HR Attrition Prediction: comparison on two different synthetic datasets

Alex Santini
National College of Ireland Dublin,
Dublin
Email: 1st x21211604@student.ncirl.ie,

Abstract

Employee turnover can be considered one of the biggest concerns for companies and HR managers.

In 2022 newspapers have covered quiet intensively the causes and recent trends related to the event known as "great resignation".

Following the COVID-19 pandemic and related changes in workforce development, unemployment benefits and working conditions (mainly remote working) companies are now facing huge resignation rates.

Recent figures for Ireland shown all time high turnover rates and employers increasing salaries to try retaining talent and staff.

Turnover can be attributed to internal and external factors to an organization, however Human resources departments are now very interested in predicting and forecasting employee turnover.

For some company; technical, security and time constraints make attrition prediction a crucial aspect for creating a good work environment and maintaining productivity levels high.

Employees can be considered one of the most valuable assets of an organization and employee retention can have huge consequences in terms of value added and quality of the process within the organization.

Therefore, companies know how much it is indispensable to maintain a stable workforce; a task now more difficult to achieve than ever.

This project represents an attempt to discover the main causes of attrition and predicting employee turnover/attrition through different machine learning technique.

In line with previous literature this project achieved excellent levels of prediction on datasets and developed a comprehensive framework for attrition prediction using two publicly available synthetics multivariate datasets both revolving around employee retention but from different point of views and employing different features.

A. Introduction

Attrition is a common occurrence in the workplace, it refers to the gradual reduction in the number of employees within a company due to retirements, resignations, or deaths and can have significant impacts on an organization.

Attrition can lead to a decline in productivity and an increase in the workload for remaining employees, it can also be costly for an organization to constantly have to recruit and train new employees to replace those who have left.

In human resources, it is important for companies to understand the causes and effects of attrition and to implement retention strategies in order to mitigate its negative impacts.

Among the many approaches that organizations are using, machine learning techniques are now being deployed attempting to better understand and prediction attrition ratio withing the organization.

By analyzing large amounts of data on employee characteristics and behaviors; machine learning algorithms can identify patterns and trends that may predict an individual's likelihood of leaving an organization.

For example, factors such as job satisfaction, salary, benefits, and performance evaluations can have a huge impact in identifying employees who are at higher risk of leaving.

By proactively addressing the factors that contribute to attrition, organizations can effectively reduce employee turnover and improve retention rates.

Recent times have shown an increasing trend of higher rates of attrition across all organizations, researchers explain how it might be due to a combination of factors such as: a highly competitive job market, availability of more attractive job opportunities, and changes in the work environment; for example, the rise of remote work and the COVID-19 pandemic have disrupted traditional work models and may have contributed to higher levels of employee turnover.

Scope of this project is to develop and compare different machine learning models on the following task: predicting attrition on two different open sources datasets both from "Kaggle.com" which will allow for covering different features and techniques in line with the related literature; the models being used were: KNN, Decision Tree, Random Forest, Naïve Bayes and AdaBoost.

The overall objective will be to compare the results achieved in both datasets to understand which features are more likely to develop better results, therefore

helping future research or industry on real-word implementation with a better understanding on what kind of dataset to use for their purpose.

B. Related work

There have been several studies that have explained the importance of Human resources management (HRM) and its relationship and positive impact to working scenarios [1] showing in fact economic effects in intensity and business's capital growth.

A huge debate is still ongoing related to the causes and effects of attrition and its recent developments as today attrition have been classified in different kinds:

1. Voluntary (employee leaves by their own will)
2. Involuntary (employee leaves due to negative forces)
3. Compulsory (employee leaves due to rules and regulation)
4. Natural (employee leaves due to natural causes)

Previous research [2] have investigated different cause for an employee leaving the workplace such as salary, promotion, transfers, workplace infrastructure, tasks, flexibility, and job security to cite a few.

The nature itself of the problem categorize the research question as a standard binary classification problem with a clear identifiable variable (dwill the employee leave? Yes or no).

With no surprise a conspicuous group of researchers in recent years have been applying machine learning models to predict attrition rates within organization on various datasets.

Several studies have been published showing promising results but also challenging aspects: in [3] researchers have developed a XGBoost algorithm, the study found that the machine learning model was able to accurately predict employee attrition, with an average accuracy of 89%; however, the lack of more in dept metrics such as F-scores or recall ratio might not have shown the true impact of the model since class imbalance issued and a small dataset size.

