VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

J Component Report

Project Title: Employee Turnover Analysis

Course Code: SWE2011

Course Name: Big Data Analytics

Faculty: Dr. Saleena B

Team Members:

Vishnupriya S – 18MIS1022

N Pranitha – 18MIS1095

B Priyanka – 18MIS1097

Abstract:

Employee turnover is predicting when and why an employee will quit their work. It is similar as Customer turnover. There are various case studies available on employee turnover in a company. If an employee leaves the company, then it is known as turnover or churn. In case studies it was found that employee turnover will be affected majorly by work satisfaction, age, salary, marital status, working environment, etc. There are some employees with higher skills are very tough to replace. This will affect the ongoing projects and work and also it affects the productivity of already existing employees in the company. And also, there will be most cost effective because we need to replace the employees for that we need to hire the employees. So, there will be high hiring cost and also training cost. And also, we need to train the newly hired employee to the same level of business and technical knowledge of an older employee. For this it will take more time. So here the company will face both cost and time effective.

Here in this project, we will be solving this problem by applying some machine techniques and also by using big data technologies for predicting the employee's turnover. And also, we visualized our findings using visualization tool. This will help the company or organization to take the necessary actions before head.

Introduction:

Employee turnover is one of the critical problems in an organization now a days. The company needs to analyze this throughout the year. As it is difficult to predict, so the managers need to consider all the situations of an employee. If there is a high rate of turnover of employees, it will affect the organization in several ways.

One way of dealing this problem is by predicting the employee's turnover using some data mining techniques and also by some big data technologies. This will help the company organization and HR for taking the necessary actions towards the employee either by hiring the new employee or by increasing the facilities and making the current employee more comfortable to work in the company.

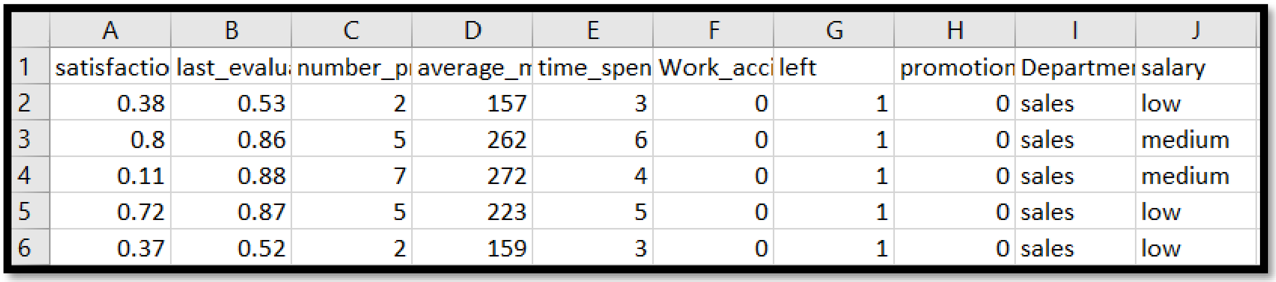
Acknowledgement:

We are profoundly grateful to Dr. Saleena B for her expert guidance and continuous encouragement throughout to see that this project reaches its target from its commencement to its completion.

Dataset Used:

Link: https://www.kaggle.com/aliu233/employee-turnover-prediction

Size: 15000 rows and 10 attributes. No null or missing values in the dataset.



	A	B	C	D	E	F	G	H	I	J
1	satisfactio	last_evalu	number_p	average_m	time_spen	Work_acci	left	promotion	Departme	salary
2	0.38	0.53	2	157	3	0	1	0	sales	low
3	0.8	0.86	5	262	6	0	1	0	sales	medium
4	0.11	0.88	7	272	4	0	1	0	sales	medium
5	0.72	0.87	5	223	5	0	1	0	sales	low
6	0.37	0.52	2	159	3	0	1	0	sales	low

The attributes in our dataset are:

satisfaction_level - It is employee satisfaction point, which ranges from 0-1.

last_evaluation - It is evaluated performance by the employer, which ranges from 0-1.

number_projects - Numbers of projects assigned to an employee working in a company.

average_monthly_hours - Average number of hours worked by an employee in a month.

time_spent_company - It is employee's working experience. The number of years spent in the company by an employee.

work_accident - It is an employee has had a work accident or not.

promotion_last_5years - Whether an employee has had a promotion in the last 5 years or not.

Departments - Employee's working department in a company.

Salary – It is the salary of the employee such as high, low and medium.

left - Here “1” means employee who left the company and “0” means employees who are working in the company and left is the target attribute for our project. It is the employee’s condition like working in the company or not.

Tools and Technologies Used:

1. Data Mining Algorithms
2. Hadoop
3. Hive
4. Tableau – Visualization

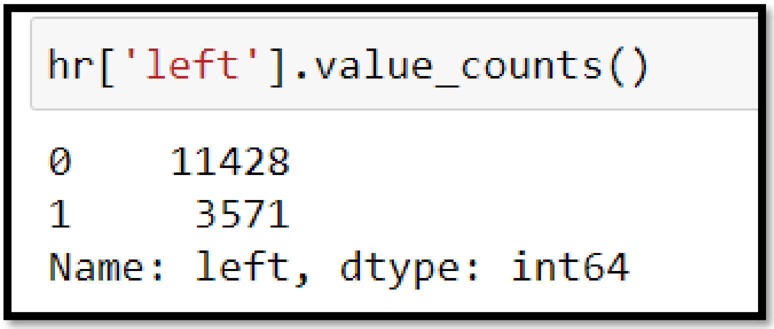
Methodologies Used:

- Feature Selection
- Logistic Regression
- Random Forest
- Support Vector Machine
- Cross Validation
- K – Means Clustering

Implementation Details:

Data Exploration:

Here we found the number of employees working in the company and the number of employees left the company from our dataset. There are 11428 employees are working in the company and 3572 employees were left the company.



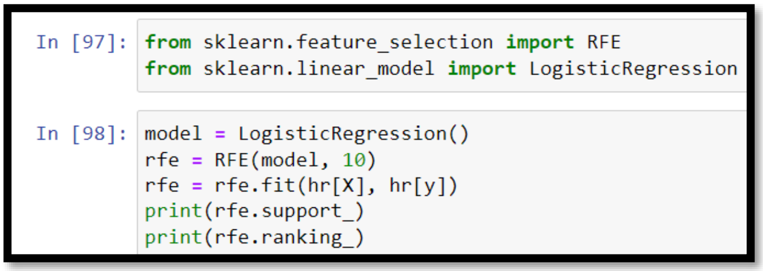
```
hr['left'].value_counts()
```

```
0    11428
1     3571
Name: left, dtype: int64
```

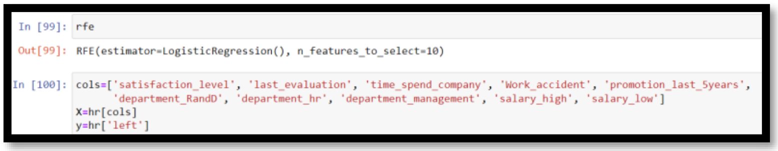
Data Mining Algorithms:

Feature Selection:

This Recursive Feature Elimination works recursively by removing the variables and with the remaining variables it will build the model. For predicting the target attribute to know which variable contribute the most, it uses the model accuracy.



```
In [97]: from sklearn.feature_selection import RFE
         from sklearn.linear_model import LogisticRegression

In [98]: model = LogisticRegression()
         rfe = RFE(model, 10)
         rfe = rfe.fit(hr[X], hr[y])
         print(rfe.support_)
         print(rfe.ranking_)
```

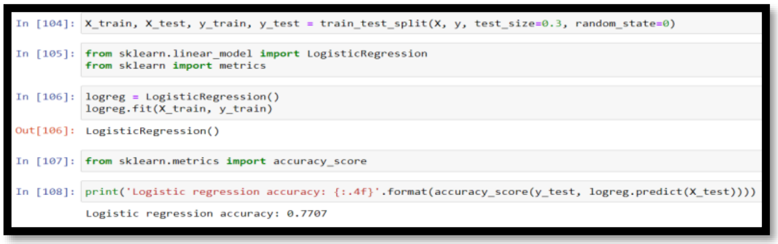


```
In [99]: rfe

Out[99]: RFE(estimator=LogisticRegression(), n_features_to_select=10)

In [100]: cols=['satisfaction_level', 'last_evaluation', 'time_spend_company', 'Work_accident', 'promotion_last_5years',
                'department_RandD', 'department_hr', 'department_management', 'salary_high', 'salary_low']
          X=hr[cols]
          y=hr['left']
```

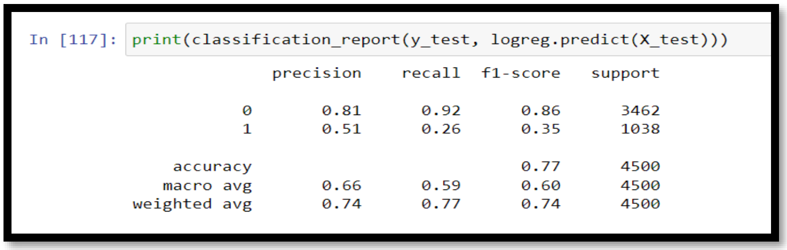
Logistic Regression:

Logistic regression is a geometrical analysis method used to anticipate a data value based on earlier perceptions of the data set. It also predicts a tentative data variable by analyzing the connection between at least one existing dependent variable.

