REPORT ON

EMPLOYEE TURNOVER PREDICTION

By:-Arghya pan

CONTENTS:

- INTRODUCTION
- WHAT IS PYTHON
- ANACONDA
- BRIEF EXPLANATION OF DATA ANALYTICS
- WHAT IS DATA ANALYTICS
- TYPES OF ANALYTICS APPLICATIONS
- DIFFERENT PROCESS OF DATA ANALYTICS
- WHAT IS DATA SCIENCE
- PROCESS OF DATA SCIENCE
- ABOUT PROJECT
- PLOTTING HISTOGRAM
- FEATURE SELECTION
- IMPLEMENTING LOGISTIC REGRESSION
- CHECKING CLASSIFICATION REPORT
- PLOTTING ROC CURVE
- CHECKING ACCURACY SCORE
- CONCLUSIONS

INTRODUCTION

1.1 PYTHON

Python is a high-level general-purpose, open source, strictly typed programming language. The language provides constructs intended to enable clear programs on both a small and large scale.

Python was created by Guido van Rossum.

The Python Software Foundation (PSF) is the organization behind Python.

PYTHON VERSIONS

- First released in 1991.
- Python 2.0 was released on 16 October 2000 □ Python 3.0 was released on 3 December 2008 □ Current Versions:
- 3.7.0
- 2.7.14

PYTHON FEATURES

Some of the features of python include

- Dynamic
- Object oriented
- Multipurpose
- Strongly typed
- Open Sourced

Python is widely used in many domains

- Web Development
- Data Analysis
- Machine Learning
- Internet Of Things

- GUI Development
- Image processing Data visualization
- Game Development

IDLE

IDLE (short for integrated development environment integrated development and learning environment is in integrated development environment for Python, which has been bundled with the default implementation of the language since Linux distributions. It is completely written in Python and the Tkinter GUI toolkit (wrapper functions for Tcl/Tk).

IDLE is intended to be a simple IDE and suitable for beginners, especially in an educational environment. To that end, it is cross-platform, and avoids feature clutter.

1.2 ANACONADA :

Anaconda is a open source Distribution for data science and machine learning using python. It includes hundreds of popular data science packages and the conda package and virtual environment manager for Windows, Linux, and MacOS. Conda makes it quick and easy to install, run, and upgrade complex data science and machine learning environments like scikit-learn, TensorFlow, and SciPy. Anaconda Distribution is the foundation of millions of data science projects as well as Amazon Web Services.

1.3 IPYHTON :

Interactive Python is a command shell for interactive computing in multiple programming languages, originally developed for the Python programming language, that offers introspection, rich media, shell syntax, tab completion, and history. IPython provides the following features:

- Interactive shells (terminal and Qt-based).
- A browser-based notebook interface with support for code, text, mathematical expressions, inline plots and other media.
- Support for interactive data visualization and use of GUI toolkits.
- Flexible, embeddable interpreters to load into one's own projects.
- Tools for parallel computing.

1.4 PACKAGES:

1.4.1 NUMPY

NumPy is the fundamental package for scientific computing with Python. It contains among other things:

- a powerful N-dimensional array object
- sophisticated (broadcasting) functions
- tools for integrating C/C++ and Fortran code
- useful in linear algebra, Fourier transform, and random number capabilities Besides its obvious scientific uses, NumPy can also be used as an efficient multidimensional container of generic data.

Arbitrary data-types can be defined. This allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

1.4.2 MATPLOTLIB

Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shell, the jupyter notebook, web application servers, and four graphical user interface toolkits.

Matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, errorcharts, scatterplots, etc., with just a few lines of code.

For simple plotting the pyplot module provides a MATLAB-like interface, particularly when combined with IPython. For the power user, you have full control of line styles, font properties, axes properties, etc, via an object oriented interface or via a set of functions familiar to MATLAB users.

1.4.3 SCIKIT-LEARN

Scikit-learn provides machine learning libraries for python.

- Simple and efficient tools for data mining and data analysis

- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

1.4.4 PANDAS

Pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.

pandas is a NumFOCUS sponsored project. This will help ensure the success of development of pandas as a world-class open-source project, and makes it possible to donate to the project.

1.4.5 SEABORN

Seaborn is a Python visualization library based on matplotlib. It provides a highlevel interface for drawing attractive statistical graphics.

DATA ANALYTICS :

Data analytics (DA) is the process of examining data sets in order to draw conclusions about the information they contain, increasingly with the aid of specialized systems and software. Data analytics technologies and techniques are widely used in commercial industries to enable organizations to make moreinformed business decisions and by scientists and researchers to verify or disprove scientific models, theories and hypotheses.

As a term, data analytics predominantly refers to an assortment of applications, from basic business intelligence (BI), reporting and online analytical processing (OLAP) to various forms of advanced analytics. In that sense, it's similar in nature to business analytics, another umbrella term for approaches to analyzing data -- with the difference that the latter is oriented to business uses, while data analytics has a broader focus. The expansive view of the term isn't universal, though: In some cases, people use data analytics specifically to mean advanced analytics, treating BI as a separate category.

Data analytics initiatives can help businesses increase revenues, improve operational efficiency, optimize marketing campaigns and customer service efforts, respond more quickly to emerging market trends and gain a competitive edge over rivals -- all with the ultimate goal of boosting business performance. Depending on the particular application, the data that's analyzed can consist of either historical records or new information that has been processed for real-time analytics uses. In addition, it can come from a mix of internal systems and external data sources.

Types of data analytics applications:

At a high level, data analytics methodologies include exploratory data analysis (EDA), which aims to find patterns and relationships in data, and confirmatory data analysis (CDA), which applies statistical techniques to determine whether hypotheses about a data set are true or false. EDA is often compared to detective work, while CDA is akin to the work of a judge or jury during a court trial -- a distinction first drawn by statistician John W. Tukey in his 1977 book Exploratory Data Analysis.

Data analytics can also be separated into quantitative data analysis and qualitative data analysis. The former involves analysis of numerical data with quantifiable variables that can be compared or measured statistically. The qualitative approach is more interpretive -- it focuses on understanding the content of non-numerical data like text, images, audio and video, including common phrases, themes and points of view.

At the application level, BI and reporting provides business executives and other corporate workers with actionable information about key performance indicators, business operations, customers and more. In the past, data queries and reports typically were created for end users by BI developers working in IT or for a centralized BI team; now, organizations increasingly use self-service BI tools that let execs, business analysts and operational workers run their own ad hoc queries and build reports themselves.

