Royal Melbourne Institute Of Technology

Predictive Analysis On Employee Turnover Rate

Student1: Shamini Puthooppallil Baby, Student Id: 3674381
Post Graduate Student At RMIT
Email: s3674381@student.rmit.edu.au

Student2: Monika Vurigity, Student Id:3675394
Post Graduate Student At RMIT
Email: s3675394@student.rmit.edu.au

Date Of Publication:22nd May, 2018

Table of Contents

Abstract

The main objective of this report is to examine the features influencing turnover rate of employees in a company which helps to predict the decision of a particular employee towards working further with the company. Information of employees from a company was collected from a Cloud Storage Electronic Employee Records done by Human Resource Department of many firms. The results point outs that the employee turnover percentage in accompany depends on many factors such as satisfaction level, work accident etc. The overall analysis summarizes that the decision of employee towards staying the company can be accurately predictable by different factors. It is suggested that the companies should improve the facilities to make employees satisfiable and thereby reduce the work accident and gain overall growth.

Introduction

"Big data can enable companies to identify variables that predict turnover in their own ranks."-Harvard Business Review, August 2017
"*Employee churn analytics is the process of assessing your staff turnover rates in an attempt to predict the future and reduce employee churn."*-Forbes, March 2016

An organization is always rated by its employees. If the employees are more efficient and productive, the organizational growth curve will always grow upward. Employees in the company are the true sources of competitive advantage against other firms. Every organization will face employee turnover and results huge cost and downward overall performance. Replacement of workers and training costs are the main challenges faced by many organizations. Before the turnover becomes a serious problem, the companies should take corrective measures to retain the employees in the company. So, it's always important to analyze the employee turnover and mainly, the cause behind it to predict the attitude of each employee. This report investigates and analyze the relationship between the employee turnover and cause of the same using different analytical models. Main focus of these methodologies is to group employees according to their decisions by analyzing different factors determining the turnover rate, and then predict the behavior of employees from the existing findings to increase the overall growth of the organizations.

Methodology

Method:

The research used a methodology which collects data from a **Cloud Storage Electronic Employee Records** done by Human Resource Department of many companies.

The initial step was to explore the dataset and identify the main classes. All the variables are visualized using appropriate graphs, also the relationship between few factors which helped in grouping the variables to target set and feature set. Then adopted classification technique to analyze the data set and predict the future trends with accuracy.

Data Overview:

There are 10 columns and 14999 rows in the datasets.
Two categorical variables:

- sales: This variable shows in which team (department) the employee is working in. It is later changed department for ease of understanding.
- salary : Indicates the salary group of the employee whether it is high, medium or low

Eight numerical variables:

- satisfaction_level: Evaluation given by the employee.
- last_evaluation: Evaluation graded by the employee's manager.
- number_project: The number of projects an employee is involved in.
- average_monthly_hours: The average of monthly hours the employee has billed.
- time_spend_company: The time spent by the employee in the company.
- Work_accident: Value to show whether an employee has met with an accident.
- promoted_last_5years: Value to show whether an employee has got promotion in last 5 years.

Data Cleansing:

The process of the data cleansing is done through the below methodologies:
1. unique(): this a function used for checking for all acceptable values and range for each variable.
2. value_counts(): This function is used for checking the typos in each column.
3. strip() : The stirp function is used for removing any whitespaces present in the value any variable these white spaces may also influence the final result by leading to the mismatch of same values.
4. isnull() : Function used to check for the null values in each column.
5. datatypes: All the data types are checked, and each data type is changed to the suitable data type.
7. lowercase(): all the values in each column are converted to the lower case so that there should not be any mismatch caused due to case sensitiveness.

Feature Selection:

Recursive feature elimination(RFE) technique used to find the factors which are not determining the final target variable ('Left').
The data is divided into two parts, the first is turnover_f which contains all the possible features except the target value. The second turnover_t which consists of the target value 'Left'.
We adopted three models to analyze and predict the turnover dataset:

1. K-Nearest Neighbour Model
2. Decision Tree Model
3. Logistic Regression Model

For all the three models, RFE listed out the following results:

```
rfe = RFE(model, 10)
rfe = rfe.fit(turnover[X],turnover[y])
print(rfe.support_)
print(rfe.ranking_)
```

```
[ True  True  True  True  True  True  True  True  True]
[1 1 1 1 1 1 1 1 1]
```

Since the RFE listed all variables except the target variable preceding the models with 9 columns as features.

Features Selected:

- Sales(Department)
- Salary
- Satisfaction Level
- Last evaluation`
- `number_project
- `average_monthly_hours`
- `time_spend_company
- `Work_accident
- `left`promoted_last_5years

Analytical Techniques:

We have used the classification techniques for this analysis. The data is classified according to attitude of employee whether they continue to work or leave the company in future. Before the model selection all the categorical variables should be replaced with the corresponding numerical values. The new dataset contains only the numerical values now.

1. K-Nearest Neighbour Model**:** In this model we are classifying the data into two sets according to the features selected. First we divide the data into train and test sets with 70% train and 30% test with a random state.

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(turnover_f,turnover_t,test_size=0.30,random_state=6)
```

After splitting we data we train the model according to K-Nearest Neighbour Classification. To improve the performance of the model we iterated the model with different sets of parameters and finalized the below features to predict the model with highest accuracy.

```
model = KNeighborsClassifier(5,metric='manhattan',p=1,leaf_size=40,weights='distance',algorithm='kd_tree')
```

Now we fit the train dataset to K-nearest Classifier and predicting the behavior of unseen data using this model. The accuracy of the model is 95.51

2. Decision Tree Model: In this model we are classifying the data into two sets according to the features selected. First we divide the data into train and test sets with 70% train and 30% test with a random state.

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(turnover_f,turnover_t,test_size=0.30,random_state=6)
```

After splitting we data we train the model according to decision tree model. To improve the performance of the model we iterated the model with different sets of parameters and finalized the below features to predict the model with highest accuracy.

```
clf = DecisionTreeClassifier()
fit = clf.fit(X_train, y_train)
```

Now we fit the train dataset to K-nearest Classifier and predicting the behavior of unseen data using this model. The accuracy of the model is 97.89

3. Logistic Regression Model: In logistic regression model we are classifying the data into two sets according to the features selected. First we divide the data into train and test sets with 70% train and 30% test with a random state.

