International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Foreseeing Employee Attritions using Diverse Data Mining Strategies

Jalpesh Vasa, Kanksha Masrani

Abstract: ***"Employee turnover is a noteworthy matter in knowledge-based companies." On the off chance that employee leaves, they carry with them tacit information, often a source of competitive benefit to the other firms. Keeping in mind the end goal, to stay in the market and retain its employees, an organization requires minimizing employee attrition. This article discusses the employee churn/attrition forecast model using various methods of Machine Learning. Model yields are then scrutinized to outline and experiment the best practices on employee withholding at different stages of the employee's association with an organization. This work has the potential for outlining better employee retention designs and enhancing employee contentment. This paper incorporates and condenses the capacity to gain from information and give information-driven experiences, choice, and forecasts and thinks about significant machine learning systems that have been utilized to create predictive churn models.***

Keywords**:** ***Machine Learning techniques, Prediction, Classification, Algorithms, Turnover, Data mining, Attrition, Data Analytics, Predictive Modelling***

I. INTRODUCTION

With the growth of cutting edge data frameworks and databases that are able to hold colossal measures of information, the requirement of analysing it has turned out to be dynamically more pertinent. "This can be found in patterns of Google search, by keywords 'Data Science', 'Big Data' and 'Machine Learning' (Fig.-1)[1]". Additionally, the way that universities have begun offering affirmation courses and master's degrees in Predictive Analytics and Big Data Analytics which reflects the growth and popularity of this field.

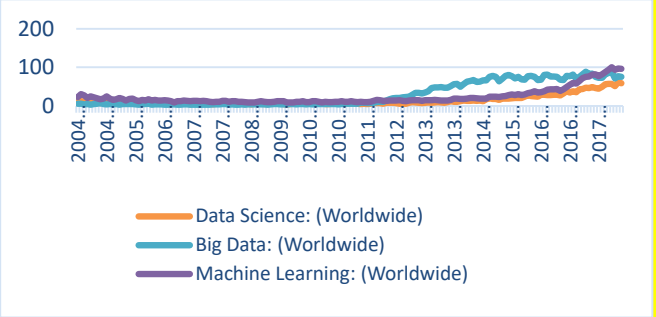


Fig. 1. Word Trends on YoY [1]

The uncooked facts itself does not bring an awful lot value without any additional processing and analysing. In Fig.-1 it is clearly noticeable that as analysing Big Data became a relevant problem for the current world, interest in Data Science and Machine Learning grew. There distinctive kinds of Data Analytics beginning from Descriptive Analytics developing to something further developed, as Predictive Analytics. The predictive analysis enables the investigator to work on authentic and current data helping to predict the likely future environment. This predictive insight promotes much better decision making and improved results.

Utilization of predictive analytics is wide, it empowers organizations to improve pretty much aspect of their business fortifying their decision making power. As Human Resources (HR) possesses enormous sums of employee data, demand for analysis is high. *"However, HR Information Systems (HRIS) are often underfunded contrasted with information systems of other domains of enterprise, which are directly connected with the main business [2]."* This prompts the way that HR data contains a lot of noise and errors. Therefore, building an accurate analytical model is challenging for HR.

One of the uses of Predictive Analytics for Human Resources (HR) is foreseeing employee turnover, attrition or retention. Employee turnover has different negative impacts including loss of enterprise knowledge, costs linked with leaving and substitution. To decisively make sense of who is leaving and what is the basic reason are key issues for HR workforce planning.Employee turnover is a piece of a quotidian business action; with life circumstances transforming they may stop their jobs. Businesses understand this and, to be sure, firms normally have whole divisions devoted to the administration of HR with a specific end goal to make the progress as smooth as workable for both administration and worker and to lessen the related enrolling and preparing costs.Employee turnover caused due to non-business related causes are generally out of the business' control. Non-business related explanations behind employee turnover are those things in the worker's near and dear life that influence their execution in the workplace.

