per month they will more likely leave the company. Employees who worked for 2 or 3 years were the ones who left the company most.

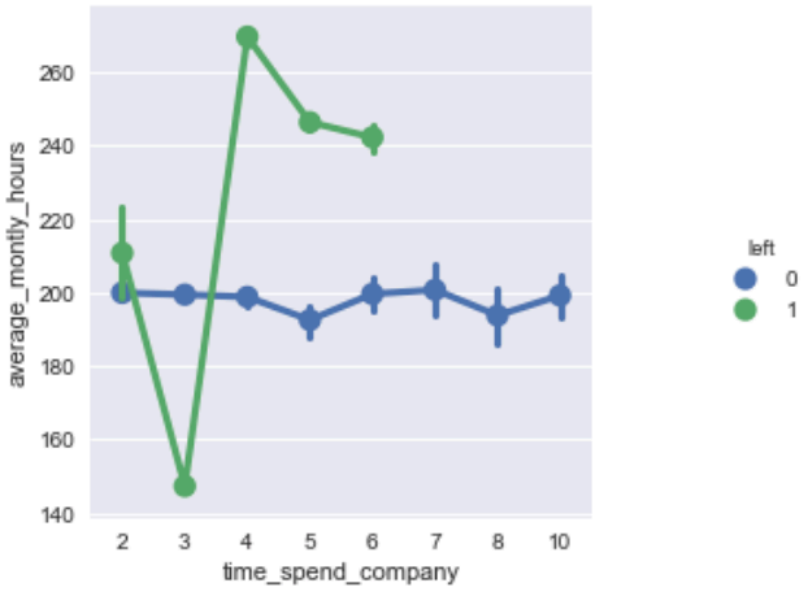


Figure 15: Time spend company and average monthly hours vs left

Time spent at company Vs Satisfaction level

Looking at the figure indicates that people who spent 3 to 6 years at the company are the ones who are least satisfied from their job. People who spent 2 years at the company are the most satisfied.

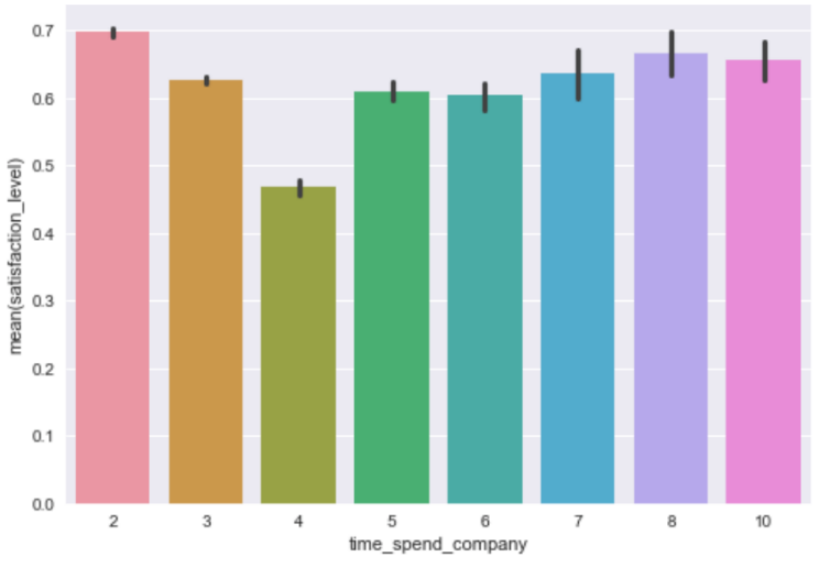


Figure 16: Time spend company vs satisfaction level

Time spend at company Vs Average monthly hours

Employees who have worked for 4 to 6 years in the company they usually put around 200 average monthly hours. People working for 4 or 5 years have highest number of monthly hours.

People who spent 3 years at the company have the lowest number of monthly hours and didn't leave the company.

