European Economic Letters
ISSN 2323-5233
Vol 13, Issue 5 (2023)
https://doi.org/10.52783/eel.v13i5.828
http://eelet.org.uk

A Case Study on HR Analytics Employee Attrition Using Predictive Analytics

[1]Pooja Singh, [2]Seema Shokeen, [3]Vandana Raghava, [4]Samiksha Garg
[1]Department of Computer Science, Maharaja Surajmal Institute, New Delhi, India
pooja.asm@gmail.com
[2]Department of Business Administration, Maharaja Surajmal Institute, New Delhi, India
seemashokeen@msijanakpuri.com
[3]Department of Management, Institute of Information Technology and Management,New Delhi, India
vandanaraghava@gmail.com
[4]Department of Computer Science, Maharaja surajmal Institute, New Delhi, India
samikshaaggrawal678@gmail.com

Abstract- In this present contemporary scenario organizations all around the world, employee attrition has been ornamented as one of the climacteric complications within the organization. Attrition denotation is the continual shrinkage in the integral value of staffs or employees uninterruptedly because of retirement, resignation, and death [1]. When an accomplished, consummate, and level-headed employee evacuates the organization because of any sort of causes and rationales, it generates an inoperativeness and desolation in an organization. It fabricates a substantial misfortune for HR staff to thong the void that has transpired. Today's world HR managers are alluring innumerable exertions to roll back the outlay of attrition in the organization and it has been a noteworthy affliction for today's HR managers. Numerous employees are already unfortified for evacuating from the job because of numerous undisclosed integrant like job insecurity, deprivation of career path, propensity to progress towards current openings, supposition of elevated salaries, supervisors not skillful enough to resolve employee's complications, and unnamed other circumstances. This case study helps us to enlightening, like what are the grounds for attrition of employees, moreover to advocates certain solutions in retaining the employees.

Keywords: Machine Learning, Random Forest, Decision Tree, AUC, ROC curve, Employee Attrition.

I. INTRODUCTION

Most businesses utilize fundamental Human Resources (HR) parameters to magistrate the execution of their HR department. Such specifications incorporate administering surveys, disseminating questionnaires, and employing scorecards for recruitment, placement, and prolongation of their employees. These approaches have been the foremost frequently pre-owned techniques even within the pinnacled organizations, we don't nominate them since they have a qualitative constituent inherent in them. That subjective facet is exceptionally fluctuating in nature. For instance, if one employee grants an outcome of 8/10 to his manager today, it's abundantly practicable that the identical employee may dispense a score of 2/10 at the later time in same day. Moreover, there's no predetermined outlook that displays absolutely the identical state of affairs within the upcoming survey phase, the employee will evaluate it as 8/10. Presently stated organizations prerequisites a vigorous algorithm that should be proficient enough of quantifying humongous aggregation HR metrics. It should have no less than the prerequisite necessity that contributes the identical insertion at two non-identical times, the equivalent output value is recorded. This philosophy of ours has been entitled because of the Theory of HR Quantification. The theory grinds in concurrence with multifold parameters. One such parameter which is the talking point of attentiveness for the ongoing learning is HR Analytics. [2]

The machine learning methodology institutes with assemblage of data-set from countless diversified sources and documentations. Eventually, the further juncture is to rectify complications associated with data-set like mislaid merits, standardization of values, so that we will utilize specification pre- processing step. The succeeding step is to envisage the data to discover the perceptions from the data. For such purpose, we will employ distinctive visualization libraries like seaborn, matplotlib, etc. Since the magnitude of real-world information operated for data analysis and modelling is comprehensive, so it enhances the burdensome for the system to compel accurate and meticulous decisions along with pronouncements from the real-world data, so the machine learning algorithms [14] are paradigm in a constructive manner that they accommodate analytical approach, and mathematical outlook such as probability and its distribution. For the exploration and investigation of the data-set and fetch important insights from prior knowledge. The subsequent stage is to test our machine learning classifier representation against our test dataset. To inspect, how skillfully our model performs in opposition to the test data we will implement performance metrics likely as precision, recall, accuracy, F1score etc. Then we have to plump for the pre-eminent machine learning algorithm for the specified data. Machine learning applications in copious disciplines like credit card artifice discernment, unsolicited e-mail, weather forecasting, etc. Our model is prosecuted in Google Collab Python for HR Analytics: Employee Attrition to acknowledge the hidden prototypes in the data. This dataset ventilates the HR Analytics data of Employees of a