Manuscript published on 30 September 2019
* Correspondence Author
Jalpesh Vasa*, Department of Information Technology, Chandubhai S. Patel Institute Of Technology(CSPIT), Faculty of technology & Engineering (FTE), Charotar University of science and Technology (CHARUSAT), Changa, Anand, India. Email: jalpeshvasa@gmail.com
Kanksha Masrani, Faculty of Applied Science, Simon Fraser University, Vancouver, Canada. Email: kanksha.masrani@gmail.com

Instances of these eventual movements, family issues and concoction mishandle. In spite of the fact that these causes are not straightforwardly inside the business' control, a few associations have supported responsive projects for the non-work related classification, for example, worker help projects and stress administration preparing those better-arranged representatives to manage individual issues that affect their work execution.

The largest rate (67.57%) of respondents left because of work-related reasons, trailed by 27.03% who indicated individual reasons to be specific family, wellbeing, additionally examines the reason and so on. "*The activity related reasons included: disappointment with retail, disappointment with compensation, better business nonretail, no professional development and scheduling issues.[3]"*

A. PROBLEMS & GOALS DEFINED

The acumen of this abstraction is to advance Predictive Analytics for HR on the case of employee turnover and to investigate variables that influence employee attrition within the organization, using Machine Learning algorithms. The point is to experiment with various Machine Learning algorithms and assess their performance on the company's data so as to choose the most accurate model. Data for this modelling problem comprises of structured data from numerous sources, therefore data pre-processing will be required. It will incorporate demographic information of employees and outcome value will be the possibility of an employee leaving the organization. Precisely forecasting of employee turnover will empower the organization to make strategic decisions regarding employee retention and take vital actions.

Terms, for example, "Data Science", "Data Analytics", "Predictive Analytics", "Predictive modelling", "Data Mining" and Machine Learning are as yet used dubiously and interchangeably. "*The term Data Science refers to everything that is related to data cleansing, preparation and analysis in order to excerpt insights and under stability information from data[4]."* It combines mathematical, statistical methodologies besides programming. Data Analytics, on the other hand, concentrates more on deriving conclusions based on raw data. "*An enormous amount of data that cannot be stored or treated within a given timeframe using standard innovations is called Big Data[5]."*

"Employee turnover can be interpreted as a leak or departure of intellectual capital from the employing organization. Maximum of the literature around turnover categorizes turnover as either voluntary or involuntary[6]."

Organizations cope with such issue by implementing machine learning systems to foresee turnover consequently giving them the imaginative and prescient to make an important move. High turnover has a few adverse consequences for an association. It is hard to supplant workers who have speciality ranges of abilities or are business space specialists. It influences progressing work and profitability of existing representatives. Obtaining new representatives as substitution has its own particular costs like employing costs, preparing costs and so forth. Additionally, new representatives will have their expectations to absorb information towards touching base at comparative levels of specialized or business skill as a prepared inside worker.

II. MACHINE LEARNING METHODS

The Analytical and Predictive systems conveyed depend on surely understood machine learning strategies, for example, Logistic Regression, Naïve Bayes, Decision Tree, Random Forest, Gradient Boosting trees and K-means. They have been utilized to represent class polarity and affiliation and potentially related inclinations.

A. CLASS IMBALANCE CORRECTION

"Class Imbalance is a typical theme on employee churn prediction [7], [8]...)," which is a similar point to the one in this paper. Standardized arrangements include subsampling systems to smoothen the ambiguities between the classes.

Due to the skewed distribution of employees who did and did not leave, this case is an example of a class imbalance problem. More skewed the class means precision breaks down. For such kinds of situation, assessing the algorithms used in light of precision is the wrong component to quantify. We might need to realise the distinctive mistakes that we consider and correct decisions. Precision does now not quantify a vital idea that should be pondered about in this type of assessment.

False Positive and False Negative errors as follows:

False Positives (Type I Error): You predict the employee is going to leave but don't leave.

False Negatives (Type II Error): You predict the employee is not going to leave but is leaving.