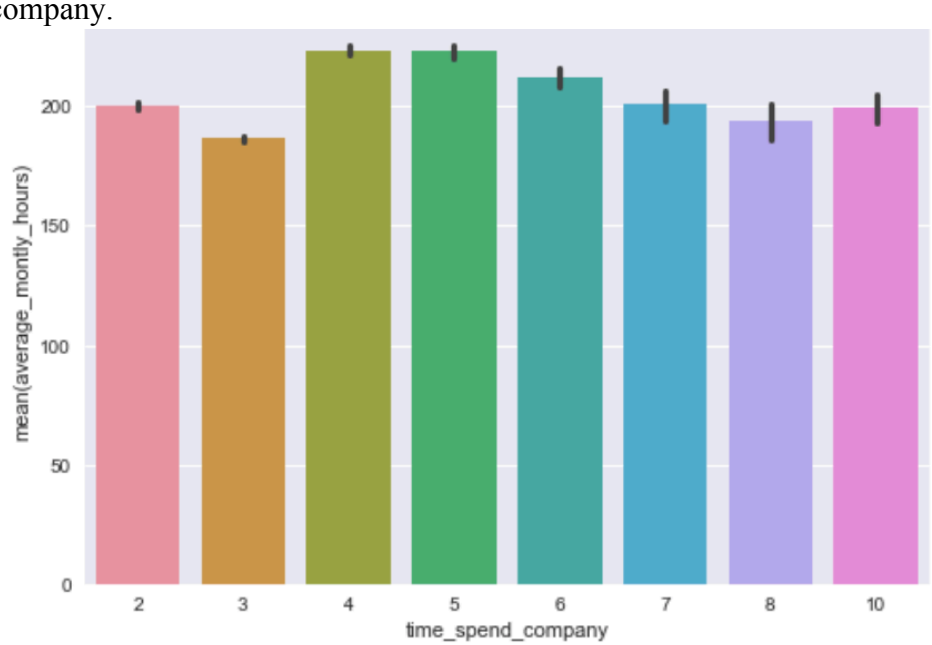


Figure 17: Time spend at company vs average monthly hours

Satisfaction level Vs left (employees leaving)

So, there are Three distribution of employee left the company based on their satisfaction level.

1. Employee left the company they have low satisfaction level (0-0.2).
2. Employee left the company with medium satisfaction level (0.3-0.5).
3. Employee left the company with high satisfaction level (0.7-1).

From the analysis of figure if an employee has satisfaction level between (0.5-1); they stay in company as comparison to the people left the company with same satisfaction level.

