EDA AND TERMINATION PREDICTION OF EMPLOYEE

Umar, Harshika, Manasa, Nitya, Siddhesh
Mentor - Dr S. Sumalatha

****

Abstract

Management of the human resources are complex and if the resources are managed in an effective way, the fallouts would be highly effective. We are driven by this motive and chose this domain to carry out the project. We use the exploratory data analysis on the dataset to find the interesting and undiscovered information from the attributes. We use machine learning libraries to train the model. The model's main functionality is to predict whether the employee will continue to be the part or decide in either way. We use every feature of the dataset to the maximum extent to put forth the accuracy of the model.

Introduction

Human resource is a department that deals with employee details from marriage, marital status, gender to employee satisfaction, special projects, last performance review, etc. The human resources data aids the whole company's ability to operate and achieve more. HR plays a critical role in fostering a positive corporate culture as well as increasing employee engagement and productivity. Employee wellbeing and personal development are also handled by the HR department.

In this case, we used EDA (exploratory data analysis) to evaluate and study the data and summarize the key characteristics using a graphical representation. The basic goal of EDA is to inspect every fine feature. It enables analysts to have a better understanding of the data before making any assumptions. EDA is primarily used to explore what data might reveal beyond the formal modeling or hypothesis testing task, and it helps to comprehend data set variables and their interactions. It can also assess whether the statistical techniques you're considering for data analysis are appropriate.

Problem survey

Human resource sector is one of the areas around the corner for data science to bloom. Human resource data is very hard to get but the magnitude of its significance is phenomenal. We did choose this because more or less every human resource department's main area of work is to maintain the workforce in the company. For an employee to work in a company, many specific factors play key roles. So pondering upon the factors as such will definitely help to understand what is happening in the brains of employees in a company. With analyzing the data, they get to know the areas they need to work on inorder to facilitate better working conditions.

Dataset description

The dataset in the consideration is last updated on 2021, April and is in csv format. There are 36 columns and nearly 300 records. The dataset revolves around a fictitious company and the core data set contains names, DOBs, age, gender, marital status, date of hire, reasons for termination, department, whether they are active or terminated, position title, pay rate, manager name, and performance score.

Dataset preprocessing

The dataset does not have many preprocessing records since it is only of 300 records. We passed the data through the standard preprocessing checks such as finding any null values, checking data format and checking for the integrating of the columns to get enhanced results.

In the first step we went through the metadata and understood the contents of the dataset. We searched to find any null values in the dataset. A null in the dataset can be of many reasons and in order to get the desired results, the data must have to be complete. Taking care of the null values is dependent on the kind of data and primarily focused on the domain of the data.

Out of 36 columns in the dataset, there is only one attribute with the null values, it is "Date of termination". Having null values in this specific attribute of the dataset can be possible because of many varied kinds of reasons. One of the most common reasons could be the error in data management. This kind of issue can only be resolved by the

persons who collected and maintained the data as one cannot fill the value of this attribute based on a certain criteria. It is also possible that the null value may represent that an employee did not leave the company. So It can be considered as an employee still in the firm. As the data in the attribute is of date kind, the nulls have to be replenished by the same kind of data, and so to do that, the original data is required which we do not suppose to have. We can replace the null values with some other data like flagging it as 1 for instance, but this will disturb and change the data homogenous nature of the data in the attribute. So we chose to not include this attribute in the exploratory data analysis part and in the model building. In real life, attributes as such would carry a phenomenal value to the dataset but due to lack of resources are left with no option to make the better of it.

We checked the dataset to find the inconsistencies of the data type. We did not find any of such.

Since the dataset has 36 columns, we tried to chop down the data attributes which are not carrying the ample significance. Correlation has been found out between each and every attribute. We studied the correlation heatmap and tried to understand the reasons to have such results for every single attribute. We found most of the attributes to contribute in its own way and did not remove any attributes. We found some attributes in the dataset irrelevant and chose to place them out of consideration. This is the data preprocessing we did on the dataset, since the dataset is small and most of it completed necessary checks beforehand there is not much preprocessing needed to be carried out.