Contingent upon the errors, distinctive expenses are evaluated keeping in mind the type of employee being taken into consideration. For instance, would any employee be treated differently depending on whether his/her salary is high or low? The price for error will vary and it ought to be evaluated as per requirement.

In our employee retention issue, in place of just anticipating whether a representative will quit the organization inside a specific time span, we would much rather have a gauge of the likelihood that he/she will quit the organization. We might rank representatives through their probability of leaving, then allocate restricted incentive finance to the highest likelihood instances.

B. SELECTED CLASSIFICATION METHODS

Machine Learning is about giving computers the ability to learn without explicit programming. As said by Baron Schwartz When you are raising funds It's Artificial Intelligence, When you are hiring It's Machine Learning When you are Implementing It's Linear Regression and when you are debugging its print(). Similarly, the hiring process was done here.

With the given dataset various machine learning algorithms are used and the best forecasting method will be chosen on the basis of its accuracy and execution time.

"One can classify human beings based on their race or can categorize products in a supermarket based on the consumers shopping choices. In general, classification involves examining the features of new objects and trying to assign it to one of the predefined set of classes[18]."

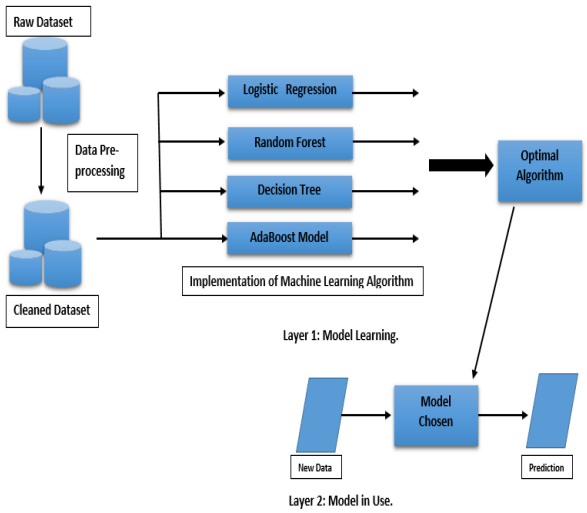


Fig. 2. Project Flow

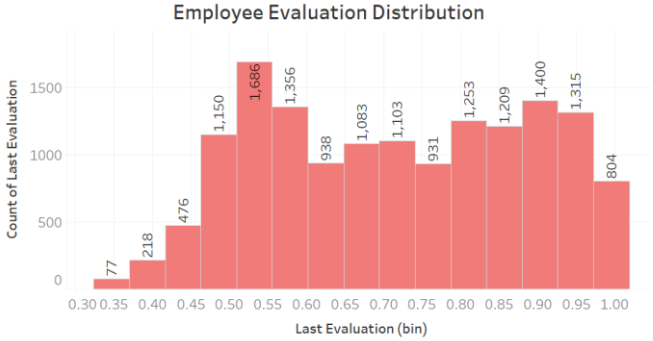
In order to describe significant data classes and to predict the forthcoming Data tendency two forms of data analysis are required- Classification and prediction which help in extracting a model for the above-mentioned purpose. The entire process is divided into two parts: The learning part and the classification part. The learning requires the training dataset to be analysed by the chosen algorithm. The classification part is where the accuracy of the created model is estimated using the test dataset.

There are several techniques which can be implemented such as a decision tree, random forest, Logistic regression and AdaBoost which is used to improve the performance of the decision tree on binary classification problems. All these classifiers have their own advantage and disadvantage, for that reason, this paper endeavour to study the techniques for human talent data.

III. METHODOLOGY

A. DATA PRE-PROCESSING

The sample dataset(Synthetic dataset-generated programmatically with proper distribution in each attribute) consists of information on the percentage of employees who left and stayed with the company. The details of 15,000 employees consisted of attributes such as the number of hours worked, the number of the project worked on, the evaluation, and salary and many other details. From the total, 24% left the company and 76% stayed. The dataset we had used is a synthetic dataset generated by a python script.



