A Machine Learning Approach for Employee Retention Prediction.

Ggaliwango Marvin, Majwega Jackson, Md. Golam Rabiul Alam
Department of Computer Science and Engineering.
Brac University
Dhaka, Bangladesh
Email: ggaliwango.marvin@g.bracu.ac.bd, rabiul.alam@bracu.ac.bd, majwega.jackson@g.bracu.ac.bd

2021 IEEE Region 10 Symposium (TENSYMP) | 978-1-6654-0026-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/TENSYMP52854.2021.9550921

***Abstract*—Massive investment in employee skills training has been adopted by lots of organizations in reaction to the rapid evolution of the global trends and technology adoption. Unfortunately, target employee retention after training unsatisfactorily gives a negative return on investment. Prediction of target candidate decision before training and understanding the features that affect the candidate decision can greatly contribute to candidate selection and decision feature optimization process for increased employee retention. The method proposed in this paper successfully models and analyses various machine learning classifiers for illustrating features that affect the target candidate decision and predict the probability of candidate retention before training. Classical metrics are used to express the results of the algorithms used and the Random Forest Classifier revealed the finest percentage in accuracy summarized as 99.1%, 84.6%, 91.8% on the training, testing and overall dataset respectively.**

Keywords— Artificial Intelligence, Human Resource Management, predictive decision making, Machine Learning, Classification Algorithms, Employee Retention Prediction, Talent Management.

I. Introduction

A. Background and Motivation

Strategic investments based on intelligent decisions are necessary for company development. Decision-making with the assistance of machine learning developed algorithms is increasingly determining our lives [6]. There are numerous areas in which the artificial intelligence adoption impacts decision-making activities in a company or organization [26, 33]. HR management is one of those where the quality of employee skills greatly influences the growth and competitiveness of companies [32]. Being adopted in marketing and sales, artificial intelligence is now applied in HR management employee decisions based on objective analysis of data in large amounts [15, 28, 29] to advance the decision-making processes, achieve set corporate objectives and improve competitiveness. [19, 10].

Since skills, knowledge and employee loyalty are fundamental for business continuity and sustainability,artificial intelligence in the HR field can utilize employee data through models that allow predictions of employee decisions, hence minimizing risks and optimizing HR activities [24, 8]. Organizations invest much resources in employee recruitment and training based on their strategic needs. [30] so, when an employee leaves the company, it loses valuable resources in terms of labor, time, money and effort that were invested in selecting and training the employee for their related tasks. Therefore, the organization must endlessly invest in selection, training, recruiting and developing new staff to fill important job positions which is a long and costly process. Companies are interested in retaining well-trained and highly motivated employees since they form the core of a productive company [23].

In today's rapidly evolving global trends, the most successful companies are those that are adapting quickly and effectively. These global trends require competent and skilled personnel who need to be selected analytically [2]. Moreover [1] predicted that 85% of the jobs to exist by 2030 have not been invented yet. This nerve-wracking uncertainty involuntarily activated lots of companies and organizations to increase the investment set aside for employee training including skills development, up skilling, compliance training and lifelong learning in order to increase employee retention, improve productivity, boost competitiveness and keep employees up-to-date. Unfortunately, despite the massive investment in employee training, Employees decide to leave the company or turn down the job offers after training.

B. Our Contributions.

In this paper, we model classifiers that illustrate the features that affect the candidates' decision and predict the probability of candidate retention for the HR department to take timely appropriate measures before selecting the candidate for training. We test the classifiers on a real dataset provided by HR analytics [18] with 14 features and about 19159 samples. By analyzing the correlations in the heat map of 14 features, we derive the features that have high correlations related to decision making for the target employees. Classical metrics are used to express the results of the algorithms used and Random Forest Classifier revealed the finest percentage in accuracy summarized as 99.1%, 84.6%, 91.8% on the training, testing and overall dataset respectively. The outcomes of the data analysis prove that the implementation of machine learning systems that use the Random Forest Classification algorithm can support the HR departments in strategic selection and

training of employees based on their retention probabilities and factors affecting their job acceptance decisions. The summary of the contributions is stated below.

1) The proposed Models predict the probability of candidate retention before the candidate is selected to be part of any company training program.
2) The Models illustrate the features that affect employee decisions to the company retention.
3) Moreover, we test the various classification models and recommend the best classifier for related Human Resource predictive applications.
4) Finally, we evaluate the performance and accuracy of the models with both synthesized and real datasets obtained for purposes of HR Analytics.

The rest of the paper is organized as follows: Literature review in section II, Problem and Solution Formulation, Data preparation, Data Processing, Training and Sampling in section III, Section IV represents simulated results and performance evaluation of the classifiers. Final Conclusion and recommendations for future work are in section V.