Implementation with code link

Exploratory data analysis is a critical process of performing initial investigations on data so as to find patterns, spot anomalies, to test hypotheses and check assumptions with the help of summary of the statistics and graphical representations. This helps to gain insights of the dataset and also maximizes the insights.

We found the unique values of each attribute. We calculated the value counts of the attributes to understand the data spread in the dataset. We used countplot to understand the spread of attributes in the perspective of the other attribute to find out the patterns and if any anomalies. We also used density plots to observe the data.

We selected the attributes on a criteria in which we thought the attributes have potential to affect the model's accuracy. We deliberately left out some attributes as they carry no value in this particular context. To not lose the data we created another dataframe of the same data with the selected attributes for the model building.

In the dataframe selected to build the model, there are some attributes which are of category data type such as "Sex" and it carry values as "Male and Female". Models only work on the numbers. So we used dummies to convert this kind of attribute to numerical data. To not face the dummy trap, we deleted some attributes for which dummies are formed. For example let us consider, there is an attribute with three categories, when this is passed to get dummies, three attributes will be formed and one among the three has to be dropped to not fall in the dummy trap which will affect the model. Values are scaled down in the dataset.

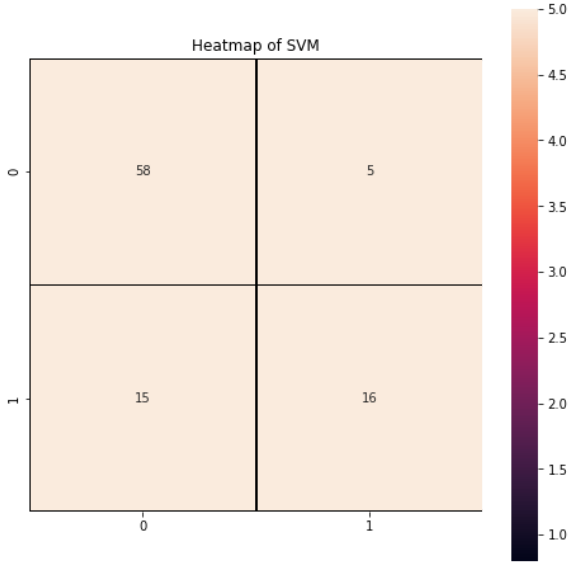
Data splitting is an essential task for model building. Data will be splitted into two portions for the cross-validatory purposes. One part will be used to train the model to develop a predictive model and the latter, testing data, is used to evaluate the model's performance. We splitted the data in a 70-30 ratio. 70% of the ratio is training data and the remaining 30% is testing data. We created the models in such a way that the result of the model's prediction is about the termination of the employee. It will tell whether an employee will leave the company or continue. We chose Random forest classifier, Decision tree. Logistic regression and Support vector machine models for the task from the sklearn library. To analyze the working efficiency of the model, every model's classification report has been generated in which the accuracy will be contrasted with the precision, recall. F1-score and support.

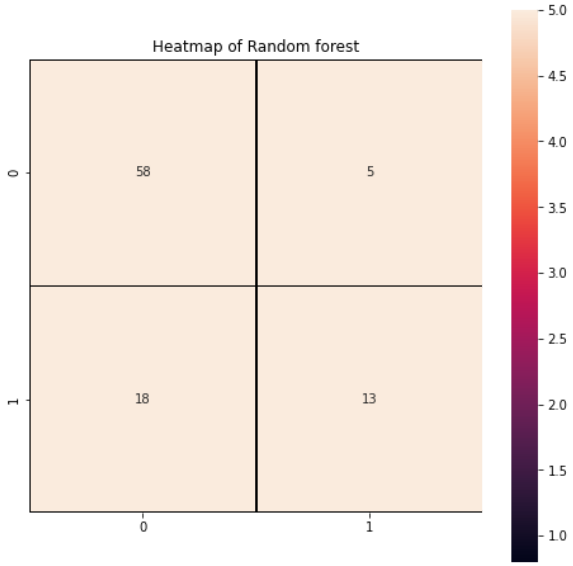
Link :