More advanced types of data analytics include data mining which involves sorting through large data sets to identify trends, patterns and relationships; predictive analytics which seeks to predict customer behavior, equipment failures and other future events; and machine learning an artificial intelligence technique that uses automated algorithms to churn through data sets more quickly than data scientists can do via conventional analytical modeling. Big data analytics applies data mining, predictive analytics and machine learning tools to sets of big data that often contain unstructured and semi-structured data. Text mining provides a means of analyzing documents, emails and other text-based content.

Data analytics initiatives support a wide variety of business uses. For example, banks and credit card companies analyze withdrawal and spending patterns to prevent fraud and identity theft. E-commerce companies and marketing services providers do clickstream analysis to identify website visitors who are more likely to buy a particular product or service based on navigation and page-viewing patterns. Mobile network operators examine customer data to forecast churn so they can take steps to prevent defections to business rivals; to boost customer relationship management efforts, they and other companies also engage in CRM analytics to segment customers for marketing campaigns and equip call center workers with up-to-date information about callers. Healthcare organizations mine patient data to evaluate the effectiveness of treatments for cancer and other diseases.

Inside the data analytics process:

Data analytics applications involve more than just analyzing data. Particularly on advanced analytics projects, much of the required work takes place upfront, in collecting, integrating and preparing data and then developing, testing and revising analytical models to ensure that they produce accurate results. In addition to data scientists and other data analysts, analytics teams often include data engineers whose job is to help get data sets ready for analysis.

The analytics process starts with data collection, in which data scientists identify the information they need for a particular analytics application and then work on their own or with data engineers and IT staffers to assemble it for use. Data from different source systems may need to be combined via data integration routines, transformed into a common format and loaded into an analytics system, such as a Hadoop cluster NoSQL database or data warehouse. In other cases, the collection process may consist of pulling a relevant subset out of a stream of raw data that flows into, say, Hadoop and moving it to a separate partition in the system so it can be analyzed without affecting the overall data set.

PROCESS OF DATA ANALYTICS :

Analysis refers to breaking a whole into its separate components for individual examination. Data analysis is a process for obtaining raw data and converting it into information useful for decision-making by users. Data are collected and analyzed to answer questions, test hypotheses or disprove theories.

Statistician John Tukey defined data analysis in 1961 as: "Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data."

There are several phases that can be distinguished, described below. The phases are iterative, in that feedback from later phases may result in additional work in earlier phases. The CRISP framework used in data mining has similar steps.

Data requirements:

The data are necessary as inputs to the analysis, which is specified based upon the requirements of those directing the analysis or customers (who will use the finished product of the analysis). The general type of entity upon which the data will be collected is referred to as an experimental unit (e.g., a person or population of people). Specific variables regarding a population (e.g., age and income) may be specified and obtained. Data may be numerical or categorical (i.e., a text label for numbers).

Data collection:

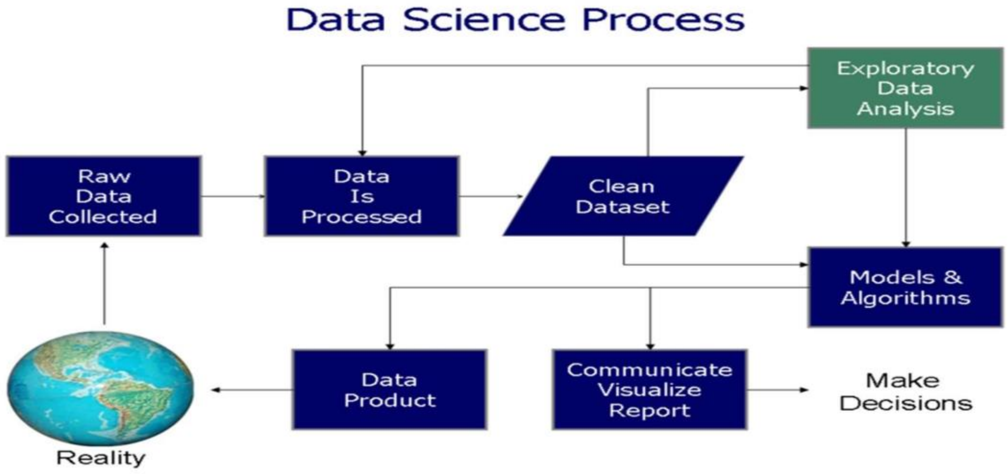
Data are collected from a variety of sources. The requirements may be communicated by analysts to custodians of the data, such as information technology personnel within an organization. The data may also be collected from sensors in the environment, such as traffic cameras, satellites, recording devices, etc. It may also be obtained through interviews, downloads from online sources, or reading documentation.

DATA SCIENCE

Data science is a multidisciplinary blend of data inference, algorithmm development, and technology in order to solve analytically complex problems.

At the core is data. Troves of raw information, streaming in and stored in enterprise data warehouses. Much to learn by mining it. Advanced capabilities we can build with it. Data science is ultimately about using this data in creative ways to generate business value.

PROCESS OF DATA SCIENCE :



Different Types Of Data Analytics:

1. Descriptive Analytics

As the name implies, descriptive analysis or statistics can summarize raw data and convert it into a form that can be easily understood by humans. They can describe in detail about an event that has occurred in the past. This type of analytics is helpful in deriving any pattern if any from past events or drawing interpretations from them so that better strategies for the future can be framed

This is the most frequently used type of analytics across organizations. It’s crucial in revealing the key metrics and measures within any business.

2. Diagnostic Analytics

The obvious successor to descriptive analytics is diagnostic analytics. Diagnostic analytical tools aid an analyst to dig deeper into an issue at hand so that they can arrive at the source of a problem.

In a structured business environment, tools for both descriptive and diagnostic analytics go hand-in-hand!

