ISSN (O) 2393-8021, ISSN (P) 2394-1588

IARJSET

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 11, November 2021

DOI: 10.17148/IARJSET.2021.81151

Predicting Employee Churn in Python

Tushar Singh[1], Dr. Anu Rathee[2]

[1]Student, Information Technology, MAIT, Delhi, India

[2]Assistant Professor, Information Technology, MAIT, Delhi, India

Abstract: For any service-providing organizations, churners have always been a big issue. It increases the company's cost and lowers down the profit rate. Commonly, customer deterioration can be identified when they instigate the process of service termination. At the same time, the people and the organizations that provide the data residing on the government databases and the agencies who sponsor the collection of such details are becoming increasingly conscious that extended analytical capabilities also deliver tools that endanger the confidentiality of data records. Nevertheless, using predictive analysis using customer's previous service usage, service performance, spending, and other behavior patterns, the possibility of whether a customer wants to discontinue service can be nailed. In this paper, the writers address the problem of churn analysis, assuming a scenario in which a organization owning confidential databases wishes to run a churn analysis technique on the union of their databases without exposing any unnecessary information. The paper aims to predict whether a customer will churn shortly or not based on the predictive analysis using billing data of a telecom company.

Keywords : Random Forest, Churn , Employee-Churn

1. INTRODUCTION

Definition

Data mining is defined as the method of finding patterns in data. The procedure must be automated or (more usually) semiautomatic. The patterns found must be noteworthy because they lead to some advantage, usually an economic benefit. The data is invariably present in significant quantities. Data mining is a domain linking the three globes of Databases, Artificial Intelligence, and Statistics. The knowledge age has allowed many organizations to assemble enormous volumes of data.

However, the effectiveness of this data is inconsequential if "significant information" or "knowledge" cannot be pulled from it. Data mining, otherwise known as knowledge discovery, tries to respond to this need. Data mining techniques search for exciting information without requiring a prior hypothesis in disparity to standard statistical techniques. As a domain, it has presented new concepts and algorithms such as association rule learning. It has also used known machine-learning algorithms such as

inductive-rule learning (e.g., by decision trees) to the setting where extensive databases are involved. Data mining methods are used in corporations and research and are becoming more and more widespread with time

A. Data mining is the procedure of pulling patterns from data. As more data are gathered,with the quantity of data doubling every three years, data mining is becoming an increasingly essential tool to convert these data into information. It is generally

used in various profiling practices, such as marketing, surveillance, fraud detection and scientific discovery. Data mining is utilized for a variety of objectives in both the private and public sectors. Industries such as banking, insurance, medicine, and

retailing commonly utilize data mining to lower costs, enhance research, and increase sales.

For example:

- The insurance and banking enterprises use data mining applications to catch forgery and help in threat assessment (e.g., credit scoring).
- Utilizing consumer data gathered over several years, corporations can design models that indicate whether a customer is a proper credit risk, or whether an accident claim may be dishonest and should be examined more closely.

• The medical community sometimes uses data mining to help predict the effectiveness of a process or medicine. The insurance and banking industries use data mining applications to detect fraud and assist in risk assessment (e.g., credit scoring). Pharmaceutical businesses use chemical compounds and genetic material data mining to support focus research on new cures for diseases.

- Retailers can utilize information accumulated through affinity programs (e.g., shoppers' club cards, regular flyer points, contests) to evaluate the efficacy of product selection and placement decisions, coupon offers, and which products are often bought jointly.

ISSN (O) 2393-8021, ISSN (P) 2394-1588

IARJSET

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 11, November 2021

DOI: 10.17148/IARJSET.2021.81151

- Organizations such as telephone service providers and music clubs can utilize data mining to make a “churn analysis” to evaluate which customers are likely to remain subscribers and which ones are likely to change to a competitor.

One special part of data mining is Churn analysis. It is the estimation of the rate of attrition in the customer base of any corporation. It involves recognizing those consumers who are most presumably to quit using a service or product. Churn analysis is especially valuable in designing a sustainable and powerful strategy for customer retention in a company. When a corporation is conscious of the percentage of customers who terminate relationships with them in a given period, they can efficiently develop a detailed analysis of the reasons for the churn rate using churn analysis. This assists in developing effective customer retention programs for the company. Churn rate typically applies to many industries especially among them are subscription services, such as long-distance phone service or magazines. Churn analysis assists in comprehending the behavior of customers who unsubscribe and move their business to a competitor and predict the likelihood of this event. Other uses vary from calculating employee attrition in any given company.