In [4] researchers have achieved great results in terms of F1 score and accuracy on a particular kind of dataset by using different machine learning techniques; their approach differentiated between departments and due to the high results achieved this project will avail of the same dataset and expand by employing more machine learning techniques without using the department specification.

Authors in [5] compared decision tree algorithm and naïve bayes and predicted the likelihood of an employee leaving their current workplace using tenfold cross-validation and a split ratio of 70-30. The results showed an accuracy of around 80% for both and F1 scores up to 50%.

A paper by Alduayj and Rajpoot [6] carried out an interesting experiment: while classifying for attrition prediction the authors experimented with sampling techniques and feature selection showing how oversampling approach performed better than under sampling reaching F1 scores of 90%: the oversampling technique used is called ADASYN.

This sampling method solves the class imbalance by creating synthetic instances based on the distribution of the class with a smaller number of samples.

A huge contribution came from [7] where authors developed a reliable approach comparing multiple ML techniques on several dataset of different sizes.

Their researcher advised some algorithms over others regarding to dataset size and explained effects on preprocessing tools such as scaling and feature selection on turnover prediction which will prove extremely useful.

Another important contribution came in [8]; the authors used decision tree classifiers such as: CHAID, CART and C4.5.

Their paper highlighted how among the various features taken in consideration salary and length of service were the prominent ones able to being used successfully in turnover prediction.

One of the limitations raised by the authors were the reluctance of organization to provide their data to external researchers, this led to a smaller dataset than expected: organization who expects to exploit the benefit of predictive tools should be ready to provide data and support.

In [9] researchers developed an approach on attrition prediction on a limited dataset from IBM (1500 employees in total) employing both decision tree and other classifiers such as Naïve Bayes, the authors emphasize the importance of turnover retention and the value added from attrition prediction techniques.

Limitations of such approach were explained in the conclusion stating how the influence of the variables could be assessed across educational background, gender and skill level which implies that results can only be generalized to the specific context the data was taken, in this case, India sales industry; but still papers like this might help enterprise lower their turnover costs by implementing better retention strategy.

Lastly, in a recent paper [10] authors were able to reach an overall accuracy of 96% with Naïve Bayes reaching a high true positive rate (TPR) of more than 80%; the paper employed the ROC curve as reliable measure performance achieving an area under the ROC curve of about 90% supporting findings from previous studies.

C. Data Understanding

The datasets have been downloaded from Kaggle website.

Both data sources consist of synthetic data specifically created to try and predict employee turnover.

Both datasets consist of more than 10000 entries and more than 10 features.

Dataset 1 has a high granularity within the data, it looks like has created as the result of questionnaire among the employees, it consists of 10 features:

Field Name	Type	Description
Satisfaction level	numeric	score from 0 to 1; with 1 being the highest satisfaction level possible
Last evaluation	numeric	score from 0 to 1; no info were revealed on the actual meaning of this field, clarance have been asked to the owner but still not answer
Number project	numeric	From 1 to 7, total number of project employee has been involved
Average montly hours	numeric	employee average monthly hours
Time spend company	numeric	from 1 to 10; period spent in the current company by the employee
Work accident	categorical	0 or 1; where 0 rapresents no work accident
Left	categorical	0 or 1; where 1 rapresent if the employee left the company
Promotion last 5years	categorical	0 or 1; where 1 rapresent if the employee got a promotion in the last 5 years
Department	categorical	text data; name of the employee's department
Salary	categorical	low, medium or high; salary range of the employee

Figure 1. Dataset 1 Features

After loading the data Frequency Distribution and trend analysis have been drawn for the first dataset.

As clear in Figure 3 it is possible to see clear clusters, for example employees with low satisfaction levels and high monthly hours are more likely to leave the organization.

