Pritam shrestha DSC_630

June 24, 2020 Prof. Fadi Alsaleem

Survey Analysis Case Study: Overview

It is believed that most companies would like to understand why their best employees voluntarily leave. This is one of the widely used case studies to find why it is happening and which factors are being responsible for that. I have found various studies and articles on the internet to solve this problem finding factors responsible for that. According to the article by Balance Articles, employees jump across jobs and roles for averagely 12 times during their lifetime career. It clears that leaving and finding new jobs wasting billions of dollars. Therefore managing their employee attrition is very important in the company. To hire a new employee, they need to go through several processes that take a long time and they have to pay a lot of money for the hiring process. Even after hiring, they have to train them about the company policy to ongoing project implementation. Hence, it is very important to implement predictive analytics for decision making to understand employee's leave to reduce turnover rate, save recruiting and training cost, and optimum use of work productivity. Thus, to stop this leaving problem we can go through the following processes.

Understanding: Defining the Problem

Understanding is itself a solution to the problems. To understand the problem we need to find the problem first. Problem finding is not easy because it could be related to anything. So first of all we need to categories the type of problem whether it is solvable or not. If it is solvable then they need to pay attention to it. It always comes with the management. It means managing employee attrition effectively and efficiently is always important. Hiring new employees is not a good

choice rather than keeping old experienced employees because a new employee will spend lots of time and money to hire and train them.

So, due to this negative impact, they should understand why it is happening, which factors are being responsible for this incident? And what are the problems and their solutions? For this, first, they have to find the possible cause of leaving such as

Work-life balance,

Bad managers/ colleague

Social pressure

Another better offer (salary, location, position, etc.)

Discrimination / harassment

Policy, etc.

Data Understanding

Data understanding is a major part of any issue because everything stored as data but we need to figure it out implementing it from a technical and logical point of view. If we deep drive into available data we can discover some sort of insights because employee expresses their satisfaction, eager, sadness, and happiness as available data indicator. For example

Satisfaction level (scale: 0 – 10)

Last evaluation (scale: 0 – 10)

Average monthly hours

Length of time with the company

Number of promotions/years with the company

Salary/ Bonus/ Paid vacations

Transportation /communication cost etc.

Besides that, maternity, paternity, and health insurance could play a vital role to decide to leave the company.

Data Preparation

I have downloaded this data from the following source:

https://raw.githubusercontent.com/VincentTatan/PythonAnalytics/master/Youtube/dataset/HR_comma_sep.csv

Employees with less than one year of tenure were removed from the dataset, as there is an overall tendency for new hires to leave at a higher-than-normal rate. No further data preparation was required, as there were no missing or incorrectly coded variables:

```
df=pd.read_csv(r"C:\Users\pritam\Desktop\employee_satisfaction_data.csv")
df.head()
```

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

And it has 14999 rows and 10 columns. To find the actual summary of the data we need description of the data which is I have got running on Jupiter notebook.

```
df.describe()
```

	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_last_5years
count	14999.000000	14999.000000	14999.000000	14999.000000	14999.000000	14999.000000	14999.000000	14999.000000
mean	0.612834	0.716102	3.803054	201.050337	3.498233	0.144610	0.238083	0.021268
std	0.248631	0.171169	1.232592	49.943099	1.460136	0.351719	0.425924	0.144281
min	0.090000	0.360000	2.000000	96.000000	2.000000	0.000000	0.000000	0.000000
25%	0.440000	0.560000	3.000000	156.000000	3.000000	0.000000	0.000000	0.000000
50%	0.640000	0.720000	4.000000	200.000000	3.000000	0.000000	0.000000	0.000000
75%	0.820000	0.870000	5.000000	245.000000	4.000000	0.000000	0.000000	0.000000
max	1.000000	1.000000	7.000000	310.000000	10.000000	1.000000	1.000000	1.000000

Hence, we can discover a possible cause of leave analyzing this data. So data is a key part of any data science project.

After seeing a full description of the data summary it is concluded that if we understand the data statistics we can solve most of the problem factors.

Here, I am going a little dipper to show the actual data of leaving versus not leaving. For this, I have used matplotlib and seaborn library.

```
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
left_counts=df['left'].value_counts()
print(leftcounts)
plt.pie(left_counts,labels=['Not leave','leave']);
```

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

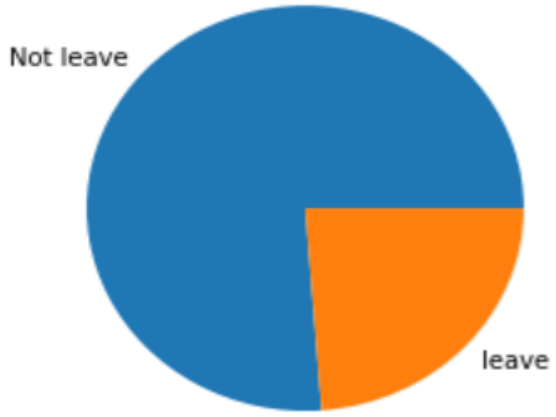


Fig: pie chart for leave and not leave.