European Economic Letters
ISSN 2323-5233
Vol 13, Issue 5 (2023)
https://doi.org/10.52783/eel.v13i5.828
http://eelet.org.uk

designated company to perceive whether the employee will restrain attrition or not. It incorporates one dependent variable and distinctive independent variables. In our research paper, we have done profound and legitimate assessment about the effectiveness of three contrasting classifier ML models like Logistic Regression [3], Random Forest [4] and Decision Tree algorithms [4] to prognosticate attrition of employees.

II. LITERATURE REVIEW

2.1 Talapatra, Pradip Kumar & Rungta, Saket & Anne, Jagadeesh. (2016).

Employee attrition and strategic retention challenges in Indian manufacturing industries: a case study. Vsrd International Journal of Business and Management Research. VI. 251-262. This research paper discusses about the obstacles and defiance's encountered by the organization in retention of their employees. It further states the considerations regarding the exceptions and challenges to retain employees in the contentious marketing. The vindications for employee attrition may be like not acquiring proper reimbursement, not having quality swotting possibilities and occasions in the organization etc.

2.2 Yadav, Sudhir & Joshiya, Vikas. (2021).

Human Resources Practices for Retention for Business Process Outsourcing Industry in National Capital Region‖. International Journal of Trade and Commerce-IIARTC.10.38-50.10.46333/ijtc/10/1/. This research paper discusses about HR implementations that are implemented to diminish the employee attrition or retaining the employees in the organization. HR job has inclined as very resistant to perpetuate their finest employees in present moment's competitive market. HR strategies incorporates work-life stability, admiration for every employee, job reliability, every now and promotions, adaptable operational liberty of working etc.

2.3 B, Senthilnayaki & M, Swetha & D, Nivedha. (2021).

Customer churn prediction. Iarjset. 8. 527-531. 10.17148/iarjset.2021.8692.This research paper is regarding anticipate the customers who are in all probability to churn from the public relations associated company. In this study we become aware of gaining new customers are more high-priced than retaining the subsisting customers. Justifications for customer churn includes: Poor customer facilitations, excessive invoices etc.

2.4 Madaan, Mehul & Kumar, Aniket & Keshri, Chirag & Jain, Rachna & Nagrath, Preeti. (2021).

Loan default prediction using decision trees and random forest: A comparative study. IOP Conference Series: Materials Science and Engineering. 1022. 012042. 10.1088/1757-899X/1022/1/012042.This research paper concerns about recognition of the substandard customers (i.e., did not paid bank loan). Because not long ago the computations of bank defaults have extended that turns out to be enormous monetary mislaying to the banks. So, this research paper grants a solution to demarcate between the defaulter and non-defaulter with the assistance of machine learning algorithms, for instance Random Forest, Decision Tree etc.

III. OBJECTIVES

Problem Statement: There is an organization named as XYZ organization. It engages more than 4000 employees. It has been delineated to the organization that every single year around 20 % of its employees evacuate the organization and because of this they require to recruit pristine talents from the job market. Management conjecture that this attrition proportion (employees quitting on their own or because they have been released) is disagreeable for the company due to the specified considerations: Firstly, the erstwhile associates or employee projects are detained, composing it arduous to light on deadlines, leading to a deprivation of reputation amidst consumers and partners. Secondly, a substantial department must be perpetuated for the contemplation of recruiting upcoming talent. Thirdly, new employees must be instructed more frequently for such placement and/or have time to acculturate in the company. So, management contracted an HR analytics firm to acknowledge what components to focus on to restrict the employee attrition. So, our objective is as follows:

1) To construct classification models to understand and apprehend what elements must be focused on, in order to restraint attrition. i.e., what substitutes should be made at workplace, consecutively to acquire most of employees to stay.