3. Predictive Analytics

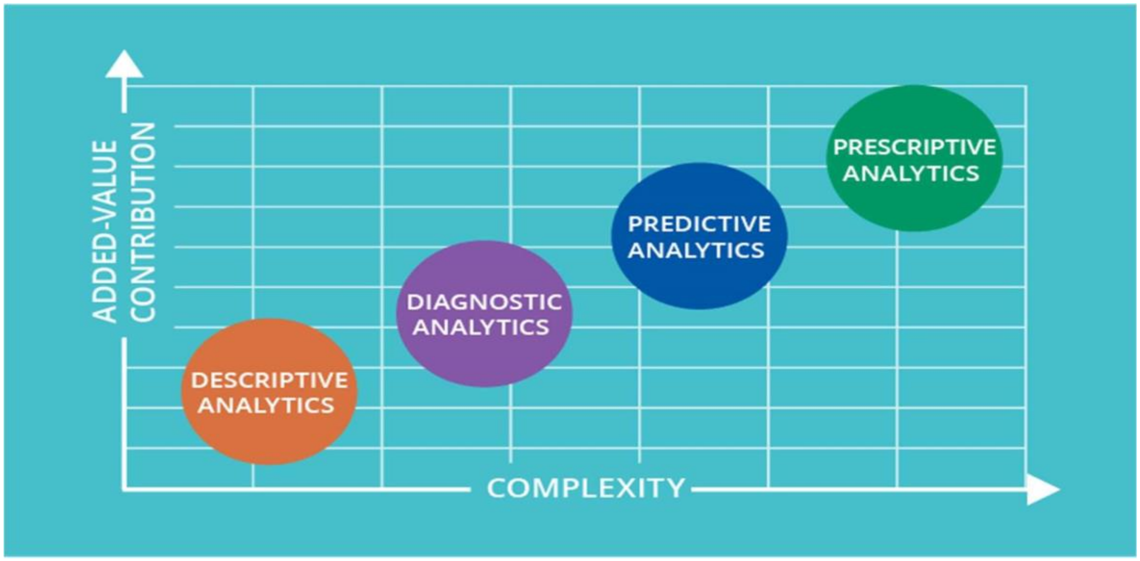
Any business that is pursuing success should have foresight. Predictive analytics helps businesses to forecast trends based on the current events. Whether it's predicting the probability of an event happening in future or estimating the accurate time it will happen can all be determined with the help of predictive analytical models.

For example, in the healthcare domain, prospective health risks can be predicted based on an individual's habits/diet/genetic composition. Therefore, these models are most important across various fields.

4. Prescriptive Analytics:

This type of analytics explains the step-by-step process in a situation. For instance, a prescriptive analysis is what comes into play when your Uber driver gets the easier route from maps. The best route was chosen by considering the distance of every available route from your pick-up route to the destination and the traffic constraints on each road. A data analyst would need to apply one or more of the above analytics processes as a part of his job.

IMAGE OF DIFFERENT TYPES OF DATA ANALYTICS :



PROJECT: EMPLOYEE TURNOVER PREDICTION

INTRODUCTION:

We are finally very glad to say that we completed our practical work on the Employee turnover prediction project and here in this report we are explained how we implement it in jupyter notebook and how it should be beneficial for market.this post presents a reference implementation of an employee turnover analysis project that is built by using Python’s Scikit-Learn library. In this article, we introduce

Logistic Regression, Random Forest, and Support Vector Machine. We also measure the accuracy of models that are built by using python analytics.

DATA PREPROCESSING:

The data was provided by NIVT. It is pretty straightforward. Each row represents an employee, each column contains employee attributes:

□ satisfaction_level (0–1)

□ last_evaluation (Time since last evaluation in years)

□ number_projects (Number of projects completed while at work)

□ average_monthly_hours (Average monthly hours at workplace)

□ time_spend_company (Time spent at the company in years)

□ Work_accident (Whether the employee had a workplace accident)

□ left (Whether the employee left the workplace or not (1 or 0))

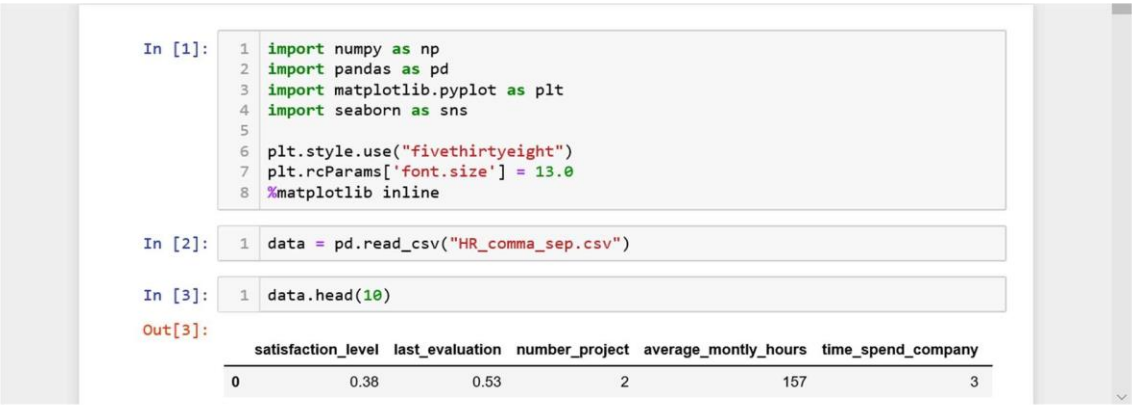
□ promotion_last_5years (Whether the employee was promoted in the last five years)

□ sales (Department in which they work for)

□ salary (Relative level of salary)

We import the packages.

Then we read the data and by using 'shape' function we find out the how many rows and columns are present in the dataset.



In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

plt.style.use("fivethirtyeight")
plt.rcParams['font.size'] = 13.0
%matplotlib inline
```

In [2]:

```
data = pd.read_csv("HR_comma_sep.csv")
```

In [3]:

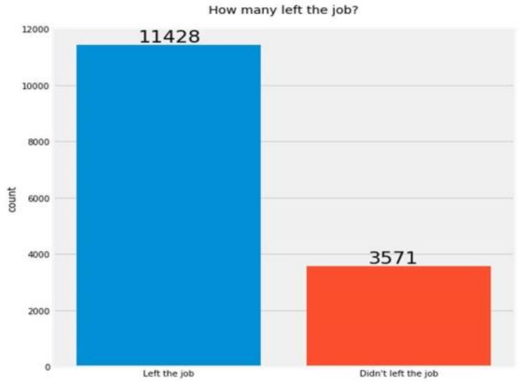
```
data.head(10)
```

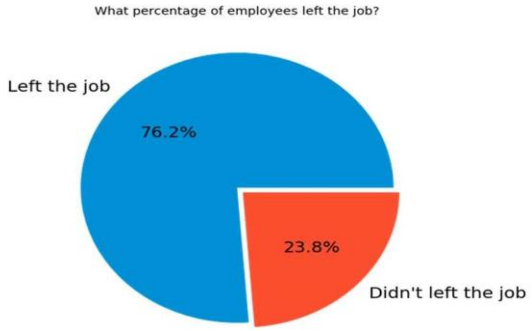
Out[3]:

	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company
0	0.38	0.53	2	157	3

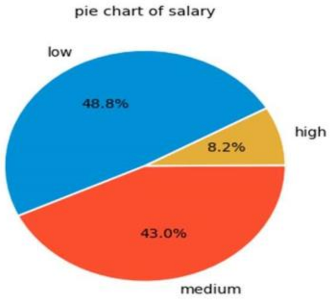
And using ‘info()’ function we find out that there should be 9 independent variables present in the dataset and then we describe the dataset.

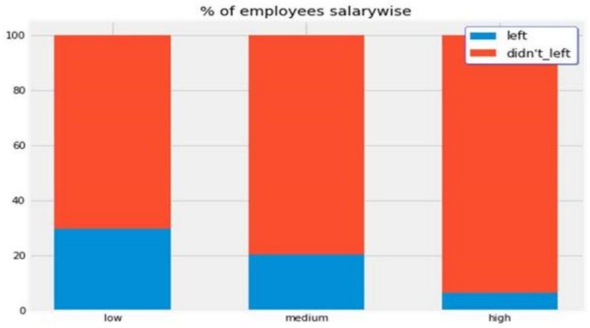
Then we will do a subplot and plot a pie chart to see how many of the employees was left the job and how many did not left the job.Here we find out that 11428 peoples left and 3571 peoples did not left the job using subplot and using piechart its more clear that 76.2% was left the job and 23.2% don’t left the job.





After we see that how many of the employee get the low salary,how many get the medium salary and how many get the high salary.