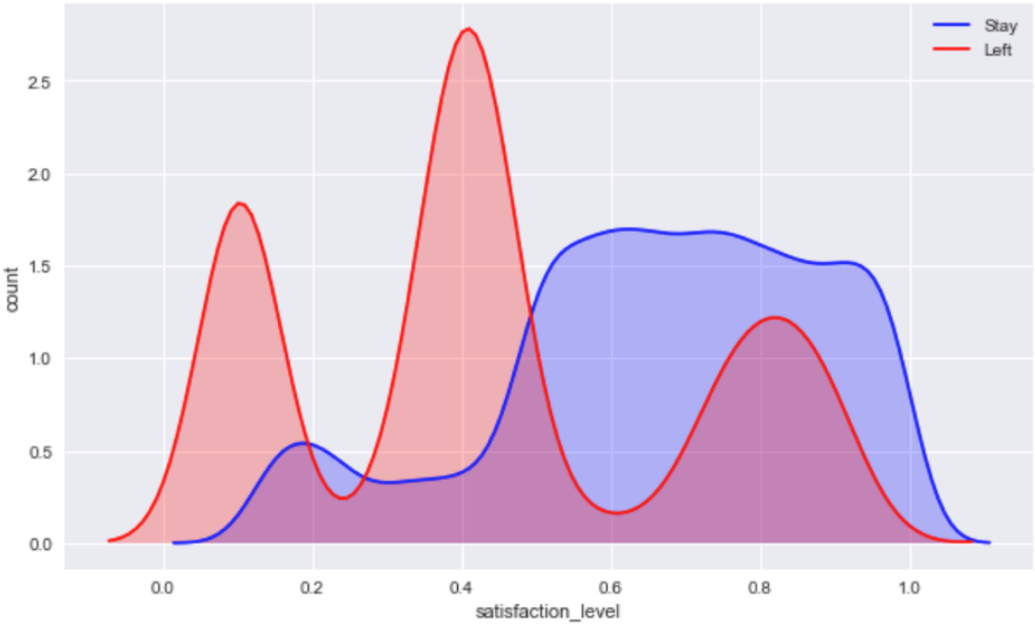


Figure 18: Satisfaction level Vs left.

Evaluations Vs left (employees leaving)

Employee with evaluation level (0.6-0.8) had the smallest number of employee left the company. So, we divided the evaluation into two parts:

1. Employee with low evaluation level (0.2-0.6), they left the company in highest count.
2. Employee with high evaluation level (0.8-1), they left the company also.

It is main concern here why are the high evaluated employee are leaving the company. There will be lot of factors responsible for it.

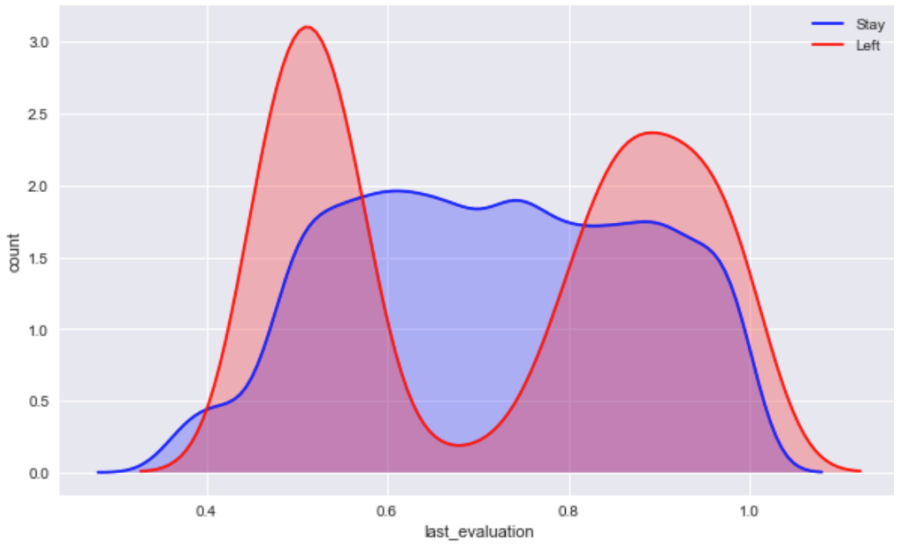


Figure 19: Employees who stayed or left depending on the last evaluation.

Average Monthly hours vs Left (employee leaving)