2) To find the foremost excellent classifier model for HR analytics employee data with regards to performance and accomplishment evaluation metrics such as AUC score, F1 score, recall and precision accuracy.

IV. MATERIALS AND METHODS USED

Dataset of employees for HR Analytics case study has been get hold of from Kaggle. In this HR analytics employee attrition dataset, there are comprehensively 4410 samples fractionated into two categories: attrition and non-attrition.

European Economic Letters
ISSN 2323-5233
Vol 13, Issue 5 (2023)
https://doi.org/10.52783/eel.v13i5.828
http://eelet.org.uk

It has 5 different CSV files named as follows: -

- The Manager Survey Data – Accumulated from a company survey database.
- The Employee Survey Data – Gathered from a company survey database.
- In-Time Data – Assembled from company's attendance Records.
- Out –Time Data – Cumulated from company's attendance Records.
- General Data – General data incorporates employee's distinctive data further incorporating education and their satisfaction level for association with XYZ org, etc.

We have scrutinized this Employee attrition dataset with the assistance of python data manipulation tools namely pandas, NumPy, seaborn, matplotlib, and more. [5] Feature engineering portray an extremely crucial role throughout the uninterrupted machine learning lifecycle. Present scenario's actuality datasets are characterized by countless characteristics to fabricate a realistic machine learning [11] model. Conventionally, humongous characters of real-world data are of no convenience to the machine learning models. In Preprocessing course of action, we unfasten the outliers and misplaced values, then to hand-pick salient attributes or features for our dataset. Here we must stipulate abundant of assorted essential variables for our dataset. We can consummate this with Recursive Feature Selection (RFE) algorithm's assistance. After perceiving the principal variables, we will try to construct the model with such variables only.

To examine the model's coherence, we will utilize the performance metrics like accuracy, F1 score, recall, precision, Receiver Operating characteristics (ROC) Curve.

Predominantly there are two distinct classifications of machine learning [18] techniques: One is the supervised i.e., superintended learning technique in which trained data is identical to the labeled data and the second one is an unsupervised learning technique in which we possess unlabeled trained data.

Common supervised techniques are namely Logistic Regression, Linear Regression, Decision Tree [12], Random Forest, etc.

Common un-supervised techniques in particular is K-Means clustering, etc.

For our employee attrition [16] dataset we designate supervised learning techniques into force. To prognosticate whether an employee is proceeding to evacuating the company or not, we have maneuvered distinctive algorithms: Logistic Regression, Random Forest, and Decision Tree. Delineation about these algorithms is along these lines:

4.1 Logistic Regression

Logistic Regression also makes an appearance under the supervised technique that is operated to anticipate a binary outcome, similar to yes or no, deployed from pre-determined scrutiny and considerations of a data set. It is commonly utilized for classification problems. In logistic regression model we observed and prospected the resemblance connecting one or more predictor attributes along with the outcome attribute of presented data for prediction. Formula for Logistic Regression:

$P = 1 / (1 + e^{-(\beta 0 + \beta 1 x)})$

$\beta 0$ is the intercept (it is a constant value) of regression line i.e., we obtain the value of $\beta 0$ when variable x is identical to zero. Here in this regression slope is $\beta 1$.

4.2 Random Forest

Random forest also undergoes the supervised learning method [4]. It can be petitioned on both regression as well as on classification problems. Random forests are assembled from a segment of data & over-fitting complication that has been sorted out by random forest algorithm because the concluding consequence is predicated on the average in occurrence of regression or majority ranking in instance of classification problems. Random forest arbitrarily picks out the observations, substantiate a decision tree and from this we lay hold on the average result.