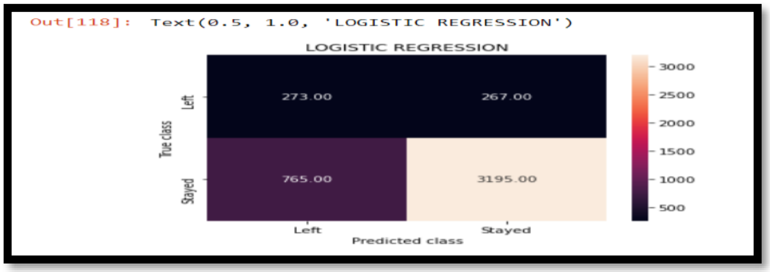
The Logistic Regression accuracy obtained is **0.771**.



```
In [104]: X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)

In [105]: from sklearn.linear_model import LogisticRegression
          from sklearn import metrics

In [106]: logreg = LogisticRegression()
          logreg.fit(X_train, y_train)

Out[106]: LogisticRegression()

In [107]: from sklearn.metrics import accuracy_score

In [108]: print('Logistic regression accuracy: {:.4f}'.format(accuracy_score(y_test, logreg.predict(X_test))))

          Logistic regression accuracy: 0.7707
```



```
In [117]: print(classification_report(y_test, logreg.predict(X_test)))
                        precision    recall  f1-score   support

                   0       0.81      0.92      0.86      3462
                   1       0.51      0.26      0.35      1038

            accuracy                           0.77      4500
           macro avg       0.66      0.59      0.60      4500
        weighted avg       0.74      0.77      0.74      4500
```

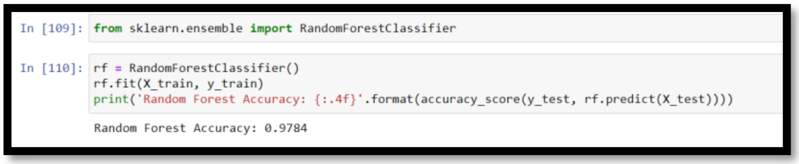


```
Out[118]: Text(0.5, 1.0, 'LOGISTIC REGRESSION')
```

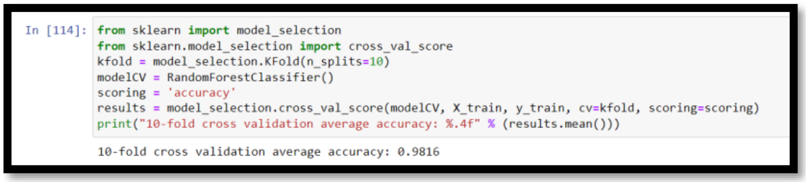
Random Forest:

Random forest joins hundreds or thousands of decision trees, prepares everyone on a unique arrangement of the observations, splitting nodes in each tree thinking about a limited number of the features. The final predictions of the random forest are made by averaging the predictions of every individual tree. We can quantify how each feature declines the pollutant of the split. For each component we can gather how on normal it diminishes the impurity. The normal over all trees in the forest is the proportion of the component importance.

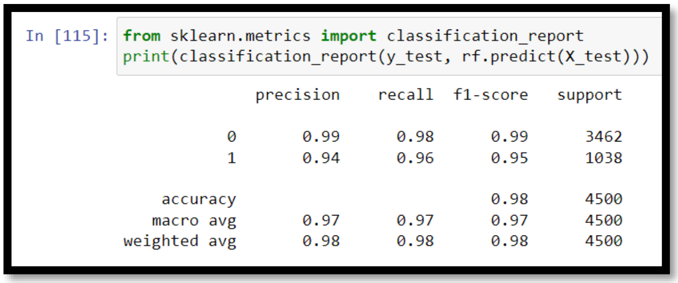
The Random Forest accuracy obtained is **0.97**.



```
In [109]: from sklearn.ensemble import RandomForestClassifier

In [110]: rf = RandomForestClassifier()
          rf.fit(X_train, y_train)
          print('Random Forest Accuracy: {:.4f}'.format(accuracy_score(y_test, rf.predict(X_test))))

          Random Forest Accuracy: 0.9784
```

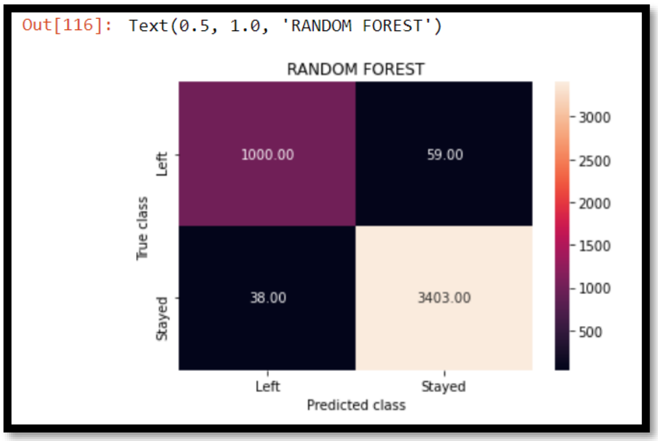


```
In [114]: from sklearn import model_selection
          from sklearn.model_selection import cross_val_score
          kfold = model_selection.KFold(n_splits=10)
          modelCV = RandomForestClassifier()
          scoring = 'accuracy'
          results = model_selection.cross_val_score(modelCV, X_train, y_train, cv=kfold, scoring=scoring)
          print("10-fold cross validation average accuracy: %.4f" % (results.mean()))

          10-fold cross validation average accuracy: 0.9816
```



```
In [115]: from sklearn.metrics import classification_report
          print(classification_report(y_test, rf.predict(X_test)))
```

```
              precision    recall  f1-score   support

           0       0.99      0.98      0.99      3462
           1       0.94      0.96      0.95      1038

    accuracy                           0.98      4500
   macro avg       0.97      0.97      0.97      4500
weighted avg       0.98      0.98      0.98      4500
```

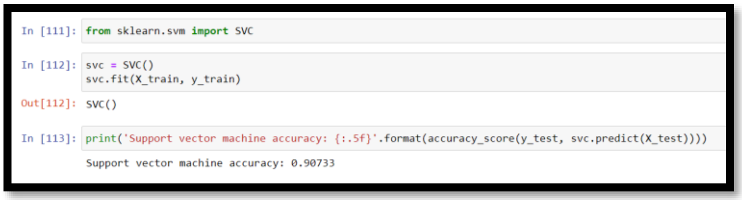


```
Out[116]: Text(0.5, 1.0, 'RANDOM FOREST')
```

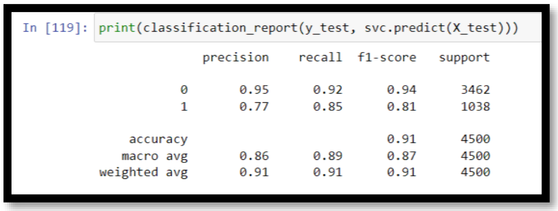
Support Vector Machine:

Support Vector Machines (SVMs) are supervised learning methods used for classification and the regression tasks that began from measurable learning theory. As an order technique, SVM is a global classification model that generates non-overlapping partitions and usually employs all attributes. When an employee left, how regularly does the classifier predict that effectively? This estimation is called “recall" and a brief glance at these diagrams can exhibit that random forest is clearly best for these criteria. The turnover “recall" of about (991 out of 1038), far better than logistic regression or support vector machines. When a classifier predicts an employee will leave, how regularly does that representative really leave? This estimation is designated as “precision". Random forest again outperforms the other two at about (991 out of 1045) precision with logistic regression at about (273 out of 540), and support vector machine at about (890 out of 1150).