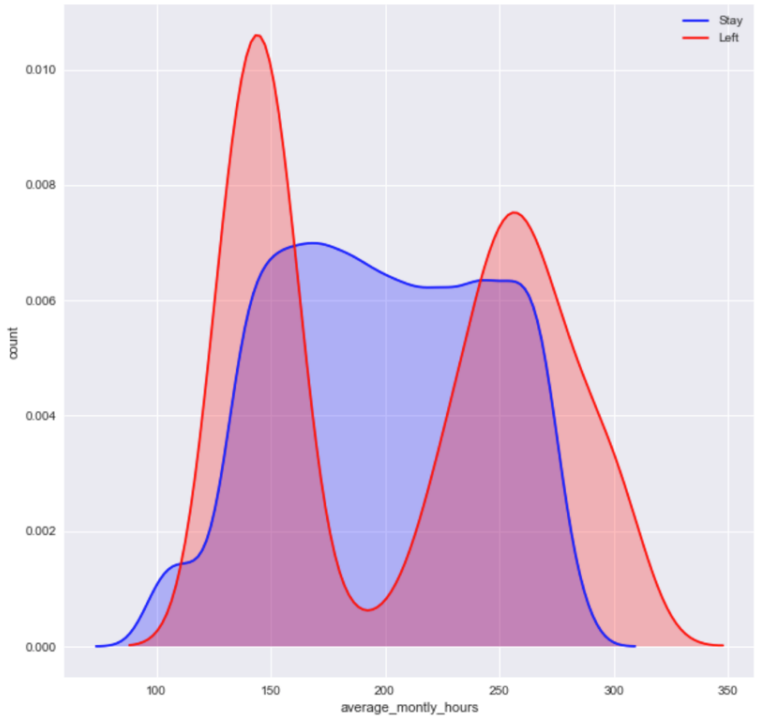


Figure 20: Employees who stayed or left depending on the average monthly hours.

1. Employee who left the company because they are not working enough were underworked and who worked more monthly work are overworked.
2. Employee who had less hours of monthly hours (120-150 hours and less) left the company more as comparison to other employee.
3. Employee who had work more hours of monthly hours (200-250 hours and more) left the company less than the less hours monthly hours employee

Methods

We are using Logistic regression to predict an outcome of variable that is categorical form of predictor variable that have continuous form.
Logistic regression deals with the discrete variable by using the logarithmic transformation on the outcome variable Which allow us to model a non-linear association in a linear way

Decision tree require relatively little effort from the user for data interpretation. Non-linear relationship between the variable doesn't affect the decision tree performance. It explains in a easy way

Random forest unexcelled in accuracy among the other algorithm. You can run this model on large data without having any problem. It can handle thousand of variable without delete anyone.
Some time generated forests can be saved for future use of any other data

Chapter 4

Experimental Setup

Data Training and Testing Split

It's good idea to calculate score for training and test, reasons are:

1. It's often a good sign that your training and test scores are relatively close
2. It's very useful for deciding how to improve your model.
3. Looking at the training score can help you prevent overfitting.
4. The higher the score is, the better the model is performing.

```
# now i spliting my Traning and testing data
from sklearn.model_selection import train_test_split
panel = mydataset[['satisfaction_level','average_montly_hours','promotion_last_5years','salary','number_project','last_
x=panel # i tried to use pop command, but itn't let me take value more than 2.
y=mydataset['left']
x_train, x_test, y_train, y_test = train_test_split(x, y, random_state=42,test_size=0.25)
print('Training set volume:', x_train.shape[0])
print('Test set volume:', x_test.shape[0])
```

```
('Training set volume:', 11249)
('Test set volume:', 3750)
```

It's good idea to calculate score for training and test, reasons are: 1-It's often a good sign that your training and test scores are relatively close 2-It's very useful for deciding how to improve your model. 3- looking at the training score can help you prevent overfitting. 4- The higher the score is, the better the model is performing. I took

Figure 21: Data splitting.

After analysis the all factor; I decide work accident have least affect in people decision making about leaving the company.

Different models that I used for my project to find out the accuracy score are as follows:

1. Logistic regression.
2. Decision tree
3. Random Forest tree
4. Gaussian naïve bayes
5. KNN

***Logistic Regression*:** Logistic regression analysis studies the association between a categorical dependent variable and a set of independent variables. For our project Accuracy score of Logistic regression is **0.764** based on the seven factor that we decided is important to calculate the best accuracy score.

```
lr=LogisticRegression()
lr.fit(x_train,y_train)
testscoreLR=accuracy_score(y_test,lr.predict(x_test))
print('logistic regression accuracy score:'+str(testscoreLR))
print(confusion_matrix(y_test,lr.predict(x_test)))
print(classification_report(y_test,lr.predict(x_test)))
```

```
logistic regression accuracy score:0.764
```

Figure 22: Logistic regression.