4.3 Decision Tree

Decision tree also subjected under the supervised learning method. It can be appertained on both regression in addition with the classification problems [4]. It commences with a root node (or source node) and it is terminated with a decision forged by the leaves.

A decision tree carves up the data into numerous sets. Then, every one of these specified sets is supplementarily divides into subsets to transpire a decision.

V. PREDICTIVE MODEL

In our propounded predictive [13] model, to acquire the streamlined out-turns we appertained data pre-processing

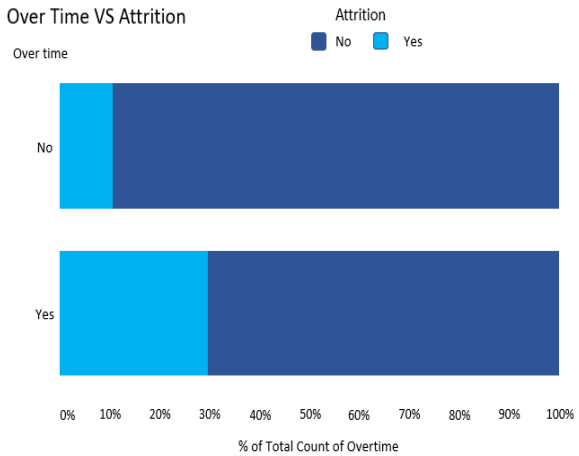
European Economic Letters
ISSN 2323-5233
Vol 13, Issue 5 (2023)
https://doi.org/10.52783/eel.v13i5.828
http://eelet.org.uk

approach on our employee attrition raw data and numerous distinctive attribute-based engineering techniques like one hot encoding etc. [5,6]. In the pre-processing of data, we eliminated the outliers that will influence our model result and the absent value estimation in the data. In our proposed predictive model, we desire for foremost characteristics or variables from the raw data (test data), so for that we implied Recursive Feature Elimination (RFE) [10] algorithm to acquire only the indispensable variables or features from our model's frame of reference and obliterating the added yet less impactful variables or features. Feature engineering on raw data has compelled the model into more efficient one, so it renders a very important role in developing well organized i.e., methodical models.

VI. RESULTS AND DISCUSSION

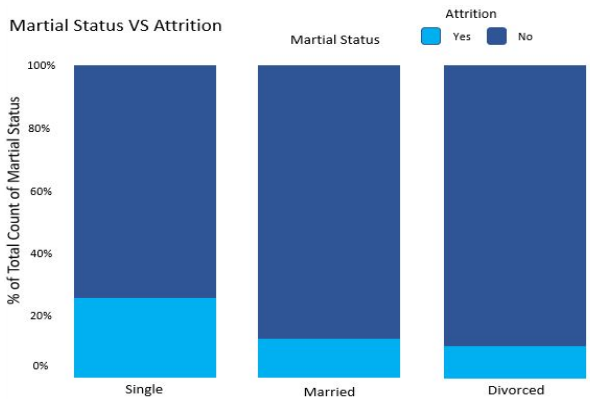
Prior analysis of Employee's HR analytics data can be helpful to fulfill the employee expectation and to prevent them from leaving the company. All techniques of the classification were analyzed in Google Collab Python notebook. First, we have done all the data preprocessing steps and then we move on to data visualization. The dataset has been split into two chunks. The first chunk is called the training data (i.e.,70% of original data) and later chunk is called the test data (i.e.,30 % remaining data of the original data). Firstly, we trained our classifier model on training data (i.e.,70% of original data) and tested it on the test data (i.e.,30 % remaining data of the original data). We established three different models using supervised learning technique to find that either the employee [19] will leave the company or not. For this purpose, we use different classification algorithms [15] like Logistic Regression, Random Forest, and Decision Tree. Here are some of the important insights or visualization that we get from out HR Analytics dataset:

Fig.1. Marital Status



Inference: Employees who are unmarried are prone to leaving the company.