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(turnover_f,turnover_t,test_size=0.30,random_state=6)
```

After splitting we data we train the model according to logistic regression model. To improve the performance of the model we iterated the model with different sets of parameters and finalized the below features to predict the model with highest accuracy.

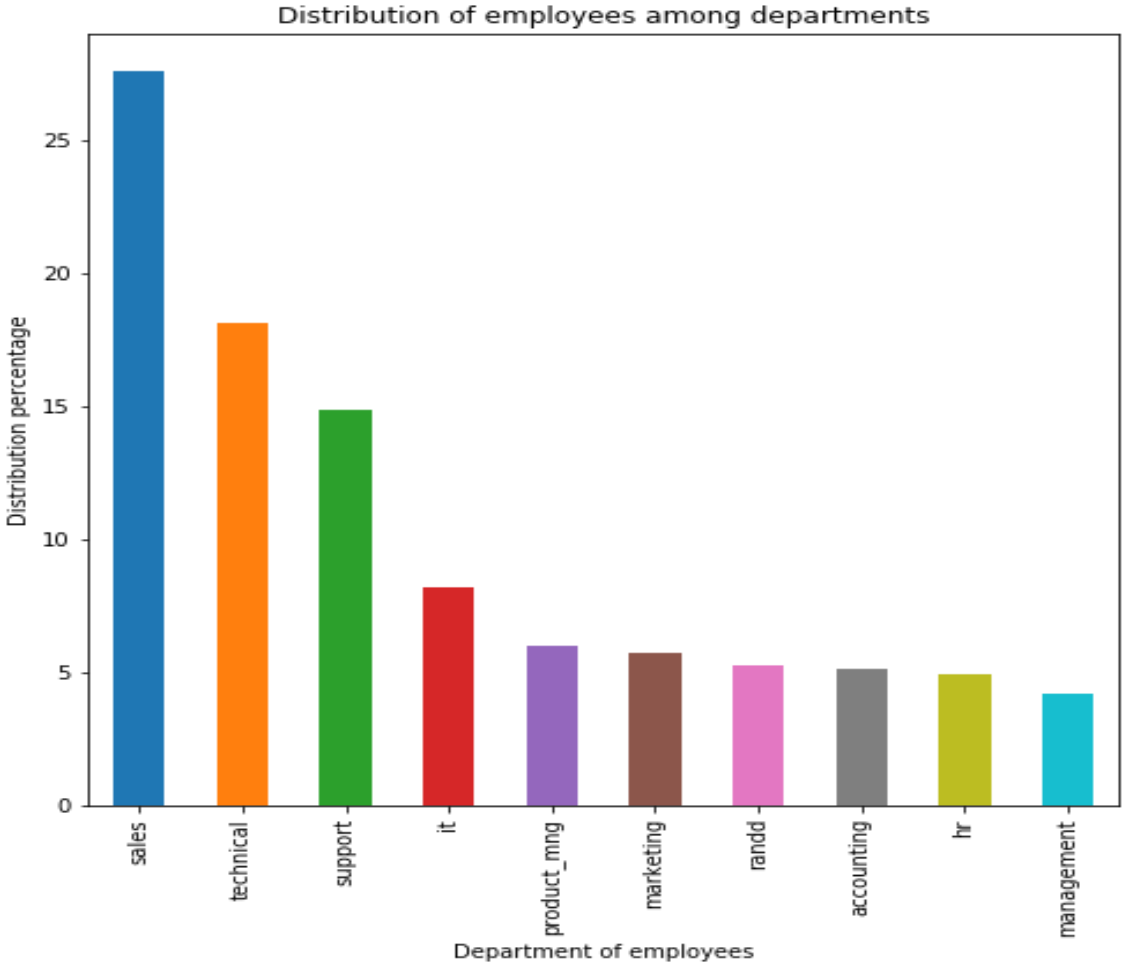
```
logreg = LogisticRegression(class_weight='balanced',max_iter=110)
logreg.fit(X_train, y_train)
```

Now we fit the train dataset to K-nearest Classifier and predicting the behavior of unseen data using this model. The accuracy of the model is 74.8.

Results

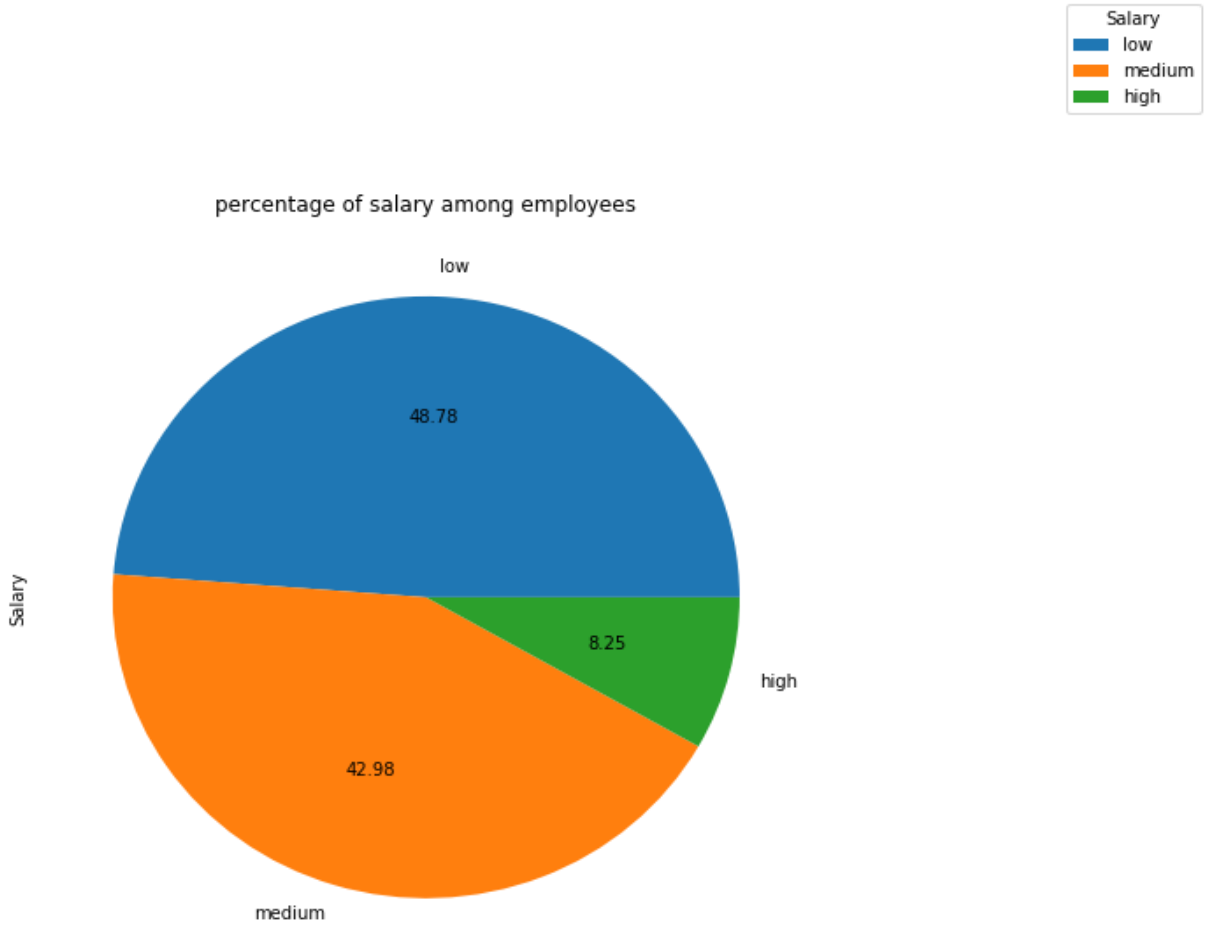
Visualizations:

Department: department is a categorical variable of different teams where employees are allocated in. Since it is a categorical variable with more number of variables so bar chart is the best suggested to visualize it graphically.



Bar-Graph For Department

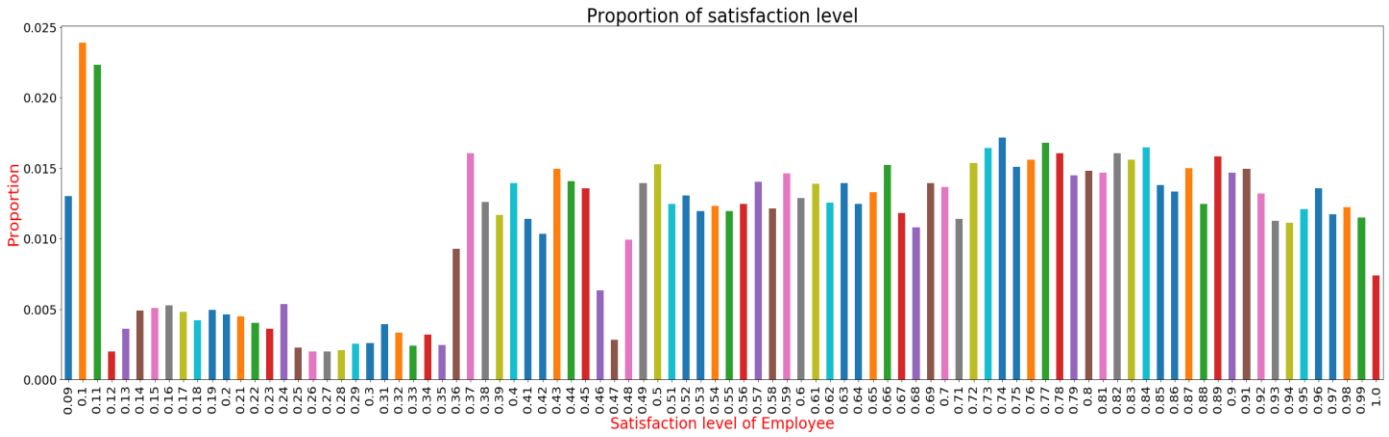
Salary**:** The categorical variable salary indicate the levels of salary which is categorized as high, low, medium. Pie chart picturizes these values graphically.



Pie-Chart For Salary

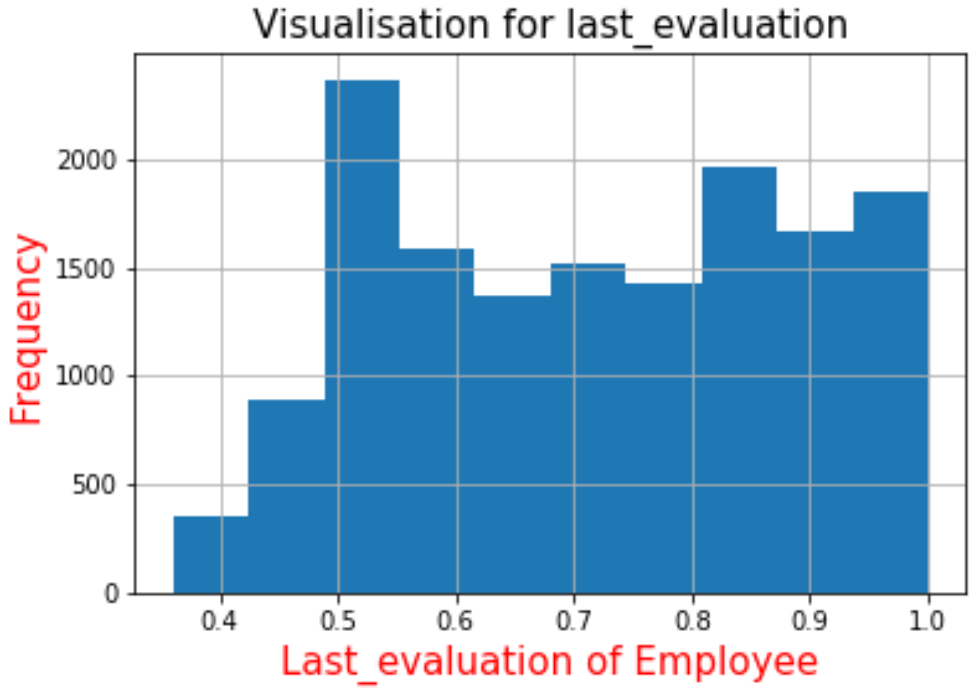
Satisfaction Level: The satisfaction level is numerical variable which is given by the employee, this value lies between 0 to 1. For this numerical variable histogram and bar chart are used to visualize the data graphically.

Histogram For Satisfaction Level

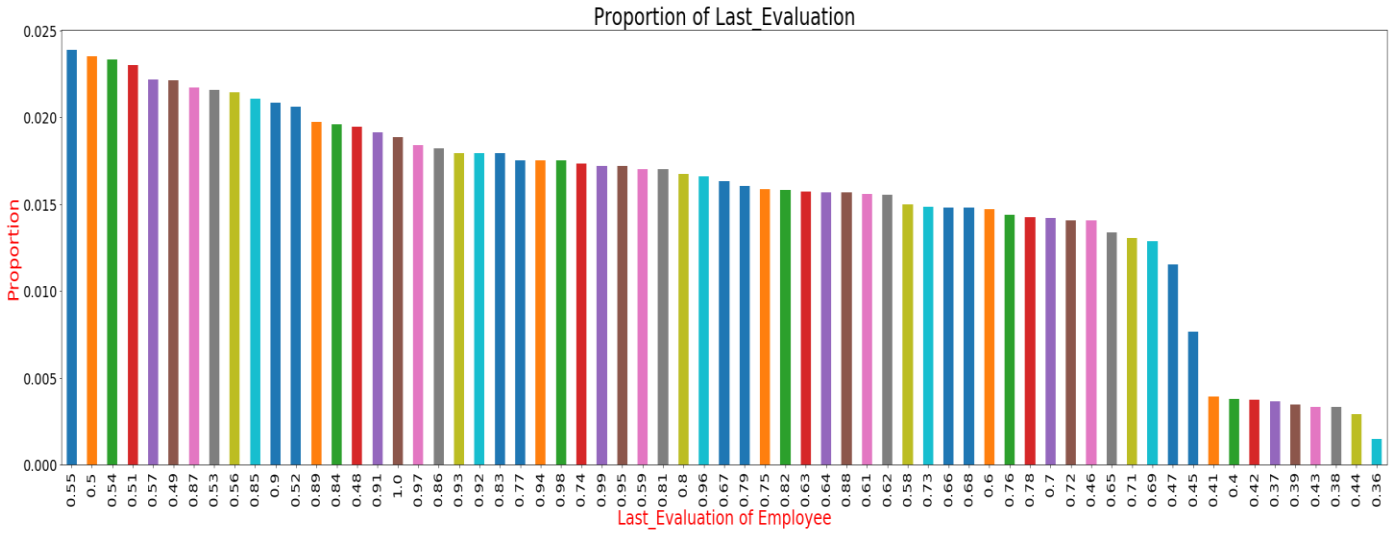


Bar-Graph For Satisfaction Level

Last Evaluation: The last evaluation is a numerical variable, this value is given to employee by the manager in charge, it ranges between 0 to 1. As this a numerical variable histogram and bar graph are used have a better vision graphically.



Histogram For Last Evaluation



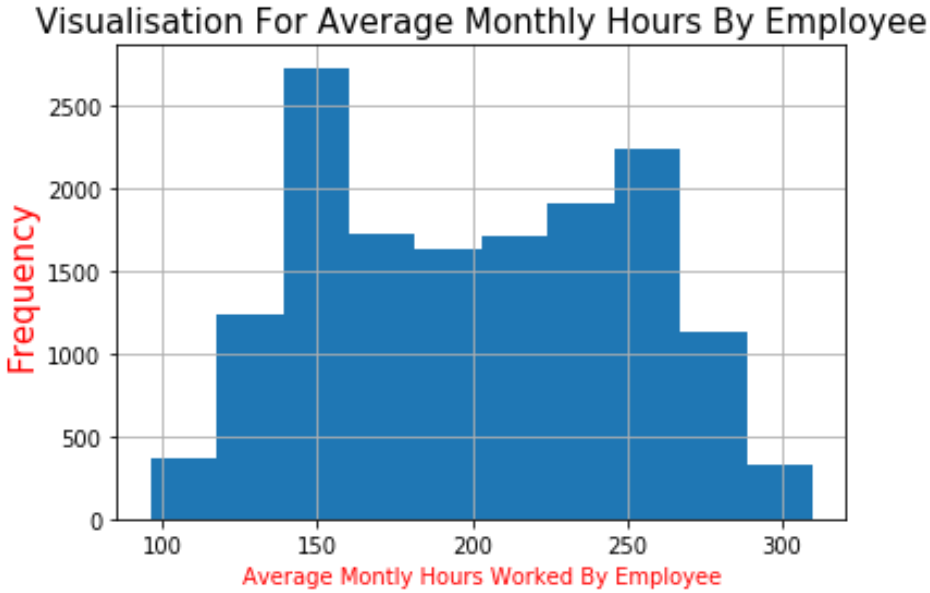
Bar-Graph For Last Evaluation