***Decision Tree*:** Decision tree make regression model like tree structure. It breaks down our dataset into smaller and smaller subset while at the same associated decision tree is incrementally developed.

```
from sklearn import tree
dt = tree.DecisionTreeClassifier(max_depth=8)
dt.fit(x_train, y_train)
testscoreDT=accuracy_score(y_test,dt.predict(x_test))
print("decision tree accuracy Rate is:"+str(testscoreDT))
print(confusion_matrix(y_test,dt.predict(x_test)))
print(classification_report(y_test,dt.predict(x_test)))
```

```
decision tree accuracy Rate is:0.976533333333
```

(A)

```
#Decision tree
from sklearn import tree
dt = tree.DecisionTreeClassifier(max_depth=3)
dt.fit(x_train, y_train)
testscoreDT=accuracy_score(y_test,dt.predict(x_test))
print("decision tree accuracy Rate is:"+str(testscoreDT))
print(confusion_matrix(y_test,dt.predict(x_test)))
print(classification_report(y_test,dt.predict(x_test)))
```

```
decision tree accuracy Rate is:0.950666666667
```

(B)

Figure 23: Decision tree; (A) maximum depth 8. (B) Maximum depth 3.

I used max depth features in this modeling. If I used max depth 8 I got better accuracy score (**accuracy score=0.97**) than the max depth 3(**accuracy score=0.95**). Max Depth define the depth of the tree. It can take any integer value. If none, then node is expanded until all leaves contain least split samples.

***Random Forest*:** Here, we create the forest with many trees. In general, the more tree in the forest the more robust the forest looks like. In the same way, random forest classifier, higher the number of tree in the forest gives the high number in accuracy score. For our project Random Tree give us **accuracy score of 0.97**.

```
from sklearn.ensemble import RandomForestClassifier
rf= RandomForestClassifier(n_estimators=100)
rf.fit(x_train, y_train)
testscoreRF=accuracy_score(y_test,dt.predict(x_test)).mean()
print("Random Tree accuracy Rate is:"+str(testscoreRF))
print(confusion_matrix(y_test,rf.predict(x_test)))
print(classification_report(y_test,rf.predict(x_test)))
```

```
Random Tree accuracy Rate is:0.949333333333
```

Figure 24: Random tree.

***Gaussian Naïve Bayes*:** Naive Bayes classifier assumes that the presence of a feature in a class is unrelated to the presence of any other feature. For this project Gaussian Naïve Bayes **accuracy score is 0.82.**

```
nb=GaussianNB()
nb.fit(x_train,y_train)
testscoreNB=accuracy_score(y_test,nb.predict(x_test))
print('GaussianNB accuracy score:'+str(testscoreNB))
print(confusion_matrix(y_test,nb.predict(x_test)))
print(classification_report(y_test,nb.predict(x_test)))
```

```
GaussianNB accuracy score:0.827466666667
```

Figure 25: Gaussian navie bayers

KNN: This algorithm uses the neighbor points information to predict the target class. The closed class can be find out using the Euclidean Distance. For this project KNN **accuracy score is 0.9432.**

```
kn=KNeighborsClassifier()
kn.fit(x_train,y_train)
testscoreKN=accuracy_score(y_test,kn.predict(x_test))
print('KNeighborsClassifier accuracy score:'+str(testscoreKN))
print(confusion_matrix(y_test,kn.predict(x_test)))
print(classification_report(y_test,kn.predict(x_test)))
```

```
KNeighborsClassifier accuracy score:0.9432
```

Figure 26: KNN.

After the analysis of all the models one must decide which model to use in prediction model. Hence, a ranking table is created based on the models testing scores as shown in figure 27 below.

	Model	Testing_Score
0	random Forest	0.976533
1	Decision Tree	0.976533
3	KNN Regression	0.943200
4	Gaussian Naive Bays	0.827467
2	Logistic Regression	0.764000

Figure 27: Ranking of all the models.