Modeling

So far, we have discovered so many insights on this case study. Now modeling is one of the best ways to validate our work. For this, we need to split the datasets into two parts called training set (usually we make it 80%) and testing set (usually we make it 20%). Basically, a training set is used to train the model, and the test set is used to test the model. This is to avoid the overfitting

the model to reach high accuracy. For example, after completing data wrangling we split the sample data following ways.

Splitting the train and test sets

X_train, X_test, y_train,y_test= cross_validation.train_test_split(X,y,test_size=0.2)

Once we have done the split, we will proceed to train and validate the model. We can use several models to achieve our goal for example

Support Vector Machine (SVM)

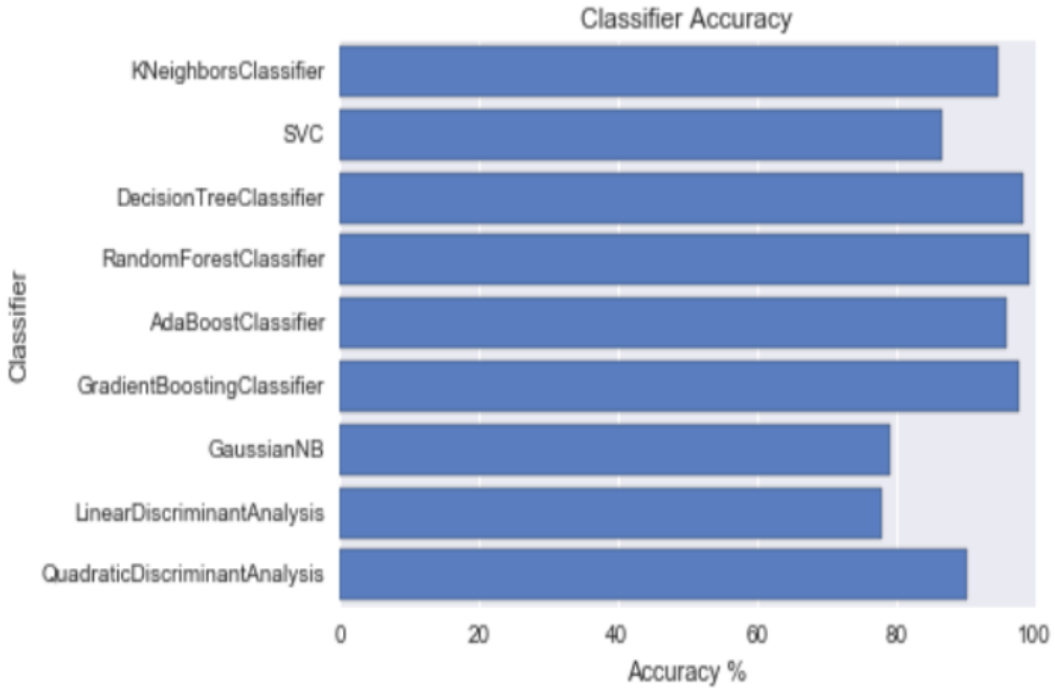
Decision Tree

Random forest

KNearest neighbor etc.

But for this case study, I will explain about SVM. It is a supervised machine learning algorithm for classification and regression. This model will process each data point in n-dimensional space where n means the number of features we have. Each of the values is the value of a particular coordinate in the n-dimensional planes. The classification will then find this optimal hyperplane which gives the largest minimum distance to the training examples.

Likewise, we can go through each model to find the best accuracy. For example, I found this picture on the internet for this case study. Let’s have a look.



Source: **https://towardsdatascience.com/whos-quitting-today-e1b0ca2fa90f**

Model Interpretation

Based on the result, we can choose the model because svc has 87% accuracy and the random forest has 99% accuracy.

Revisit Models

To achieve the best model accuracy, we can perform training on several other classifications models than just 1 or 2.

Deployment

The main purpose of this model is finding why their best employee voluntarily leaves their job. After walking through all the above steps it is concluded that we need to find which factors are being responsible to make them that decision and HR should take action accordingly to fix these problems lower the company's attrition rate.

Summary and Conclusions

The main purpose of this case study is to find the causes and associated factors to leave employees voluntarily from their job. This is a very serious matter for many companies because it wastes millions of dollars for hiring and training a new employee. Thus only the salary and position can't stop this problem. It is also related to the relationship of the manger, colleagues, health insurance, and so on. Hence it can be minimized but not 100% controllable. Even if they minimize this problem, they can save millions of dollars. For this study, I have downloaded the datasets from Vincent Tatan GitHub repository. In fact, these datasets have a full explanation of the factors and variables so we can go each and a specific point to extract more meaningful insights.

References:

[1] https://towardsdatascience.com/whos-quitting-today-e1b0ca2fa90f

[2] Wiley, Applied Predictive Analytics

[3] Eric S., Willey, Predictive Analytics