II. Existing Works

A. Predictive decision making

Advanced analytical techniques are becoming popular in resolving complex classification type decision problems in many fields. Various researchers have evidenced the worth of human resource management (HRM) in many settings for example, working situations, production, management and relationships identification with productivity[22, 3]. The impact of HRM on productivity is evidenced to have positive outcomes on business intensity and capital growth [9]. Most studies focus on analyzing and monitoring customers and their behavior [14, 20] but often overlook the employees. Other studies have focused on analyzing employee attrition but none on retention before hiring. An existing research [11] focused on understanding the driving factors and critical inputs to improve talent management and reducing turnover.[12] Even warns that, "Should the voluntary turnover trend continue, more than one in three workers will voluntarily quit by 2023". In [27], researchers only focused on work-specific factors that influence employee turnover. Generally, most of the research focuses on employee attrition and retention for already hired employees. No work has been done about prediction of employee retention before getting hired or trained and that is the research gap we are tackling.

B. Machine Learning Classifiers

Machine Learning Algorithms are of three types namely;

Supervised Learning: They comprise of a target / outcome variable (or dependent variable) to be predicted from a given set of predictors (independent variables). Using these set of variables, a function that maps inputs to desired outputs is created. The training process continues until the model achieves a preferred accuracy level on the training data. Decision Tree, Random Forest, KNN, Logistic Regression among others are some of this type of learning.

Unsupervised Learning: There is no target or outcome variable to predict / estimate in this algorithm. It used to group population in different clusters, which is widely used for segmenting different groups of customers for specific intervention. Apriori algorithm, K-means are the examples of this algorithm.

Reinforcement Learning: Here the machine is trained to make specific decisions. The machine is open to an environment where it trains itself continually using trial and error. For this machine to learn, it will consider past experience and will absorb the best knowledge to make accurate business decisions. Markov Decision Process is one of the examples of this algorithm.

To predict the possibility of an employee leaving the company, authors in [31] made a comparison between a decision tree algorithm J48 Naive Bayes classifier and Naïve Bayes classifier while [21] concentrated on using Machine Learning Techniques to predict Employee attrition. We focused on using Machine learning techniques and analysis of machine learning classification algorithms for retention predictive decision making amongst target employee candidates and the features that affect these retention decisions.

C. Imbalanced Datasets for accurate classification

The other major issue we are handling is classification of imbalanced data since majority of the machine learning algorithms never consider the distribution of the data sample[7, 16] by default. Results are often skewed towards the majority sample class distribution if the model is built using an Imbalanced dataset [18] hence extremely misleading both in practice and theory [7, 16] if the imbalanced data is not handled, therefore we applied Synthetic Minority Oversampling Technique (SMOTE) and Adaptive Synthetic (ADASYN) Sampling Approach to independently handle Imbalanced data.

III. Proposed Method

A. Problem and Solution: Employee retention predictive decision making.

Machine learning is appealing to support to strategic company investments through predictive HR applications. Our proposed ML application is typically limited to objective prediction of decisions and interpretation of features that affect the after training target employee decisions.

We propose ML classification of potential features that affect the retention decisions of employees based on predictive analysis of employee data. |The proposed automated and intelligent decision making is performed through the application of machine learning classification algorithms to implement a probabilistic approach for predicting target candidate decision making with a clear interpretation of all the models that illustrate the features which affect the candidates' decision.

B. Dataset Description and Features.

The dataset[18] used had been designed to recognize the issues that lead a person to leave their current job for HR researchers. But using models with the current qualifications, demographics and experience data we can predict the probability that a target candidate will decide to be retained and interpret factors that affect the candidate's decision. The whole data is divided to train and test. Target is not included in the test' but the test target values data file is available for the related tasks therefore its usable for us to demonstrate our idea. The dataset is imbalanced and most features [Table 1] are categorical (Nominal, Ordinal, Binary), some with high cardinality and possibilities of missing imputation.

TABLE I
DATASET FEATURES.

enrollee_id	Unique ID for candidate
city	City code
city_ development _index	Development index of the city (scaled)
gender	Gender of candidate
relevent_experience	Relevant experience of candidate
enrolled_university	Type of University course enrolled if any
education_level	Education level of candidate
major_discipline	Education major discipline of candidate
experience	Candidate total experience in years
company_size	No of employees in current employer's company
company_type	Type of current employer
lastnewjob	Difference in years between previous job and current job
training_hours	training hours completed
2target	0 – Not looking for job change
	1 – Looking for a job change

Data Exploration.

We first explored the data types in our dataset, the significance of NULL data contained and the frequency of each category visually separated by a clear labels as illustrated in [Fig1].

