IJARCCE

ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 7, July 2021

DOI 10.17148/IJARCCE.2021.10776

Machine Learning to Predict Employee Satisfaction v/s Exit: Regression and Classification Algorithms

Ravi Chandra L[1], Mithun Manjunath[2]

BE (CSE), VTU, India [1]

BE (CSE), VTU, India [2]

Abstract: Enterprise culture is the soul of an enterprise, which is the key to obtain sustainable competitive advantage. For enterprise survival and development, enterprise culture is not the direct factor, but the most lasting decisive factor. In this paper, given the important role of enterprise culture in the process of human resource management practice, combining cultural construction with recruiting ,training, utilizing, and retaining talent to improving the level of human resource management to achieve benign interaction between culture construction (of the company) and human resource management. An effort is made in realizing a long-term sustainable competitive goal to obtain an invincible position in the present competitive market. Human Resource Management can be defined as planning, organizing, directing and compensating human resources resulting in the creation and development of human relation with a view to contribute proportionately to the organizational and individual goals. The examination of raw or crude data and drawing conclusions out of it is called data analytics. In this paper we will be analysing the employee turnover pattern and the factors contributing to it. Efforts will be made to create a model that can predict if a certain employee will leave the company or not. The goal is to create or improve different retention strategies on targeted employees. The first step in data analytics- data pre-processing is presented in the paper. Data pre-processing techniques convert crude data into useful format. Real world data are generally incomplete- noisy, inconsistent and contains many errors. Removing these factors improves the quality of analysis and prediction. The focus of data analytics lies in inference, the process of deriving conclusions. In this paper 2 out of top 3 strategies affecting employee turnover are being analysed and graphs plotted. The 3 top features include evaluation v/s exit, average monthly income v/s exit and satisfaction v/s exit. In this paper we have taken up the challenge of predicting the exit vs. evaluation trends of the company, first using Logistic Regression method, and later using Random forest or Random decision forest method. We also implement SVM (Support Vector Machine) and KNN (K Nearest Neighbour) classification algorithm on the dataset and compare the accuracies of the model. Best model is selected on the basis of ROC graph and Accuracy Percentage.

Keywords: Logistic Regression, Random decision forest, Human Resource Management, examination of raw or crude data and drawing conclusions- data analytics, employee turnover pattern, data pre-processing, evaluation v/s exit, average monthly income v/s exit, satisfaction v/s exit, planning, organizing, directing and compensating, competitive advantage, SVM (Support Vector Machine), KNN (K Nearest Neighbour) classification algorithm, ROC graph.

I. INTRODUCTION

In this information era, huge amount of data is being stored, exchanged and conditioned. The volume of data that one has to deal with has exploded to unimaginable levels. Most of the data exists in its crude form and needs to be converted to useful format before analysis. This process of converting raw data into useful format is called data pre-processing. Real world data is [1]

- Incomplete: consists of missing attribute values or consists of only aggregate data.
- Noisy: containing errors or outliers.
- Inconsistent: containing discrepancies in code.
- Non Numerical values

II. MOTIVATION

Encourage the commitment of employees to increase their performance and also be loyal to the organization as a whole. Emphasis on the quality of employees engaged in organizations goes a long way in producing quality goods and services, which is of great benefit both to the customers and the organization. Ensuring flexibility plays an important part in the way employees are organized, this makes them to be adaptive and receptive to all forms of changes in all aspects of their jobs such as work hours; working methods. Integrating organizational goals into strategic planning in order to make these policies cut across ranks and files of organization and ensuring that they are gladly accepted and implemented on daily routine by line managers. [2] Unfortunately remuneration and designation are the chief factors

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 7, July 2021

DOI 10.17148/IJARCCE.2021.10776

determining the longevity of an employee in an organization. Paucity in opportunities is one of the factors that lead to attrition of the employees. Also, the employee's relationship with their supervisor plays a key role. Prejudicing and suppressing the growth of an individual leads the employee to search for an alternative.

III. PROBLEM STATEMENT

In the previous paper titled "*Human Resource Management: Big Data Analytics*" we considered a company doing business in this 21st century world. This company has employed large number of employees working in different departments. The size of the company is huge and has several departments. The company is existent in business from a very long period of time and several employees have already left the company. The company's historical data is well tabulated and records maintained. The company wants to understand what factors contributed most to employee turnover and to create a model that can predict if a certain employee will leave the company or not. The goal is to create or improve different retention strategies on targeted employees. In this paper we have taken up the challenge of predicting the exit vs. evaluation trends of the company, first using Logistic Regression method, and later using Random forest or Random decision forest method. KNN and SVM have to be applied on the dataset and accuracy to be noted and conclusions to be drawn.

