Employee Future Prediction

IFELOLUWA ESTHER BAKARE
B1239012@tees.ac.uk
Module Leader: Dr. Alessandro Di Stefano

Department of Data Science Teesside University, England, United Kingdom.

***Abstract*— The ability to solve real life situation by creating highly efficient and accurate model is the beauty of every machine learning project, the paper is a report of a research to create a model that predict whether an employee would leave work or not within two years, this is achieved by modelling past dataset gotten from Kaggle, the dataset was tuned to get the best result, major classification algorithms were performed, after which the dataset was oversampled to cater for biasness and the algorithms were performed again and compared, although all models performed well generally, the decision tree classifier and the adaptive boosting classifier performed best and showed to be the most efficient and accurate. The most important feature of a model is unbiasedness and this was majorly tackled in this research.**

***Keywords*— oversampling, exploratory data analysis, classification, boosting, knn, random forest, logistic regression.**

I. INTRODUCTION

Machine learning algorithms builds a model based on data in order to make predictions or decisions, with the overall data split in train and test data. Technology is constantly evolving making it possible to solve complex big data issues. The aim of this research is to build a highly efficient machine learning model to predict employee's future in a company. To do this past data were trained and modelled. Data for this study is obtained from Kaggle. This dataset contains 4653 records with 9 columns, it provides information about employees and whether or not they left the organization, the information provided includes Education, Joining Year, City, Payment Tier, Age, Gender, Ever Benched, Experience and Leave or not. The data would be modelled as a classification problem with Leave or not column as the target variable.

II. DATA EXPLORATION AND FEATURE SELECTION

A. Univariate Analysis

The dataset contains 4653 employee records with 9 columns

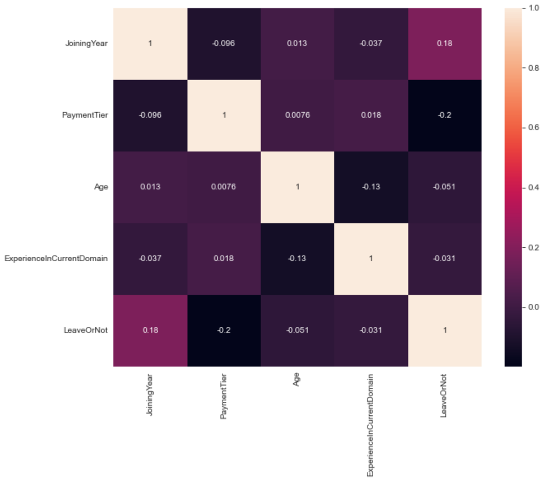


Figure 1 Correlation Heat Map

The correlation is weak between the features.

In this session, each column was visualized with a graph to draw inference and make conclusions on the data. The data column is listed below:
['Education', 'Joining Year', 'City', 'Payment Tier', 'Age', 'Gender', 'Ever Benched', 'Experience in Current Domain', 'Leave or Not']