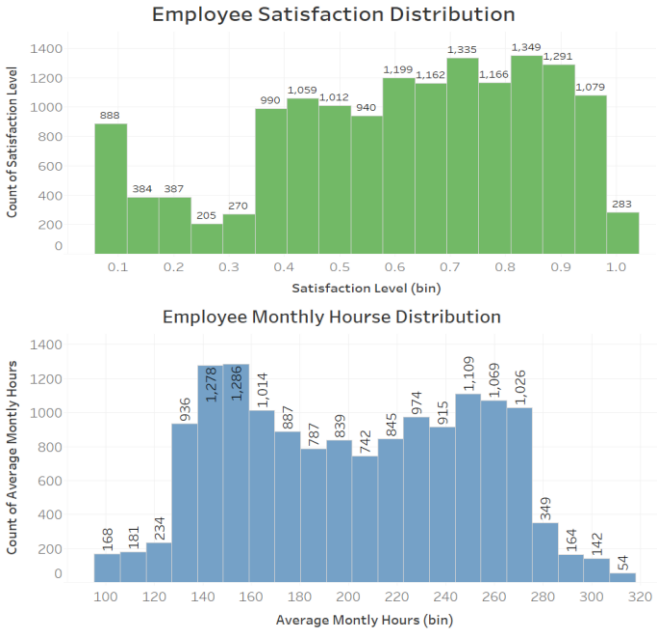


Fig. 3. Frequency Distribution

Upon more detailed it was seen that the greater part the employees with 2, 6 and 7 assignments left the organization. The greatest number of employees who did not leave had 3, 4 and 5 assignments and every one of the employees with 7 projects left the organization. Another bi-modular distribution of employee that turnover portrayed that employee who worked for a minimum and maximum hours of worked left. The conclusion arrived was that there was an increase in the turnover if they were underworked or overworked.

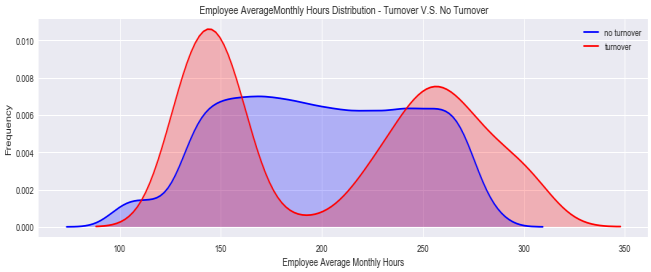


Fig. 4. Average monthly Hours v/s Turnover

Maximum number of Employees who left had a greater performance evaluation and the ones who stayed were in the range of 0.6-0.8.

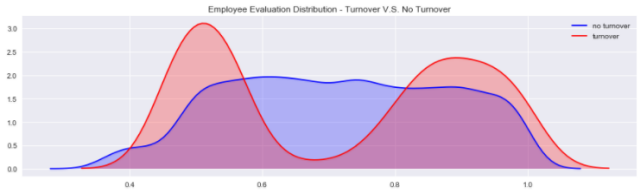


Fig. 5. Employee Evaluation v/s Turnover

Barely any employee with high salary left. Majority of employees left were a part of the low and medium salary range.

Fig. 6. Employee Salary Turnover Distribution

When the analysis was done department wise the sales, technical and support department had the highest employee turnover whereas the management department has the least percentage of turnover.

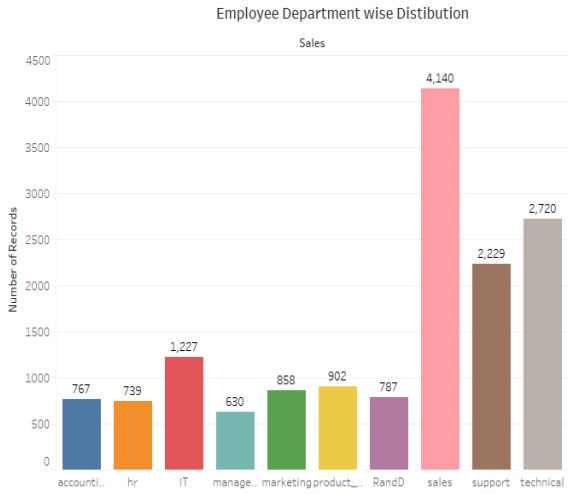


Fig. 7. Employee Department wise Distribution

Another bi-modal distribution depicted that employees working for a minimum and maximum hours of the work left. A tri-modal distribution of employee turnover depicted that employees having very low-low satisfaction levels left the company.