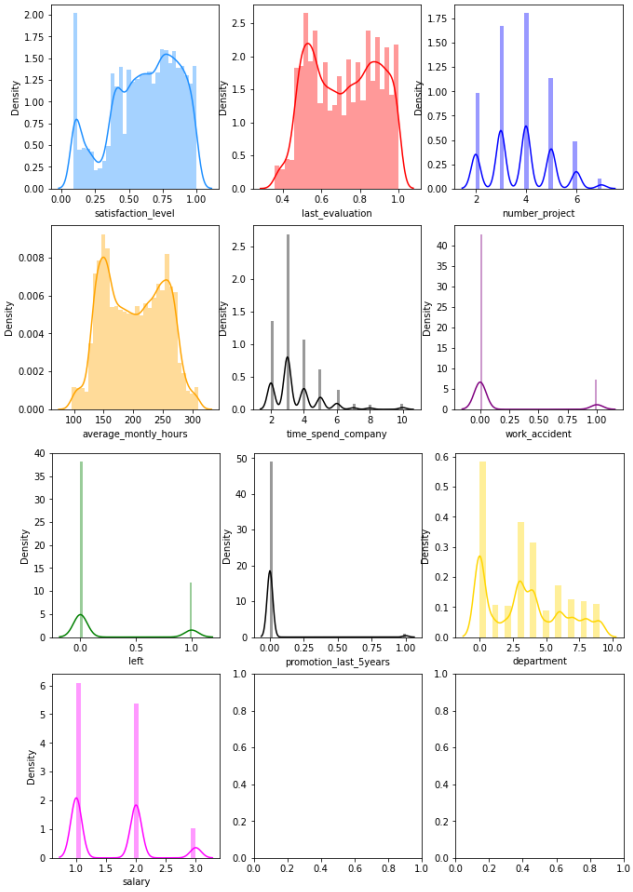

Figure 2. Dataset 2 frequency distribution

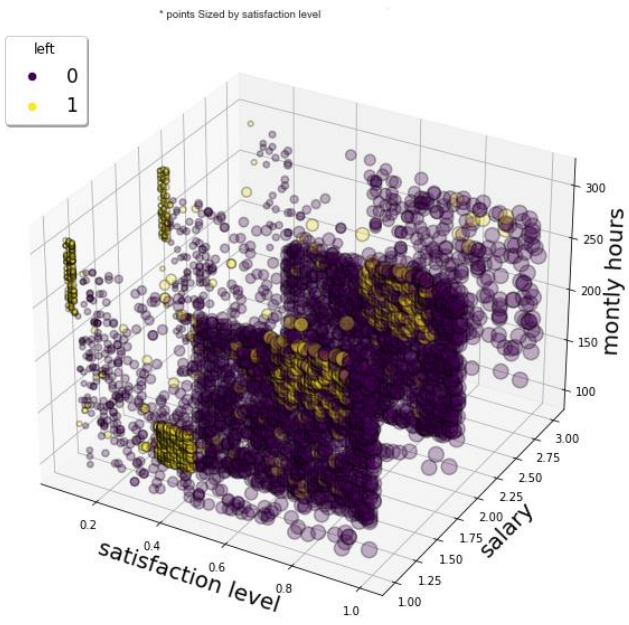

Figure 3. 3D plot "satisfaction level", "salary", "monthly hours"

Differently from the first Dataset, the second contains mostly categorical data, which are intuitively

easier to collect (department name, store, age) from a HR perspective.

Row data contains eighteen features, however only eight are useful to the intent of the project.

Field Name	Type	Description
age	numeric	employee age
Length of service	numeric	How long the employee was employed within the organization
City name	categorical	text, City of employment
Department name	categorical	text, Department of employment
Job title	categorical	text, Job title
Store name	categorical	text, Store of employment
Gender	categorical	text, employee Gender
Status	categorical	Active or Inactive; employment Status
Business Unit	categorical	Store or Headoffice

Figure 4.Dataset 2 features

Dataset 2 however does not contain such well-defined clusters, even though some trends can be drawn: for example, age impacts turnover in two different ways, the turnover decreases as age increases after 20 years old until it increases vertiginously from 60 years old going forward due to retirement as shown in fig.5.