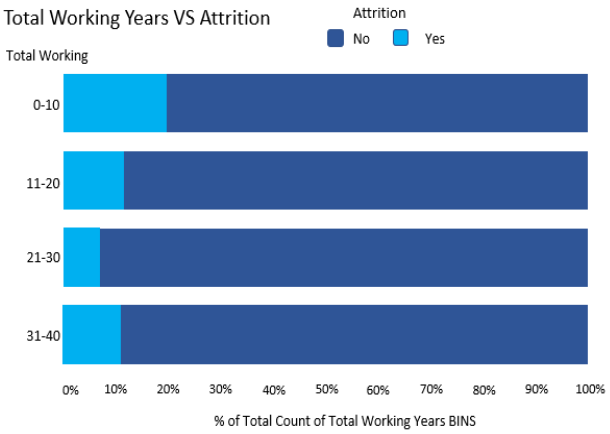
Fig.2. Over Time



Inference: - The more an employee [20] works overtime on an average the more are the chances that he/she will leave the company.

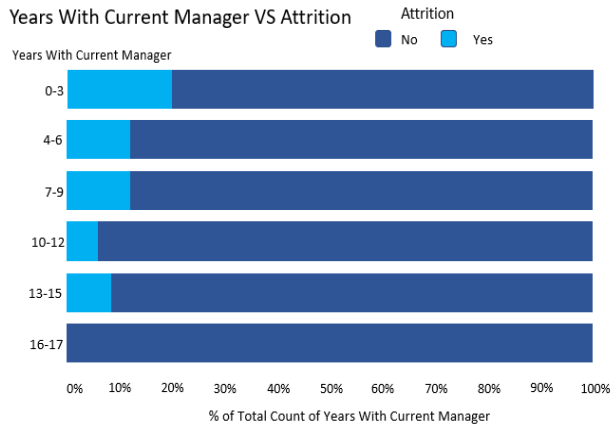
European Economic Letters
ISSN 2323-5233
Vol 13, Issue 5 (2023)
https://doi.org/10.52783/eel.v13i5.828
http://eelet.org.uk

Fig.3. Total Working Years



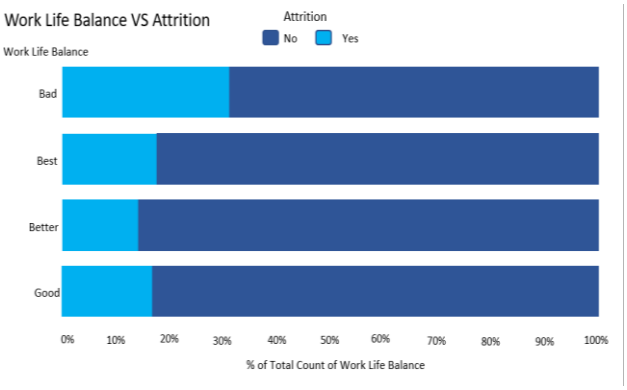
Inference: Hire people with more experience as they are less likely to leave the company.

Fig.4. Years with Current Manager



Inference: If an employee works for a longer duration of time with the same manager, the lesser are the chances that employee will leave the company.

Fig.5. Work Life Balance



Inference: Work life balance, the better these are for employees the less are their chances of leaving the company.

European Economic Letters
ISSN 2323-5233
Vol 13, Issue 5 (2023)
https://doi.org/10.52783/eel.v13i5.828
http://eelet.org.uk

Table 1. Assessment criterion of different model

S.NO.	Predictive Models	Assessment criterion				
		Recall	AUC	F1 score	Precision	Accuracy
1.	Logistic Regression	0.73	0.62	0.66	0.64	0.77
2.	Random Forest	0.90	0.90	0.93	0.97	0.97
3.	Decision Tree	0.95	0.95	0.94	0.94	0.97

For finding employee attrition [17] or not with the help of HR Analytics data, we evaluated the performance of all three different classification models on some criterion such as accuracy, precision, AUC-ROC curve [7], recall and F1 score (Table 1). Accuracy specifies that how much our classifier is correctly able to predict that whether the employee is going to leave the company or not. Precision specifies the classifier's capability of providing true positive forecasts of employee attrition. Recall specifies the classifier's capability of providing the segment of genuine positive cases of employee attrition. F1 score gives a thorough stability among the precision & recall. F1 score value near to 1 of a classifier model is called as the finest model.