And using barplot we see that how many % of employee get the low,medium and high salary.Here we see that maximum low salary employees are left the job then comes medium and then comes high.Here 48.8% who got low salary was left the job and 43.0% high salary persorns left the job and 8.2% high salary persons left the job.

Plotting Histogram:

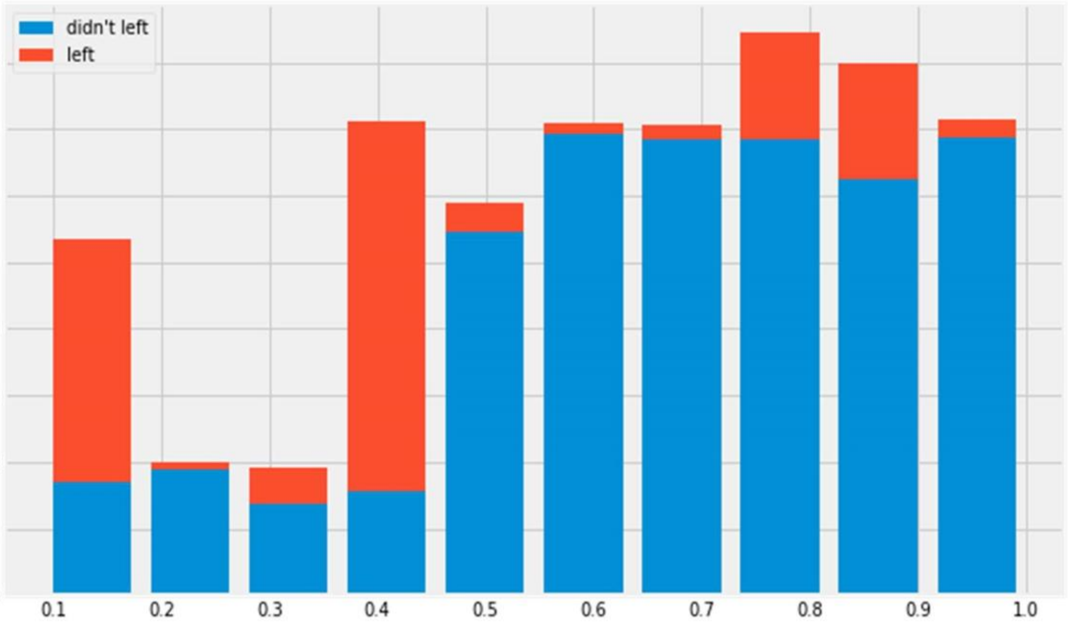
Histogram is a graphical representation of the Distribution of data

Bins: the intervals used in a histogram. The data must be separated into mutually exclusive and exhaustive bins

Cutpoints: the values that define the beginning and the end of the bins

Frequency: the count of the number of the data values in each bin The

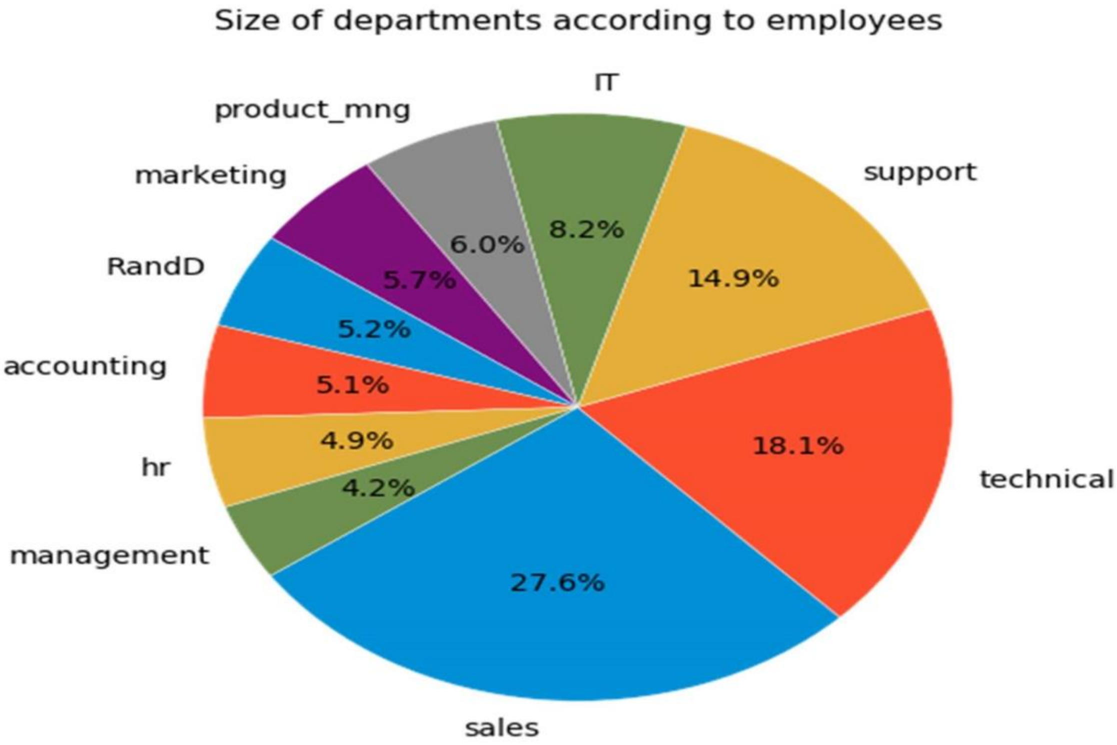
peaks in the distribution are called modes.

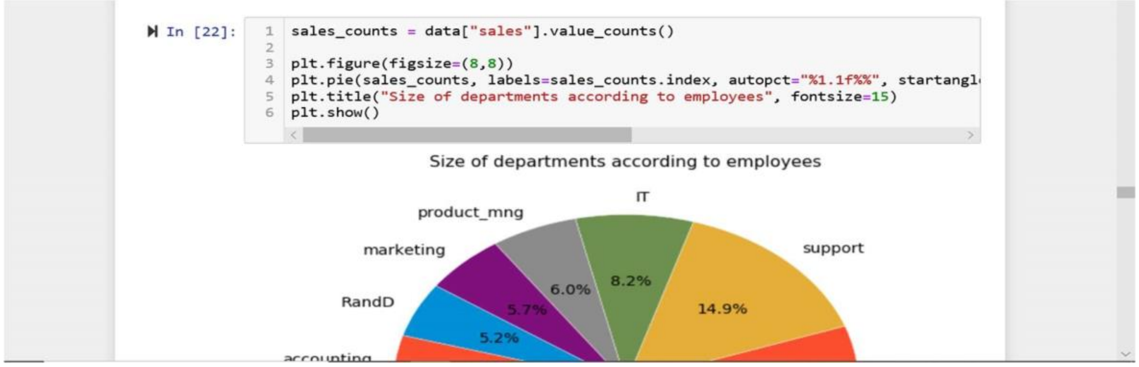


Then we count the employees and see that

This number of employees present in each department.

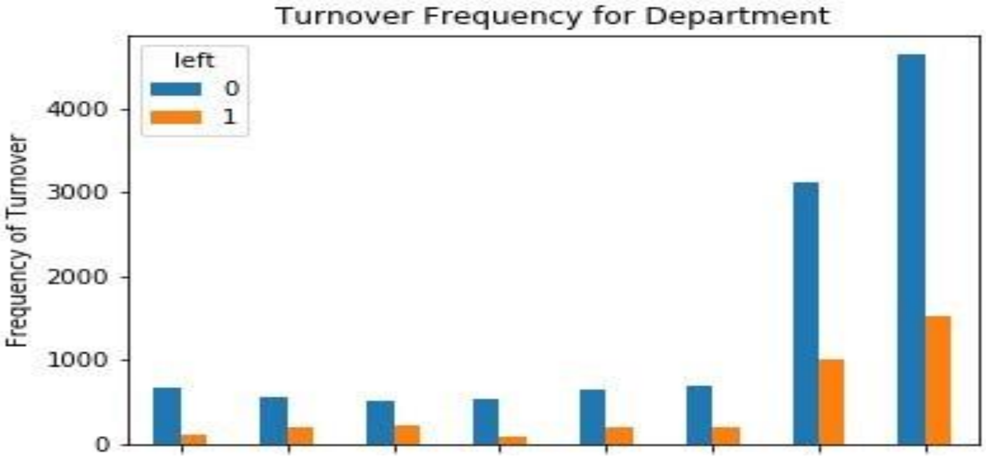
Then we import LogisticRegression model in X we take ‘left’ and in y we take all the columns.




In [22]:

```
sales_counts = data["sales"].value_counts()

plt.figure(figsize=(8,8))
plt.pie(sales_counts, labels=sales_counts.index, autopct="%1.1f%%", startangl
plt.title("Size of departments according to employees", fontsize=15)
plt.show()
```

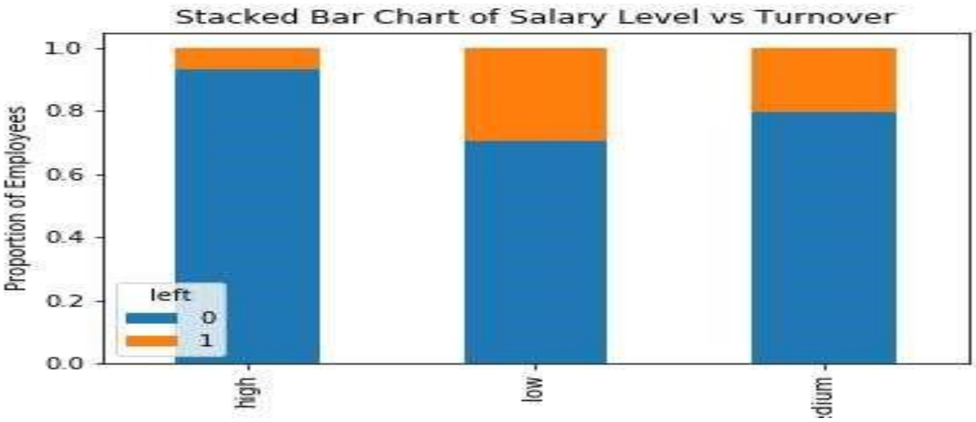
Let us visualize our data to get a much clearer picture of the data and the significant features.



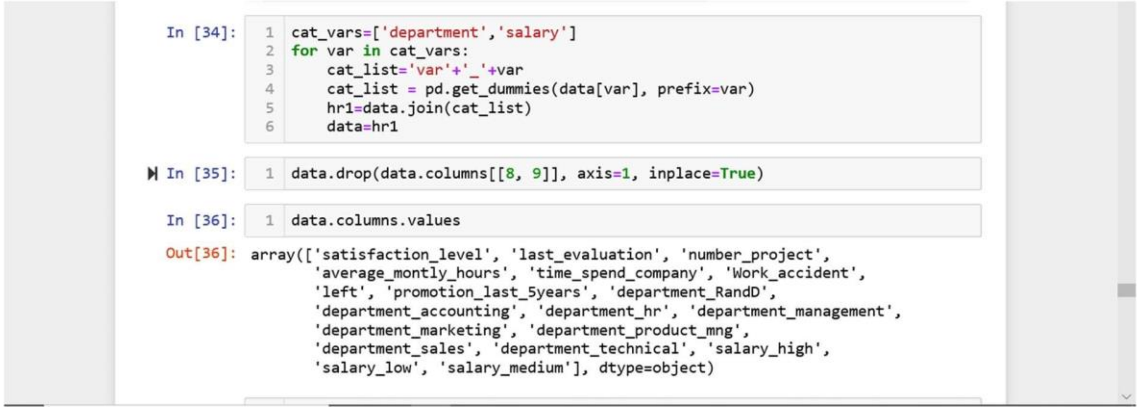
BAR CHART FOR THE FOLLOWING GRAPH :

table=pd.crosstab(hr.salary, hr.left) table.div(table.sum(1).astype(float), axis=0).plot(kind='bar', stacked=True) plt.title('Stacked Bar Chart of Salary Level vs Turnover') plt.xlabel('Salary Level') plt.ylabel('Proportion of Employees')

plt.savefig('salary_bar_chart')



CREATING DUMMY VARIABLES :



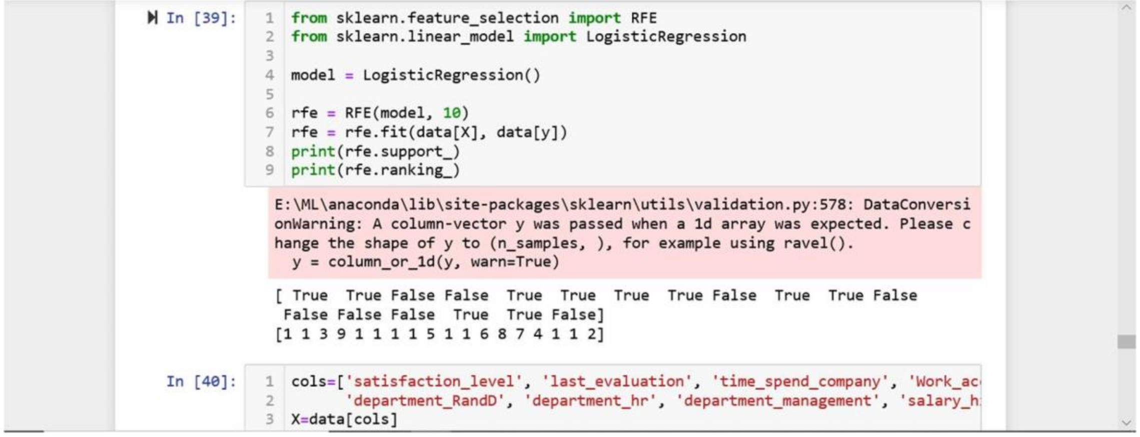
```
In [34]: cat_vars=['department','salary']
         for var in cat_vars:
             cat_list='var'+'_'+var
             cat_list = pd.get_dummies(data[var], prefix=var)
             hr1=data.join(cat_list)
             data=hr1

In [35]: data.drop(data.columns[[8, 9]], axis=1, inplace=True)

In [36]: data.columns.values

Out[36]: array(['satisfaction_level', 'last_evaluation', 'number_project',
                'average_montly_hours', 'time_spend_company', 'Work_accident',
                'left', 'promotion_last_5years', 'department_RandD',
                'department_accounting', 'department_hr', 'department_management',
                'department_marketing', 'department_product_mng',
                'department_sales', 'department_technical', 'salary_high',
                'salary_low', 'salary_medium'], dtype=object)
```

Here the actual categorical variable needs to be removed once the dummy variables have been created.

FEATURE SELECTION:

The Recursive Feature Elimination (RFE) works by recursively removing variables and building a model on those variables that remain. It uses the model accuracy to identify which variables (and combination of variables) contribute the most to predicting the target attribute.

Let’s use feature selection to help us decide which variables are significant that can predict employee turnover with great accuracy. There are total 18 columns in X, how about select 10?



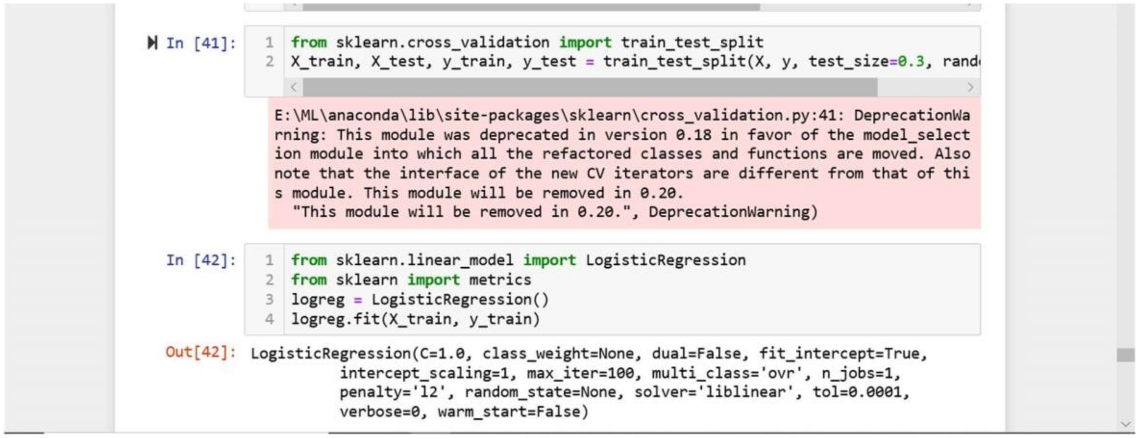
```
In [39]: from sklearn.feature_selection import RFE
         from sklearn.linear_model import LogisticRegression

         model = LogisticRegression()

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

E:\ML\anaconda\lib\site-packages\sklearn\utils\validation.py:578: DataConversi
onWarning: A column-vector y was passed when a 1d array was expected. Please c
hange the shape of y to (n_samples, ), for example using ravel().
  y = column_or_1d(y, warn=True)