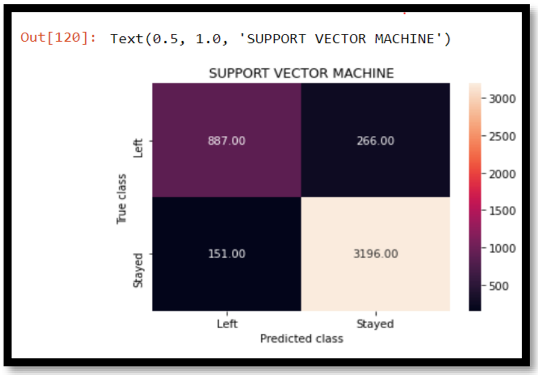
The Support Vector Machine accuracy obtained is **0.91**.



```
In [111]: from sklearn.svm import SVC

In [112]: svc = SVC()
          svc.fit(X_train, y_train)

Out[112]: SVC()

In [113]: print('Support vector machine accuracy: {:.5f}'.format(accuracy_score(y_test, svc.predict(X_test))))

          Support vector machine accuracy: 0.90733
```



```
In [119]: print(classification_report(y_test, svc.predict(X_test)))

                        precision    recall  f1-score   support

                   0       0.95      0.92      0.94      3462
                   1       0.77      0.85      0.81      1038

            accuracy                           0.91      4500
           macro avg       0.86      0.89      0.87      4500
        weighted avg       0.91      0.91      0.91      4500
```

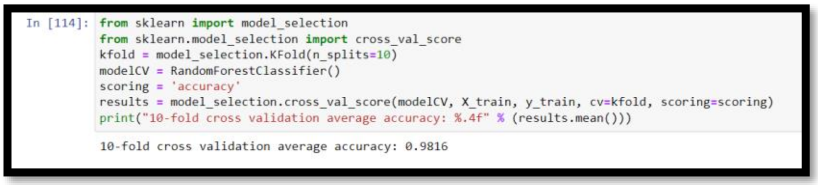


```
Out[120]: Text(0.5, 1.0, 'SUPPORT VECTOR MACHINE')
```

Cross Validation:

Cross validation endeavors to avoid overfitting while as yet creating an expectation for very perception dataset. We are utilizing 10-fold Cross-Validation to prepare our Random Forest model. The average accuracy remains very much close to the Random Forest model accuracy. Hence, we can conclude that the model generalizes well.

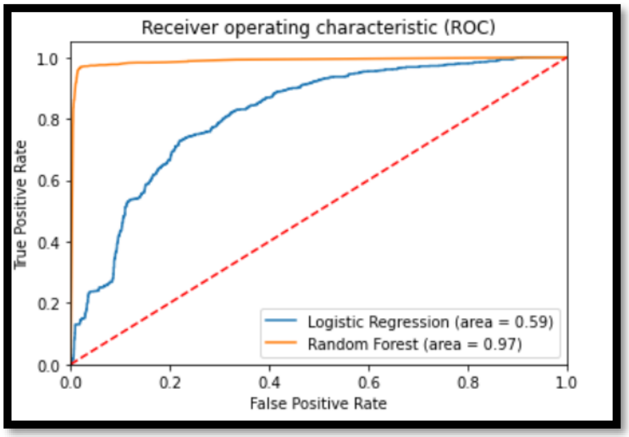
Cross validation average accuracy is **0.977**.



```
In [114]: from sklearn import model_selection
          from sklearn.model_selection import cross_val_score
          kfold = model_selection.KFold(n_splits=10)
          modelCV = RandomForestClassifier()
          scoring = 'accuracy'
          results = model_selection.cross_val_score(modelCV, X_train, y_train, cv=kfold, scoring=scoring)
          print("10-fold cross validation average accuracy: %.4f" % (results.mean()))

          10-fold cross validation average accuracy: 0.9816
```

ROC Curve:

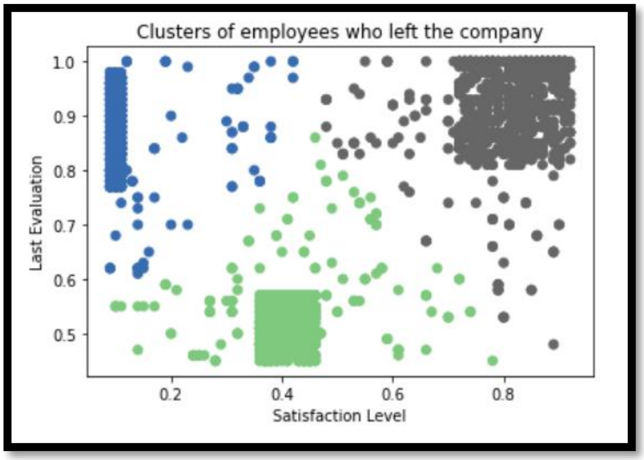
ROC curve, is a graphical plot that shows the analytic capacity of a double classifier framework as its separation edge is varied. The strategy was created for administrators of military radar beneficiaries, which is the reason it is so named.



Clustering:

K Means Clustering:

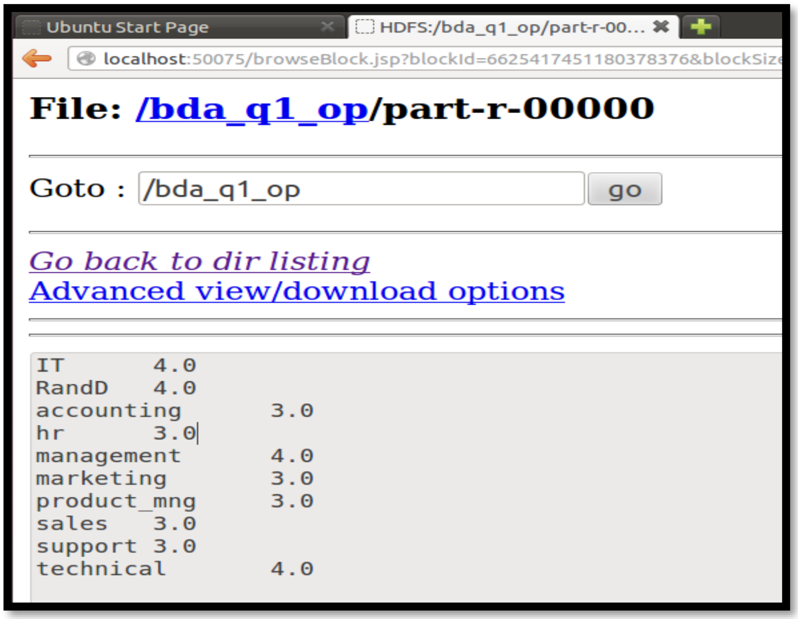
K-Means clustering plans to partition n objects into k clusters in which each item has a place with the cluster with the closest mean. It is an unsupervised learning algorithm.



Here we found the clusters for employees who left the company. We analysed that the most important for the employee to work or leave the company is their satisfaction level and their performance in the company. So we have done cluster based on those attributes using k means cluster analysis.

Hadoop:

1) Finding the average number of projects for the departments who left the company.