One model: Four models were used to measure the accuracy of the data. The best model based on the accuracy score and ROC curve is Random Forest. But predicting future value we must understand the past value very well.

Base model is a simple model used as reference point for comparing how well a model is performing. In this dataset, the majority class that will be predicted will be zero, which are employee who did not leave the company.

Precision and Recall

Precision:

- Of those classified as *Will stay*, what proportion did?
- True positive / (True positive + False positive)

Recall:

- Recall: Of those that in fact *left*, what proportion were classified that way?
- True positive / (True positive + False negative)

Better models have higher values for precision and recall.

- You can imagine a better model with 94% precision and 97% recall.
- A weaker model might have 95% precision but 50% recall.
- Or maybe the model has 60% precision and 60% recall.

After analysis the all regression and classifier precision and recall Random Forest have the higher value of precision and recall.

```
logistic regression accuracy score:0.764
[[2641  212]
 [ 673  224]]
             precision    recall  f1-score   support

          0       0.80      0.93      0.86      2853
          1       0.51      0.25      0.34       897

avg / total       0.73      0.76      0.73      3750
```

Figure 28: Confusion matrix and classification report of logistic regression.

```
decision tree accuracy Rate is:0.976533333333
[[2829   24]
 [  64  833]]
             precision    recall  f1-score   support

          0       0.98      0.99      0.98      2853
          1       0.97      0.93      0.95       897

avg / total       0.98      0.98      0.98      3750
```

Figure 29 : Confusion matrix and classification report of decision tree.

```
Random Tree accuracy Rate is:0.949333333333
[[2846    7]
 [  32  865]]
             precision    recall  f1-score   support

          0       0.99      1.00      0.99      2853
          1       0.99      0.96      0.98       897

avg / total       0.99      0.99      0.99      3750
```

Figure 30: Confusion matrix and classification report of Random tree.

```
GaussianNB accuracy score:0.827466666667
[[2478  375]
 [ 272  625]]
             precision    recall  f1-score   support

          0       0.90      0.87      0.88      2853
          1       0.62      0.70      0.66       897

avg / total       0.84      0.83      0.83      3750
```

Figure 31: Confusion matrix and classification report of gaussianNB.

```
KNeighborsClassifier accuracy score:0.9432
[[2707  146]
 [  67  830]]
             precision    recall  f1-score   support

          0       0.98      0.95      0.96      2853
          1       0.85      0.93      0.89       897

avg / total       0.95      0.94      0.94      3750
```

Figure 32: Confusion matrix and classification report of KNN.

Chapter 5

Result and Discussion

Features Importance: After analysis everything I decided to use random forest and decision tree model to find out the important features of used for the prediction model. It helps as in creating our model for random Forest and Decision tree because it gives us more idea what is going on in our project. In first figure I used Random Forest and 2nd Figure I used Decision Tree. Both model have same accuracy among in all model.

Important features:

```
importances=rf.feature_importances_
f=np.argsort(importances)[::-1]
print ('feature ranking:')
for i in range(x.shape[1]):
     print ("feature no. {}: {} ({})".format(i+1,x.columns[f[i]],importances[f[i]]))
```

```
feature ranking:
feature no. 1: satisfaction_level (0.334497954026)
feature no. 2: number_project (0.188536952156)
feature no. 3: time_spend_company (0.18638798068)
feature no. 4: average_montly_hours (0.150783966431)
feature no. 5: last_evaluation (0.127619506421)
feature no. 6: salary (0.0107167088575)
feature no. 7: promotion_last_5years (0.00145693142702)
```

Figure 33: Important features using random forest.

```
importances=dt.feature_importances_
f=np.argsort(importances)[::-1]
print ('feature ranking:')
for i in range(x.shape[1]):
     print ("feature no. {}: {} ({})".format(i+1,x.columns[f[i]],importances[f[i]]))
```

```
feature ranking:
feature no. 1: satisfaction_level (0.52546172206)
feature no. 2: last_evaluation (0.150481782448)
feature no. 3: time_spend_company (0.146132274565)
feature no. 4: number_project (0.103759601444)
feature no. 5: average_montly_hours (0.0719158964222)
feature no. 6: salary (0.00224872306004)
feature no. 7: promotion_last_5years (0.0)
```

Figure 34: Important features using decision tree.