Samples of the graphs are shown below

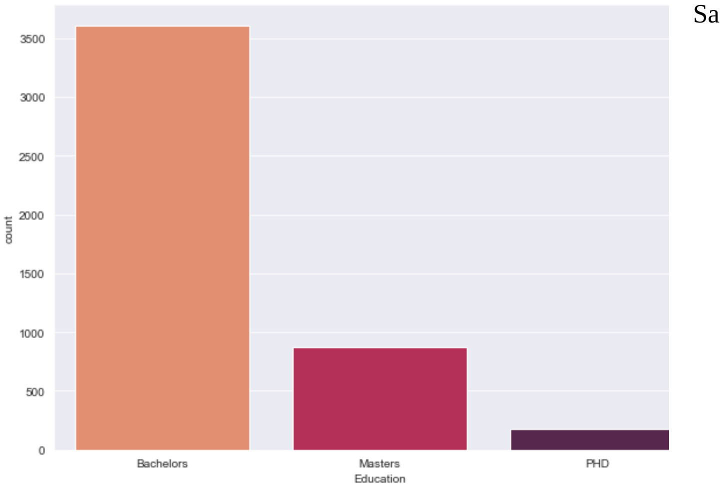


Figure 2 Education bar plot

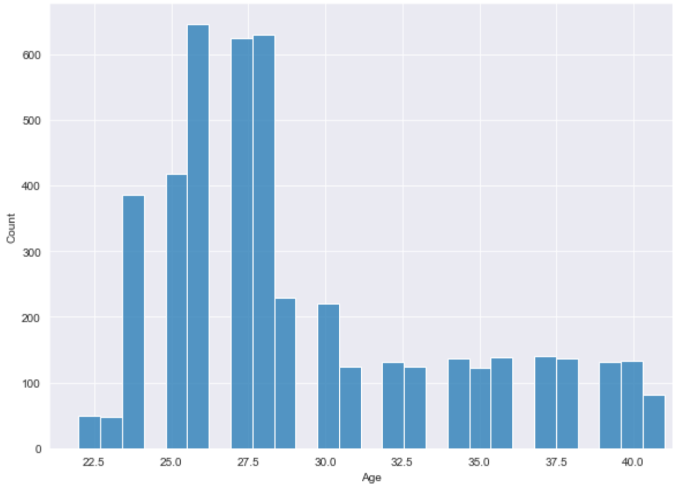


Figure 3 Age bar plot

The following observation were made from the analysis.

- Most of the employer had a bachelor's degree, followed by a Master's degree and very few of the employee had a PhD.
- Most employee joined in year 2017.
- A larger number of the employee were from Bangalore.
- A larger number of the employee belong to the 3rd payment tier.
- A larger number of the employee were male.
- Most employees are between age 25 to 28.
- Most employee have not been benched from a project.
- A higher number of employees did not leave.
- Most employee has experience between 0 to 2 years

B. *Bivariate Analysis*

In this session, each column was visualized against the target variable "Leave or not" with a graph to draw inference and make a more absolute conclusion on the data.

Samples of the graphs are shown below

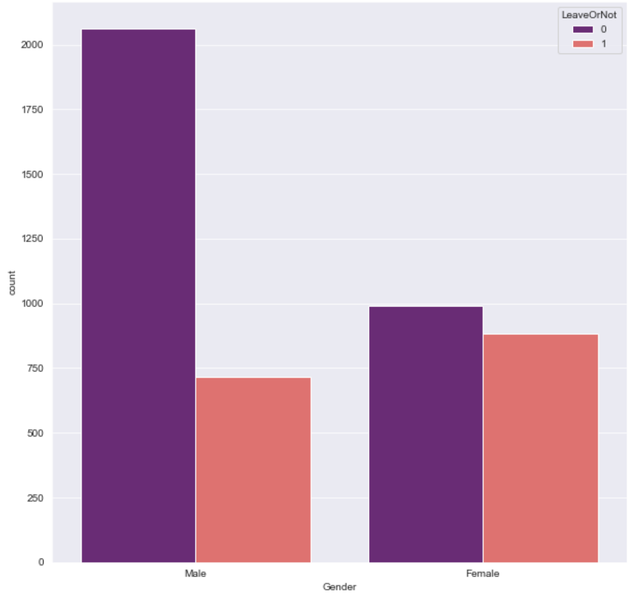


Figure 4 Proportion of gender that left and stayed

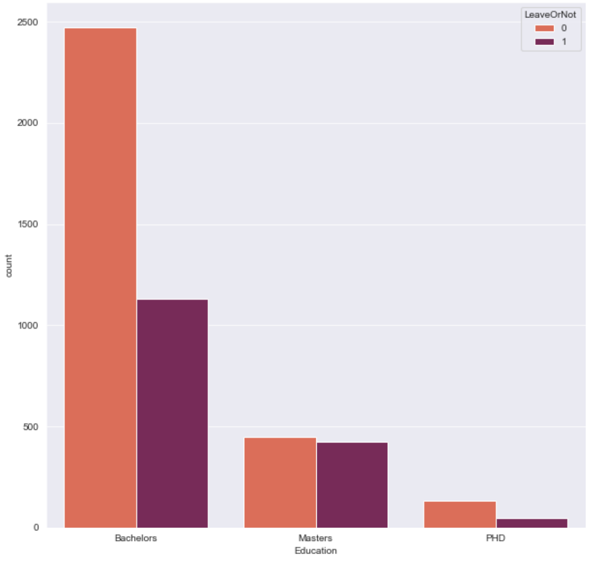


Figure 5 Education Level to Staying or Leaving

The following observation were made from the analysis.