Ubuntu Start Page
HDFS:/bda_q1_op/part-r-00...
localhost:50075/browseBlock.jsp?blockId=6625417451180378376&blockSize

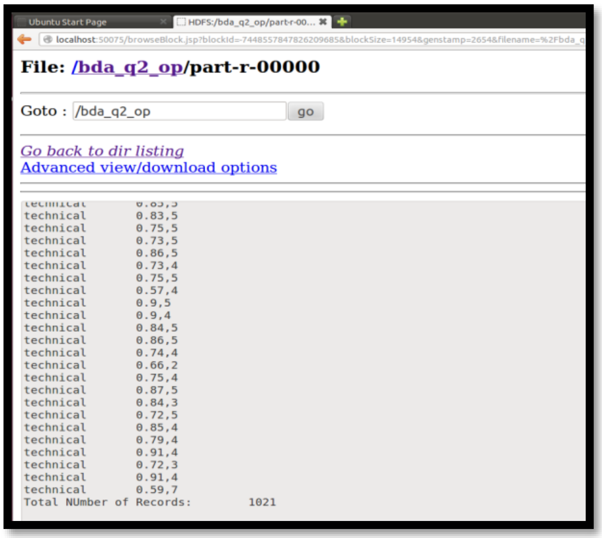
File: /bda_q1_op/part-r-00000

Goto : /bda_q1_op go

Go back to dir listing
Advanced view/download options

```
IT      4.0
RandD   4.0
accounting      3.0
hr      3.0
management      4.0
marketing       3.0
product_mng     3.0
sales   3.0
support 3.0
technical       4.0
```

2) Finding the details of the employees with the count who left the company(left==1) with the satisfaction level > 0.5.



Ubuntu Start Page
HDFS:/bda_q2_op/part-r-00...
localhost:50075/browseBlock.jsp?blockId=-7448557847826209685&blockSize=14954&genstamp=2654&filename=%2Fbda_q

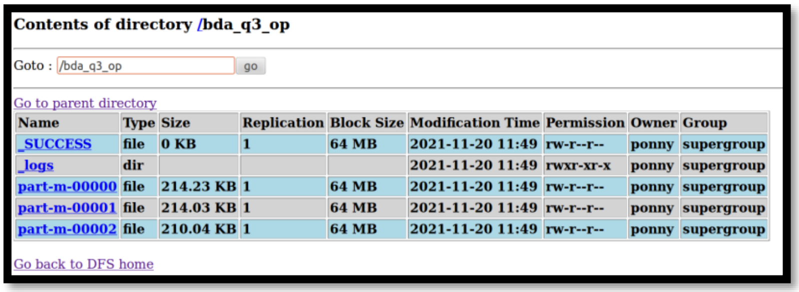
File: /bda_q2_op/part-r-00000

Goto : /bda_q2_op go

Go back to dir listing
Advanced view/download options

```
technical       0.83,5
technical       0.75,5
technical       0.73,5
technical       0.86,5
technical       0.73,4
technical       0.75,5
technical       0.57,4
technical       0.9,5
technical       0.9,4
technical       0.84,5
technical       0.86,5
technical       0.74,4
technical       0.66,2
technical       0.75,4
technical       0.87,5
technical       0.84,3
technical       0.72,5
technical       0.85,4
technical       0.79,4
technical       0.91,4
technical       0.72,3
technical       0.91,4
technical       0.59,7
Total NUmber of Records:        1021
```

3) Using NLineinput Format for our dataset and dividing into 5000 lines in each mapper.



Contents of directory /bda_q3_op

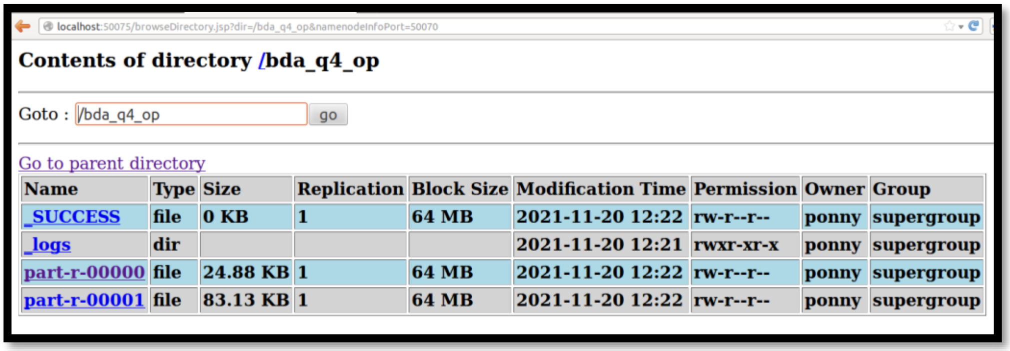
Goto : /bda_q3_op go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
_SUCCESS	file	0 KB	1	64 MB	2021-11-20 11:49	rw-r--r--	ponny	supergroup
_logs	dir				2021-11-20 11:49	rwxr-xr-x	ponny	supergroup
part-m-00000	file	214.23 KB	1	64 MB	2021-11-20 11:49	rw-r--r--	ponny	supergroup
part-m-00001	file	214.03 KB	1	64 MB	2021-11-20 11:49	rw-r--r--	ponny	supergroup
part-m-00002	file	210.04 KB	1	64 MB	2021-11-20 11:49	rw-r--r--	ponny	supergroup

Go back to DFS home

4) Partitioning our dataset based on the left attribute. Creating 2 partitions with left = 1(who left the company) and left = 0 and writing the salary details along with the left attribute value. Also displaying the count of the records satisfying the above conditions.



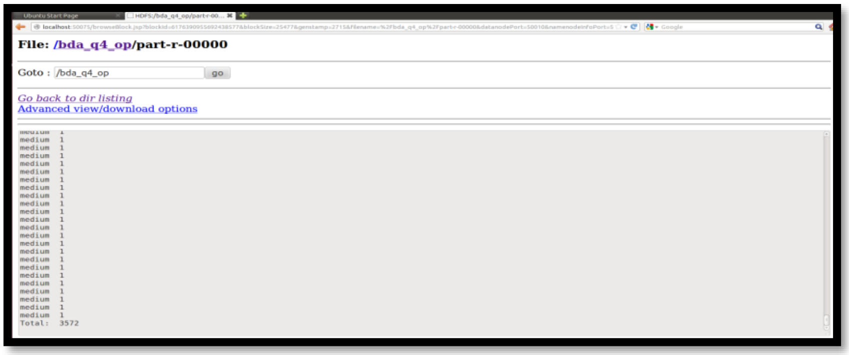
localhost:50075/browseDirectory.jsp?dir=/bda_q4_op&namenodeInfoPort=50070

Contents of directory /bda_q4_op

Goto : /bda_q4_op go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
_SUCCESS	file	0 KB	1	64 MB	2021-11-20 12:22	rw-r--r--	ponny	supergroup
_logs	dir				2021-11-20 12:21	rwxr-xr-x	ponny	supergroup
part-r-00000	file	24.88 KB	1	64 MB	2021-11-20 12:22	rw-r--r--	ponny	supergroup
part-r-00001	file	83.13 KB	1	64 MB	2021-11-20 12:22	rw-r--r--	ponny	supergroup


Ubuntu Start Page
HDFS:/bda_q4_op/part-r-00...
localhost:50075/browseBlock.jsp?blockId=6176390955692438577&blockSize=25477&genstamp=2715&filename=%2Fbda_q4_op%2Fpart-r-00000&datanodePort=50010&namenodeInfoPort=5
Google

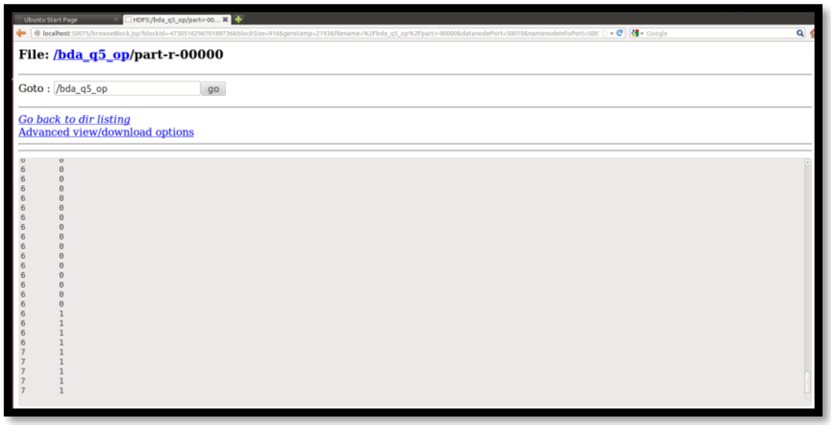
File: /bda_q4_op/part-r-00000

Goto : /bda_q4_op go

Go back to dir listing
Advanced view/download options

```
medium  1
medium  1
medium  1
medium  1
medium  1
medium  1
medium  1
medium  1
medium  1
medium  1
medium  1
medium  1
medium  1
medium  1
medium  1
medium  1
medium  1
medium  1
medium  1
medium  1
medium  1
medium  1
medium  1
Total:  3572
```

5) Using the user defined counter's function, we printed the details of the employees whose satisfaction level is greater than 0.5 and number of projects greater than 5 along with the left attribute.