Figure 5.Turnover on age, Dataset 2

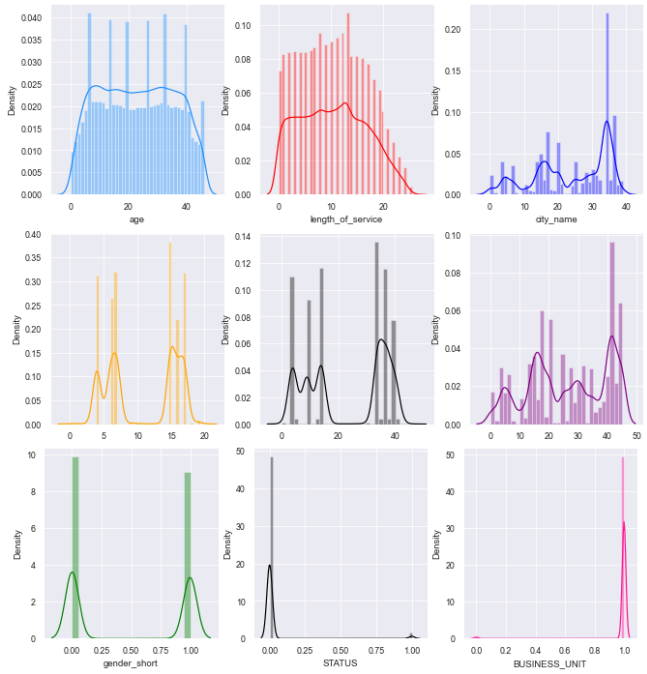

Figure 6. Dataset 2 Frequency distribution

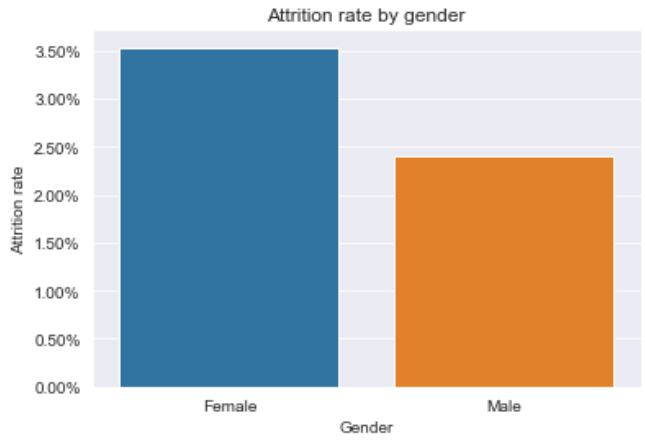

Figure 6. Attrition rate by gender, Dataset 2

D. Data cleaning and data transformation

For both datasets data cleaning tools have been used such as: general cleaning and renaming of wrong data.

Data sources have been checked for null values; both datasets seem to have been previously cleaned before being loaded on Kaggle and no null values have been found.

Even if most of the data consisted in categorical data; outliers have been checked and none could be found in the few numeric features available.

Various encoders have been tested throughout the project, both for plotting and modelling.

E. Machine Learning Methodology

As presented, the overall question remains: is it possible to predict employee's attrition? If so, which attributes are the most valuables in terms of turnover prediction.

The question presented itself as a binary classification problem.

As stated, the two databases do not contain real-word data, the overall approach is looking to realize strong results by modelling, which in the future might be used for real-word implementation with suggested features for analysis.

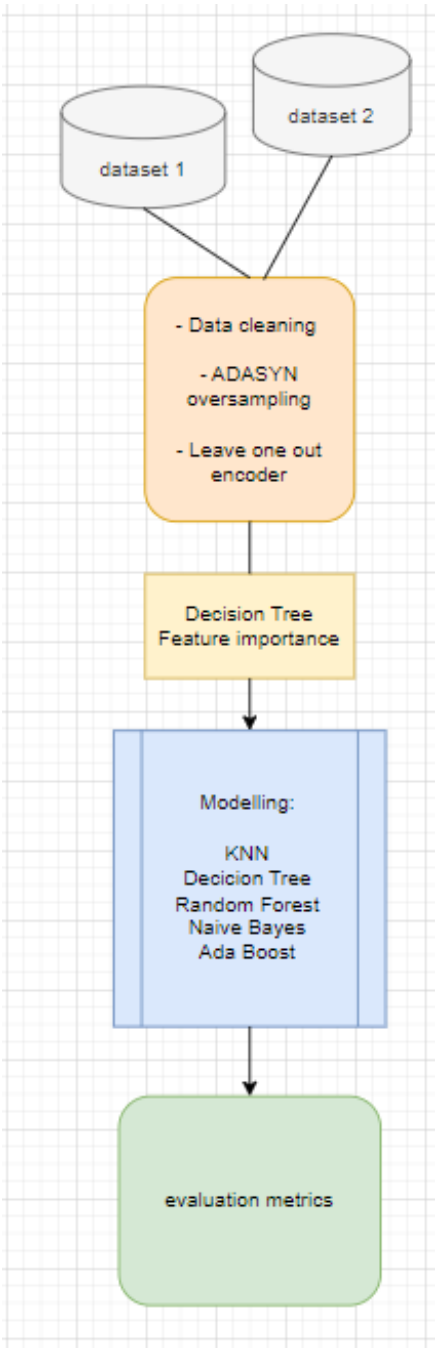

Figure 7. Overall classification methodology

In terms of modelling as shown by Zhao [7] tree-based ensembles seem to perform best: "For medium and large HR datasets, the data variance decreases, and a more reliable model may be built. Best practice would be using tree-based ensemble methods such as extreme gradient boosting and gradient boosted trees. Extreme gradient boosting is preferred due to its superior predictive power and speed. This approach requires the least data preparation—it does not need data scaling and type conversions—and is likely to result in decent, if not the best performance".