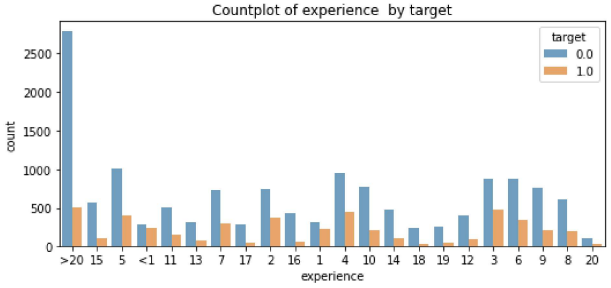

Fig 1. Count plot Histogram.

From the histogram charts plotted, there was no special correlation between the variables with the target function observed to distinguish the value of the target. Furthermore, categorical variables could not determine the correlation factor between these variables and the target functions.

C. Data Cleaning

Correlation cannot be calculated for categorical variables so we removed numerical columns and assigned each categorical variable a value number like [P, Q, P, R] such that the named values are mapped to [4,6,4,8] accordingly. Then we examine the results, considering the correlation between "pseudo categorical variables" and the "target" objective function.

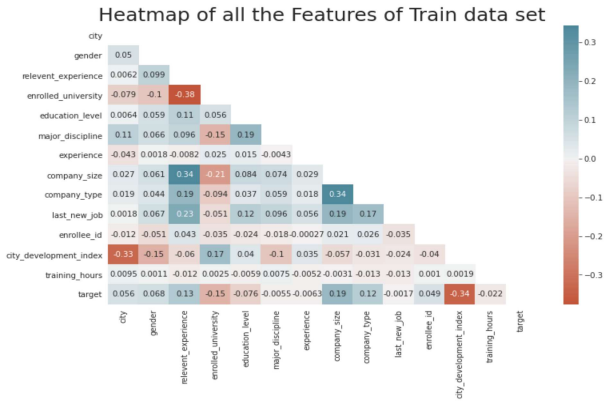

Fig 2. Correlation matrix Heat map of all features in train dataset.

The heat map [Fig2] illustrates the correlations among all variables as Grey fields representing no correlation, while the relative intensity of the red and blue colors represents an increase in correlation, red reveals a positive or direct correlation and blue reveals a negative or indirect correlation.

The results in [Fig2] are not bad, but we need to limit some of variables with low correlation results by setting limits by choosing the correlation above 0.01 or below -0.01 as shown in [Fig3]. The target has a high dependence on the city_development_index.

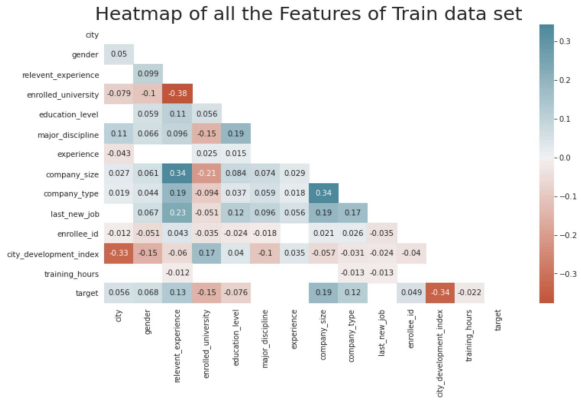

Fig 3. Correlation matrix Heat map of all features after limiting variables with low correlation results.

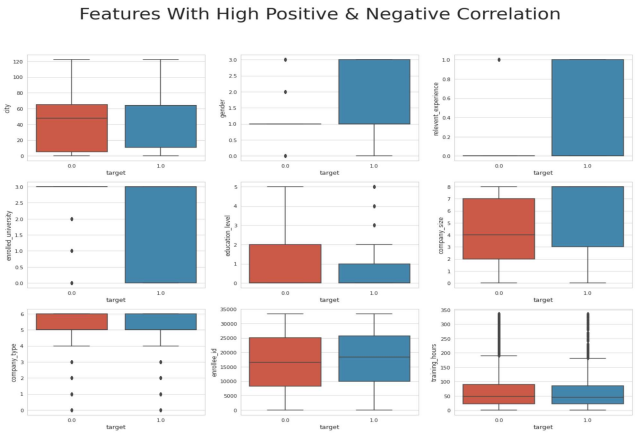

Fig 4. Features with high positive and negative correlation in train data.

Since the correlation between the variables in the train and test data in [Fig5] is nearly the same, good predictive results could be expected if a good train result is obtained.