Ubuntu Start Page
HDFS:/bda_q5_op/part-r-00...
localhost:50075/browseBlock.jsp?blockId=4730516296701897368&blockSize=916&genstamp=2743&filename=%2Fbda_q5_op%2Fpart-r-00000&datanodePort=50010&namenodeInfoPort=500
Google

File: /bda_q5_op/part-r-00000

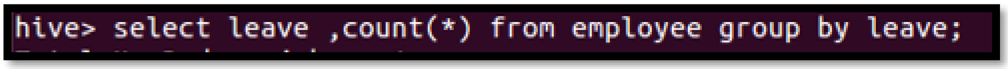
Goto : /bda_q5_op go

Go back to dir listing
Advanced view/download options

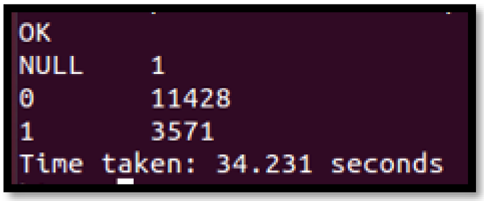
```
6       0
6       0
6       0
6       0
6       0
6       0
6       0
6       0
6       0
6       0
6       0
6       0
6       0
6       0
6       0
6       1
6       1
6       1
6       1
7       1
7       1
7       1
7       1
7       1
```

Hive:

1) Query to count number of employees "left" from the company.

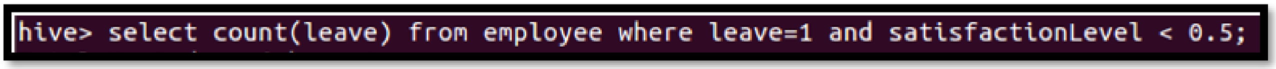


```
hive> select leave ,count(*) from employee group by leave;
```

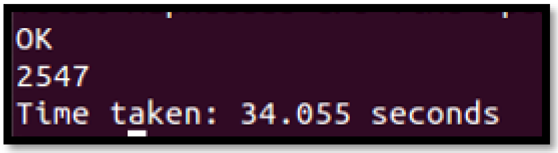


```
OK
NULL    1
0       11428
1       3571
Time taken: 34.231 seconds
```

2) Based on the satisfaction level, checking the number of employees left the company.

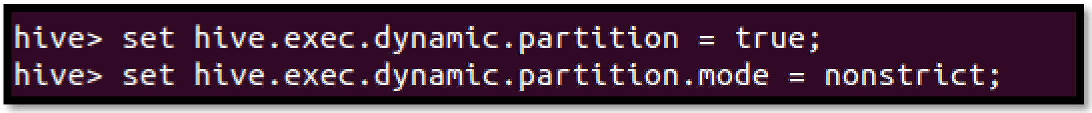


```
hive> select count(leave) from employee where leave=1 and satisfactionLevel < 0.5;
```

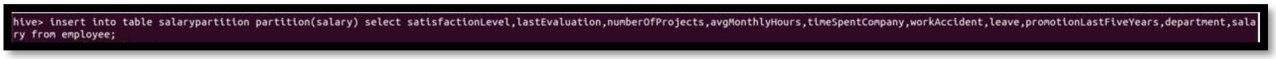


```
OK
2547
Time taken: 34.055 seconds
```

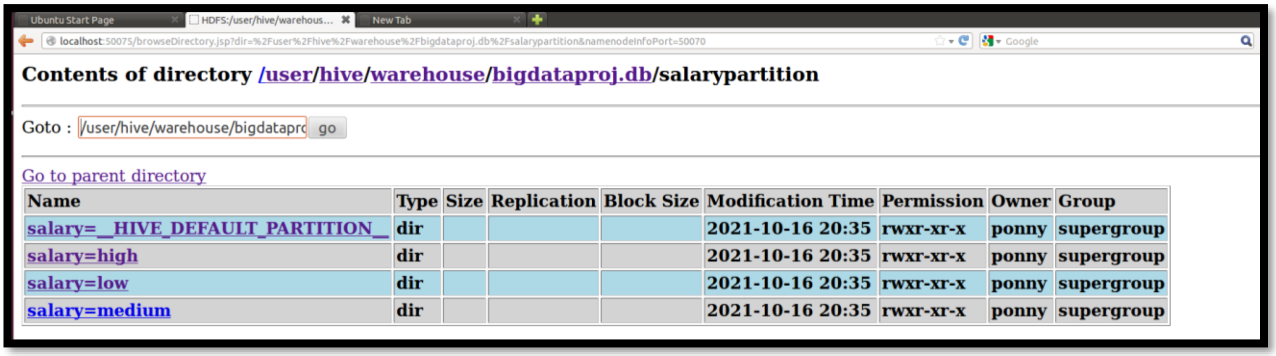
3) Dynamic Partitioning



```
hive> set hive.exec.dynamic.partition = true;
hive> set hive.exec.dynamic.partition.mode = nonstrict;
```



```
hive> insert into table salarypartition partition(salary) select satisfactionLevel,lastEvaluation,numberOfProjects,avgMonthlyHours,timeSpentCompany,workAccident,leave,promotionLastFiveYears,department,sala
ry from employee;
```



Ubuntu Start Page
HDFS:/user/hive/warehous...
New Tab
localhost:50075/browseDirectory.jsp?dir=%2Fuser%2Fhive%2Fwarehouse%2Fbigdataproj.db%2Fsalarypartition&namenodeInfoPort=50070
Google

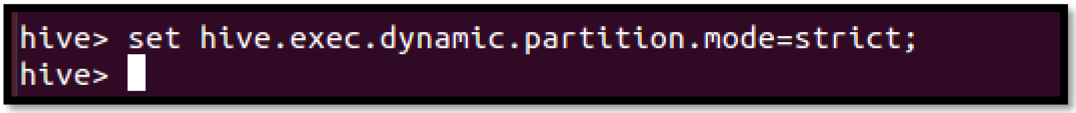
Contents of directory /user/hive/warehouse/bigdataproj.db/salarypartition

Goto : /user/hive/warehouse/bigdatapro go

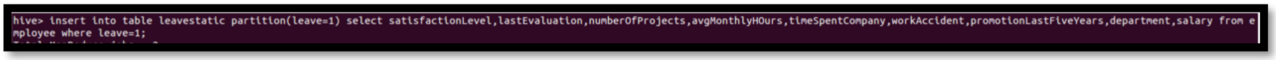
Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
salary=__HIVE_DEFAULT_PARTITION__	dir				2021-10-16 20:35	rwxr-xr-x	ponny	supergroup
salary=high	dir				2021-10-16 20:35	rwxr-xr-x	ponny	supergroup
salary=low	dir				2021-10-16 20:35	rwxr-xr-x	ponny	supergroup
salary=medium	dir				2021-10-16 20:35	rwxr-xr-x	ponny	supergroup

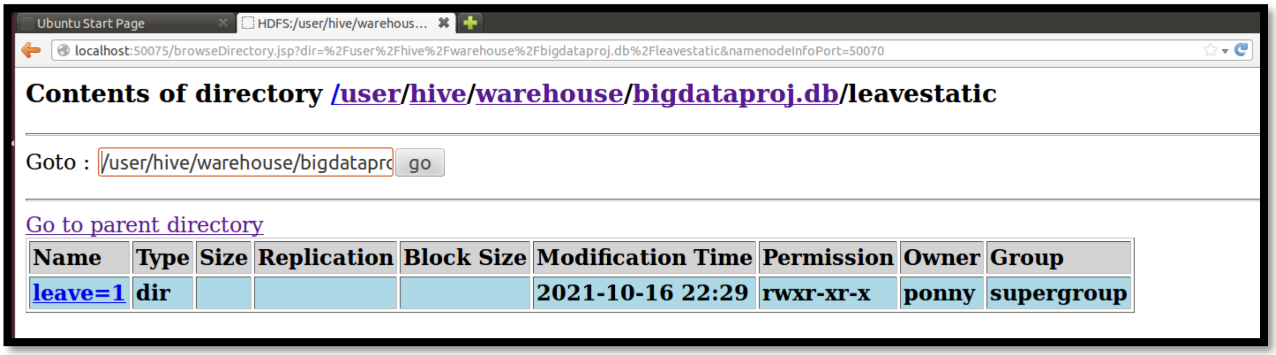
4) Static Partitioning:



```
hive> set hive.exec.dynamic.partition.mode=strict;
hive> 
```



```
hive> insert into table leavestatic partition(leave=1) select satisfactionLevel,lastEvaluation,numberOfProjects,avgMonthlyHOurs,timeSpentCompany,workAccident,promotionLastFiveYears,department,salary from e
mployee where leave=1;
```