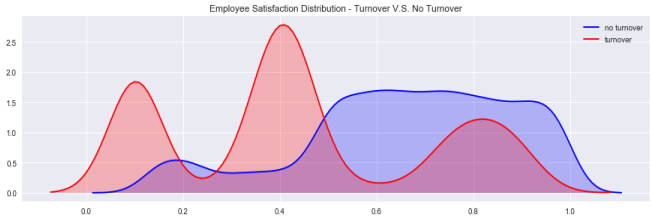


Fig. 8. Employee Satisfaction v/s Turnover

When a box-plot graph was generated between the project count and average monthly hours the employees who had consistent monthly hours did not leave the company whereas the ones that had an increase in the monthly hours left.

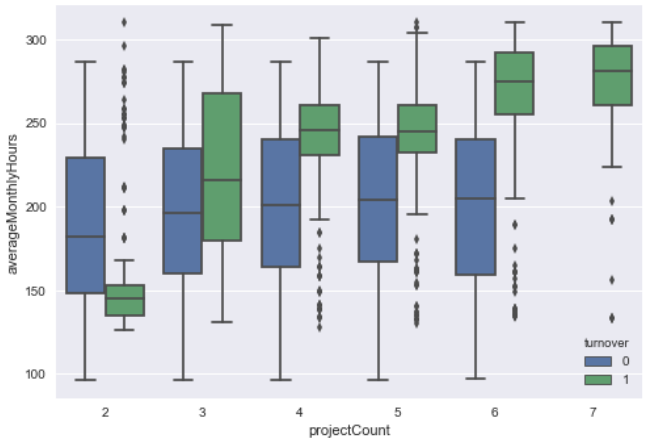


Fig. 9. Project count v/s Average Monthly Hours

Similarly, project count and evaluation also showed some kind of relationship. The box-plot here depicted that employees who had more work to do and high evaluations left the company whereas, employees who had a consistent evaluation score irrespective of the increase in projects stayed showing that the ones with higher evaluation would have left because of the better job opportunity.

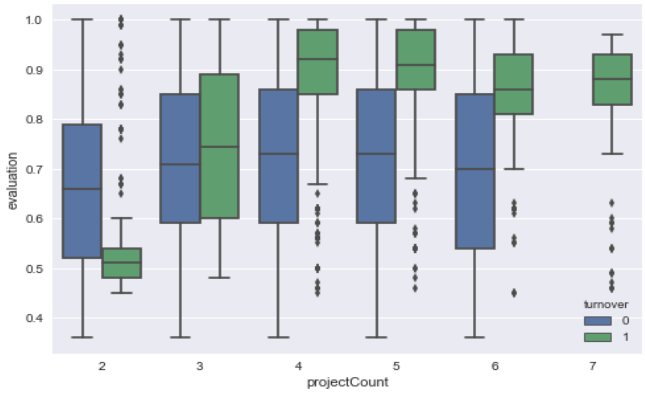


Fig. 10. Project Count v/s evaluation

After plotting graphs between the attributes depicting years at the company to the turnover most employees left after working for 4-5 years. An interesting insight was drawn from plotting of satisfaction and evaluation attributes by creating clusters. It can be depicted in the table below:

Table 1: Satisfaction v/s Evaluation Cluster

Satisfaction	Evaluation	Cluster
< 0.2	>0.75	Overworked
0.35~0.45	~0.58	Under-performed
0.7~0.1	>0.8	Ideal

Table 2: Cluster conclusion

Cluster	Conclusion
Overworked	Good employees left who felt horrible at work.
Under-performed	Badly evaluated and felt horrible at work.
Ideal	Loved their work and had a great performance. Left due to better job opportunities.