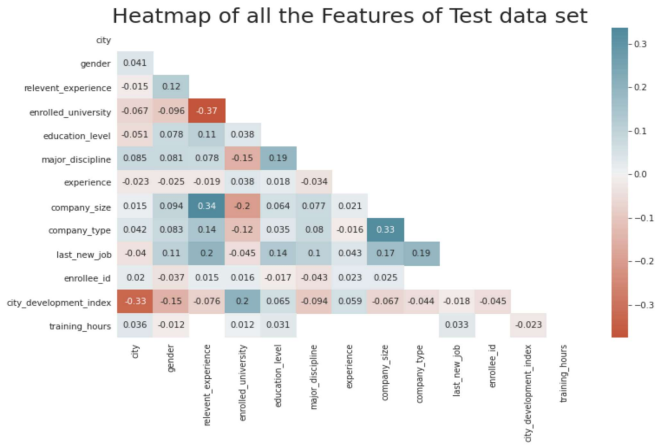

Fig 5. Correlation matrix Heat map of all features in test dataset.

D. Sampling and Training.

On checking for data balance, target 0 had the majority samples and this would affect our model. So we chose good sampling techniques that to enable us obtain sufficient synthetic data for training our Models for optimum results, they included: SMOTE (Synthetic Minority Over-sampling Technique)[4] to help us to create more synthetic data for the minority class 1, Borderline-SMOTE which is a new Over-Sampling Method for Imbalanced Data Sets Learning [25],Borderline-SMOTE SVM to help in identifying misclassified examples on the decision boundary [13] and Adaptive Synthetic Sampling (ADASYN) to generate more synthetic instances in regions of the feature space where the density of minority is low and fewer or none where the density is high [17].

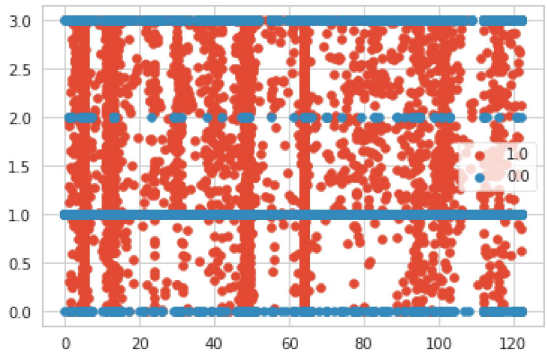

Fig 6. Adaptive Synthetic Sampling (ADASYN) scatterplot

IV. Results and Discussion

A. Classification and Prediction.

Logistic regression: This is a classification technique which is part of the linear classifiers' group and is almost similar to polynomial and linear regression. It is fast and relatively basic, and it's fit for result interpretation. Besides being an essential method for binary classification, it can also be handy for multiclass problems.

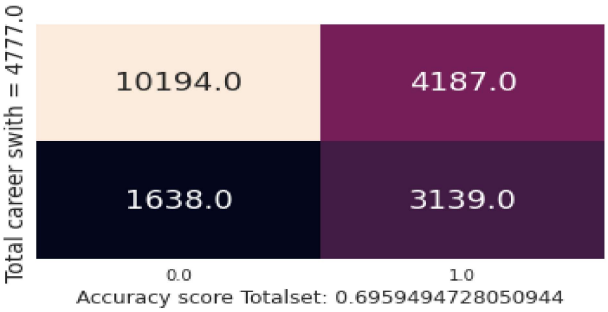

Fig 7. LogisticRegression using SMOTESVM

Naive Bayes (Gaussian Naive Bayes): It is a known classification technique based on Bayes' theorem which assumes independence between predictors, it assumes that the presence of a particular feature in a class does not depend onto the presence of any other feature. These models are easy to build and particularly useful for very large data sets, they are known to outperform even highly sophisticated classification methods, extremely fast, simple and often suitable for very high-dimensional datasets.

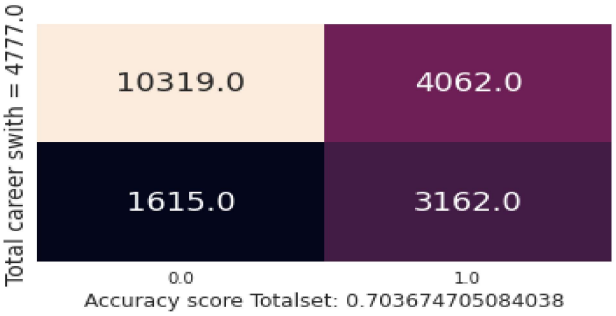

Fig 8. GaussianNB using SMOTESVM

kNN (k- Nearest Neighbors): Although it is more widely used for classification than regression, modest, comprehensible, adaptable and one of the topmost machine learning algorithms, kNN is computationally expensive. Variable normalization is required to prevent bias from high range variables and results if same scale data normalization between 0 and 1 is done. kNN is inappropriate for the large dimensional data otherwise dimension need reduction to improve its performance. Handling missing values can also improve results.