Number Of Project: It is a numerical variable which shows the number of projects each employee is involved. For this numerical variable histogram and bar chart are used for visualizing them graphically.

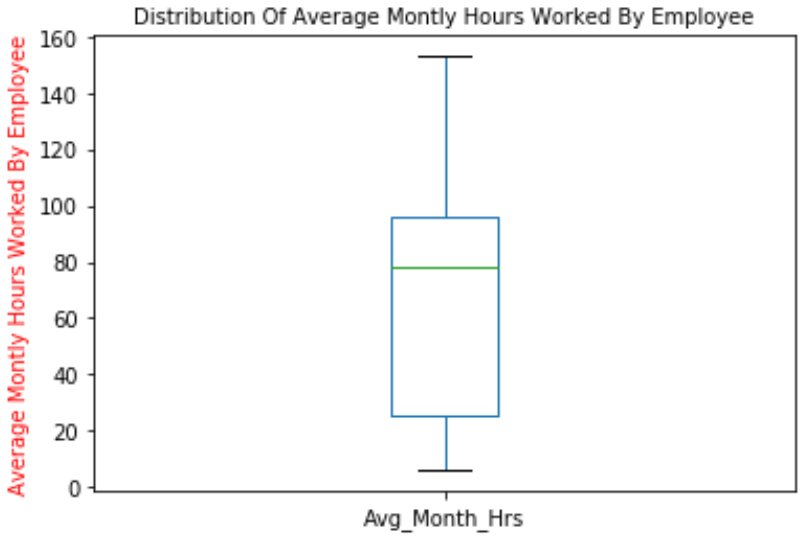
Histogram For Number Of Projects Worked By Each Employee

Bar-Graph For Number Of Projects Worked By Each Employee

Average Monthly Hours: The average of each month's worked hours the employee billed. This numerical variable is shown in a histogram and box plot for a better visualization.

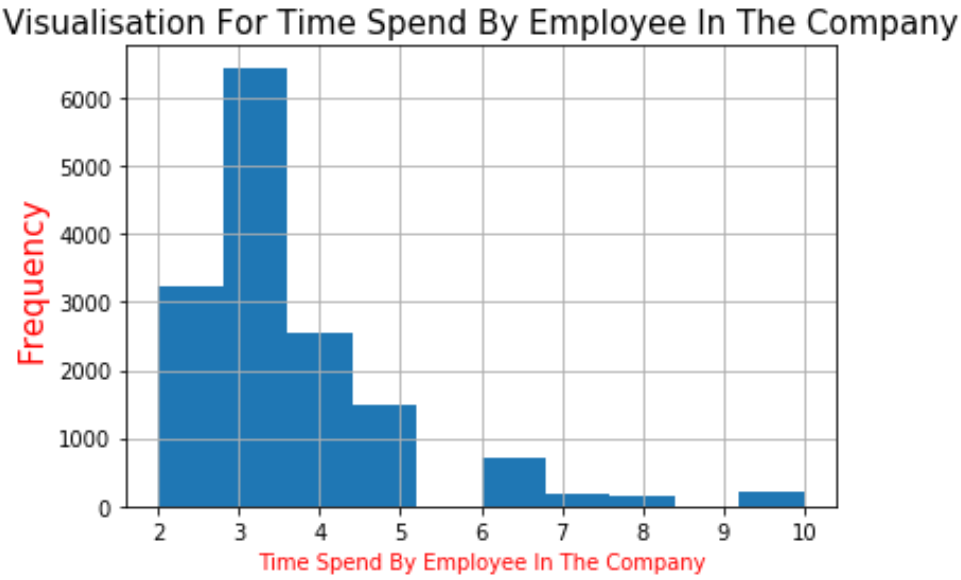


Histogram For Average Monthly Hours The Employee Billed

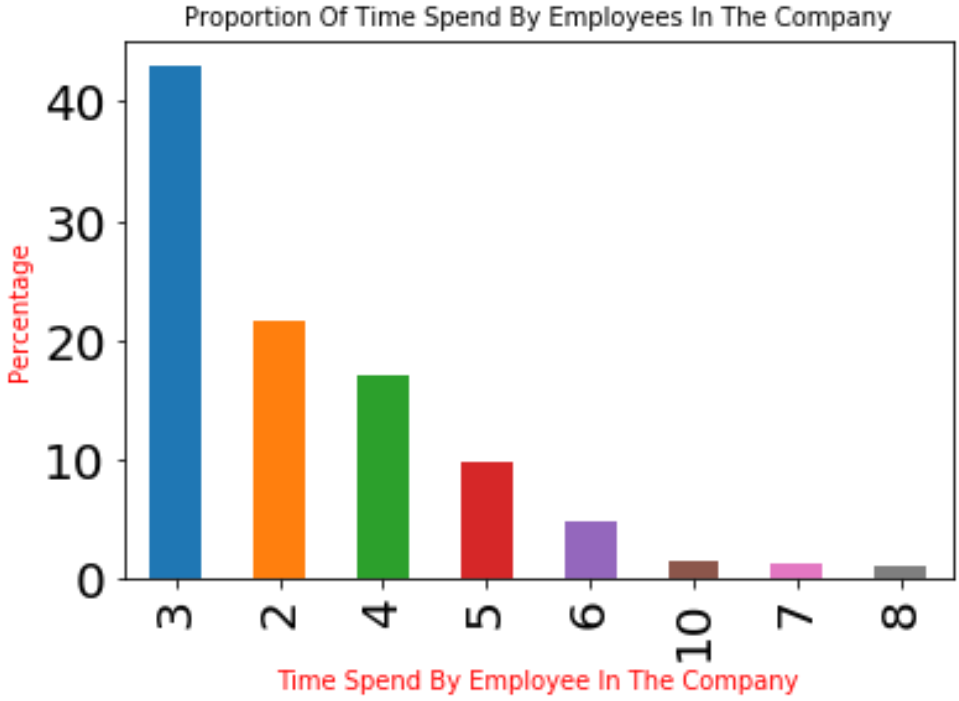


Box Plot For Average Monthly Hours The Employee Billed

Time Spend Company: It is a numerical value showing the service of an employee in the company. This numerical value is shown in histogram and bar chart.

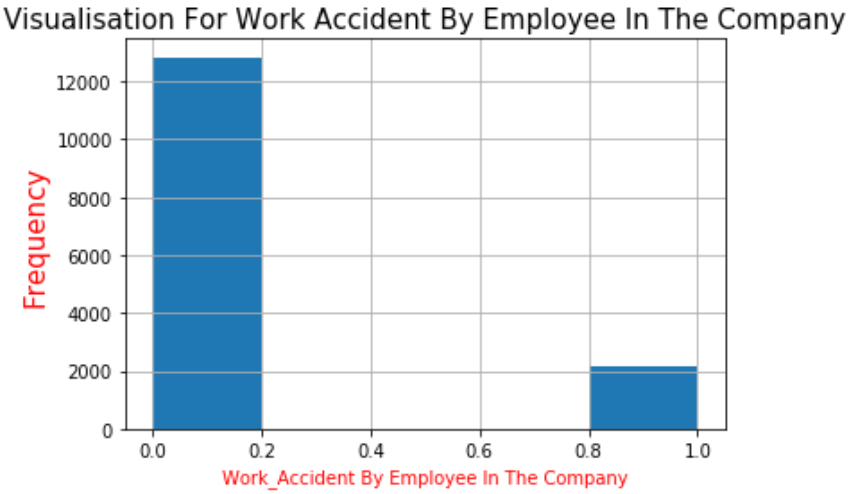


Histogram For Time Spend By Employees In The Company

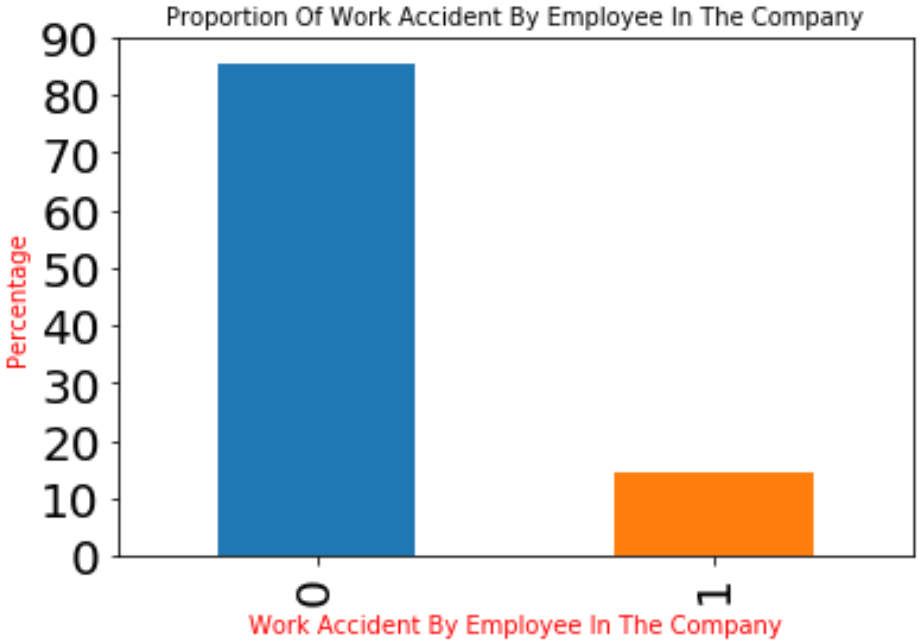


Bar-Graph For Time Spend By Employees In The Company

Work Accident: It is numerical variable showing 1 if the employee has met met with any accident while and 0 if not. It is shown in histogram and bar chart.

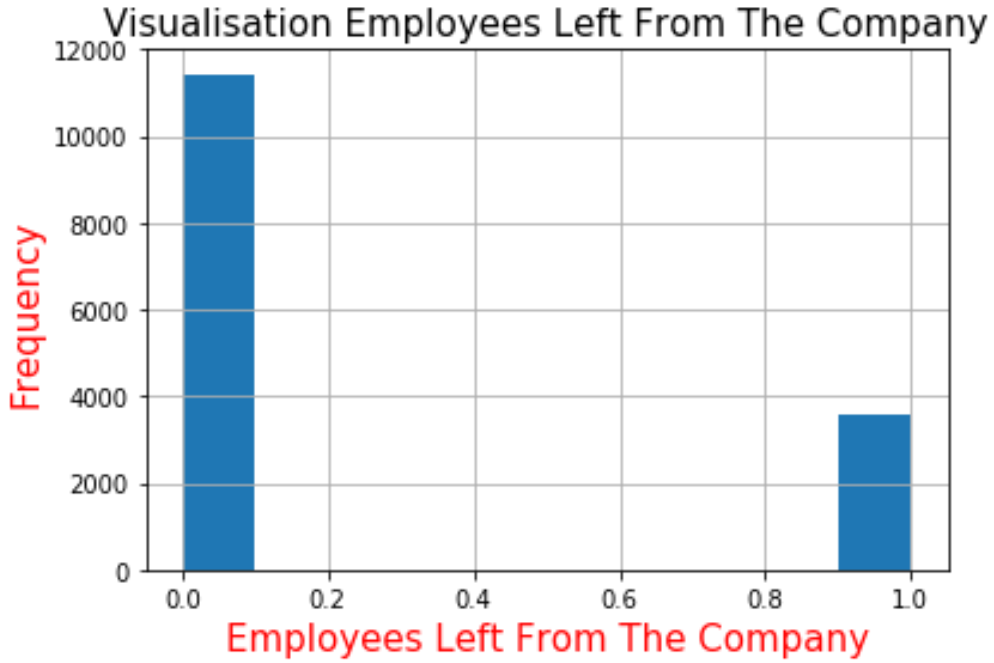


Histogram For Work Accident Met By The Employee

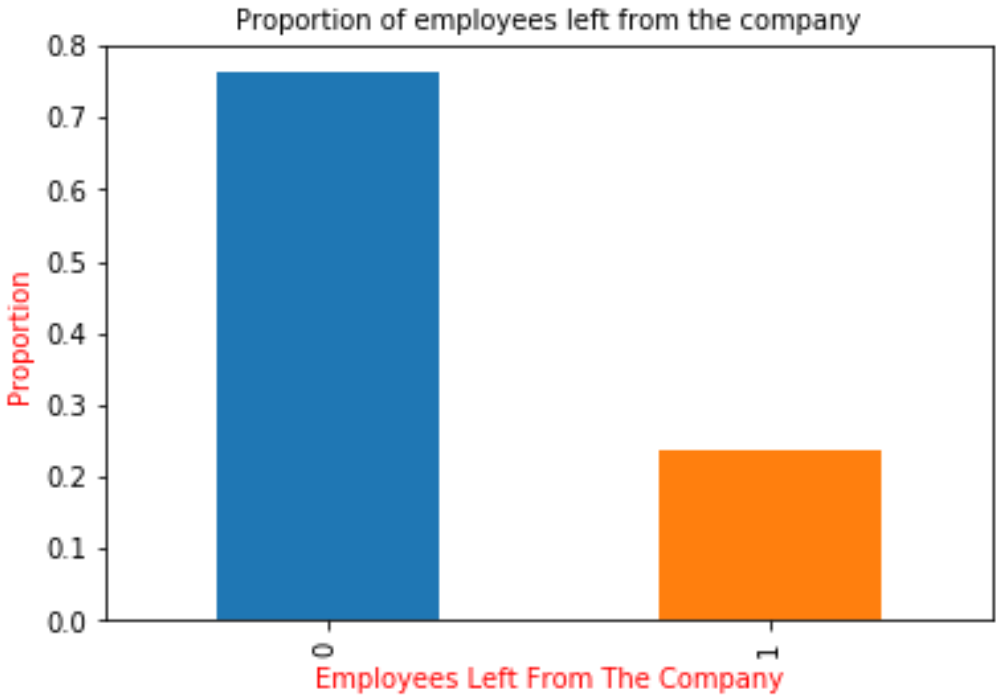


Bar-Graph For Work Accident Met By The Employee

Left : Left is a numerical variable which is 1 if the employee has left the company and 0 if not. This numerical value is shown in a histogram and bar graph.

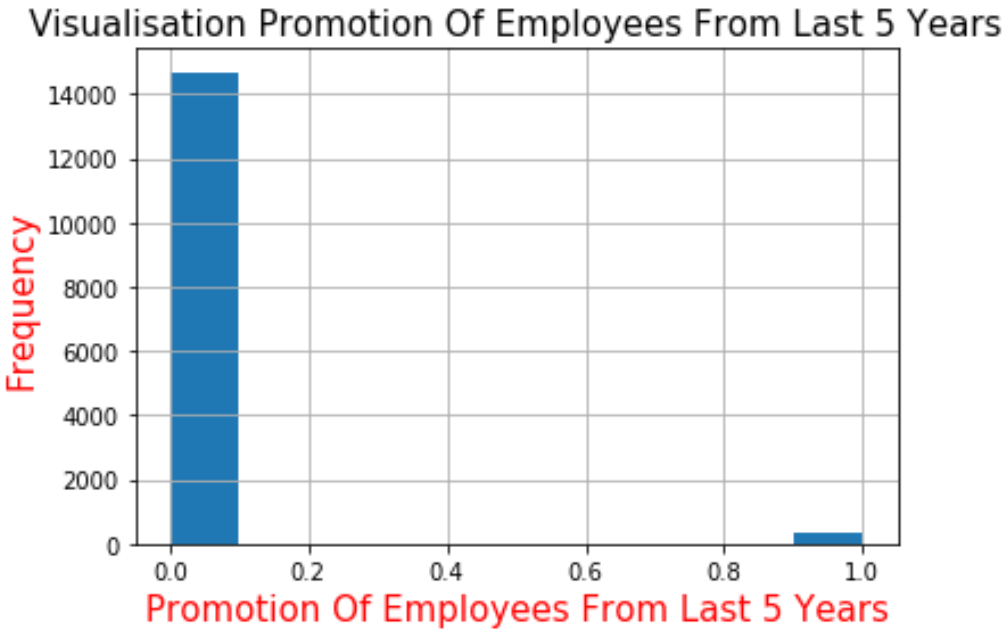


Histogram For Employees Left The Company

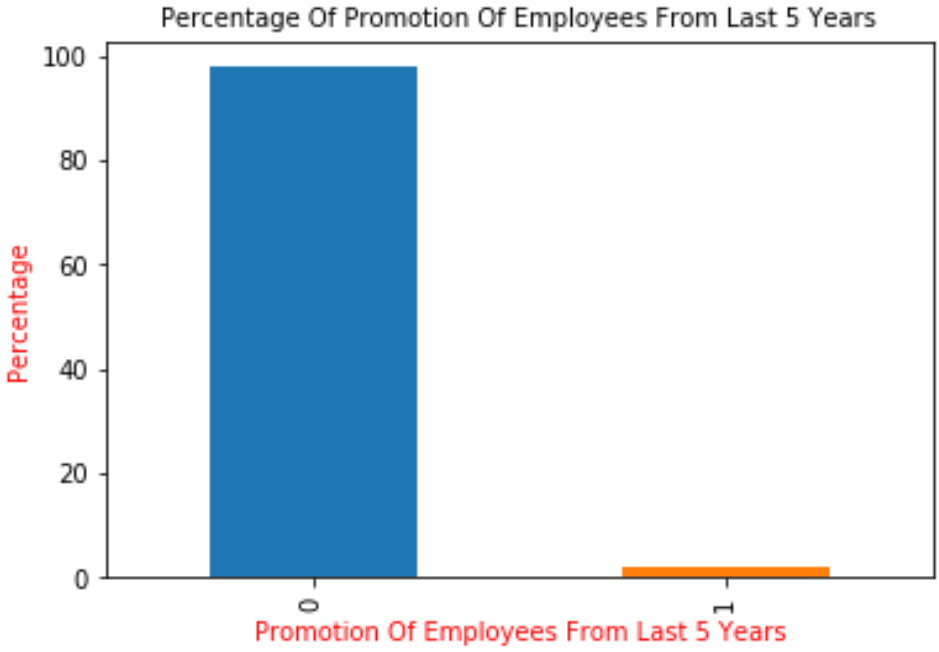


Bar-Graph For Employees Left The Company

Promotions: This is a numerical variable which shows whether the employee has got any promotion in the last five years. This is shown in histogram and bar graph for a better visualization.



Histogram For Promotions Of Each Employee