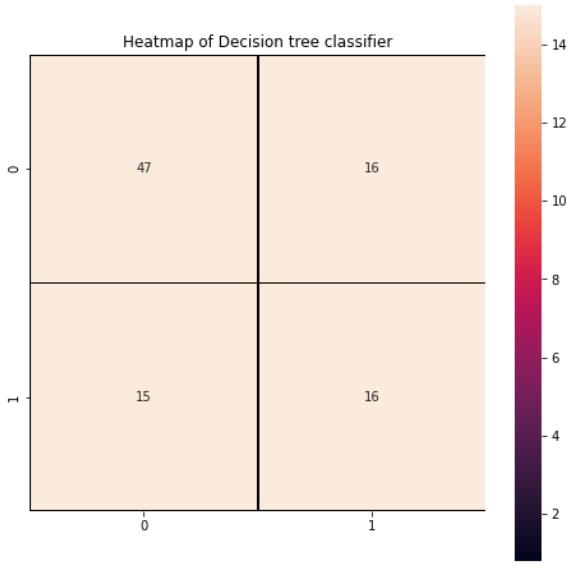
https://colab.research.google.com/drive/1T4kWNIVARTprORDdipj9CG0Kg2ask58i?usp=sharing

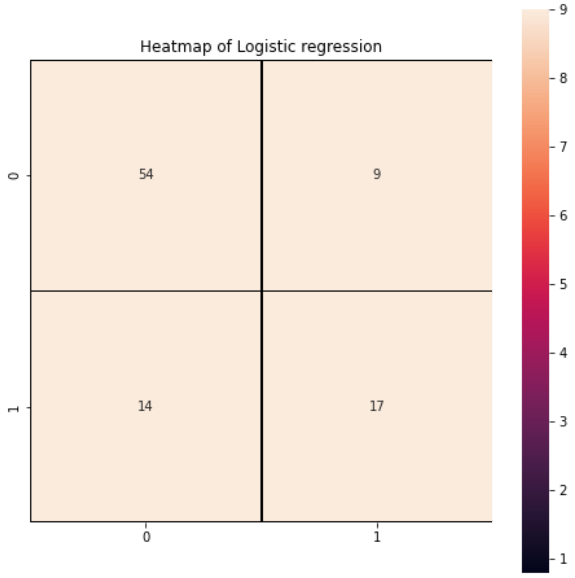
Resultant Graphs and Reports

Following are the heatmaps of the models which are tested against the testing data.









Classification reports of the models are as follows:

1. SVM

```
Classification report of SVM :
              precision    recall  f1-score   support

           1       0.76      0.52      0.62        31
           0       0.79      0.92      0.85        63

    accuracy                           0.79        94
   macro avg       0.78      0.72      0.73        94
weighted avg       0.78      0.79      0.77        94
```

2. Decision tree

```
Classification report of Decision tree classifier mo
              precision    recall  f1-score   suppor

           1       0.50      0.52      0.51        3
           0       0.76      0.75      0.75        6

    accuracy                           0.67        9
   macro avg       0.63      0.63      0.63        9
weighted avg       0.67      0.67      0.67        9
```

3. Logistic regression

```
Classification report of Logistic regression :
              precision    recall  f1-score   suppo

           1       0.65      0.55      0.60        3
           0       0.79      0.86      0.82        6

    accuracy                           0.76        9
   macro avg       0.72      0.70      0.71        9
weighted avg       0.75      0.76      0.75        9
```