B. Motivation Our work is motivated by the need to both protect privileged information and enable its use for commercial or other purposes

2. RELATED WORK

We have studied to discover out how much job is done in the domain of churn analysis. We discovered some works that discuss churn analysis in detail in our search.

We are conveying few of the results here :

In [5] the authors offer predictive modeling for churners based on data mining techniques. The paper also concerns how to use the decision tree analysis model in detail. The paper mainly examined customer churning from a enterprise perspective. In the end of the paper, it also concerned case studies along with process flows and modeling techniques.
In [6] Teemu Mutanen depicted a case study on customer decay. The paper explained in detail the techniques used for the prediction, data utilized, and the achieved outcome. The author expressed two ways for churn analysis. The first one is logistic
regression. Logistic regression is utilized to indicate a discrete effect based on continuous and categorical variables. In this way, only one dependent variable can exist. This technique applies highest likelihood estimation after converting the dependent variable into a logistic variable. The second technique examines the estimation
results of the logistic regression. It is known as the lift curve. This curve is related to the ROC curve of signal detection theory and the precision-recall curve. The lift measures the predictive model calculated as the ratio between the results obtained with and without the predictive model.
In [7] Shyam V. Nath depicts a case study in which an Oracle-based database of fifty
thousand customers of wireless telecommunication industry was studied to indicate
churners. The study utilized JDeveloper tools, and the analysis was accomplished using Naïve Bayes algorithm with supervised learning.
Marco Richeldi and Alessandro Perucci [8] wrote a paper on case study of churn analysis. This paper examines the usage of Mining Mart, an churn analysis instrument. It primarily concerns the preprocessing of data to study with Mining Mart.

3. METHODOLOGY

A. Our Data: The data we gathered are of a renowned telecom company. To load the dataset we use pandas csv function.

- ○ This dataset has 14,999 samplings and 10 attributes (6 integer, 2 float, and 2 objects)
- ○ No variable column has null/missing values

The 10 attributes are as follows:

- satisfaction level: It is employee happiness point, which ranges from 0-1.
- last evaluation: It is evaluated by performance of the employer, which also ranges from 0-1.
- number projects: How many projects were assigned to an employee?
- average_monthly_hours: How many average numbers of hours worked by an employee in a month?
- time_spent_company: time_ How much average hours worked by an worker in a month? spent_company means employee experience i.e. The number of years spent by an employee in the company.
- work_accident: Whether an employee has ever had a work mishap or not.
- promotion_last_5years: Whether an employee has had a promotion in the previous 5 years or not.
- Departments: Employee's working department.

IARJSET

ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 11, November 2021

DOI: 10.17148/IARJSET.2021.81151

- Salary: The employee's salary level such as low, medium, and high.
- Left: Whether the employee has left the company or not.