- More women left their jobs.
- Employees with masters are more likely to leave their jobs compared to those with Bachelor's and Phd.
- Most of the employees that joined in 2018 left.
- Employees with 2 years' experience are most likely to leave.
- Employee between the ages 24 to 30 are the ones who leave the most.

The target variable is the column “Leave or not” while the remaining columns are the training features, there are no missing values in the data.

III. Experiments

Some features of the data were encoded to a numerical variable, these features include City, Education, Ever Benched and Gender. With all data types being integers, test and train split is performed with a test size of 0.3 and random state of 101.

A. *Models Used.*

The following algorithms were carried out K Neighbors Classifier, Random Forest Classifier, Logistic Regression, Decision Tree Classifier and Adaptive Boost Classification Algorithm.

B. *Evalation Metrics*

- **Precision** shows the proportion of true positives to the sum of true positives and false positives.
- **Recall** is the proportion of true positives to the sum of true positives and false negatives.
- **Accuracy** is the proportion of the correct prediction to all prediction

C. *Observation*

From the classification report of each algorithm given in tables below, the recall is poor, which could lead to missing out employees that would leave the organization, to address this oversampling is carried out on the data and the results are compared.

	K Neighbors Classifier		
	Precision	**Recall**	**F1 score**
0	0.77	0.97	0.86
1	0.88	0.44	0.58

Table 1 K Neighbors Classifier *report*

	Logistic Regression		
	Precision	**Recall**	**F1 score**
0	0.77	0.91	0.83
1	0.73	0.45	0.55

Table 2 Logistic Regression report

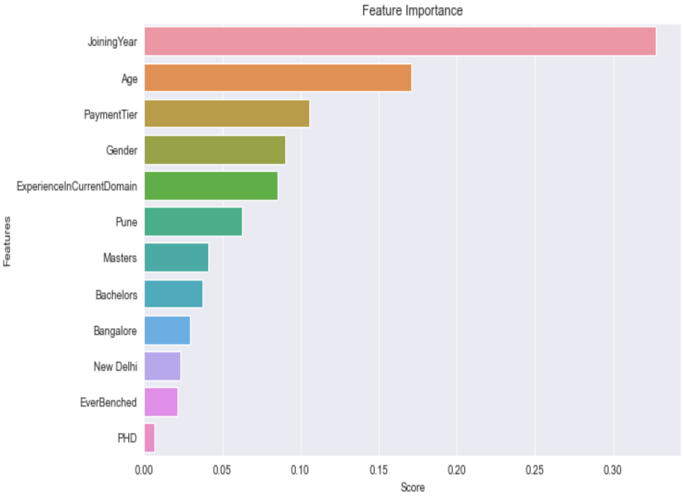
	Decision Tree Classifier		
	Precision	**Recall**	**F1 score**
0	0.83	0.86	0.84
1	0.70	0.64	0.67

Table 3 Decision Tree Classifier *report*

	Adaptive Boost Classification Algorithm		
	Precision	**Recall**	**F1 score**
0	0.82	0.97	0.89
1	0.90	0.60	0.72

Table 4 Adaptive Boost Classification Algorithm *report*

D. *Features Importance*



E. *Oversampling*

SMOTE (synthetic minority oversampling technique) is one of the most commonly used oversampling techniques today to solve data imbalance problem. It balances minority and majority distribution by randomly increasing the minor class by reproducing them, this way both the majority and minority are balanced up to each other and the problem of biasness is taken care of. After Oversampling, the classification algorithms are performed again and the precision value, recall value , accuracy and F ratio of the model before and after oversampling are compared .

IV. Results

The classification report gotten after oversampling improved the models greatly, the results are shown in the table below

	K Neighbors Classifier		
	Precision	**Recall**	**F1 score**
0	0.77	0.88	0.82
1	0.86	0.74	0.80

Table 5 K Neighbors Classifier *result*

	Logistic Regression		
	Precision	**Recall**	**F1 score**
0	0.67	0.67	0.67
1	0.67	0.67	0.67

Table 6 Logistic Regression result

	Decision Tree Classifier		
	Precision	**Recall**	**F1 score**
0	0.81	0.86	0.83
1	0.85	0.79	0.82

Table 7 Decision Tree Classifier *result*

V. DISCUSSION

I will briefly justify the algorithms that performed best from the result above.