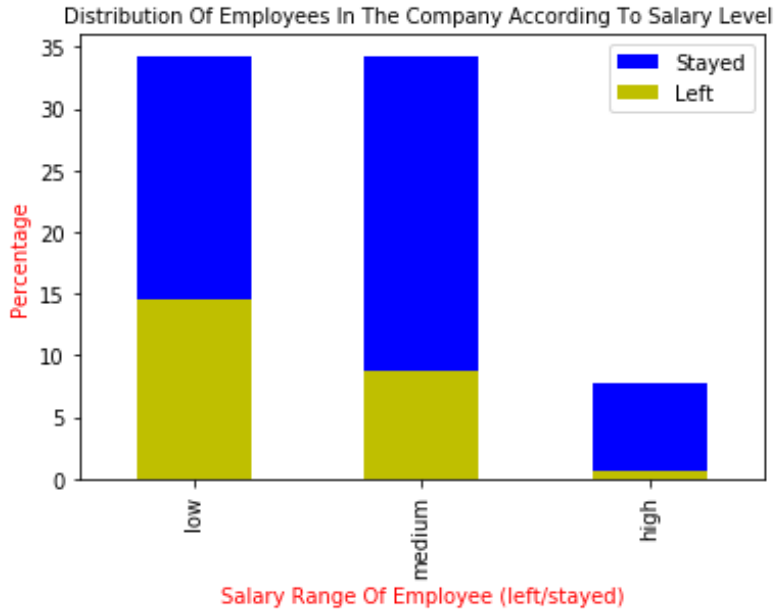


Bar-Graph For Promotions Of Each Employee

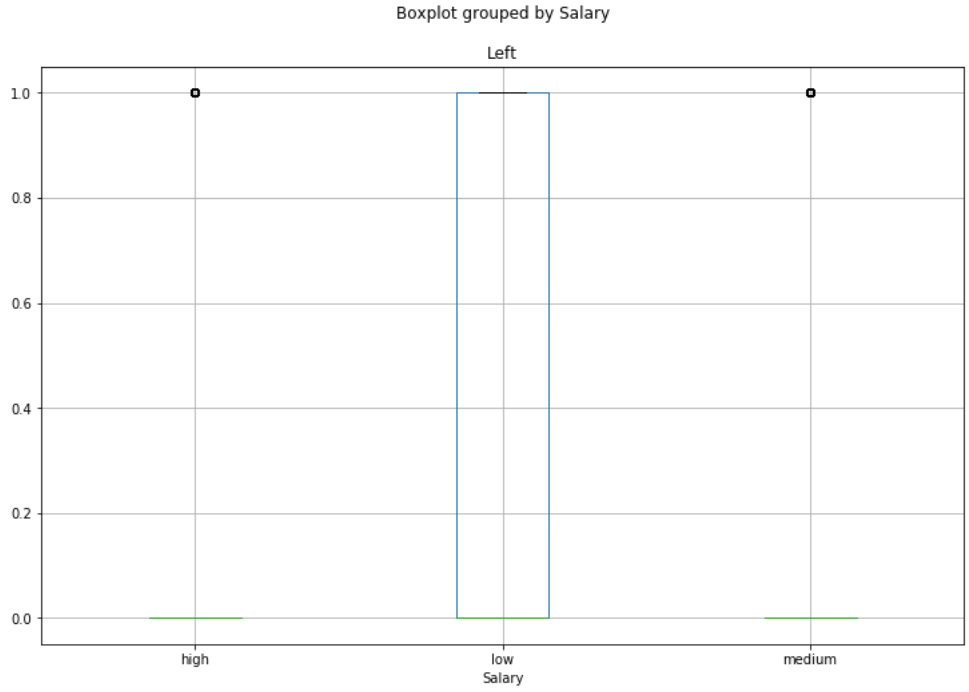
Visualizations For Relations Between Features:

Salary Vs Left:

From the visualization we can understand that in the low salary category the employees who left the company are almost half of those who stayed. In the medium range of salary the employees who left the company are almost quarter of those who stayed. At last its clear that in high salary category the employees left are very low when compared to those who stayed. Overall the employees who get the high salary are in the least proportion in the company.

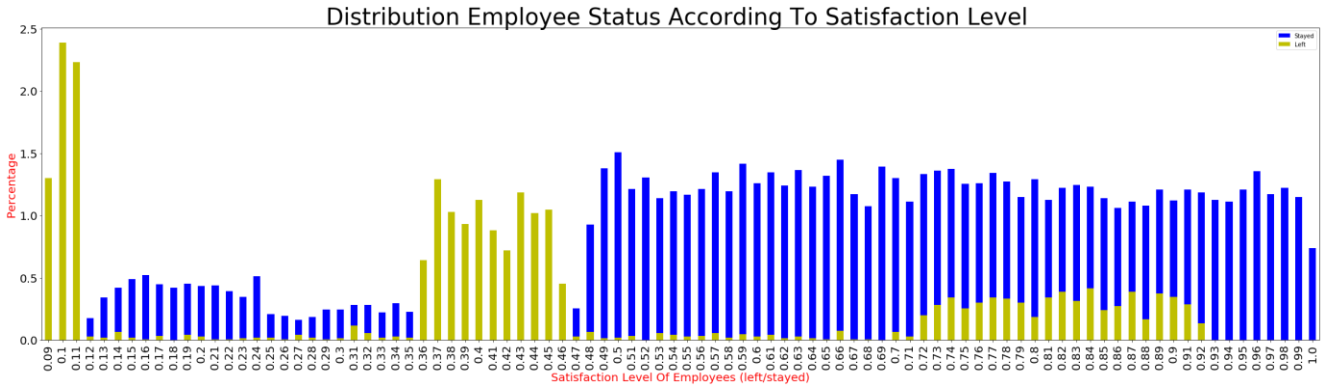


For the better understandability of descriptive statistics of salary range and employee left we plotted the box plot below.



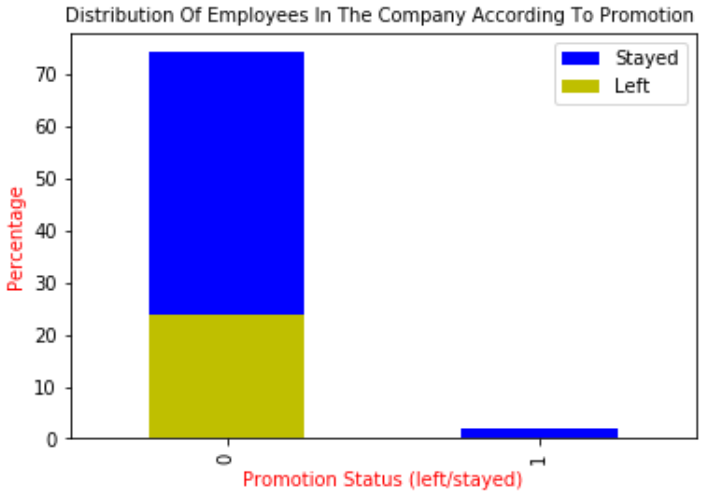
Satisfaction Level Vs Left:

The employees who satisfaction level less than 0.12 have left the company as expected however majority of employees who have satisfaction level between 0.12 and 0.35 have stayed. Ot the other hand those who have the satisfaction level within 0.36 to 0.46 have left the organization. Leaving percentage has decreased afterwards, up to 0.71. Not surprisingly, the employees whose satisfaction level is above 0.93 have decided not to leave the company. From this we can observe that there is no relation between the satisfaction level and the employees left.



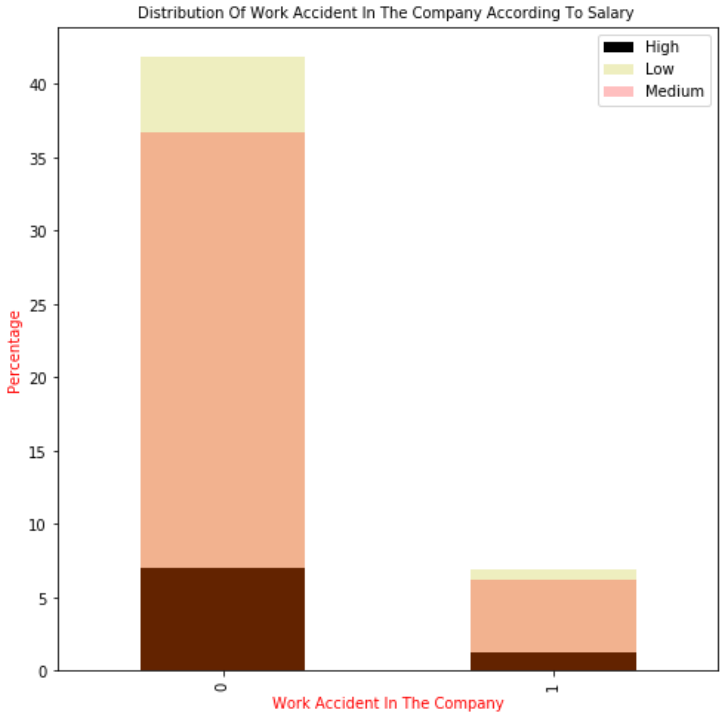
Left Vs Promotions:

The promotion is not affecting the turnover rate is not much dependent on each other as we can see that approximately 75% of the employees who did not get the promotion stayed back in the company and the rest left the company. The percentage of employees who got promotion is less then 5% and this is very low when compared to the other group. These depicts that the employees are not bothered about the promotion much.



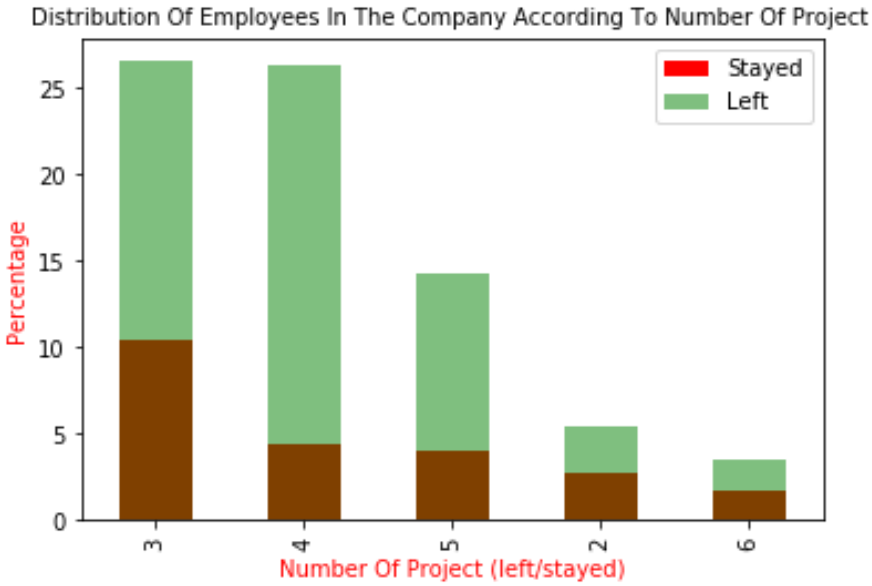
