HR ANALYTICS

MAYANK
PRASHAR

Steps to develop the model

1. Understanding the data
2. Data cleaning
3. Data visualization
4. Model building

Understanding the data

- Import various libraries required
- Read the data
- Check the various attributes of data:-
  - Name of columns
  - How the data is distributed
  - What kind of data types we have

Observations made:-

- The data had 10 columns and 15000 entries.
- Most of the data was in numerical form and there was less of the categorical data.

Steps to develop the model

1. Understanding the data
2. Data cleaning
3. Data visualization
4. Model building

Data cleaning and preparation

- ‘Average monthly income’ was spelled wrong, got it done right.
- There were duplicate values ,found them and finally removed them.
- Checked for null values , there were none ,so did not had to do anything with that.
- Lastly converted all the column names to lower case,to avoid any case difference error.

Visualising the data

Categorical Data

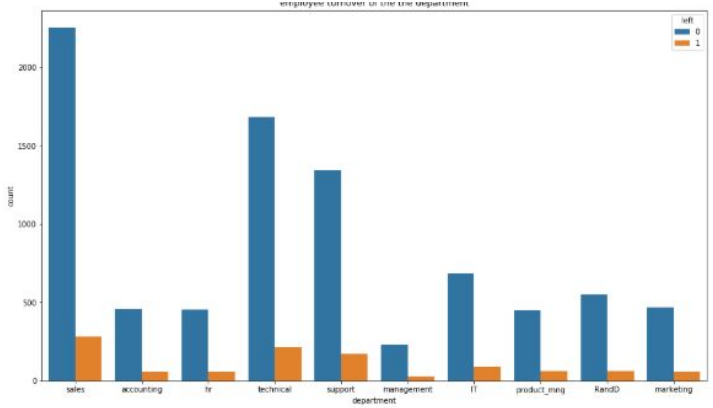
- Salary
- Department

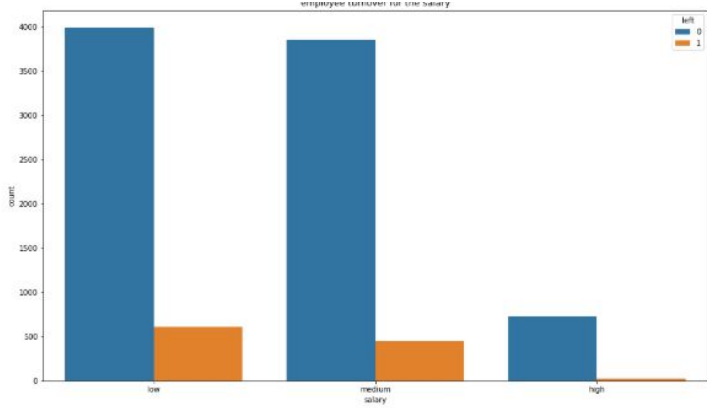
Numerical Variables

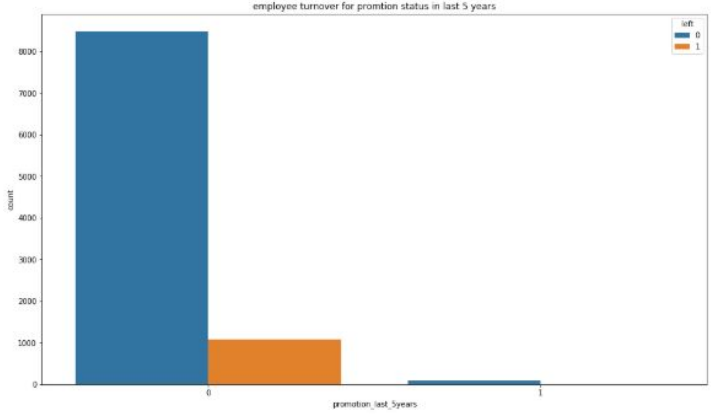
- Satisfaction level
- Last evaluation
- Number of project
- Average monthly hours
- Time spend in company
- Work accidents
- Promotions in last 5 years

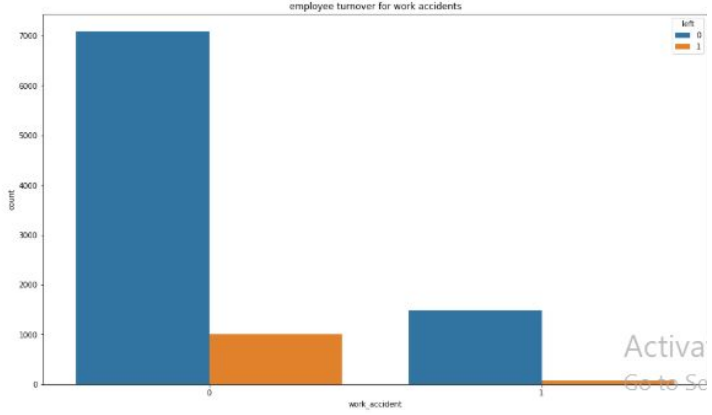
- Plotted a set of various graphs representing each categorical data as well as the numerical data and their respective relation with the ‘left’ to understand there relation with the people leaving the organisation.
- Graphs used here were **mostly countplot from the seaborn library**, which gave a good representation of the data

Fig1. The relationship between the department,salary,promotions and work accidents with the people leaving organisation.