IV. METHODOLOGY

A. Importing Libraries

We have used the following libraries: [3]

- NumPy is the fundamental package for scientific computing with Python.
- Pandas is for data manipulation and analysis. Panadas is an open source, BSD- licenced library providing easy-to-use data structures and data analysis tools.
- Matplotlib is a python 2D plotting library. It can be used in Python scripts, Jupyter notebook, web application servers and IPython shells.
- Seaborn is a Python data visualization library based on matplotlib for attractive and informative statistical graphics.

B. Importing data

C. Checking for missing values

It is very essential in data pre-processing to check for missing values. Figure 1 shows the Python code to check for missing values. In this attempt no missing values were found.

D. Renaming and rearranging the columns

It is essential to rename the columns so that analysis is effective. Figure 2 shows the process of renaming the columns and figure 3 shows an effort to move the column 'exit' to the end as it has to be predicted.

E. Exit rate

Exit rate of the employees need to be checked. Figure 4 shows the exit ratio calculation. 76% of the employees stayed and 24% of employees exited.

F. Logistic Regression

Since our dataset consists of data with categorical values and a graph showing sinusoidal behaviour is obtained using Logistic regression. Figure 5 shows the logistic regression approach.

G. Random forest method and finding accuracy

Figure 6 shows the random forest method to predict the exit ratio. Figure 7 shows the accuracy, precision and F1 score of the 2 models.

H. Plotting an ROC graph of the assignment. Figure 8 shows the ROC graph.

```
In [5]:
#Checking whether our data contains any missing value or not
df.isnull().any()

Out[5]:
satisfaction_level        False
last_evaluation           False
number_project            False
average_montly_hours      False
time_spend_company        False
Work_accident             False
left                      False
promotion_last_5years     False
sales                     False
salary                    False
dtype: bool
```

Figure 1 shows the Python code to check for missing values.

IJARCCE

ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 7, July 2021

DOI 10.17148/IJARCCE.2021.10776

V. EVALUATION V/S EXIT

- There is a bimodal distribution for those that had an exit. [4][5][6]
- Employees with low performance tend to leave the company more.
- Employees with high performance tend to leave the company more.
- The sweet spot for employees that stayed is within 0.6-0.8 evaluation.

VI. AVERAGE MONTHLY HOURS V/S EXIT

- Another bimodal distribution for employees that exited.
- Employees who had less than 150 hours of work left the company more.
- Employees who had more than 250 hours of work left the company more.

```
In [7]:
#Renaming the columns
df = df.rename(columns={'satisfaction_level': 'Satisfaction',
                        'last_evaluation': 'Evaluation',
                        'number_project': 'ProjectCount',
                        'average_montly_hours': 'AverageMonthlyHours',
                        'time_spend_company': 'YearsAtCompany',
                        'Work_accident': 'WorkAccident',
                        'promotion_last_5years': 'Promotion',
                        'sales' : 'Department',
                        'left' : 'Exit'
                        })
df.head()

Out[7]:
```

Figure 2 shows the process of renaming the columns

```
In [6]:
#Moving the column 'Exit' to the end which is to be predicted
front = df['Exit']
df.insert(0, 'Exit', front)
df.head()

Out[6]:
```

Figure 3 shows an effort to move the column ‘exit’ to the end as it has to be predicted.

```
Exit_Rate = df.Exit.value_counts()/len(df)
Exit_Rate

Out[9]:
0    0.761917
1    0.238083
Name: Exit, dtype: float64
```

Figure 4 shows the exit ratio calculation.

```
from sklearn.linear_model import LogisticRegression


print ("---Base Model---")
base_roc_auc = roc_auc_score(y_test, base_rate_model(X_test))
print ("Base Rate AUC = %2.2f" % base_roc_auc)
print(classification_report(y_test, base_rate_model(X_test)))

# NOTE: By adding in "class_weight = balanced", the Logistic Auc increased by about 10%! This adju
sts the threshold value
logis = LogisticRegression(class_weight = "balanced")
logis.fit(X_train, y_train)
print ("\n\n ---Logistic Model---")
logit_roc_auc = roc_auc_score(y_test, logis.predict(X_test))
print ("Logistic AUC = %2.2f" % logit_roc_auc)
print(classification_report(y_test, logis.predict(X_test)))
```

Figure 5 shows the logistic regression approach.