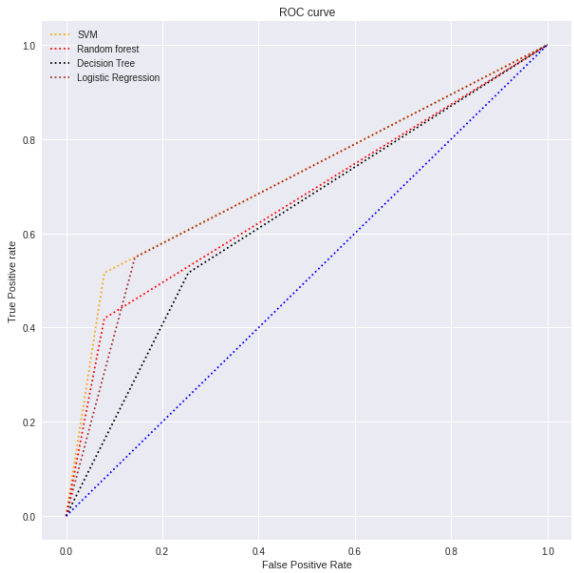
4. Random forest

```
Classification report of Random forest :
              precision    recall  f1-score   sup

           1       0.72      0.42      0.53
           0       0.76      0.92      0.83

    accuracy                           0.76
   macro avg       0.74      0.67      0.68
weighted avg       0.75      0.76      0.73
```

ROC CURVE of the models is as follows:

An Roc curve is a graph which shows the performance of the model at all classification thresholds with two parameters, True positive and False positive rate.

The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.



Conclusion

As said earlier, Management of the human resources are complex and if the resources are managed in an effective way, the fallouts would be highly effective. The exploratory data analysis. The exploratory data analysis that has been carried out has provided some significant results. Patterns have been found out and the insights also maximized. Based on the data models have been built and they can be used to predict the outcome with a good value of accuracy.

References

1.https://en.wikipedia.org/wiki/Exploratory_data_analysis
2.https://en.wikipedia.org/wiki/Human_resources
3.https://www.analyticsvidhya.com/blog/2021/06/understanding-random-for

est/#:~:text=Random%20forest%20is%20a%20Supervised,average%20in%20case%20of%20regression.
4.https://en.wikipedia.org/wiki/Decision_tree#:~:text=A%20decision%20tree%20is%20a%20flowchart-like%20structure%20in%20which,taken%20after%20computing%20all%20attributes).

Literature review

1. https://www.ijitee.org/wp-content/uploads/papers/v8i12/L359110812 19.pdf
2. https://adsabs.harvard.edu/pdf/2005ASPC..347...91B
3. https://www.dlr.de/sc/portaldata/15/resources/dokumente/pyhpc2011/submissions/pyhpc2011_submission_9.pdf
4. https://d1wqtxts1xzle7.cloudfront.net/65736020/ijatcse391012021-with-cover-page-v2.pdf?Expires=1648272451&Signature=dzxnK2pkvzz3bub~BMuQBB98D6DPn2Gdb0wDLMl-Eyn3HK83Lu-8iphly2XzjmhDturmlj-J3FykBAhGlUs0xknIMRFT~J7TIj5ztBt53rBjaCs9okbRo8xzCCm3u1~y82qiJ60QUOM38SyEMR0Unm900wWxS~PQwfOdVmjdbKoeA~d8YsQytjyi74PDdYblBMhAi~LLkKHmQf-MJGXh~kqF1j8yrlYewGWpfZtQL5m3xWaRSalmO56s~uYVYcMMKc~YwL-5-AU7ByEgX31qIqUI0dUcjpKvqZQB5pLiHys-DyarHodksfLghkqewIsY9oW0vidmPGbXBnLO5DF0AA__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA
5. https://www.kem.edu/wp-content/uploads/2012/06/9-Principles_of_correlation-1.pdf
6. https://www.mdpi.com/2227-7080/9/3/52
7. https://www.sciencedirect.com/science/article/pii/S1877042813046429