Ubuntu Start Page
HDFS:/user/hive/warehous...
localhost:50075/browseDirectory.jsp?dir=%2Fuser%2Fhive%2Fwarehouse%2Fbigdataproj.db%2Fleavestatic&namenodeInfoPort=50070

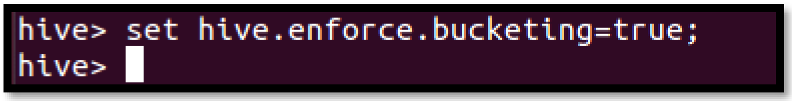
Contents of directory /user/hive/warehouse/bigdataproj.db/leavestatic

Goto : /user/hive/warehouse/bigdatapro go

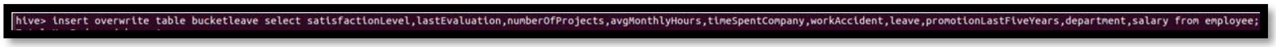
Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
leave=1	dir				2021-10-16 22:29	rwxr-xr-x	ponny	supergroup

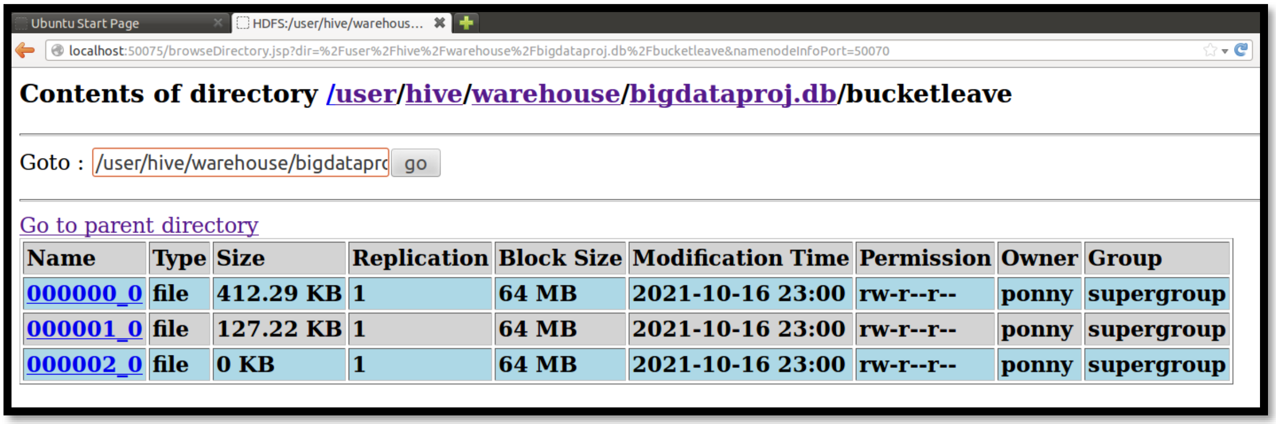
5) Bucketing:



```
hive> set hive.enforce.bucketing=true;
hive> 
```



```
hive> insert overwrite table bucketleave select satisfactionLevel,lastEvaluation,numberOfProjects,avgMonthlyHours,timeSpentCompany,workAccident,leave,promotionLastFiveYears,department,salary from employee;
```



Ubuntu Start Page
HDFS:/user/hive/warehous...
localhost:50075/browseDirectory.jsp?dir=%2Fuser%2Fhive%2Fwarehouse%2Fbigdataproj.db%2Fbucketleave&namenodeInfoPort=50070

Contents of directory /user/hive/warehouse/bigdataproj.db/bucketleave

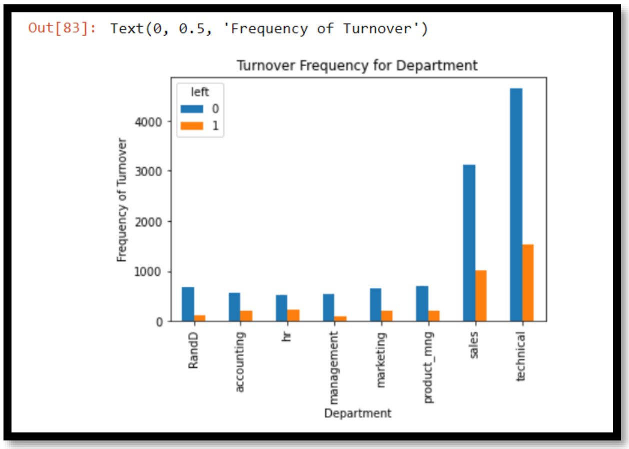
Goto : /user/hive/warehouse/bigdataprc go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
000000_0	file	412.29 KB	1	64 MB	2021-10-16 23:00	rw-r--r--	ponny	supergroup
000001_0	file	127.22 KB	1	64 MB	2021-10-16 23:00	rw-r--r--	ponny	supergroup
000002_0	file	0 KB	1	64 MB	2021-10-16 23:00	rw-r--r--	ponny	supergroup

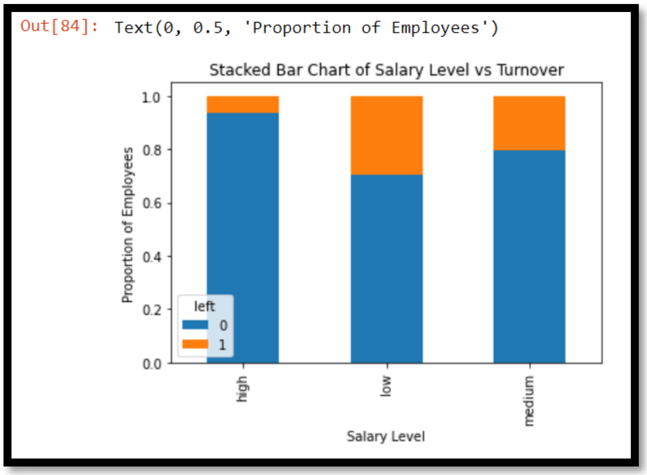
Visualization:

1) Python using Matplotlib:



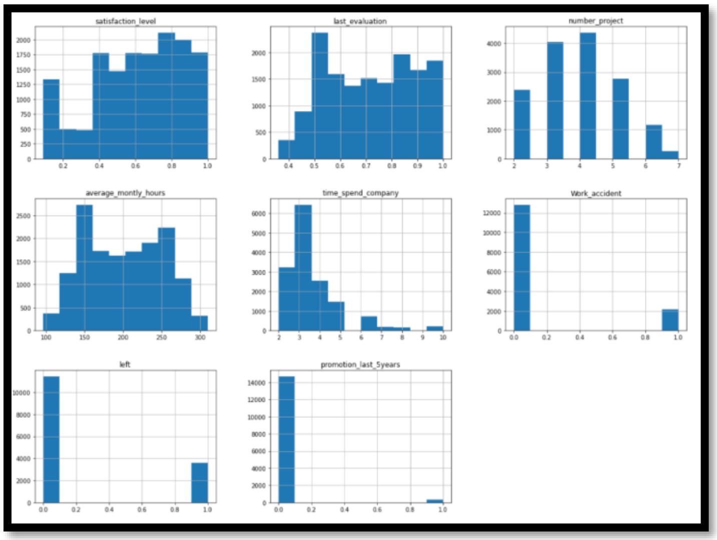
```
Out[83]: Text(0, 0.5, 'Frequency of Turnover')
```

Here the frequency of the employee turnover depends on the department they are working on. Thus, department attribute will be a good prediction attribute in predicting the outcome.