[ True  True False False  True  True  True  True False  True  True False
 False False False  True  True False]
[1 1 3 9 1 1 1 1 5 1 1 6 8 7 4 1 1 2]

In [40]: cols=['satisfaction_level', 'last_evaluation', 'time_spend_company', 'Work_ac
              'department_RandD', 'department_hr', 'department_management', 'salary_h
         X=data[cols]
```

Logistic Regression In Python:

It is a technique to analyse a data-set which has a dependent variable and one or more independent variables to predict the outcome in a binary variable, meaning it will have only two outcomes.

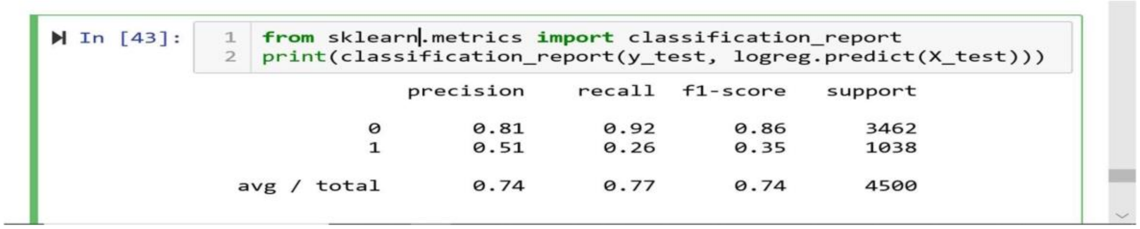
Logistic Regression equation: $p = 1 / 1 + e-(\beta0 + \beta1X1 + \beta2X2 \dots. + \beta nXn)$

Here using sklearn we split the the data into x_train,y_train,x_test,y_test and then we fitting the data for finds the coefficients for the equation specified via the algorithm being used.



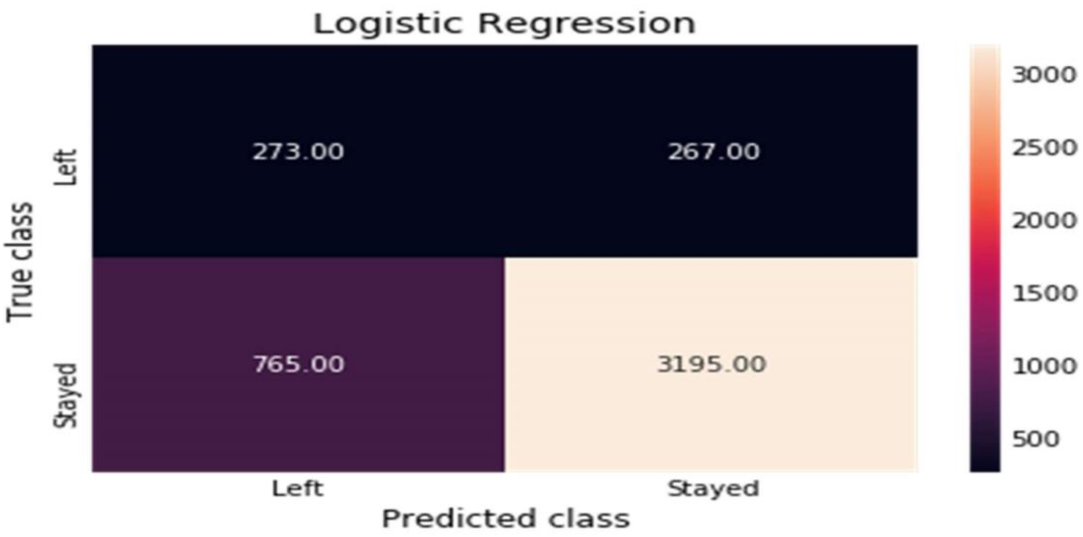
```
In [41]:
from sklearn.cross_validation import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, rand

E:\ML\anaconda\lib\site-packages\sklearn\cross_validation.py:41: DeprecationWa
rning: This module was deprecated in version 0.18 in favor of the model_select
ion module into which all the refactored classes and functions are moved. Also
note that the interface of the new CV iterators are different from that of thi
s module. This module will be removed in 0.20.
  "This module will be removed in 0.20.", DeprecationWarning)

In [42]:
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
logreg = LogisticRegression()
logreg.fit(X_train, y_train)

Out[42]: LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,
          penalty='l2', random_state=None, solver='liblinear', tol=0.0001,
          verbose=0, warm_start=False)
```

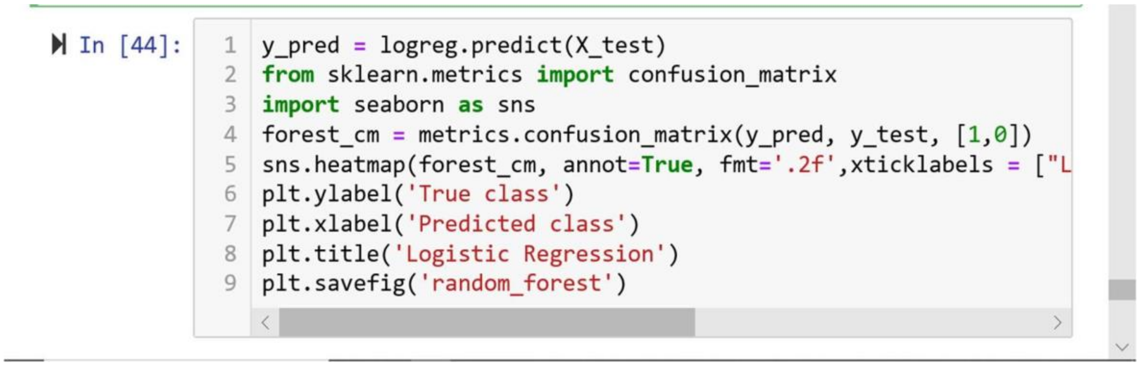
CHECKING CLASSIFICATION REPORT :