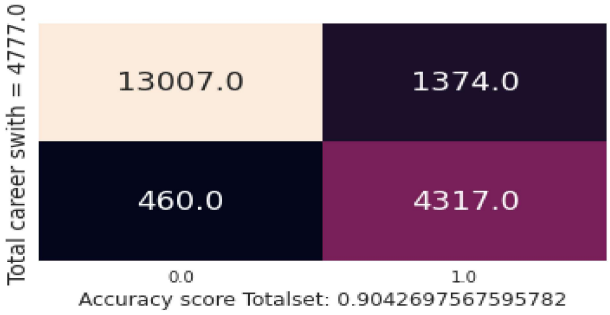

Fig 9. kNN results using BSMOTE.

Decision Tree: It is a white box type of supervised Machine learning algorithm that is mostly used for classification problems that works for both categorical and continuous dependent variables with a training time faster compared to that of neural network algorithms. It is a distribution-free or non-parametric technique which does not depend on probability dissemination assumptions and it can resolve high dimensional data with decent accuracy.

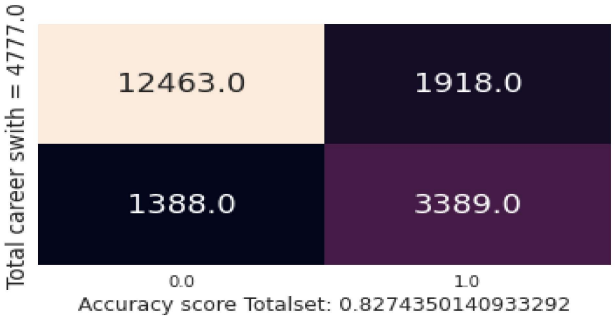

Fig 10. DecisionTreeClassifier using ADASYN

Random Forest Random: Is a supervised machine learning algorithm that ensembles learning where dissimilar types of algorithms or similar algorithm many times are merged to form a more powerful prediction model which can be used for both regression and classification tasks. Although this algorithm is very stable, effective with a mixture of categorical and numerical features efficient when data has missing values or it has not been scaled well, they require much more computational resources, due to the large number of decision trees joined together and their complexity makes them more time consuming to train than other comparable algorithms.

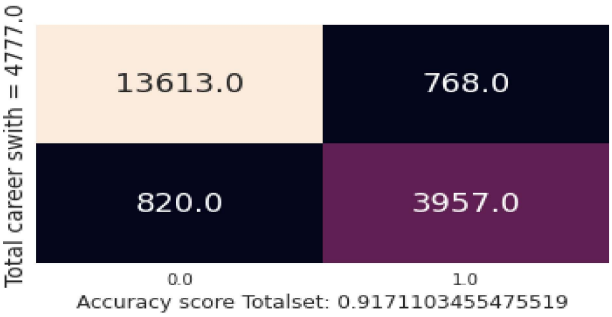

Fig 11. RandomForestClassifier/SMOTESVM

Extra Trees Algorithm: Is another ensemble machine learning algorithm that works by generating numerous of unpruned decision trees from the training dataset to facilitate predictions' making through average the prediction of the decision trees in the case of regression or using majority voting in the case of classification.

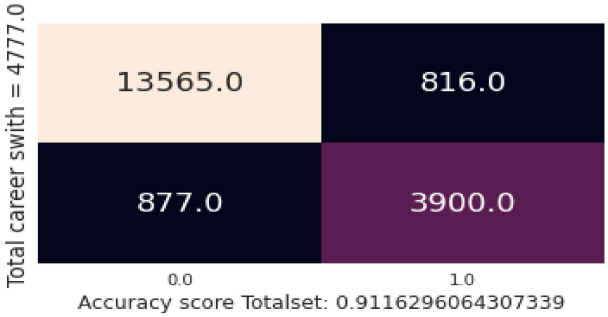

Fig 12. Extratreesclassifier using ADASYN

Gradient Boosting Algorithms GBM: It is an ensemble boosting algorithm for both classification and regression tasks used when we deal with plenty of data to make a high prediction by combining the prediction of several base estimators in to expand robustness over a single estimator. Since we interested in the classifier's accuracy on the validation set, we can assess the classifier by checking its accuracy and creating a confusion matrix then specify the best learning rate based on what we discovered[5].