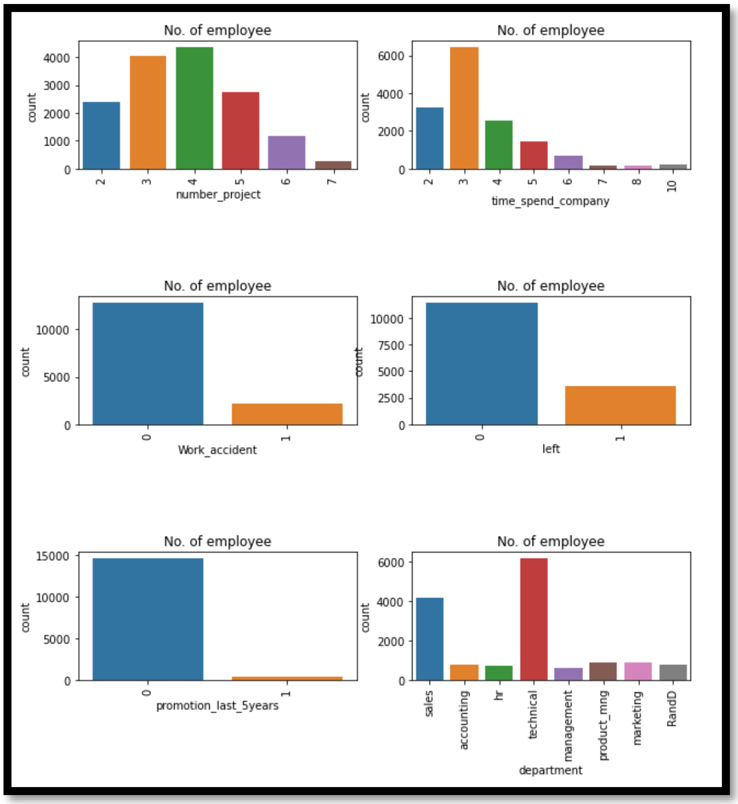


```
Out[84]: Text(0, 0.5, 'Proportion of Employees')
```

Here, the proportion of the employee turnover depends on the salary level of the employee. Hence, salary level also be a good prediction attribute in predicting the outcome.

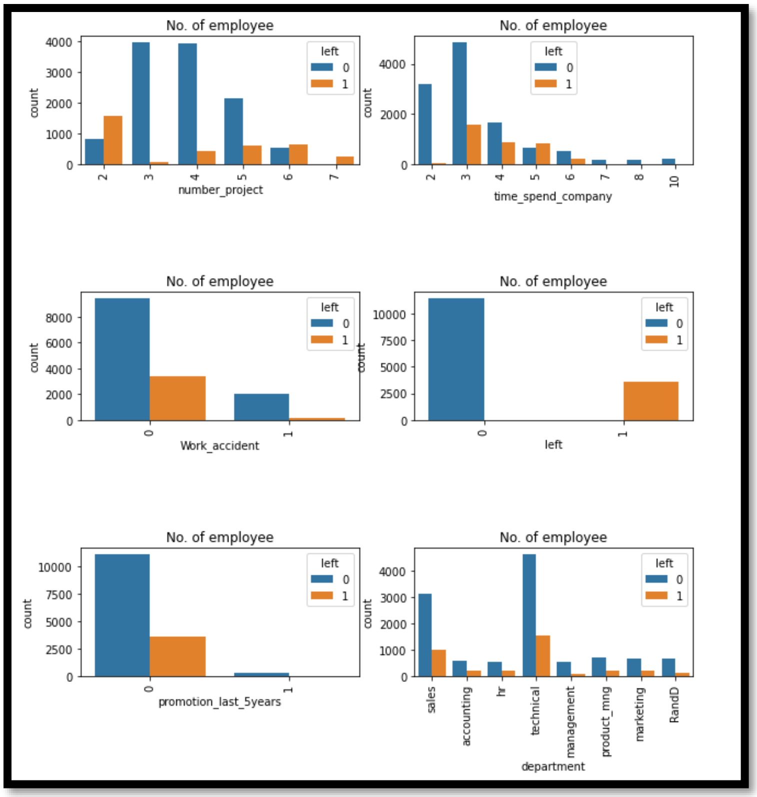


Histograms are one of the most important plot where we can use for numeric variables for the data exploration.



By observing the above visualization, we can analyze:

- Many of the employees are doing the project between the count of 3-5.
- There is a dropdown between the employees having 3 or 4 years experienced. Very less employees got the promotions in the last five years.
- 23 % of employee left from the total employment.
- Among all the departments sales had the maximum number of employees followed by technical department and support department.
- The employees with the salary medium and low are more in organization

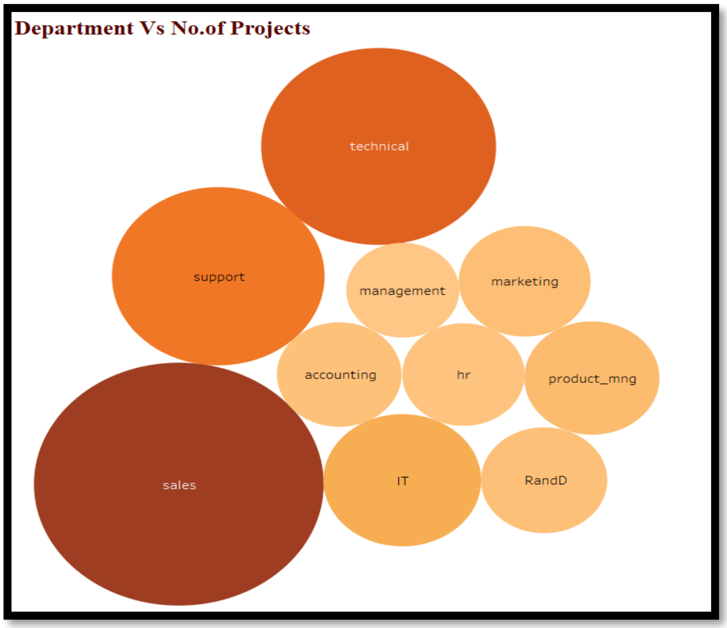


By observing the above visualization, we can analyze:

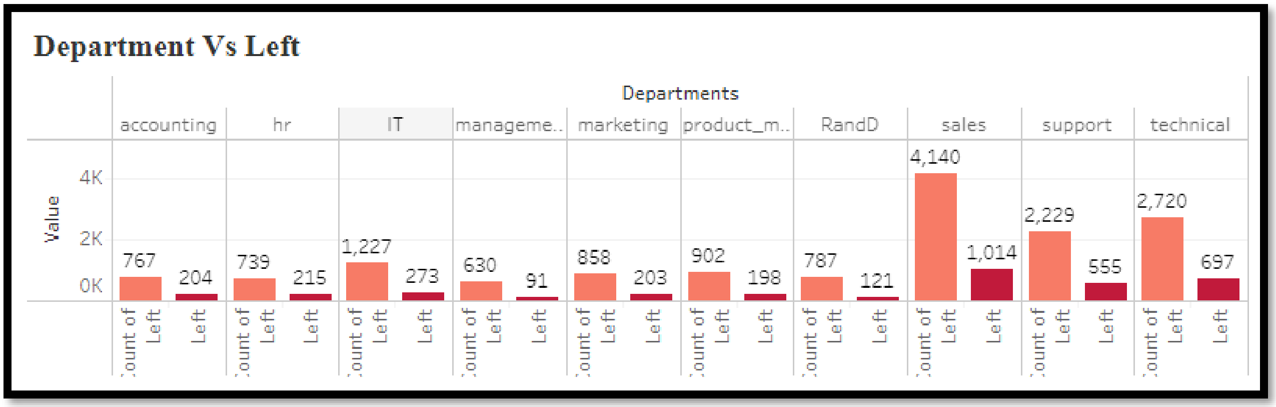
- Number of projects more than 5 those employees were left from the company.
- The employees who had done 6 and 7 projects, left the organization like can't help thinking that they were over-burden with work.
- The employees with five-year experience is leaving more in view of no promotions in most recent 5 years and over 6 years' experience are not leaving because of affection with the company.
- Those who promotion in most recent 5 years they didn't leave, i.e., every one of those left they didn't get the promotion in the past 5 years.