IJARCCE
ISSN (O) 2278-1021, ISSN (P) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 10, Issue 7, July 2021
DOI 10.17148/IJARCCE.2021.10776

```
# Random Forest Model
rf = RandomForestClassifier(
    n_estimators=1000,
    max_depth=None,
    min_samples_split=10,
    class_weight="balanced"
    #min_weight_fraction_leaf=0.02
    )
rf.fit(X_train, y_train)
print ("\n\n ---Random Forest Model---")
rf_roc_auc = roc_auc_score(y_test, rf.predict(X_test))
print ("Random Forest AUC = %2.2f" % rf_roc_auc)
print(classification_report(y_test, rf.predict(X_test)))
```

Figure 6 shows the random forest method to predict the exit ratio.

```
 ---Logistic Model---
Logistic AUC = 0.74
             precision    recall  f1-score   support

          0       0.90      0.76      0.82      1714
          1       0.48      0.73      0.58       536

avg / total       0.80      0.75      0.76      2250
```

```
 ---Random Forest Model---
Random Forest AUC = 0.97
             precision    recall  f1-score   support

          0       0.99      0.98      0.99      1714
          1       0.95      0.96      0.95       536

avg / total       0.98      0.98      0.98      2250
```

Figure 7 shows the accuracy, precision and F1 score of the 2 models.

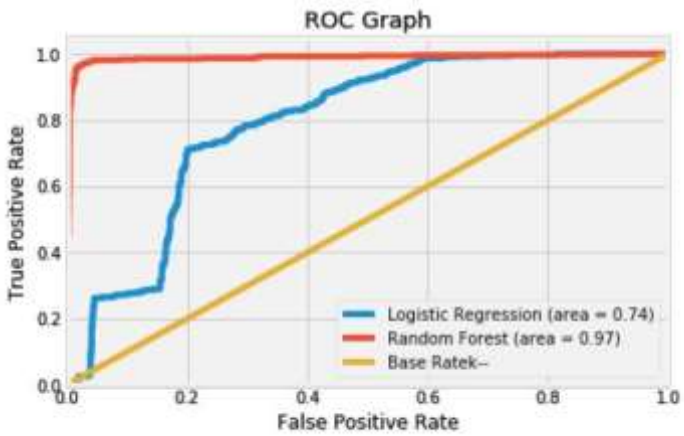

Figure 8 shows the ROC graph.

VII. SVM

Support Vector Machine (SVM) is a Classification algorithm used to create a supervised Machine Learning model. It uses a technique called the kernel trick to transform your data and then based on these transformations it finds an optimal boundary between the possible outputs. Figure 9 shows SVM modelling on the dataset. The accuracy of the model was measured to be nearly 87%. This will be later compared with the accuracies of other ML models for drawing conclusions.

IJARCCE
ISSN (O) 2278-1021, ISSN (P) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 10, Issue 7, July 2021
DOI 10.17148/IJARCCE.2021.10776

```
In [30]:
1  from sklearn.svm import SVC
2  svm = SVC(random_state = 1)
3  svm.fit(x_train.T, y_train.T)
4
5  acc = svm.score(x_test.T,y_test.T)*100
6  accuracies['SVM'] = acc
7  print("Test Accuracy of SVM Algorithm: {:.2f}%".format(acc))

Test Accuracy of SVM Algorithm: 86.89%
```

Figure 9 shows SVM modelling on the dataset.

VIII. KNN

K Nearest Neighbors (KNN) is a Classification algorithm used to create a supervised Machine Learning model. The neighbors are taken from a set of objects for which the class (for *k*-NN classification) or the object property value (for *k*-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required. Figure 9 shows KNN modelling on the dataset. The accuracy of the model was measured to be nearly 77%. Figure 11 shows the accuracy measure of all the classification algorithms.

```
In [28]:
1  # KNN Model
2  from sklearn.neighbors import KNeighborsClassifier
3  knn = KNeighborsClassifier(n_neighbors = 2)  # n_neighbors means k
4  knn.fit(x_train.T, y_train.T)
5  prediction = knn.predict(x_test.T)
6
7  print("{} NN Score: {:.2f}%".format(2, knn.score(x_test.T, y_test.T)*100))

2 NN Score: 77.05%
```

Figure 10 shows KNN modelling on the dataset.