- **K Neighbors Classifier** can compete with the most accurate models because it makes high and accurate predictions. Therefore, you can use the KNN algorithm in cases where high precision and accuracy is required
- **Decision Tree Classifier** is a type of supervised learning algorithms. Compared to other classes supervised learning algorithms, the decision tree classifier algorithm can be used for modelling and then predict both regression and classification problems and it is very efficient.
- **Adaptive Boost Classification Algorithm** can be used to boost the performance of any machine learning model. It is best used with weak algorithms. Boosting Algorithm are models that can be used to achieve more accuracy on a classification problem. The most common boosting classification algorithm is adaptive boost classification algorithm.

VI. CONCLUSION

Two classes of solutions were performed on the data, in the first stage the raw data was used without any form of sampling and major machine learning algorithms were performed, although the accuracy scores were high, the classification result depicts the model to be a biased one favoring the majority class, if this is overlooked, errors would be made in prediction in this case employee that are likely to Leave would be overlooked. Introducing Oversampling increased the model performance all round, the precision, recall and F1 score for both employee staying and employee leaving is way above 60% which makes the model efficient and accurate to an extent, although from the exploratory data analysis performed, so many factors are likely to affect whether and employee would stay or leave the company but the algorithms performed did some justice to the research question. Although, Decision Tree Classifier performed best after oversampling, the other algorithms were closely behind inefficiency. The Adaptive Boosting algorithm also greatly boosted the model. Overall, the models would perform more efficiently if the data sets were larger and more balanced.

ACKNOWLEDGMENT

My sincere thanks goes to my module leader Dr. Alessandro Di Stefano. I am highly indebted to laboratory coordinator Zainab Alalawi and all the entire team of Lecturers who have impacted me with Knowledge and the right skill set, for their direction and constant help as well as them always providing necessary information regarding the project and also for their help in completing the project, I am grateful. I am also thankful for the various Kaggle notebooks and paper referenced which has guided me in the project.

REFERENCES

1. M. Gagan Chandra. [Online] Kaggle [11classification Algos, Eda, Queries,Visualization | Kaggle]()
2. Marwan. [Online] Kaggle [Employee Future Prediction-Mn Md | Kaggle]()
3. Chatterjee, A. *et al.* (2021) 'OL13 - Machine learning with imbalanced clinical data: does synthetic minority oversampling help?', *Physica Medica*, 92, pp. S7. doi: https://doi-org.ezproxy.tees.ac.uk/10.1016/S1120-1797(22)00021-7.
4. Chen, J. *et al.* (2022) 'Machine learning-based classification of rock discontinuity trace: SMOTE oversampling integrated with GBT ensemble learning', *International Journal of Mining Science and Technology*, 32(2), pp. 309-322. doi: https://doi-org.ezproxy.tees.ac.uk/10.1016/j.ijmst.2021.08.004.
5. Kaisar, S. and Chowdhury, A. (2022) 'Integrating oversampling and ensemble-based machine learning techniques for an imbalanced dataset in dyslexia screening tests', *ICT Express*, doi: https://doi-org.ezproxy.tees.ac.uk/10.1016/j.icte.2022.02.011.
6. Liang, D. *et al.* (2022) 'Exploring ensemble oversampling method for imbalanced keyword extraction learning in policy text based on three-way decisions and SMOTE', *Expert Systems with Applications*, 188, pp. 116051. doi: https://doi-org.ezproxy.tees.ac.uk/10.1016/j.eswa.2021.116051.
7. Mylona, E. *et al.* (2020) 'PO-1535: Machine Learning and Oversampling techniques to predict urinary toxicity after prostate cancer RT', *Radiotherapy and Oncology*, 152, pp. S829. doi: https://doi-org.ezproxy.tees.ac.uk/10.1016/S0167-8140(21)01553-X.
8. Tao, X. *et al.* (2022) 'SVDD-based weighted oversampling technique for imbalanced and overlapped dataset learning', *Information Sciences*, 588, pp. 13-51. doi: https://doi-org.ezproxy.tees.ac.uk/10.1016/j.ins.2021.12.066.
9. Viloria, A., Pineda Lezama, O.B. and Mercado-Caruzo, N. (2020) 'Unbalanced data processing using oversampling: Machine Learning', *Procedia*

Computer Science, 175, pp. 108-113. doi: https://doi-org.ezproxy.tees.ac.uk/10.1016/j.procs.2020.07.018.