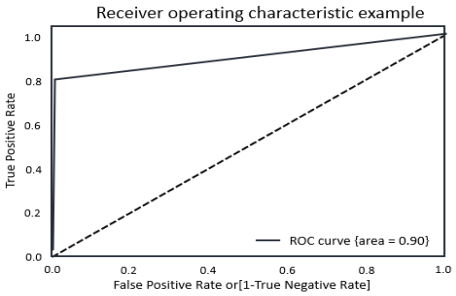


Fig.6. ROC Curve for Logistic Regression (Fig.1)

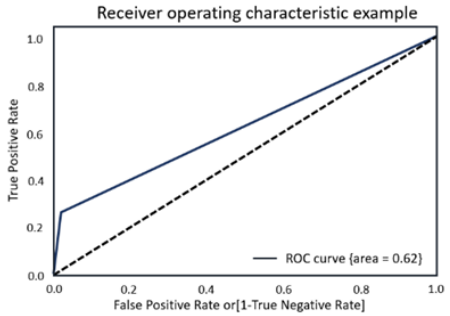


Fig.7. ROC Curve for Random Forest (Fig. 2)

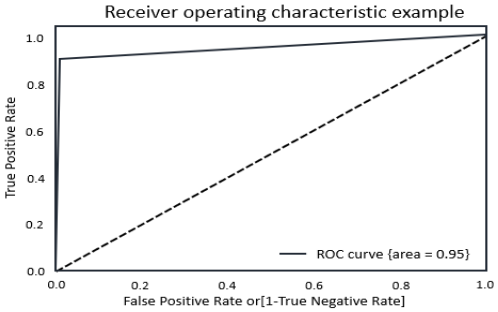


Fig.8. ROC Curve for Decision Tree (Fig. 3)

European Economic Letters
ISSN 2323-5233
Vol 13, Issue 5 (2023)
https://doi.org/10.52783/eel.v13i5.828
http://eelet.org.uk

The ROC curve is a very common method that is widely used to find the performance of our binary classifier model on HR analytics employee dataset. The ROC curve is a plot of sensitivity (or true positive rate) against false positive rate as the threshold for allocating observance are diverse to a specific category. The value of AUC should lie between 0.5 and 1 for a classifier model [7]. If its value is less than 0.50 then it specifies that it could not distinguish between true and false on some set of arbitrary data. A good or finest classifier should have a value of AUC near to 1.0. From Table 1 we get that the logistic regression model accuracy is 0.77. The accuracy of Random Forest model is 0.97 and accuracy of Decision Tree model is 0.97. Recall for the logistic regression model is 0.73 and for Random Forest it is 0.90 and for Decision Tree, it is 0.95. The precision of logistic regression is 0.64, for Random Forest it is 0.97, and for Decision tree it is 0.94. The F1 score of logistic regression is 0.66, while F1 score of Random Forest and Decision tree models are recognized to be 0.93, 0.94 respectively. Here we computed the AUC values to find the efficiency of three different classification models. From above- mentioned table1 we get the values of AUC for logistic regression, Random Forest and Decision Tree model. AUC value for logistic regression model is 0.62, for Random Forest model it is 0.90, and for Decision Tree its value is 0.95. So, from the above studies, it can be said that on the justification of all the criterion, Random Forest and Decision Tree are the two finest models to discover that either the employee is going to leave the company or not. In addition, we can see it very clearly that Random Forest model scores higher in accuracy and precision than the Decision Tree model. Whereas, Decision Tree model scores higher in F1 score and recall than the Random Forest model. We cannot rely only on accuracy as our performance metric in case of a dichotomous classification problem. In the event of unequal class distribution, an important performance evaluation metric called as F1 score will give a cut above intuition into the execution of the classifier because the F1 score provides thorough stability among recall and precision. So, in this instance we will use the F1 score. In addition, we can observe that the AUC value for Random Forest model is 0.90 (Fig 7) and that of Decision Tree model is 0.95 (Fig 8). As the AUC value for both classifier models i.e., Random Forest and Decision Tree are higher, so we can say that these two models are the finest model classifiers for the HR analytics employee data-set.