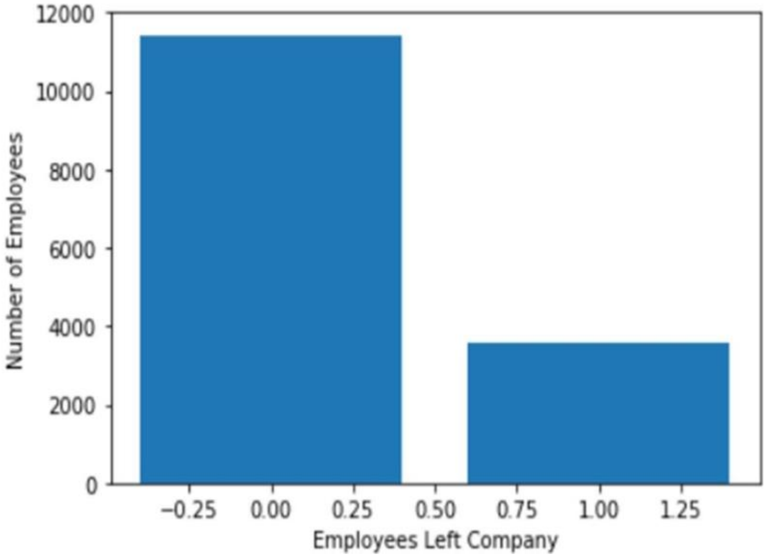
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14999 entries, 0 to 14998
Data columns (total 10 columns):
satisfaction_level       14999 non-null float64
last_evaluation          14999 non-null float64
number_project           14999 non-null int64
average_montly_hours     14999 non-null int64
time_spend_company       14999 non-null int64
Work_accident            14999 non-null int64
left                     14999 non-null int64
promotion_last_5years    14999 non-null int64
Departments              14999 non-null object
salary                   14999 non-null object
dtypes: float64(2), int64(6), object(2)
memory usage: 1.1+ MB
```

B. Data Visualization

Employees Left: to find how many employee left we can plot a bar graph using Matplotlib.

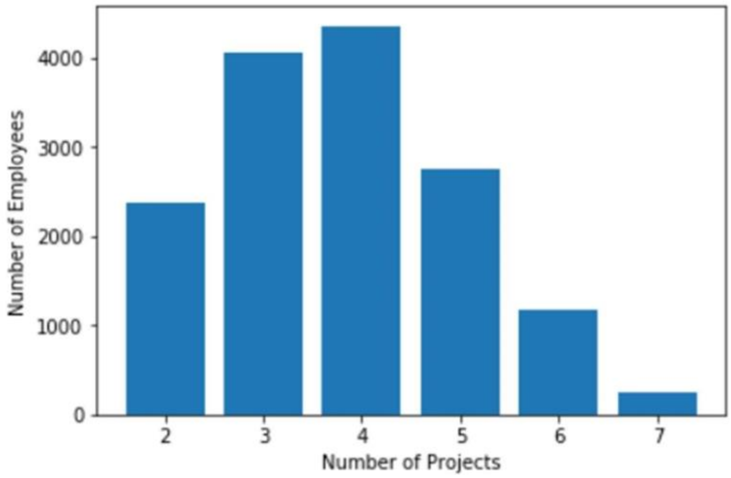
```
left_count=data.groupby('left').count()
plt.bar(left_count.index.values, left_count['satisfaction_level'])
plt.xlabel('Employees Left Company')
plt.ylabel('Number of Employees')
plt.show()
```



ISSN (O) 2393-8021, ISSN (P) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Vol. 8, Issue 11, November 2021
DOI: 10.17148/IARJSET.2021.81151

Number of Projects
Similarly we can find number of projects

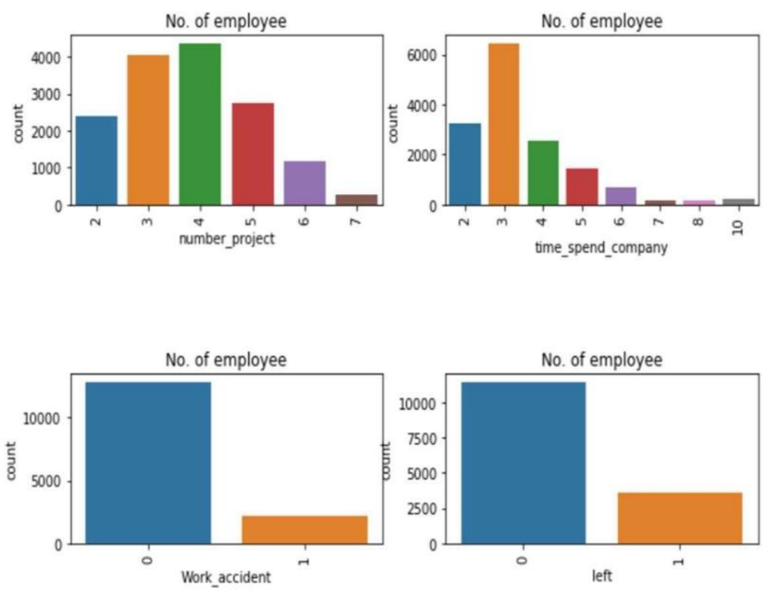
```
num_projects=data.groupby('number_project').count()
plt.bar(num_projects.index.values, num_projects['satisfaction_level'])
plt.xlabel('Number of Projects')
plt.ylabel('Number of Employees')
plt.show()
```



○ Most of the employee is doing the project from 3-5.

Subplots using Seaborn
It could be time consuming to plot graphs by 1 single attribute so we use Seaborn library and plot all the graphs in a single run using subplots.