```
In [43]:
from sklearn.metrics import classification_report
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

avg / total       0.74      0.77      0.74      4500
```

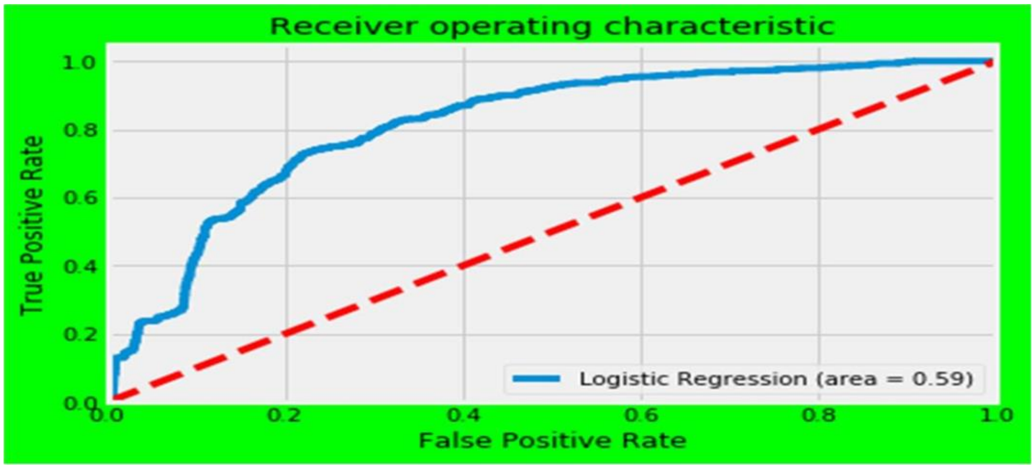
Using heatmap its clear that how many employees are left and how many stayed.So here 3195 employeees are staying for the job and 273 employees left the job.





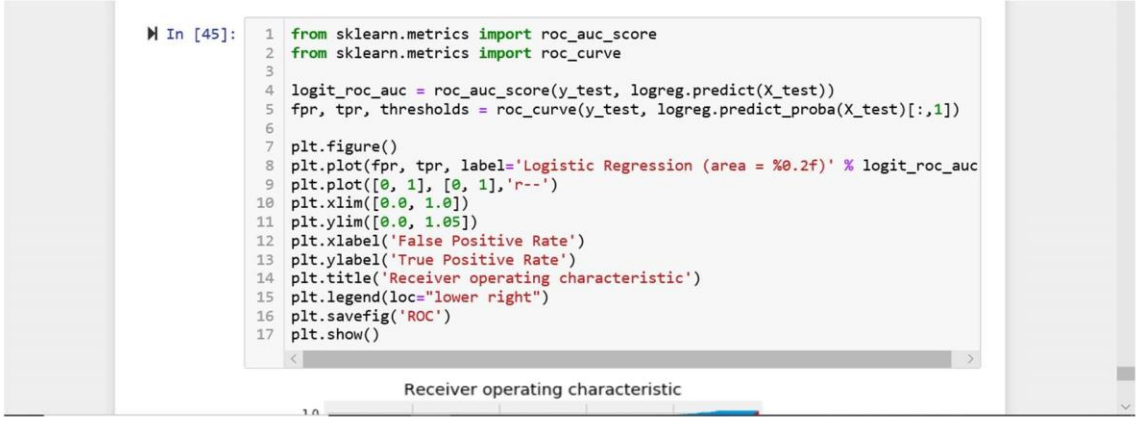
```
y_pred = logreg.predict(X_test)
from sklearn.metrics import confusion_matrix
import seaborn as sns
forest_cm = metrics.confusion_matrix(y_pred, y_test, [1,0])
sns.heatmap(forest_cm, annot=True, fmt='.2f',xticklabels = ["L
plt.ylabel('True class')
plt.xlabel('Predicted class')
plt.title('Logistic Regression')
plt.savefig('random_forest')
```

THE ROC CURVE:



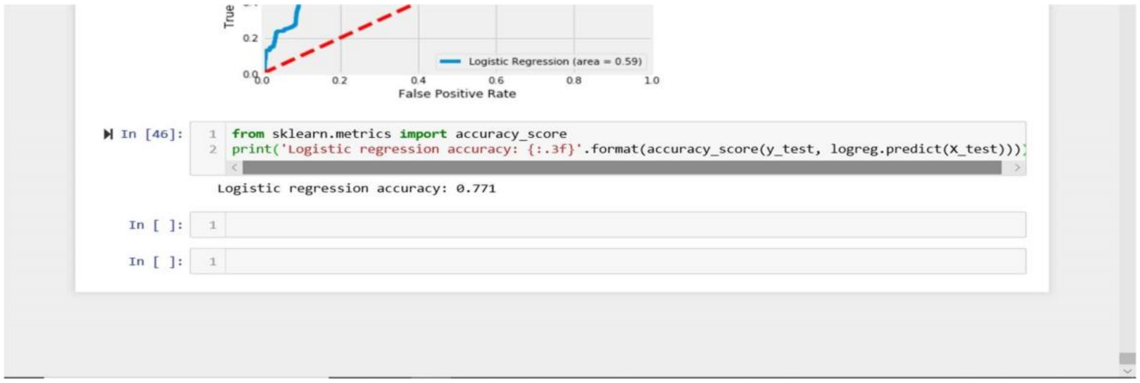
The receiver operating characteristic (ROC) curve is another common tool used with binary classifiers. The dotted line represents the ROC curve of a purely random classifier; a good classifier stays as far away from that line as possible.

CODE FOR ROC CURVE :



```
In [45]:
from sklearn.metrics import roc_auc_score
from sklearn.metrics import roc_curve

logit_roc_auc = roc_auc_score(y_test, logreg.predict(X_test))
fpr, tpr, thresholds = roc_curve(y_test, logreg.predict_proba(X_test)[:,1])

plt.figure()
plt.plot(fpr, tpr, label='Logistic Regression (area = %0.2f)' % logit_roc_auc
plt.plot([0, 1], [0, 1],'r--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver operating characteristic')
plt.legend(loc="lower right")
plt.savefig('ROC')
plt.show()
```

CHECKING THE ACCURACY SCORE :



```
In [46]:
from sklearn.metrics import accuracy_score
print('Logistic regression accuracy: {:.3f}'.format(accuracy_score(y_test, logreg.predict(X_test))))

Logistic regression accuracy: 0.771

In [ ]:

In [ ]:
```

Now we predict and check the accuracy score for see how much my prediction is correct for this model by using Logistic Regression and we find out that our prediction is 77.1% accurate.

CONCLUSIONS:

After doing all the project we finally come on conclusion. Here using logistic regression model we show that there we see that use analytics to find the main reasons behind employee turnover and it will help an organization to work on those factors to keep employees happy and satisfied and keep them from turning the job down legal issues and mistrust from employees,and use them in conjunction with employee feedback,to make best decisions possible.