VII. MODEL COMPARSON

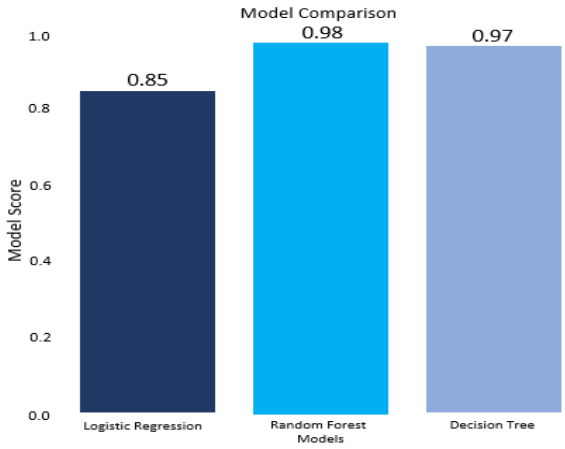


Fig. 9The above figure is the comparison of the machine learning classifier model that we built on HR analytics dataset.

As we can see from the figure that logistic regression has a model score of 85%, random forest has a model score of 98% and decision tree has a model score of 97%. So, on the basis of the above figure, we can conclude that both random forest and decision tree are the best models for HR analytics dataset.

VIII. CONCLUSION

The innovative outcomes suggested that all the models attained satisfactory outcomes; the Random Forest model yields the foremost accuracy of 0.97 and foremost precision value of 0.97 for the forecasting of attrition of employees in comparison to other models used. Additionally, the Decision tree model provided the foremost recall of 0.95 and foremost F1 score of 0.94. F1 score gives a thorough stability among the precision & recall. In addition, we observed that the AUC values of the Random Forest are 0.90 and AUC value of Decision Tree is 0.95. As the AUC values of both Decision Tree classifier and Random Forest classifier models are higher, so we can say that these two models are the finest model classifiers for the HR analytics employee data-set. So, from the aforementioned exploration, we can

European Economic Letters
ISSN 2323-5233
Vol 13, Issue 5 (2023)
https://doi.org/10.52783/eel.v13i5.828
http://eelet.org.uk

conclude that Random Forest classifier and Decision Tree classifier are the two finest machine learning (ML) models to figure out regardless if the employee is going to leave the company or not.

Important insights that we got from the Employee HR analytics data are as follows: Firstly, better the work life balance of employees, the less are their chances of leaving the company. Secondly, the more an employee works overtime on an average the more are the chances that he/she will leave the company. Thirdly, if an employee works for a longer duration of time with the same manager, the lesser are the chances that employee will leave the company. Fourth point is that hire people with more experience as they are improbable to quit the company. And the last point is that employees who are unmarried are prone to leaving the company.

IX. REFRENCES

[1] Sridhar, Guru Vignesh & Venugopal, Sarojini & Vetrivel, S. (2018). Employee Attrition and Employee Retention-Challenges & Suggestions.

[2] Çelik, Ö., Altunaydın, S.S. (2018). A Research on Machine Learning Methods and Its Applications.Journal of Educational Technology & Online Learning, 1(3), 25-40.

[3] Peng, Joanne & Lee, Kuk & Ingersoll, Gary. (2002). An Introduction to Logistic Regression Analysis and Reporting. Journal of Educational Research - J EDUC RES. 96. 3-14. 10.1080/00220670209598786.