Due to relevant literature the models taken in consideration for this project will be relevant tree-based algorithms [7] [3] [8] (AdaBoost, Random Forest, Decision Tree) and other algorithms which performed good in previous literature [6] (KNN, Naïve Bayes).

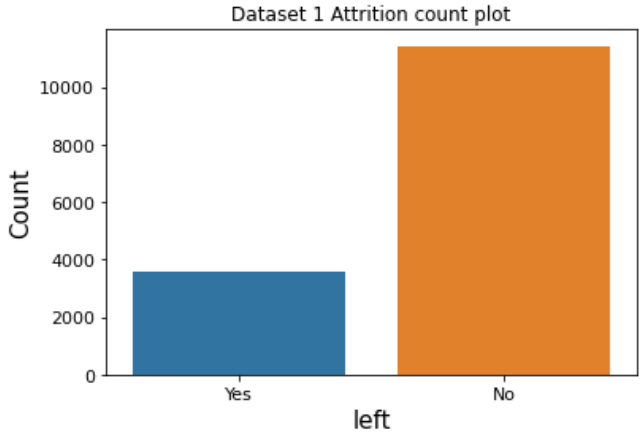

Figure 8. Dataset1 Countplot

As expected, data from both datasets shown, to different degrees, a class imbalance problem, since the number of employees leaving the company is notably less than the number of workers remaining within the company which is especially noticeable in the second dataset.

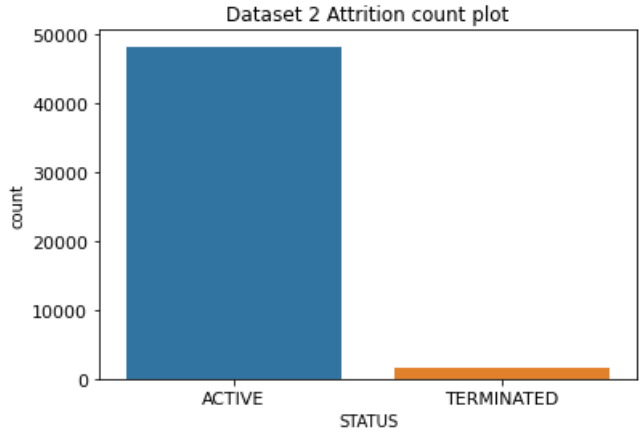

Figure 9. Dataset 2 Countplot

Luckily this problem was previously overcome by Alduayj and Rajpoot in [6] by employing ADASYN sampling algorithm "The ADASYN algorithm solves the class imbalance problem by creating new synthetic

instances based on the density distribution of the minority class. ADASYN will accomplish this by using adaptive learning to change the weights for the minority class instances. As a result, it will shift the decision boundary, which will make it easier to learn from difficult instances." The same over sampling technique has been employed in this project as well.

In:

```
print("Before undersampling: ", Counter(y))
from imblearn.over_sampling import ADASYN
ada = ADASYN(random_state=42)
X_res, y_res = ada.fit_resample(x, y)
print("After undersampling: ", Counter(y_res))
```

Out:

```
Before undersampling:  Counter({0: 11428, 1: 3571})
After undersampling:  Counter({0: 11428, 1: 11418})
```

Figure 10. ADASYN sampling code

To proceed with feature selection encoding has to be completed; due to the enormous number of categories within categorical data simple dummy features couldn't have been selected without running into memory error during the modelling phase.

To avoid memory issues, Leave One Out encoding method was utilized, which involves calculating the average value of the target variable for all data points that share the same value for the categorical feature being considered.

This encoding technique differs between the training and test datasets, with the record under examination being excluded from the calculation for the training dataset.

Utilizing this type of encoding also helps our model to avoid overfitting.

For feature selection Decision Tree classifier feature selection has been used, this allowed to select the most important features per each dataset by calculating the importance scores of each feature.

Features with scores greater than 0.1 have been selected for each dataset.