Steps to develop the model

1. Understanding the data
2. Data cleaning
3. Data visualization
4. Model building

Choosing the variables for the regression

- Finding the correlation of all the variables with respect to price and sort them as per their relations.

	left	correlation
satisfaction_level	-0.305	moderate negative
last_evaluation	0.023	weak positive
number_project	0.041	weak positive
average_monthly_hours	0.071	weak positive
time_spend_company	0.235	moderate positive
work_accident	-0.094	weak negative
left	1.000	strong positive
promotion_last_5years	-0.028	weak negative

- Finding the correlation of all the variables with respect to price and sort them as per their relations.

Checking the OLS Regression Results

OLS Regression Results

Dep. Variable:	left	R-squared:	0.231
Model:	OLS	Adj. R-squared:	0.230
Method:	Least Squares	F-statistic:	321.1
Date:	Thu, 25 Jul 2019	Prob (F-statistic):	0.00
Time:	10:37:01	Log-Likelihood:	-1865.0
No. Observations:	9653	AIC:	3748.
Df Residuals:	9644	BIC:	3813.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
satisfaction_level	-0.3376	0.012	-28.059	0.000	-0.361	-0.314
last_evaluation	0.0933	0.017	5.332	0.000	0.059	0.128
number_project	-0.0116	0.003	-4.136	0.000	-0.017	-0.006
average_monthly_hours	0.0005	6.13e-05	8.512	0.000	0.000	0.001
time_spend_company	0.0577	0.003	20.739	0.000	0.052	0.063
work_accident	-0.0717	0.008	-8.833	0.000	-0.088	-0.056
promotion_last_5years	-0.0592	0.031	-1.924	0.054	-0.120	0.001
department_code	0.0021	0.001	2.118	0.034	0.000	0.004
salary_code	0.0116	0.005	2.513	0.012	0.003	0.021

Omnibus:	3372.163	Durbin-Watson:	0.279
Prob(Omnibus):	0.000	Jarque-Bera (JB):	9342.147
Skew:	1.907	Prob(JB):	0.00
Kurtosis:	5.946	Cond. No.	2.12e+03

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.12e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

Interpreting the OLS Regression Results

All the p values are less than 0.05 except for the ‘promotions in last five years’, so we can avoid this value for the regression analysis, as it does not have significant correlation with the people leaving the organisation.

Model building

import the train_test_split from sklearn.

Split the data into train and test taking 25% test and 75% train data.

Now apply the logistic regression(as the data is binary) and predict the value based on the given data.

Check for the accuracy of the data.comes out to be 86%.

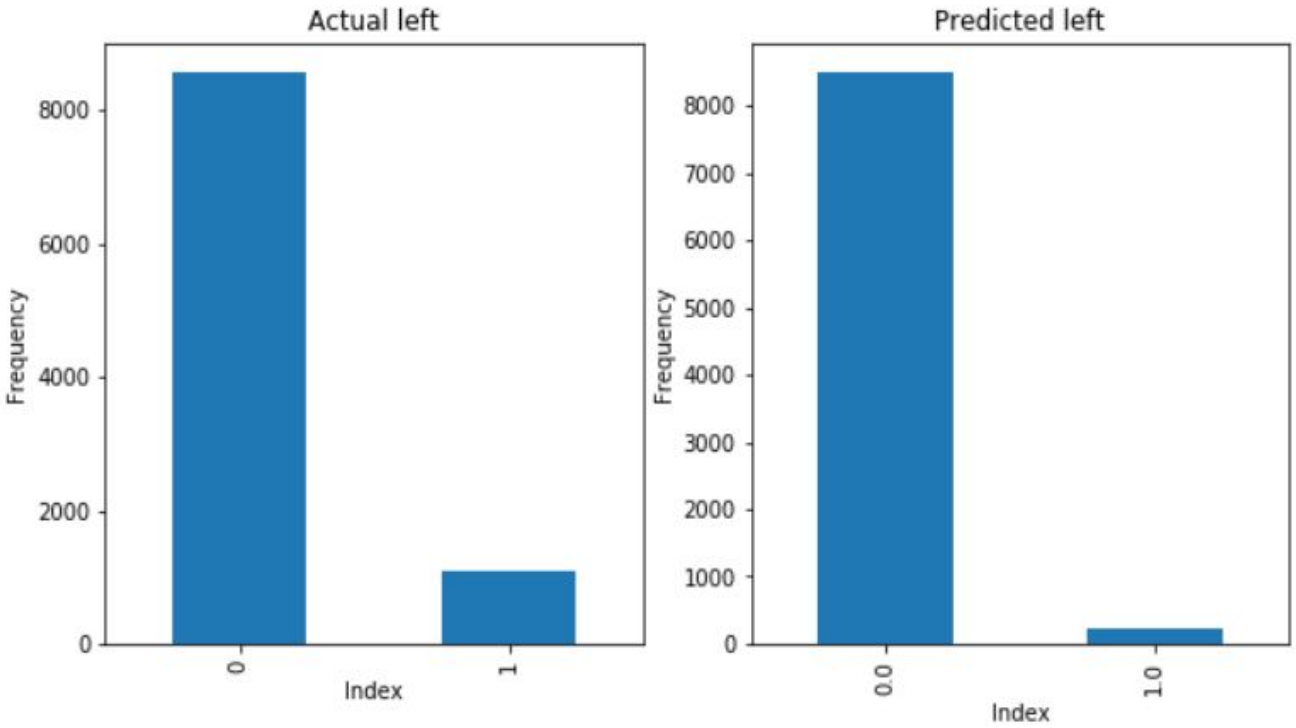
Applying random forest for better accuracy

import the RandomForestClassifier from sklearn.

Fitting the model into random forest.

Check for the accuracy of the data.comes out to be 96%.

VISUALIZE THE RESULT




thank
you