In this project Random forest and Decision tree model were used to find out the important features to calculate the predication model. Using this method, we know better our factor before we used them. It gives us better understanding of factors. This is a unique way to use features in prediction model.

Five important features obtained from Random Forest are:

1. Satisfaction level
1. Number of project
2. Time spent in company
3. Average monthly hours
4. Last Evaluation

Five important features obtained from Decision Tree top are:

1. Satisfaction level
2. Last evaluation
3. Time spent at company
4. Number of project
5. Average Monthly hours

On comparing both features ranking difference can be seen. Random tree gives second rank to number of project, whereas decision tree gives last evaluation second number. The ranking of Number of project Decision tree is four. But if you take look at the satisfaction level probability in both model, there is a huge difference between them. Decision tree satisfaction level have better probability than the random forest probability. After analysis of all these factors I decided to use Decision tree to calculate prediction probability.

Predicting Probability: On analysing everything, it can be said that 957 employees are about to leave the company. Then I calculated probability based on 45% leaving chance. The percentage of leaving can be increased or decreased. So, it will give you full prediction model of which employee is going to leave the company soon.

Using the decision tree, we find out the 957 employees are about the leave the company. We have decided the leaving probability. We decided 45% leaving probability and gives us all the employee information which have 45% chance of leaving the company. I attached some picture you see employee number and their other features probability. It gives us a perfect model of calculate the leaving probability of any employee in near future.

```
]: # Load our actual value for prediction
   dataf = pd.get_dummies(mydataset)

2]: # we need to provide information of
   left = dataf[dataf['left'] == 1]
   left1 = pd.get_dummies(leave)

   c = left1
   a = c['left'].values
   c = c.drop(['left'],axis=1)
   b= c.values
   pred = dt.predict_proba(q[:, :7]) # we have 8 model input so, i selected 7 input
   ## i used dicison tree for my data
   # number of employees that definitely are leaving
   sum(pred[:,1]==1)

2]: 957
```

Figure 35: Probability model prediction.

```
In [103]: left['Will leave the job'] = pred[:,1]
          # you can change this leaving prob,but i select 45%.
          left[left['Will leave the job']>=0.45]
```

	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_last_5years	sales_IT	sales_RandD	...	sales_mar
2	0.11	0.88	7	272	4	0	1	0	0	0	...	
6	0.10	0.77	6	247	4	0	1	0	0	0	...	
11	0.11	0.81	6	305	4	0	1	0	0	0	...	
18	0.45	0.51	2	160	3	1	1	1	0	0	...	
20	0.11	0.83	6	282	4	0	1	0	0	0	...	
22	0.09	0.95	6	304	4	0	1	0	0	0	...	
30	0.09	0.62	6	294	4	0	1	0	0	0	...	
35	0.10	0.94	6	255	4	0	1	0	0	0	...	
38	0.11	0.89	6	306	4	0	1	0	0	0	...	
43	0.10	0.80	7	281	4	0	1	0	0	0	...	

Figure 36: Probability model prediction.

In upper fig employee at 6 have satisfaction level is 0.10, number of project 6, average monthly hours are 247, time spend at company is 4.

So, we have idea what kind of factor probability we need to analyse the probability of leaving the employee.