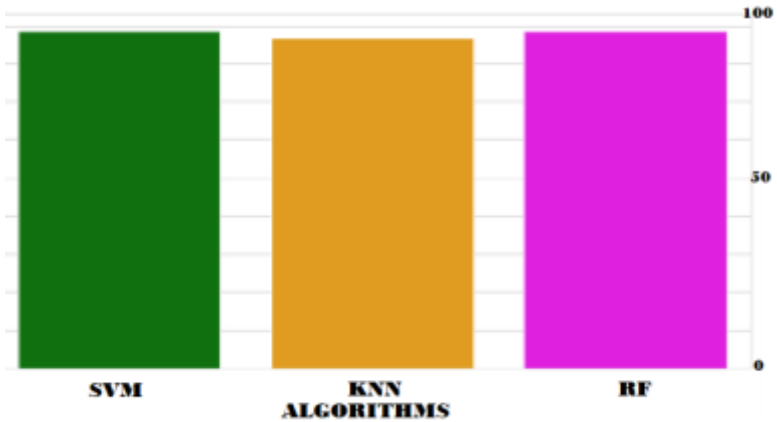

Figure 11 shows the accuracy measure of all the classification algorithms.

IX. CONCLUSIONS

Divergence between the nature of the job and the posting is controversial in most of the organizations. Organizations often treat an employee as a gofer and fail to feed the competence within the employee, which strongly needs to be condemned. This is primarily because employers think of employees as machinery parts that can be used as required depending on the requirement. As a result, the employee's attitude towards work suffers which in turn affects the productivity of the organization. A company proactive in business in this 21st century world had many workers leaving the company.[7] Data analytics had to be carried out on the data –both historical and present trend to draw inference. The goal was to create or improve different retention strategies on targeted employees working in different departments

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 10, Issue 7, July 2021
DOI 10.17148/IJARCCE.2021.10776

of the company. A python code was written and executed to analyse and draw conclusions. The first step in data analytics- data pre-processing was successfully carried out and exit ratio calculated. 2 out of top 3 strategies affecting employee turnover are being analysed and graphs plotted. The 3 top features include evaluation v/s exit, average monthly income v/s exit and satisfaction v/s exit. ROC graph [8] which can easily indicate the difference in the performance of the 2 models was plotted. The accuracy, precession, and F1 score was obtained and compared. The Accuracy of SVM and KNN were found to be nearly 87% and 77% respectively. Random Forest is the best fit for this assignment.

REFERENCES

[1] Principles of data mining- DJ Hand - Drug safety, 2007 – Springer
[2] Human Resource Management: Theory and Practice- July 2012- ISBN-978-978-50666-8-5
[3] Python for Data Analysis- Python data science libraries- Boston University-
https://pdfs.semanticscholar.org/407c/05a87c5eb47af7e7cddf414a23d2b4dfdac1.pdf
[4]Experimental data and statistical models for bimodal EM failures- 2000 IEEE International Reliability Physics Symposium Proceedings. 38th Annual (Cat. No.00CH37059)- DOI: 10.1109/RELPHY.2000.843940
[5]The need for new processes, methodologies and tools to support big data teams and improve big data project effectiveness
Published in: 2015 IEEE International Conference on Big Data (Big Data)- DOI: 10.1109/BigData.2015.7363988
[6] Bad Big Data Science- 2016 IEEE International Conference on Big Data (Big Data)- DOI: 10.1109/BigData.2016.7840935
[7] Study on Human Resource Management Practice from the Perspective of Enterprises Culture- Jibin, M., Zhenping, L., &Yizhe, Z. (2010). Study on human resource management practice from the perspective of enterprises culture. 2010, 2nd IEEE International Conference on Information Management and Engineering. DOI:10.1109/icime.2010.5477913
[8] An Introduction to ROC Analysis- Tom Fawcett, Elsevier-DOI:10.1016/j.patrec.2005.10.010

OUR GUIDE

VISHESH S (BE, MBA) born on 13th June 1992 hails from Bangalore (Karnataka) and has completed B.E in Telecommunication Engineering from VTU, Belgaum, Karnataka in 2015. He also worked as an intern under Dr. Shivananju BN, former Research Scholar, Department of Instrumentation, IISc, Bangalore. His research interests include Embedded Systems, Wireless Communication, BAN and Medical Electronics. He is also the Founder and Managing Director of the corporate company Konigtronics Private Limited. He has guided over a thousand students/interns/professionals in their research work and projects. He is also the co-author of many International Research Papers. He has recently completed his MBA in e-Business and PG Diploma in International Business. Presently Konigtronics Private Limited has extended its services in the field of Software Engineering and Webpage Designing. Konigtronics also conducts technical and non-technical workshops on various topics. Real estate activities are also carried out under the guidance of Siddesh B S BE (civil). Vishesh S along with his father BS Siddesh has received various awards and applauses from the scientific and entrepreneurial society. He was appointed as the MD of Konigtronics Pvt Ltd (INC. on 9th Jan 2017) at an age of 23 years. His name is indexed in various leading newspapers, magazines, scientific journals and leading websites & entrepreneurial forums. He is also the guide for many international students pursuing their Masters.