Average Hours Spent Vs Salary:

The low salaried employees are high in both the work accident categories(those who met with accident or not) the medium range salaried people fall just below the low. The high salaried people are at the bottom line among both categories.

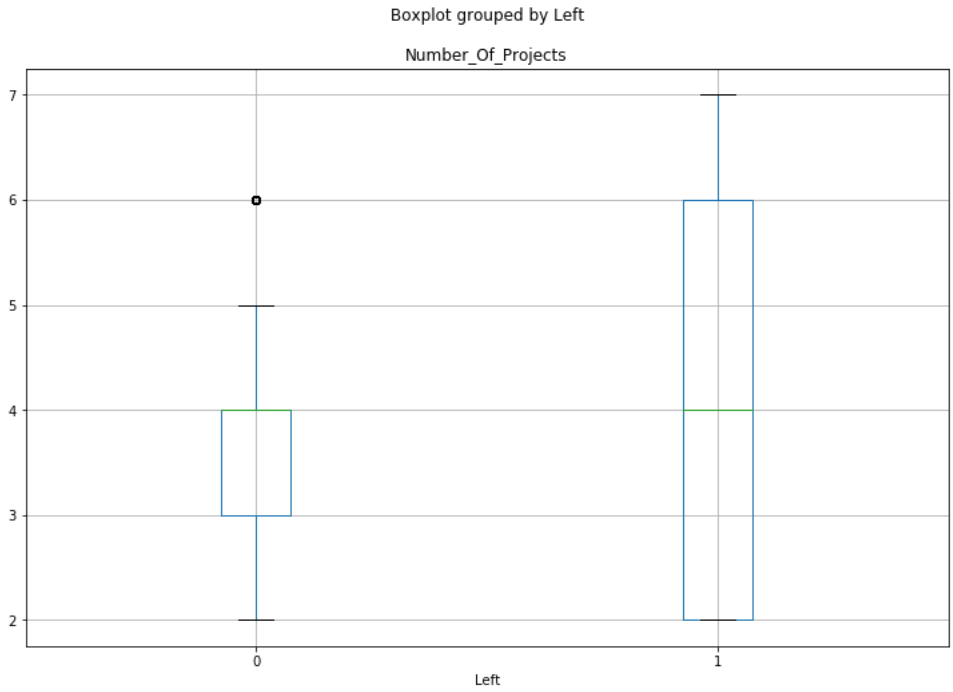


Number Of Projects Vs Left:

From the below plot its clear that the proportion of employees who left the organization is high irrespective of the number projects they are involved in. Also we cant in the category who stayed in the company does not show a proper correlation between number of projects

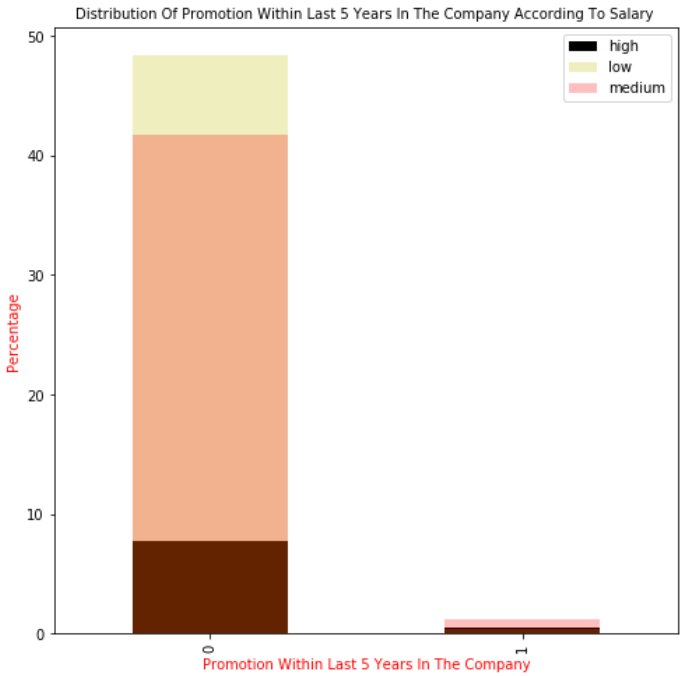


For the better understandability of descriptive statistics of number of projects the employee is involved and employees left we plotted the box plot below. From the below visualization it is clear that the mean line for both the categories are almost same (mean = 4).



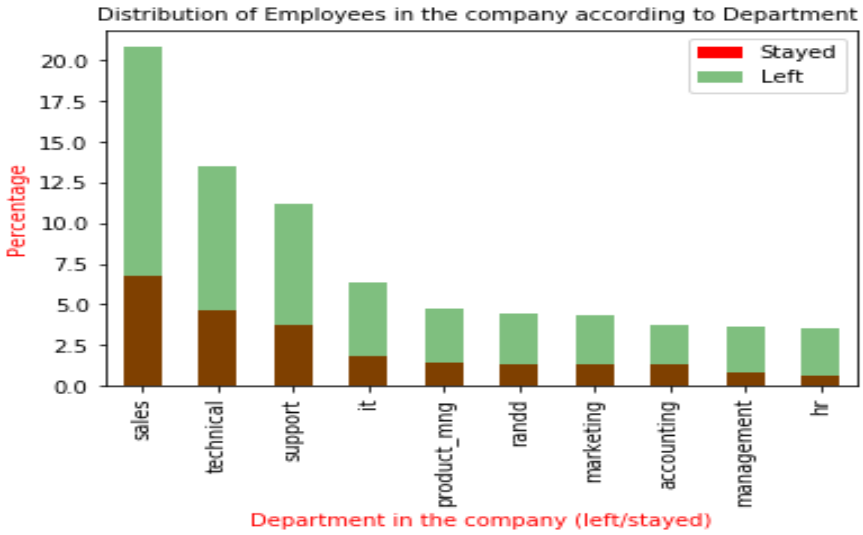
Promotions Vs Salary:

From the graph its clear that more than 90% of employees are not getting any promotion. In the promotion category medium range salaried employees are high compared to high level salaried workers. The high salaried people are at the bottom line among both categories.

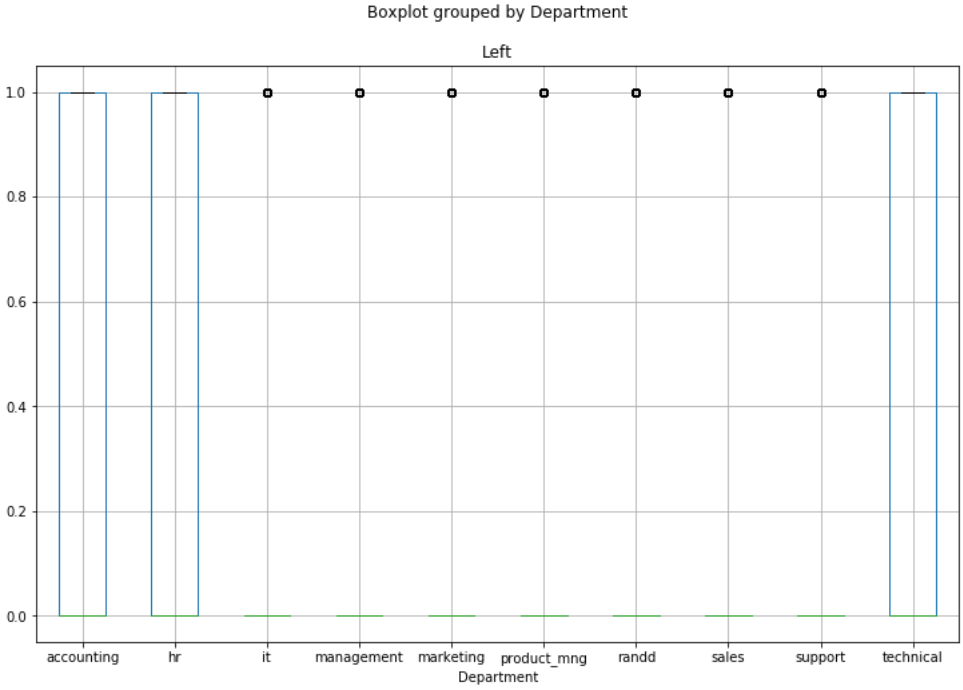


Department vs Left:

In all the departments the left percentage of employees is greater than those who stayed. The sales department has got the maximum number of employees, whereas the HR department has the least.

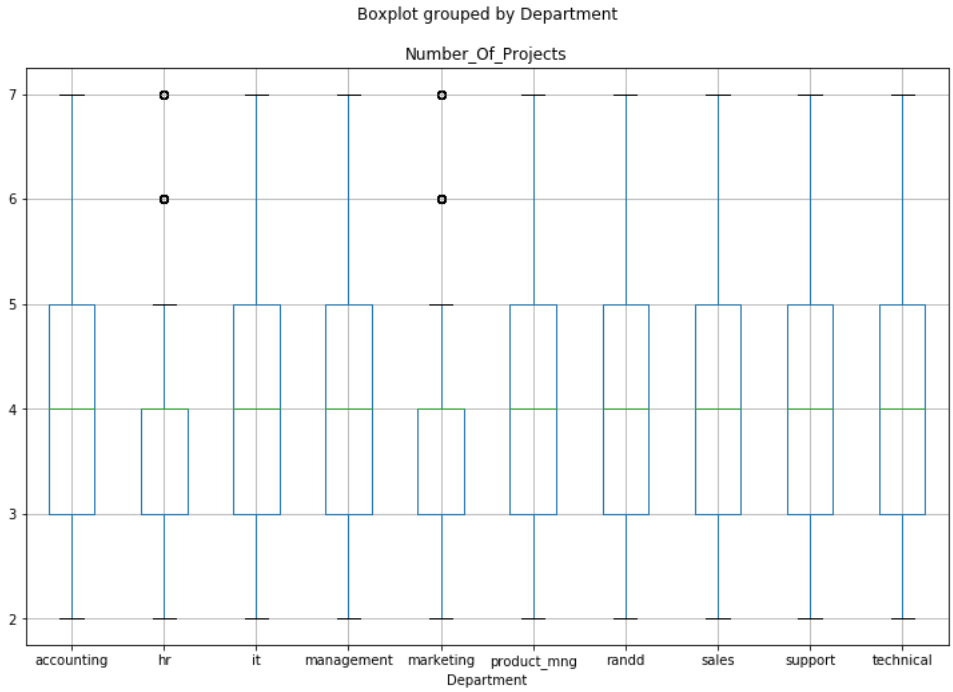


For the better understandability of descriptive statistics of number of projects the employee is involved and employees left we plotted the box plot below. From the below visualization it is clear that the mean line for all the categories are almost same (mean = 0).



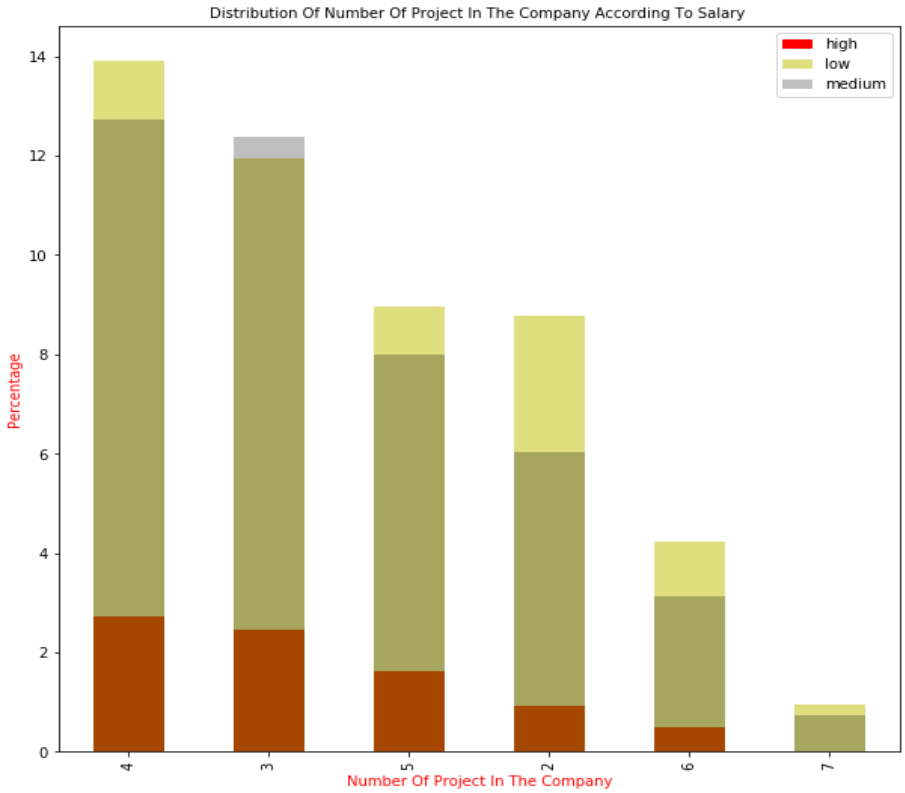
Number Of Projects Vs Department:

For all the department the mean of the number of projects is 4 and the first inter quartile is 3, while all the departments except for HR and Management have their upper quartile range which is equal to 5. The HR and Management department shows the same upper quartile range which shows 4 that is same as the mean.

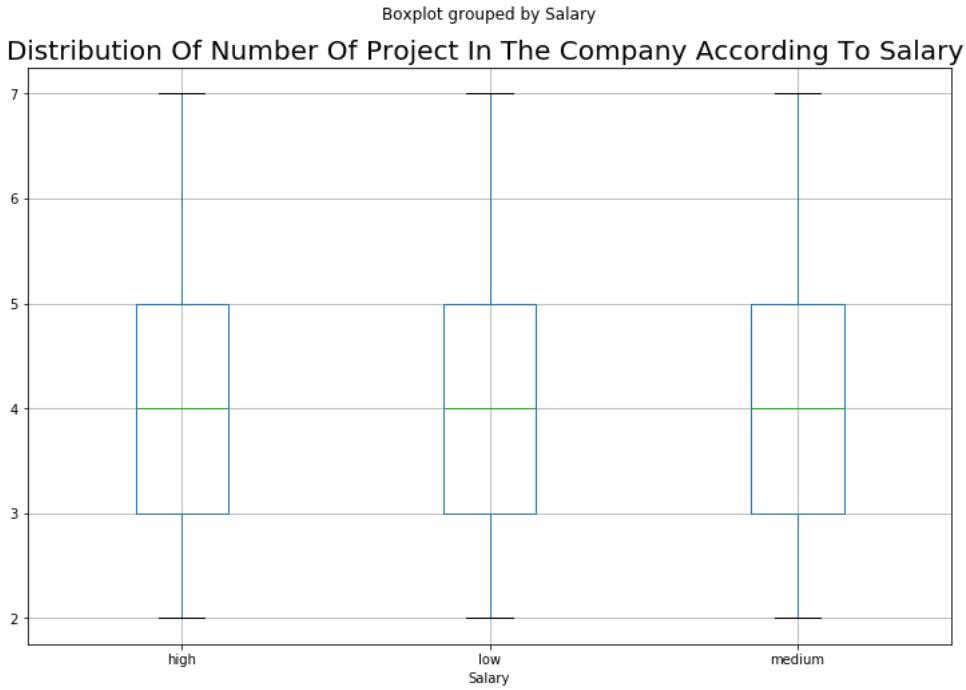


Number Of Projects Vs Salary:

From the below graph we can say that there are no high salaried employees involved in 7 projects at the same time. Also for the other different number of projects the high salaried people are less when compared to the medium and low salaried employees. Unlike other groups the medium salary ranged employees are higher the group which is involved in three number of projects.

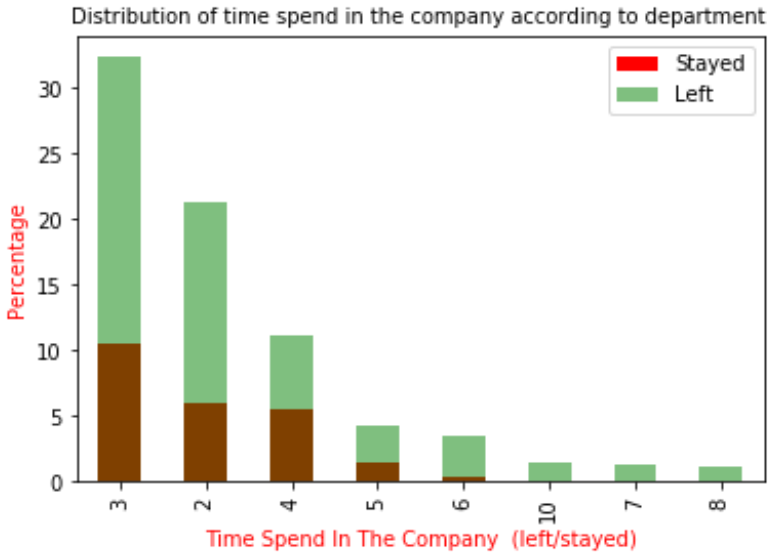


For the better understandability of descriptive statistics of number of projects the employee is involved from each salary group we plotted the box plot below. From the below visualization it is clear that all three categories show same behavior in mean median etc.



Left Vs Time Spend Company:

For all the different groups the left percentage of employees is greater than those who stayed. The employee in 3 working hours group has got the maximum proportion employees, whereas the 10,7,8 working hours groups has the least. From this we can state that those who worked for more hours has left the company and those who worked for 2 to 5 hours has the tendency to stay in the company.



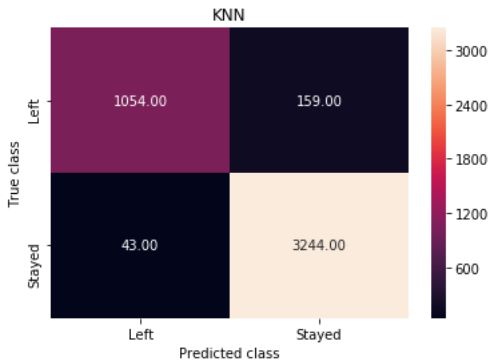
Descriptive Statistics Table For The Employee Turnover:

	count	mean	std	min	25%	50%	75%	max
Satisfaction_Level	14999.0	0.612834	0.248631	0.09	0.44	0.64	0.82	1.0
Last_Evaluation	14999.0	0.716102	0.171169	0.36	0.56	0.72	0.87	1.0
Number_Of_Projects	14999.0	3.803054	1.232592	2.00	3.00	4.00	5.00	7.0
Avg_Month_Hrs	14999.0	201.050337	49.943099	96.00	156.00	200.00	245.00	310.0
Time_Spend_Company	14999.0	3.498233	1.460136	2.00	3.00	3.00	4.00	10.0
Work_Accident	14999.0	0.144610	0.351719	0.00	0.00	0.00	0.00	1.0