```
features=['number_project','time_spend_company','Work_accident','left', 'promotion_last_5years','Departments ','salary']
fig=plt.subplots(figsize=(10,15))
for i, j in enumerate(features):
    plt.subplot(4, 2, i+1)
    plt.subplots_adjust(hspace = 1.0)
    sns.countplot(x=j,data = data)
    plt.xticks(rotation=90)
    plt.title("No. of employee")
```



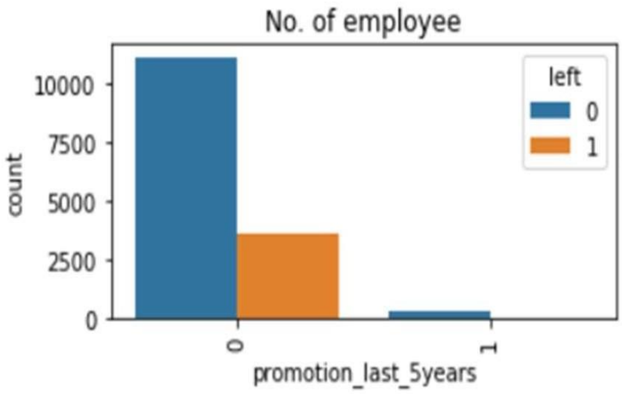
IARJSET

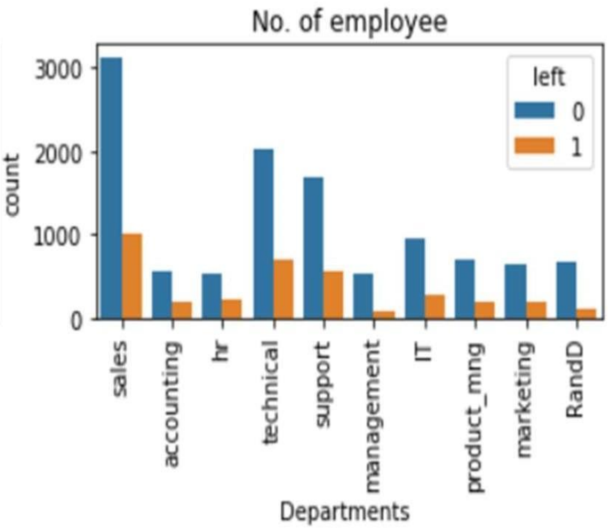
ISSN (O) 2393-8021, ISSN (P) 2394-1588

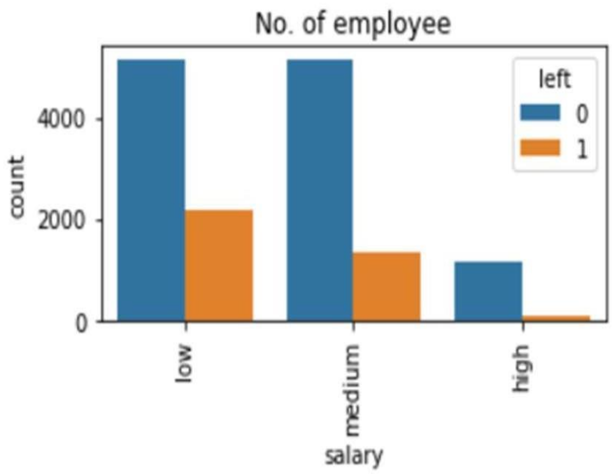
International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 11, November 2021

DOI: 10.17148/IARJSET.2021.81151







We can observe that:

○ Those employees with the number of projects greater than 5 leave the enterprise.

○ The employee who had accomplished 6 or 7 projects, left the corporation. It seems to like that they were overfilled with the assignment.

○ The employee with five-year experience is exiting more because of no advancements in last 5 years and greater than 6 years background are not exiting because of attachment with the company

○ Those whose promotion in last five years they didn't leave, i.e., all those left they didn't get the advancement in the last five years.

C. Data Analysis and Visualization Summary:

The following characteristics most force a person to exit the company:

○ Promotions: Workers are far more likely to leave their job if they haven't obtained a promotion in the last five years.

○ Time with Company: Here, The three-year mark looks like a time to be a crucial point in an employee's career. Most of them leave their job near the three-year mark. Another critical point is 6-years end, where the employee is very doubtful to leave.

○ Number Of Projects: Employee engagement is another essential factor influencing the worker to leave the company. Employees with 3-5 projects are less likely to quit the company. The employee with less and more projects are likely to exit.

○ Salary: Most of the employees quit among the mid or low-salary groups.

Cluster Analysis:

IARJSET

ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 11, November 2021

DOI: 10.17148/IARJSET.2021.81151