[4] Ali, Jehad & Khan, Rehanullah & Ahmad, Nasir & Maqsood, Imran. (2012). Random Forests and Decision Trees. International Journal of Computer Science Issues (IJCSI). 9.

[5] Vartak, Shubham. (2020). An Overview of Predictive Analysis: Techniques and Applications. International Journal for Research in Applied Science and Engineering Technology.

[6] Kumar, Vaibhav & L., M. (2018). Predictive Analytics: A Review of Trends and Techniques. International Journal of Computer Applications.

[7] M.E. Rice, G.T. Harris, Comparing effect sizes in follow-up studies: ROC area,Cohen's d, and r, Law Hum. Behav. 29 (5) (2005) 615–620.

[8] Singh, R., Rajpal, N., & Mehta, R. (2020). Wavelet and kernel dimensional reduction on arrhythmia classification of ECG signals. EAI Endorsed Transactions on Scalable Information Systems: Online First. 10.4108/eai.13-7-2018.163095

[9] Ritu Singh & Navin Rajpal & Rajesh Mehta, 2021. "Application-Specific Discriminant Analysis of Cardiac Anomalies Using Shift-Invariant Wavelet Transform," International Journal of E-Health and Medical Communications (IJEHMC), IGI Global, vol. 12(4), pages 76-96, July.

[10] Singh, R., Rajpal, N., & Mehta, R. (2021). An Empiric Analysis of Wavelet-Based Feature Extraction on Deep Learning and Machine Learning Algorithms for Arrhythmia Classification. Int. J. Interact. Multim. Artif. Intell., 6, 25-34.

[11] Mahesh, Batta. (2019). Machine Learning Algorithms -A Review. 10.21275/ART20203995.

[12] Alqahtani, Abdullah & Alsubai, Shtwai & Sha, Mohemmed & Vilcekova, Lucia & Javed, Talha. (2022). Cardiovascular Disease Detection using Ensemble Learning. Computational Intelligence and Neuroscience. 2022. 1-9. 10.1155/2022/5267498.

[13] Ghosal, Anindya & Singh, Jyostna & Singh, Ashutosh. (2022). Predicting Growth in Tourism Industry Using Machine Learning Methods. 317-321. 10.1109/COM-IT-CON54601.2022.9850491.

[14] R, Praba & G, Darshan & T, Roshanraj & B, Surya. (2021). Study On Machine Learning Algorithms. International Journal of Scientific Research in Computer Science, Engineering and Information Technology. 67-72. 10.32628/CSEIT2173105.

[15] Rostam Niakan Kalhori, Sharareh. (2022). Towards the Application of Machine Learning in Emergency Informatics. 10.3233/SHTI220003.

[16] Kavitha, V. & Kumar, G. & Kumar, S. & Harish, M. (2020). Churn Prediction of Customer in Telecom Industry using Machine Learning Algorithms. International Journal of Engineering Research and. V9. 10.17577/IJERTV9IS050022.

[17] Dutta, Shawni & Bose, Payal & Bandyopadhyay, Samir & Janarthanan, Midhunchakkaravarthy. (2022). A Hybrid Machine Learning Model for Bank Customer Churn Prediction. International Journal of Engineering Trends and Technology. 70. 13-23. 10.14445/22315381/IJETT-V70I6P202.

[18] Aggarwal, Riya & Goyal, Anjali. (2022). Anxiety and Depression Detection using Machine Learning. 141-149. 10.1109/COM-IT-CON54601.2022.9850532.

[19] Hanaysha, Jalal. (2016). Improving employee productivity through work engagement: Evidence from higher education sector. Management Science Letters. 6. 61-70. 10.5267/j.msl.2015.11.006.

[20] Zongjun, Lan, Research on Factors Affecting Employee Productivity in Shanghai (December 2019). International Journal of Management (IJM), 10 (6), 2019, pp. 147–160.