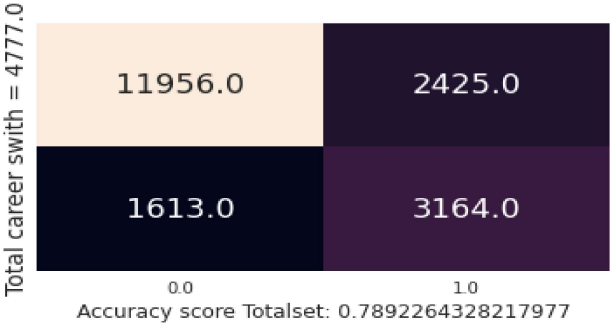

Fig 13. GradientBoostingClassifier using SMOTESVM

XGBoost: It is an algorithm with a massive predictive power making it the best choice for accuracy. This is because it has both linear model and the tree learning algorithm, making the algorithm almost 10x faster than already existing gradient booster techniques. It has become the advanced machine learning algorithm to deal with structured data since it tolerates

several objective functions besides regression, classification and ranking[21].

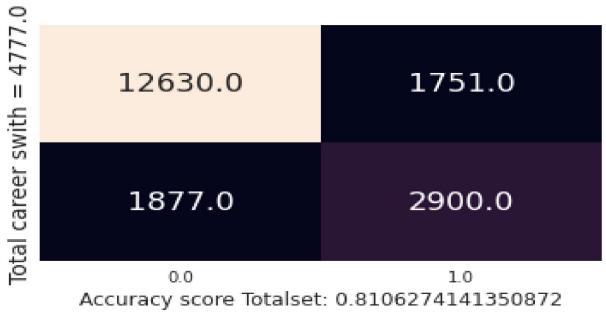

Fig 14. XGBClassifier using BSMOTE

LightGBM: It is a gradient boosting framework that deploys tree based learning algorithms with a quicker training speed and higher efficiency, Lower memory usage, better accuracy, parallel and GPU learning supported, capable of handling large-scale data. It's leaf-wise algorithm decreases extra loss compared to the level-wise algorithm and therefore results are more accurate, this can hardly be achieved by any of the remaining boosting algorithms. [**40-Bisong2019-zw**].

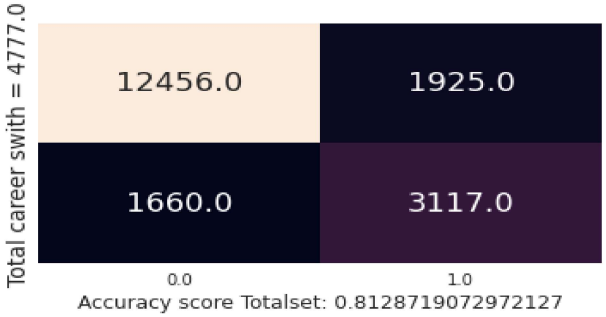

Fig 15. LGBMClassifier using ADASYN

We used the best learning algorithm to predict the true positive and negative rates necessary for plotting the Receiver Operating Characteristic curve for Model accuracy comparison.

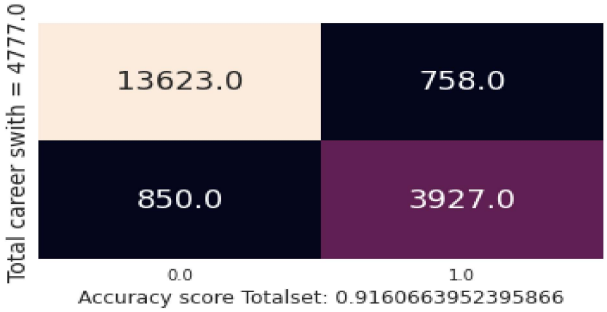

Fig 16. RandomForestClassifier using the SMOTE.

Receiver Operating Characteristic curve to compare accuracy of the predictive the predictive model.

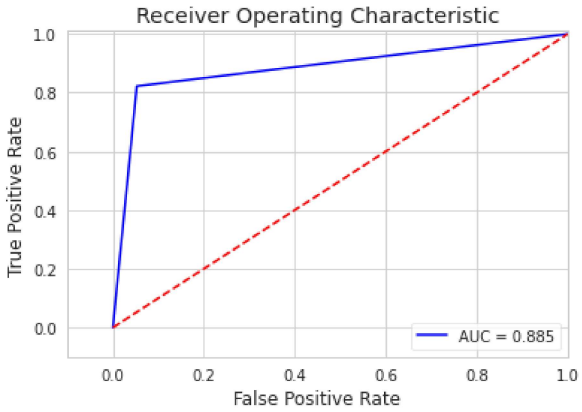

Fig 17. Receiver operating characteristic (roc) curve.