Left	14999.0	0.238083	0.425924	0.00	0.00	0.00	0.00	1.0
Promotions	14999.0	0.021268	0.144281	0.00	0.00	0.00	0.00	1.0

Discussions

K-Nearest Neighbor:

The model gives following confusion matrix, in which the classified 3244 employees in the "Stayed" category and 1054 employees in the "left" category.



The figure table indicates that we predicted that for the employees who continued to stay in the company with precision of 99% and recall of 95%. Also, for the employees who left the company the precision and recall values are predicted as 87% and 96% respectively. Since the overall predictions were above 85%,we can say that our K-Nearest Neighbor model functions well.

	precision	recall	f1-score	support
0	0.99	0.95	0.97	3403
1	0.87	0.96	0.91	1097
avg / total	0.96	0.96	0.96	4500

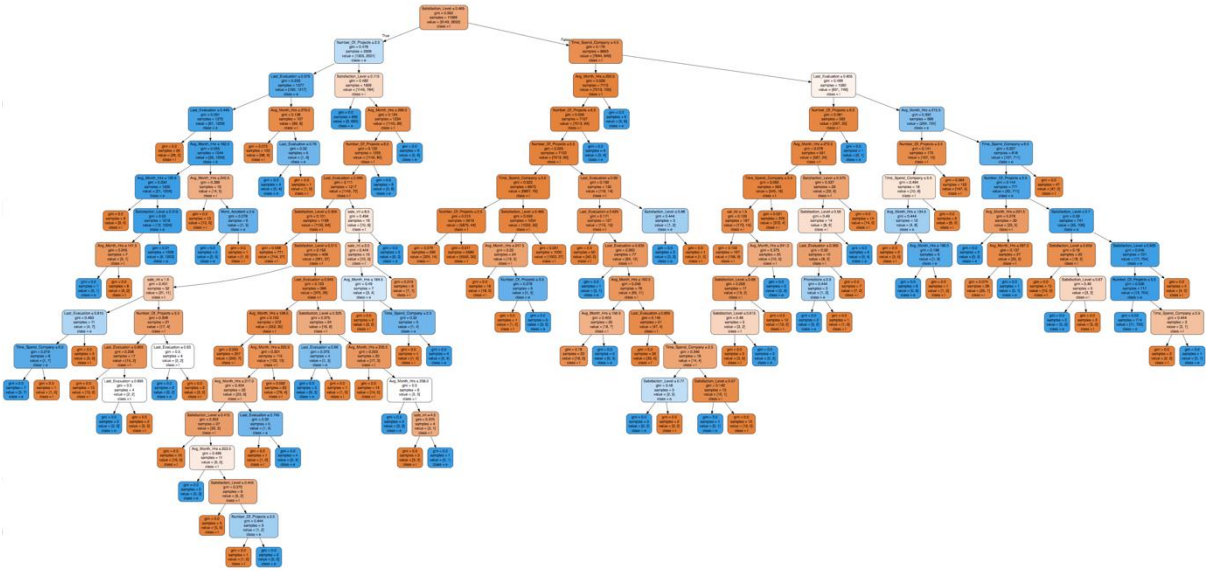
Decision tree:

The model gives following confusion matrix, in which the classified 2264 employees in the "Stayed" category and 679 employees in the "left" category.

The figure table indicates that we predicted that for the employees who continued to stay in the company with precision of 98% and recall of 99%. Also, for the employees who left the company the precision and recall values are predicted as 98% and 94% respectively. Since the overall predictions were above 90%,we can say that our Decision tree model functions well.

	precision	recall	f1-score	support
0	0.98	0.99	0.99	2279
1	0.98	0.94	0.96	721
avg / total	0.98	0.98	0.98	3000

The decision tree diagram is given below.



Logistic Regression:

The model gives following confusion matrix, in which the classified 1651 employees in the "Stayed" category and 594 employees in the "left" category.

The figure table indicates that we predicted that for the employees who continued to stay in the company with precision of 93% and recall of 72%. Also, for the employees who left the company the precision and recall values are predicted as 49% and 82% respectively.