```
# Add new column "label" annd assign cluster labels.
left_emp['label'] = kmeans.labels_
# Draw scatter plot
plt.scatter(left_emp['satisfaction_level'], left_emp['last_evaluation'], c=left_emp['label'],cmap='Accent')
plt.xlabel('Satisfaction Level')
plt.ylabel('Last Evaluation')
plt.title('3 Clusters of employees who left')
plt.show()
```

Most of the employees quit among the mid or low-salary groups:

○ High Satisfaction and High Evaluation(Shaded by green color in the graph), you can also call them Winners.
○ Low Satisfaction and High Evaluation(Shaded by blue color(Shaded by green color in the graph), you can also call them Dissatisfaction.
○ Moderate Satisfaction and moderate Evaluation (Shaded by grey color in the graph), you can also call them Poor match'.

D. Building a Prediction Model

Pre-Processing Data

Lots of machine learning algorithms need numerical input data, so we need to describe categorical columns in a numerical column

To encode this data, we can map each value to a number like a salary column's value can be described as low:0, medium:1, and high:2.

```
# Import LabelEncoder
from sklearn import preprocessing
#creating labelEncoder
le = preprocessing.LabelEncoder()
# Converting string labels into numbers.
data['salary']=le.fit_transform(data['salary'])
data['Departments ']=le.fit_transform(data['Departments '])
```

This procedure is known as label encoding, and sklearn conveniently will do this for us using Label Encoder.

Split Train and Test Set

To comprehend model performance, splitting the dataset into a training set and a test set is a sound strategy

Let's divide dataset by using function train_test_split(). We need to give 3 parameters features, target, and test_set size. Also, we can utilize random_state to choose records randomly.

IARJSET

ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 11, November 2021

DOI: 10.17148/IARJSET.2021.81151

```
#Spliting data into Feature and
X=data[['satisfaction_level', 'last_evaluation', 'number_project',
        'average_montly_hours', 'time_spend_company', 'Work_accident',
        'promotion_last_5years', 'Departments ', 'salary']]
y=data['left']
```

```
# Import train_test_split function
from sklearn.model_selection import train_test_split

# Split dataset into training set and test set
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)  # 70% training and 30% test
```

```
#Import Gradient Boosting Classifier model
from sklearn.ensemble import GradientBoostingClassifier

#Create Gradient Boosting Classifier
gb = GradientBoostingClassifier()

#Train the model using the training sets
gb.fit(X_train, y_train)

#Predict the response for test dataset
y_pred = gb.predict(X_test)
```

```
#Import scikit-learn metrics module for accuracy calculation
from sklearn import metrics
# Model Accuracy, how often is the classifier correct?
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
# Model Precision
print("Precision:",metrics.precision_score(y_test, y_pred))
# Model Recall
print("Recall:",metrics.recall_score(y_test, y_pred))
```

```
Accuracy: 0.9715555555555555
Precision: 0.958252427184466
Recall: 0.9207089552238806
```

```
from sklearn.ensemble import RandomForestClassifier
Rf = RandomForestClassifier()
Rf.fit(X_train, y_train)
y_pred = Rf.predict(X_test)
```

```
#Import scikit-learn metrics module for accuracy calculationfrom sklearn import metrics
from sklearn import metrics
# Model Accuracy, how often is the classifier correct?
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
# Model Precision
print("Precision:",metrics.precision_score(y_test, y_pred))
# Model Recall
print("Recall:",metrics.recall_score(y_test, y_pred))
```