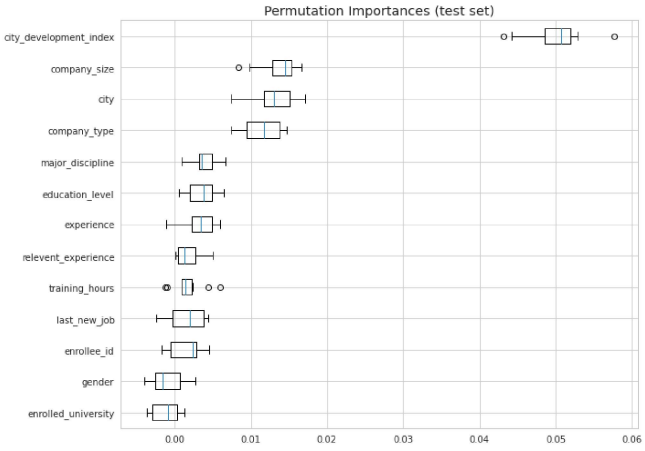

Fig 18. Permutation Importance of the test set.

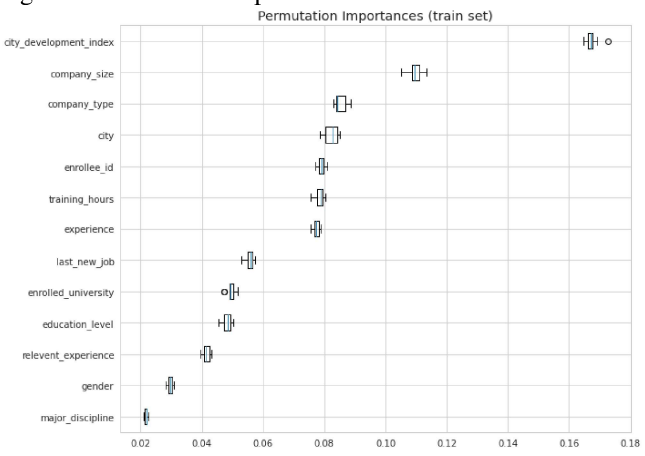

Fig 19. Permutation of importance of the train set.

Then we illustrated the most outstanding features that affect candidate decision making based on what the classifier learnt.

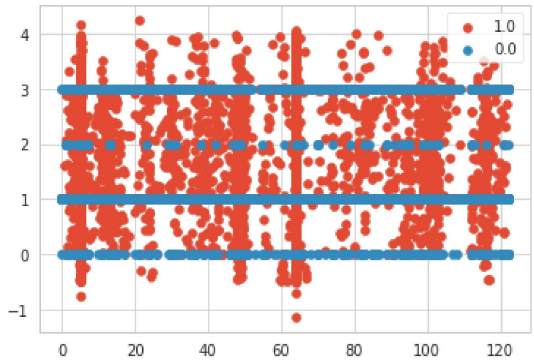

Fig 20. Illustration of features that affect decision making.

TABLE II
ACCURACY RESULTS.

Model	Training dataset	Testing Dataset	Overall
Random Forest Classifier	99.1	84.6	91.8
K Neighbors Classifier	99.1	80.4	90.5
Logistic Regression	72.8	72.1	70.2
Gaussian Naïve Bayes	72.3	71.1	70.9
Decision Tree Classifier	89.5	81.8	83.1
Extra Trees Classifier	99.1	82.5	91.0
Gradient Boosting Classifier	87.5	85.7	80.8
XGB Classifier	84.1	83.7	79.0
LGBM Classifier	87.7	86.1	81.5

V. CONCLUSION

In this work we predicted the probability of target employee retention after training and interpreted ML models that illustrate the features that affect the candidate decision and analyzed their performance in terms of classification accuracy.

The proposed method will be improved by undertaking a survey for categorizing more features that influence candidate decision making before accepting a job offer. Besides, the study of (BOCR) Benefit, Opportunity, Cost and Risk will boost the accuracy of the prediction and increase the robustness of the ML Models used.

REFERENCES

[1] https : / / www . delltechnologies . com / content / dam / delltechnologies / assets / perspectives / 2030 / pdf / SR1940_IFTFforDellTechnologies_Human - Machine_070517_readerhigh-res.pdf. Accessed: 2021-5-20.

[2] M Ayub, M J Kabir, and M G R Alam. "Personnel selection method using Analytic network Process (ANP) and fuzzy concept". In: 2009.

[3] Nicholas Bloom and John Van Reenen. "Human resource management and productivity". In: *Handbook of Labor Economics*. Elsevier, 2011, pp. 1697–1767.

[4] N V Chawla et al. "SMOTE: Synthetic minority oversampling technique". In: *J. Artif. Intell. Res.* 16 (2002), pp. 321–357.

[5] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: (2016). eprint: 1603.02754.

[6] I Cockburn, R Henderson, and S Stern. *The Impact of Artificial Intelligence on Innovation: An exploratory analysis*. 2018.