	precision	recall	f1-score	support
0	0.93	0.72	0.81	2279
1	0.49	0.82	0.61	721
avg / total	0.82	0.75	0.77	3000

Out of the three models the Decision tree model shows best precision and recall percentage. From we can summarize that the Decision tree model predicts the employee behavior towards the organization accurately.

Conclusions

Employee turnover of organization depends on many internal and external factors, there are many reasons influencing the attitude of an employee towards continuing with the company or leaving it. From the responses recorded and analysis done it can be noted that the satisfaction level of employee contributes to a large percentage of their decision to continue or leave the company. Not only the satisfaction itself but also the other factors influence the employee turnover.

The research conducted using the existing data set concluded that we can predict whether an employee will stay or leave in the company from analyzing the given factors. The proper analysis and findings will help to predict the employee turnover properly and the organizations can take appropriate actions to satisfy the employee and make them stay in the firm.

Recommendations

Employee turnover rate can be reduced by many ways. Such as by providing competitive salary benefits, promotions, flexible working hours and reduce the work pressure. Employees prefer to stay in a work friendly environment with considerable salary. Organizations should always provide better working conditions to the employees to increase the satisfaction level and thereby increase the overall productivity.

References

1. https://www.python.org/doc/
2. http://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html
3. http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
4. http://scikit-learn.org/stable/modules/neighbors.html
5. https://www.kaggle.com/manojvijayan/
6. https://rmit.instructure.com/courses/15536/files/903695?module_item_id=671666