Prediction model:

I have other model which just tell you which employee is going to leave the company. Using this predication model, you have to decide how many employee predictions you need. And it will give you randomly prediction of which employee is going to leaving the company. For this feature, I used random Forest to calculate this.

```
In [120]: test_predict = x_test.iloc[0:10,:]
          realpr = dt.predict(test_predict) # we are using real prediction value
          for i in realpr:
              print (i)

          0
          0
          0
          1
          0
          0
          0
          0
          0
          0
```

This Predicted value for staying. Now, we will apply same method on real data.

```
In [119]: y_test.iloc[0:10] # It randmly select 10 employee.

Out[119]: 6723    0
          6473    0
          4679    0
          862     1
          7286    0
          8127    0
          3017    0
          3087    0
          6425    0
          2250    0
          Name: left, dtype: int64
```

Figure 37: Prediction model.

I calculated the prediction value of staying for 10 employees. It gives you the random employee which one is going to stay or leave. So, employee 862 is going to leave the company.

This Prediction model doesn't give you the probability; that's why I calculated the predication probability.

Chapter 6

Conclusion

This paper analysis the different method done on employee dataset and usage of many different model to make prediction model to calculate the employee is going to stay or left the company. This model can be applied throughout many department of the company and used the company HR department to make better decision to make a decision about the employees before they left the company. You can make this model more perfect by adding more features by at any time.

- Employee satisfaction, years at company, number of project, average monthly hours and evaluation are top important features to determining employee left the company.
- Employees that have 2,5 and 7 projects was at risk of they are about to leave the company.
- Employees with low, medium salaries are the large number who left the company.
- Employee satisfaction is the main features or factor responsible for employee left the company.
- Employee Who generally left the company when they are worked more than 220-250 hours/month. They feel that they are doing overworked.
- Employee generally left the company when they are work 150 hours/month. They feel underworked.
- Employee who has high and low evaluations; they should consider that they will leave the company soon also.
- Employee who spend 4 and 5 years at company should consider also for high leaving.

Future work:

To have a more accurate finding more data have to be collected and more experiments should be performed. If there are some more variables included from the database that would have more impact on finding out the employee turnover such as age, gender, distance from home, etc.

Chapter 7
References

Anon, Datasets | Kaggle. Available at: https://www.kaggle.com/datasets [Accessed December 2, 2017].

Dayton, C.M., 1992. LOGISTIC REGRESSION ANALYSIS. Available at: https://www.researchgate.net/profile/C_Dayton/publication/268416984_Logistic_Regression_Analysis/links/550312ff0cf2d60c0e64c8ca/Logistic-Regression-Analysis.pdf [Accessed December 2, 2017].

Dziuban, C.D. & Shirkey, E.C., 1974. When is a correlation matrix appropriate for factor analysis? Some decision rules. *Psychological Bulletin*, 81(6), pp.358–361. Available at: http://content.apa.org/journals/bul/81/6/358 [Accessed December 2, 2017].

Inese, P., Igor, T. & Arkadijs, B., 2010. Decision Tree Classifiers in Bioinformatics. *Rigas Tehniskas Universitates Zinatniskie Raksti*, 44, p.118. Available at: http://hec.summon.serialssolutions.com/2.0.0/link/0/eLvHCXMwzV3dT9swED-tnTTxwgYbGRtIeY9SkcROnAceyjo-VCEklAfES-U452JaUpaE_f2z89GmmuCZp8iyosS-u5_vTuf7AdiMCsZQsDAKueDSI740lRehyDwmJDfJtuk0Pp-Q8QW92NAnzl-qpqhp9IBvXy95x8KdtMw4TlIgNsSWShpSa5O-OFOrth1q1ZWydx241Zy.

Silahtaroglu, G. & Donertasli, H., 2015. Analysis and prediction of E-customers' behavior by mining clickstream data. In *2015 IEEE International Conference on Big Data (Big Data)*. IEEE, pp. 1466–1472. Available at: http://ieeexplore.ieee.org/document/7363908/ [Accessed December 2, 2017].

Zhang, S., Zhang, C. & Yang, Q., 2003. Data preparation for data mining. *Applied Artificial Intelligence*, 17(5–6), pp.375–381. Available at: http://www.tandfonline.com/doi/abs/10.1080/713827180 [Accessed December 2, 2017].