[7] Richmond Addo Danquah. "Handling Imbalanced data: A case study for binary class problems". In: (2020). eprint: 2010.04326.

[8] *Data Science LandscapeData science landscape : Tracking the Ecosystem*. https : / / www . springerprofessional.de/en/data-science-landscapedata-science - landscape - tracking - the - ecosys / 15501828 ? fulltextView=true. Accessed: 2021-5-23.

[9] Deepak K Datta, James P Guthrie, and Patrick M Wright. "Human resource management and labor productivity: Does industry matter?" en. In: *Acad. Manage. J.* 48.1 (2005), pp. 135–145.

[10] Y Duan, J S Edwards, and Y K Dwivedi. "Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda". In: *Int. J. Inf. Manage* 48 (2019), pp. 63–71.

[11] *Employee retention case study*. en. https://edjanalytics.com/project/employee-retention-case-study/. Accessed: 2021-5-24. July 2018.

[12] *Employee retention report - work institute*. en. https://workinstitute.com/retention-report/.. Accessed: 2021-5-27. May 2020.

[13] Francesca Fallucchi et al. "Predicting employee attrition using machine learning techniques". en. In: *Computers* 9.4 (2020), p. 86.

[14] Niccolò Gordini and Valerio Veglio. "Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry". In: *Ind. Mark. Manag.* 62 (2017), pp. 100–107.

[15] P Gupta, S F Fernandes, and M Jain. "Automation in recruitment: A new frontier". In: *J. Inf. Technol. Teach. Cases* 8.2 (2018), pp. 118–125.

[16] Haibo He and E A Garcia. "Learning from imbalanced data". In: *IEEE Trans. Knowl. Data Eng.* 21.9 (2009), pp. 1263–1284.

[17] Haibo He et al. "ADASYN: Adaptive synthetic sampling approach for imbalanced learning". In: *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. IEEE, 2008.

[18] *Human resources analytics: Think like an engineer. North Charleston, SC: Createspace Independent Publishing Platform:* en. 2017.
[19] M H Jarrahi. "Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making". In: *Bus. Horiz* 61.4 (2018), pp. 577–586.
[20] A Keramati et al. "Improved churn prediction in telecommunication industry using data mining techniques". In: *Appl. Soft Comput.* 24 (2014), pp. 994–1012.
[21] *lightgbm.LGBMClassifier — LightGBM 3.2.1.99 documentation*. https://lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.LGBMClassifier.html. Accessed: 2021-5-26.
[22] M Marchington et al. *Human resource management at work: The definitive guide*. 7th. London, England: Kogan Page, 2020.
[23] L Martin. "How to retain motivated employees in their jobs?" In: *Econ. Ind. Democr* 41.4 (2020), pp. 910–953.
[24] S Mishra, Dev Raghvendra Lama, and Yogesh Pal. "Human resource predictive analytics (HRPA) for HR management in organizations". In: *undefined* (2016).
[25] Hien M Nguyen, Eric W Cooper, and Katsuari Kamei. "Borderline over-sampling for imbalanced data classification". In: *Int. J. Knowl. Eng. Soft Data Paradig.* 3.1 (2011), p. 4.
[26] D Paschek, C T Luminosu, and A Draghici. "Automated business process management – in times of digital transformation using machine learning or artificial intelligence". In: *MATEC Web Conf* 121 (2017), p. 04007.
[27] Evy Rombaut and Marie-Anne Guerry. "Predicting voluntary turnover through human resources database analysis". en. In: *Manag. Res. Rev.* 41.1 (2018), pp. 96–112.
[28] Sajeena. "Recruitment and selection strategies of public sector undertakings". en. In: *Int. J. Res. Granthaalayah* 5.2 (2017), pp. 333–337.
[29] N Syam and A Sharma. "Waiting for a sales renaissance in the fourth industrial revolution: Machine learning and artificial intelligence in sales research and practice". In: *Ind. Mark. Manag* 69 (2018), pp. 135–146.
[30] *There are significant business costs to replacing employees*. https://www.americanprogress.org/issues/economy/reports/2012/11/16/44464/there-are-significant-business-costs-to-replacing-employees/. Accessed: 2021-5-21.
[31] P M Usha. and N V Balaji. *Analysing employee attrition using machine learning*. http://karpagampublications.com/wp-content/uploads/2020/03/Karpagam-Sep-Oct-2019-Article-6.pdf. Accessed: 2021-5-25.
[32] P Vardarlier and C Zafer. "Use of artificial intelligence as business strategy in recruitment process and social perspective". In: Cham: Springer International Publishing, 2020, pp. 355–373.
[33] H Varian. "Artificial intelligence, economics, and industrial organization". In: *NBER Chapters* (2018), pp. 399–419.

LINK TO THE CODE