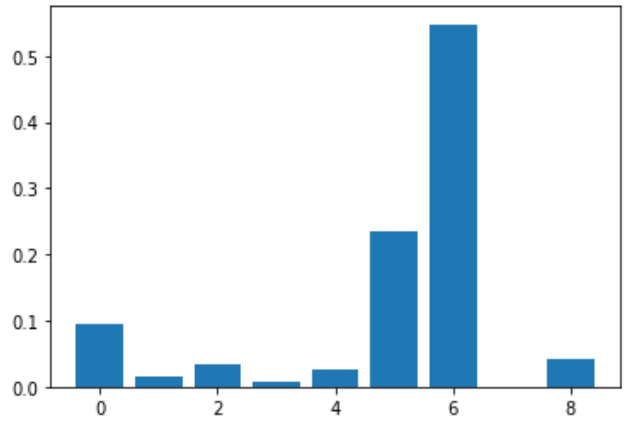

Figure 11. Dataset 1 Feature selection

For the first dataset the feature selected were:

- Promotion last 5 years (0.54)
- Work accident (0.24)
- Satisfaction level (0.10)

Which as shown in the following figure do not contain multicollinearity.

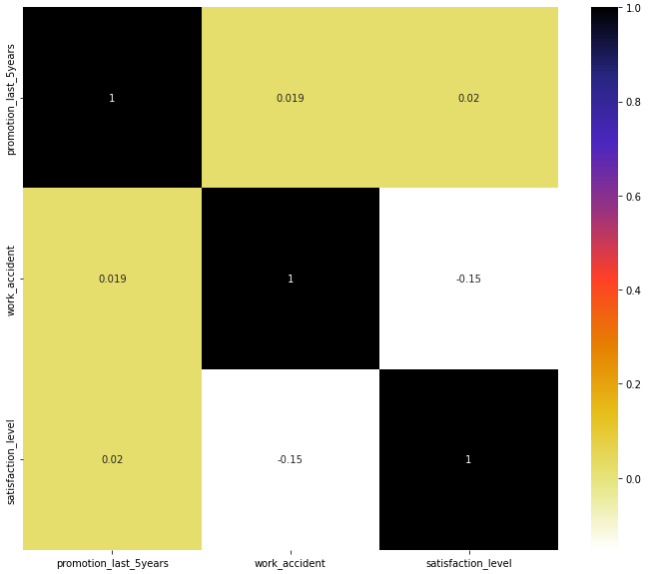

Figure 12. Correlation matric Dataset 1

For the second dataset the selected features were selected:

- Age (0.15)
- Gender (0.12)
- Business unit (0.65)

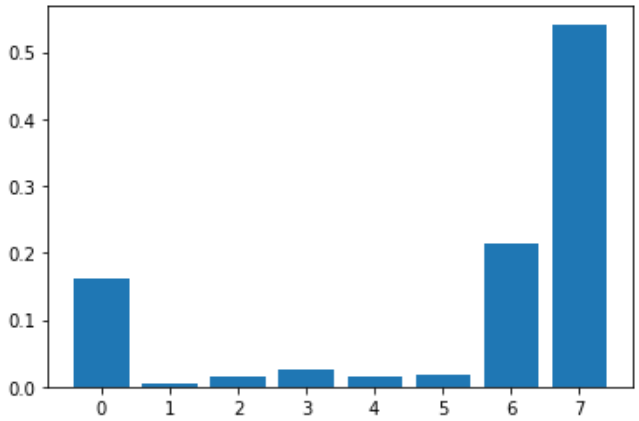

Figure 13. Dataset 2 Feature selection

Again, by looking at the correlation heatmap the selected features do not contain multicollinearity.

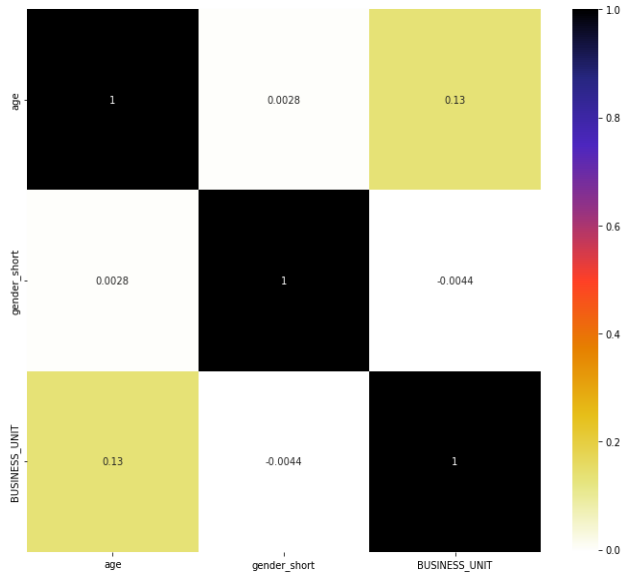

Figure 14. Dataset 2 Correlation matrix

After having selected the most important features, encoded the data and solved class imbalance the modelling phase started with the following results.