```
Accuracy: 0.9833333333333333
Precision: 0.9911330049261083
Recall: 0.9384328358208955
```

Model Building

Let's build employees a churn prediction model.
Here, we are going to anticipate churn using Gradient Boosting Classifier.
Foremost, import the GradientBoostingClassifier module and make Gradient Boosting classifier object using GradientBoostingClassifier() function.
Similarly import RandomForestClassifier() module and create random forest classifier object using RandomForestClassifier() function.
Then, we model on train set using fit() and execute prediction on the test set utilizing predict().

ISSN (O) 2393-8021, ISSN (P) 2394-1588

IARJSET

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 11, November 2021

DOI: 10.17148/IARJSET.2021.81151

Evaluating Model Performance

We got a classification rate of 97% in gradient boost and 98.3% in random forest.
PrecisionPrecision is about being accurate, i.e., how accurate your model is. In other words, we can say, when a model makes a projection, how often it is correct. In our prediction case, when our Gradient Boosting model predicted an employee will leave, that employee actually left 95% of the time or 99% in the case of the random forest model.

RecallIf there is an employee that left present in the test set and our Gradient Boosting model can identify it 92% of the time or 93.8% in case of random forest model
Tool used: jupyter notebook

4. OBSTACLES FACED

1) Incomplete Data: The data that we worked with was incomplete. Therefore, the result we achieved is not accurate. However, it gives us a picture of how churn analysis works.
2) Time: Churn analysis is a lengthy and complex process. Therefore, the time frame was not enough to do a complete churn analysis and obtain a viable result.

Confidentiality Issues: A key problem that arises in any mass grouping of data is confidentiality. The need for privacy is occasionally due to law (e.g., medical databases) or inspired by business interests. However, there are cases where data sharing can lead to joint gain. A key utility of large databases today is study, whether it is scientific or economic and market oriented. Thus, for example, the medical field has much to gain by pooling data for research; as can even competing businesses with mutual interests. Despite the potential gain, this is often not possible due to the confidentiality issues which arise.

5. FUTURE SCOPE

The result that we got is not promising. This is due to the fact that we had a dataset that was incomplete. It is our failure that we could not manage a proper dataset to do churn analysis. In future there are many things that we want to do for this project.
Some of them are listed below –

- Use a complete dataset to implement churn analysis method.
- Implement more methods to analyse churn.
- Compare different methods to find the optimal one.

What will we try to do next is to find how to mask the data to preserve privacy. The factors related to the choice of the privacy preserving algorithm are: characteristics of a good masking technique, disclosure risk, and, minimum disclosure risk.

6. CONCLUSION

We learned What is Employee Churn?, How it is different from customer churn, Exploratory data analysis and visualization of employee churn dataset using matplotlib and seaborn, model building and evaluation using python scikit-learn package. Churn analysis in data mining activities is a very important issue in many applications. Different techniques are likely to play an important role in this domain. However, this paper illustrates some of the challenges that these techniques face in churn analysis. It showed that under certain conditions it is relatively easy to breach the privacy protection offered by the different techniques. It provided extensive experimental results with different types of data and showed that this is really a concern that we must address. In addition to raising this concern the paper offers a model churn analysis technique that may find wider application in developing a new perspective toward developing better churn analysis techniques. It is interesting for a company's perspective whether the churning customers are worth retaining or not. And also, in marketing perspective what can be done to retain them. How long a time period for the data should be is also a matter of interest

7. REFERENCES

1. ps://s3.amazonaws.com/assets.datacamp.com/blog_assets/Employee+Churn+in+Python/HR_comma_s ep.csv
2. http://en.wikipedia.org/wiki/Supervised_learning
3. http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html

ISSN (O) 2393-8021, ISSN (P) 2394-1588

IARJSET

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 11, November 2021

DOI: 10.17148/IARJSET.2021.81151

4. K. B. Oseman, S.B.M. Shukor, N. A. Haris, F. Bakar, "Data Mining in Churn Analysis Model for Telecommunication Industry", Journal of Statistical Modeling and Analytics, Vol. 1 No. 19-27, 2010.
5. T. Mutanen, Customer churn analysis – a case study, Technical Report, Retrieved from, http://www.vtt.fi/inf/julkaisut/muut/2006/customer_churn_case_study.pdf , April 12, 2014.
6. S.V. Nath, Customer Churn Analysis in the Wireless Industry: A Data Mining Approach, Technical Report, retrieved from http://download.oracle.com/owsf_2003/40332.pdf , April 14, 2014.
7. M. Richeldi and A. Perrucci, "Churn Analysis Case Study", Technical Report, Telecom Italia Lab, Italy, retrieved from http://sfb876.tudortmund.de/PublicPublicationFiles/richeldi_perrucci_200 2b.pdf , April 12, 2014.