2) Tableau:

Bubble Plot for Department Vs Number of Projects:

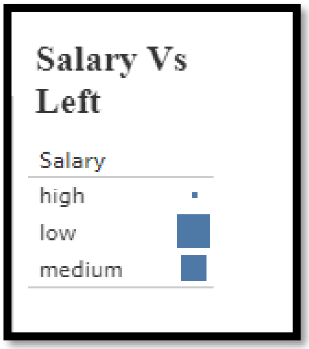


Bar Graph for Department Vs Left:



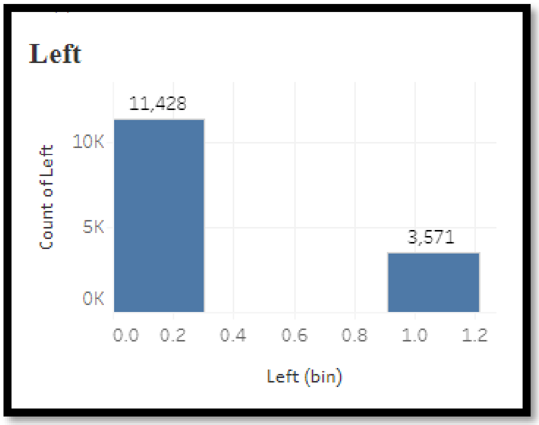
Here we analyzed by seeing **Bubble plot** and **Bar graph** is from sales department there is more chance for the employees to leave the company because in sales department the total number project count is high i.e.,15,634 and count of left is high i.e.,1,014.

Heat Map for Salary Vs Left:

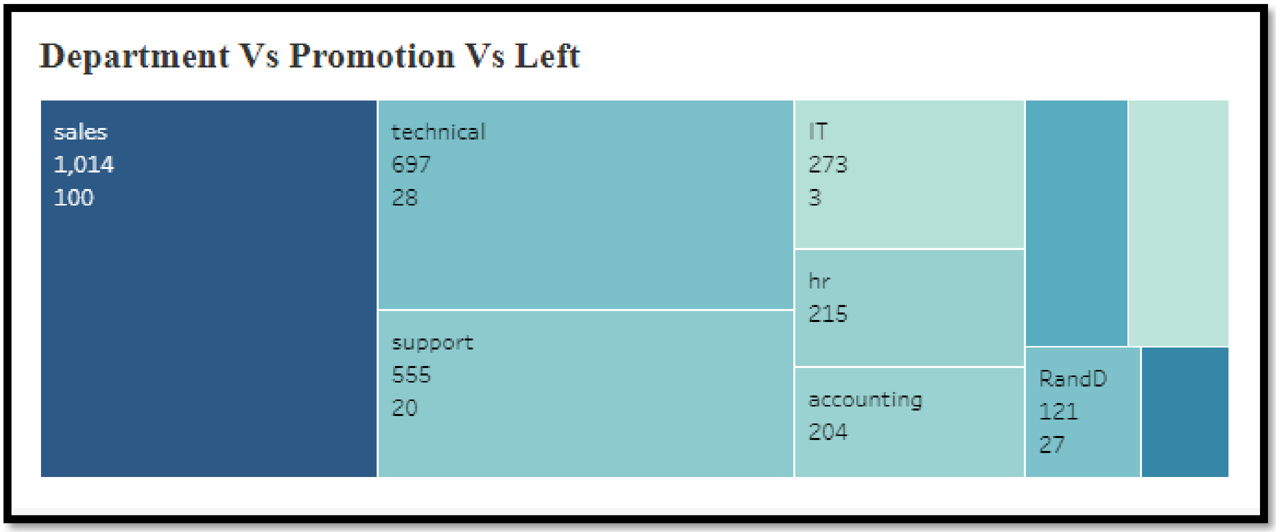


Here we analyzed that the employees with low salary had highest left count i.e., 2,172

Histogram for Left:

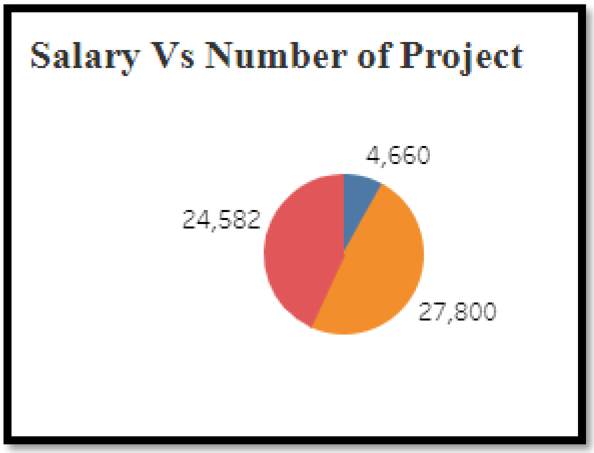


Tree Map for Department Vs Promotion Vs Left:



Here we analyzed the sales have the highest count of employees who left the company and they only got the highest promotion in last 5 years

Pie Chart for Salary Vs Number of Project:



Here we analyzed that the employees who had low salary they have done the highest number of projects i.e., 27,800

Literature Survey:

In **[1]** using machine learning techniques he predicted the employee turnover. Here they used different techniques like decision tree, gradient boosting trees, SVM, Neural Networks, Naïve Bayes, logistic regression, and random forest to do numerical experiments for real resources datasets of varying complexity and size. And also, they used statistical methods for analysing and establishing each of the above machine learning methods for employee turnover prediction.

In **[2]** Employee turnover is viewed as a significant issue for some associations and ventures. To address this need, this review plans to upgrade the capacity to figure employee turnover and present another technique dependent on an improved random forest algorithm. The proposed weighted random forest algorithm is applied to representative turnover information with high-dimensional lopsided attributes. In the space of employee turnover anticipating, contrasted and the irregular timberland, C4.5, Logistic, BP, and different calculations, the proposed calculation shows huge improvement as far as different execution markers, explicitly review and F-measure. Among them, month to month pay and additional time were the two most significant variables. The review offers another logical technique that can help human asset offices foresee representative turnover all the more precisely and its exploratory outcomes give further bits of knowledge to lessen worker turnover aim.

Inferences Analyzed:

The analysis done for the employees who left the company.

1) Number of Projects:

The employees who are doing less and a greater number of projects there is a more chance to leave the company. The employees who are doing the projects between 3 to 5 less chance to leave the company.

2) Salary:

Most of the employees who got medium and low salary have more chance to leave the company.

3) Time with Company:

Here, the three-year mark resembles a chance to be an essential point in a worker's profession. The greater part of them quit their place of employment around the three-year mark. Another significant point is 6-years point, where the worker is probably not going to leave.

4) Satisfaction level:

The average satisfaction level for the employees who are working in the company is greater than the employees who left the company.

5) Average monthly hours:

Average monthly hours for employees who left the company is greater than the employees who are currently working in the company.

6) Workplace accidents:

The work accidents are less for employees who left the company than who doesn’t have work accident.

7) Promotion last 5 years:

The employees who got promotions in last five years are less and they won't like to leave the company.

8) Department:

The employees in the sales department have more chance of leaving the company because in that department the count of projects is high.

Unique Features:

We used Big data technology like Hadoop Distributed File System (HDFS) to store our dataset. It provides easier access to data. It makes the application or dataset available to parallel processing.

Results:

Algorithm	Accuracy
Logistic Regression	0.771
Random Forest	**0.978**
Support Vector Machine	0.909

From the above table, we found that Random Forest Algorithm is the best suitable algorithm for our data with the accuracy of 98%.

References:

[1] Yue Zhao, Maciej K. Hryniewicki, Francesca Cheng,Boyang Fu,and Xiaoyu Zhu, "Employee Turnover Prediction with Machine Learning: A Reliable Approach", Intelligent System Conference (Intellisys), IEEE,Conference: 2018.

[2] Xiang Gao, Junhao Wen and Cheng Zhang, "An Improved Random Forest Algorithm for Predicting Employee Turnover", IEEE,Conference, 17 April 2019.

[3] K. Tamizharasi et al, Dr. UmaRani, "Employee Turnover Analysis with Application of Data Mining Methods", International Journal of Computer Science and Information Technologies, Vol. 5(1), 2014.

[4] Pankaj Ajit, Rohit Punnoose, "Prediction of Employee Turnover in Organizations using Machine Learning Algorithms", International Journal of Advanced Research in Artificial Intelligence, Vol. 5, No. 9, 2016.