F. Evaluation

In terms of evaluation metrics have been selected in accordance to previous literature, one important aspect of turnover retention would be to minimize the impact of false negatives; as reported in [5] "Recall was identified as the most important performance metric to ensure the minimum number of false negatives (employees who may potentially leave the company but are not classified as such) to a lack of precision resulted in greater numbers of false positives (employees who do not meet the conditions for potentially leaving but are classified as such)."

Obviously, this project has used accuracy scores as measure for first approaches and overall wellbeing of the models but it's important to notice that, As reported in [4] "The only accuracy can yield misleading results if the dataset is unbalanced and hence can be unreliable. A classification report would depict the precision, recall, and F1-Score for the model."

So the chosen metrics for evaluate the models were: accuracy, recall, precision and F1 score and Roc Auc score alongside with training and test confusion matrix.

Dataset 1	0	1
0	3462	0
1		1038

Dataset 2	0	1
0	14391	0
1	0	505

Figure 15. AdABoost Test confusion matrix

As expected tree-based and ensemble learnings algorithms (Decision Tree, AdaBoost, Random Forest) performed best among the others.

Dataset 1	0	1
0	3462	0
1		1038

Dataset 2	0	1
0	14391	0
1	0	505

Figure 16. Decision Tree Test confusion Matrix

Dataset 1	0	1
0	3462	0
1		1038

Dataset 2	0	1
0	14391	0
1	0	505

Figure 17. Random Forest Test confusion Matrix

KNN also created excellent results while Naïve Bayes remains the list in terms of performances.

Dataset 1	0	1
0	3457	5
1	4	1034

Dataset 2	0	1
0	14390	1
1	5	500

Figure 18. KNN Test Confusion Matrix

Dataset 1	0	1
0	2632	830
1	503	535

Dataset 2	0	1
0	151	14240
1	23	482

Figure 19.Naive Bayes Test confusion matrix

The modelling has achieved marvelous results with four models achieving almost or 100% f1 score on both datasets.

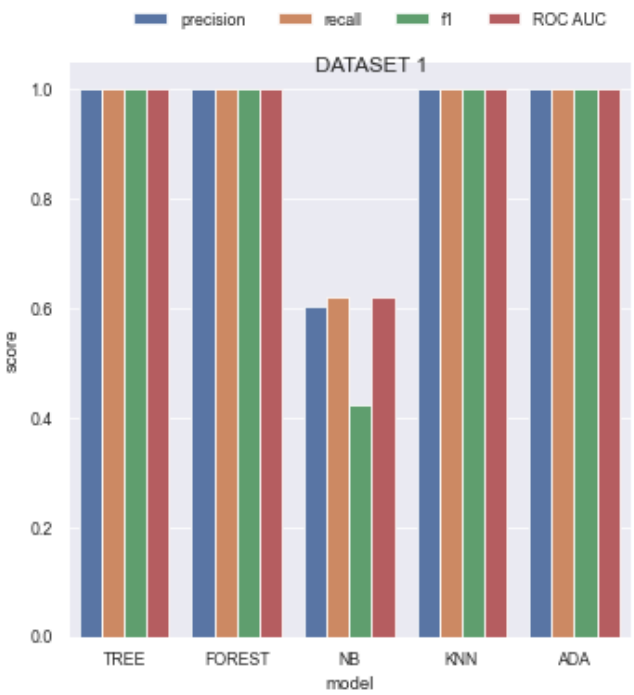

Figure 20. Dataset 1 metrics

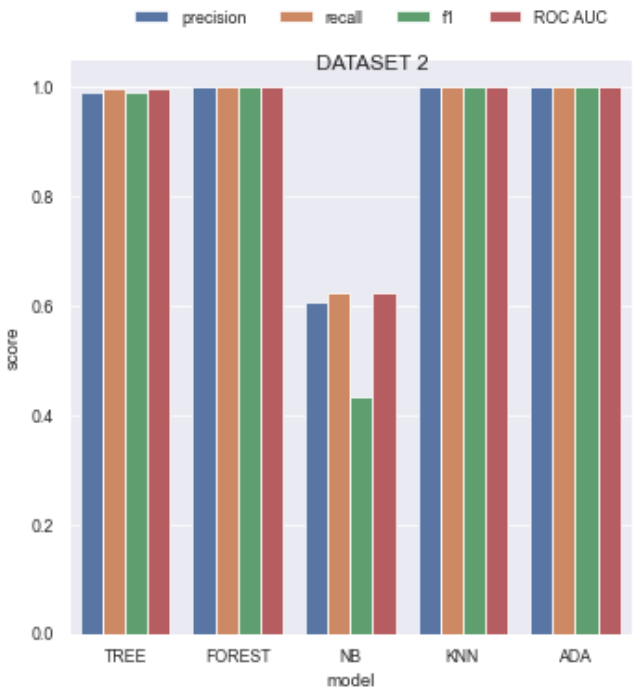

Figure 21. Dataset 2 metrics

G. Conclusion

The results provided were impressive, the processes taken were in line with previous literature [6][7][8].

Both datasets achieved results up to 100% scores in F1 and ROC-AUC score; clearly the impressive results are explained since the data used was not real-word data.

In fact, the greatest limitation of this project have been the lack of large real datasets upon which testing the findings which are working so well in a synthetic environment.

The features selected for modelling were "Promotion last 5 years, Work accident, Satisfaction level" for dataset 1 and "Age, Gender, Business unit" for the Dataset 2.

The overall results describe two different paths for retention of employees which organization could adapt; on one side the features which defined the turnover rate on the first Dataset (Promotion last 5 years, Work accident, Satisfaction level) can be summarized under an umbrella term such as "work condition".

If an organization manages to employ better working condition, minimizing work accident, maximizing satisfaction levels this might lead to less attrition and different retention rates.

Ah shown in the second datasets however, age and type of work (with different seniority) imply different retention rates; the data suggested how females are more likely to leave the organization and that managerial roles have less attrition than others.

Age impacts attrition rates in two different ways: turnover rates are higher among young people and steadily decrease with time; up until retirement age where almost all the employees leave the organization.

H. Future work

As presented the machine learning tools work with high efficiency; next step would be to adapt the concept shown in this project to real-word data.

In terms of algorithms to use KNN, AdaBoost, Decision Tree and Random Forest have achieved great results and can be put to the test on real data.

One of the challenges for implementing this approach would be data collection which can be overcome in two ways.

Data can be collected in a questionnaire form among the employees and then registering if an employee had left the company, this would act as base for predicting future leavers.

Alternatively, through organization CRM system data could be collected and creating a dataset like the second in use in this project.

Overall, the results in this project reveal how a mature machine learning techniques are in predicting HR attrition and how tools are ready for real-word implementation.

REFERENCES

[1] N. Bloom, J. Reenen, "Chapter 19 - Human Resource Management and Productivity",in Handbook of Labor Economics, Elsevier, Volume 4, Part B, 2011, Pages 1697-1767.

[2] Negi, Gayatri. "Employee attrition: Inevitable yet manageable." In International Monthly Refereed Journal of Research In Management & Technology 2.1 (2013).

[3] R. Jain, A. Nayyar, "Predicting Employee Attrition using XGBoost Machine Learning Approach," 2018 International Conference on System Modeling & Advancement in Research Trends (SMART), 2018, pp. 113-120.

[4] Jain, P.K., Jain, M. & Pamula, R. "Explaining and predicting employees' attrition: a machine learning approach." SN Appl. Sci. 2, 757 (2020).

[5] F. Fallucchi, M. Coladangelo, R. Giuliano, W. De Luca E. "Predicting Employee Attrition Using Machine Learning Techniques." Computers. 2020.

[6] S. S. Alduayj and K. Rajpoot, "Predicting Employee Attrition using Machine Learning," *2018 International Conference on Innovations in Information Technology (IIT)*, 2018, pp. 93-98.

[7] Yue Zhao , Maciej K. Hryniewicki, Francesca Cheng, Boyang Fu, and Xiaoyu Zhu, "Employee Turnover Prediction with Machine Learning: A Reliable Approach", Department of Computer Science, University of Toronto, Toronto, Canada

[8] D. Alao, A. Adeyemo "ANALYZING EMPLOYEE ATTRITION USING DECISION TREE ALGORITHMS," Department of Computer Science University of Ibadan Ibadan

[9] R. Chakraborty, K. Mridha, R. N. Shaw and A. Ghosh, "Study and Prediction Analysis of the Employee Turnover using Machine Learning Approaches," 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON), 2021, pp. 1-6.

[10] LearningXiaojun Ma , Shengjun Zhai , Yingxian Fu , Leonard Yoonjae Lee , Jingxuan Shen "Predicting the Occurrence and Causes of Employee Turnover with Machine", ComEngApp J