B. ANALYSIS

Eventually, we employed four different models for prediction upon whom we based and tested the data. This method has several possible applications in Data Science of which we chose to predict annual employee turnover.
The methods employed for the above are:
1. Logistic Regression
2. Decision Tree
3. Random Forest and
4. AdaBoost algorithm

(B)(I). Logistic Regression

Logistic regression predicts the likelihood of a result which can have only two esteems (i.e. a dichotomy). The forecast depends on the use of one or several(numerical and categorical) indicators. For the accompanying reason, linear regression is not suitable to anticipate the estimation of a binary variable.
Linear regression predicts values beyond the appropriate range (e.g. foreseeing probabilities beyond the range 0 to 1). Since the dichotomous tests can have just one of two conceivable values for each test, the residuals will not typically be distributed over the anticipated line.
Then again, a logistic regression produces a logistic curve, which is restricted to values in the vicinity of 0 and 1. Logistic regression is close to linear regression, however, the curve is developed using the common logarithm of objective variable "chances", as opposed to the likelihood. Furthermore, the indicators in each group do not need to be distributed normally or have equal variance.

(B)(II). Decision Tree

Decision trees are graphical portrayals of elective decisions that can be made by a business, which empower the chief to distinguish the most reasonable alternative in a specific situation. Decision trees will be trees that group examples by arranging them based on feature values. Every node in a decision tree speaks to an element in an instance to be classified, and each branch speaks to a value that the node can assume. Cases are characterized beginning at the root hub and arranged in light of their feature value.
A greedy algorithm is a fundamental calculation for decision tree induction that develops the tree in a top-down recursive divide-and-conquer way. This algorithm is generally utilized on the grounds that they are proficient and simple to execute, yet they usually lead to sub-optimal models. Another alternative approach can be a bottom-up approach.
Decision trees utilized as a part of data mining are of two principal types:

- Classification tree analysis is when the predicted outcome is the class to which the data belongs.
- Regression tree analysis is when the predicted outcome can be considered a real number (e.g. the price of a house, or a patient's length of stay in a hospital).

In order to elude both the above approaches CART – Classification and Regression Tree analysis was used, an umbrella term, presented by Breiman et al. Trees utilized for regression and classification have a few similitudes - yet additionally a few contrasts, for example, the technique used to figure out where to split. Numerous particular decision tree algorithms exist. Notable ones include:

- *C4.5 algorithm[9]*
- *C5.0 algorithm*
- *ID3 algorithm*
- *Chi-squared Automatic Interaction Detector (CHAID). Performs multi-level splits when computing classification trees*[10]."

(B)(III). Random Forest

"Random Forest calculation is a popular tree-based ensemble learning technique. The kind of 'ensembling' utilized here is bagging[13]." In bagging, progressive trees don't rely upon prior trees - each is independently constructed utilizing an alternate bootstrap sample of the data set. At last, a basic greater part vote is taken for the forecast. Random forest is not quite the same as standard trees in that for the latter each node is split utilizing the best split among all factors. "*In a random forest, each node is split utilizing the best among a subset of indicators randomly picked at that node*[14] *this extra layer of randomness makes it robust against over-fitting*[15]."

(B)(IV). Adaptive Boosting

"Boosting refers to the general problem of producing a very accurate prediction rule by combining rough and moderately inaccurate rules-of-thumb[11]." This includes fitting a sequence of weak learners on changed data. The predictions from every one of them are then combined through a greater part vote (or sum) to deliver the last prediction. The data alteration at each progression comprises of doling out higher weights to the training examples that were misclassified in the previous iteration. As iteration continues, cases that are hard to foresee get regularly expanding influence. "*This forces the weak learner to focus on the cases that are missed by its antecedent. AdaBoost is an algorithm of the boosted tree. It follows the gradient boosting principle[12].*" In comparison to gradient boosting, it makes use of a more standardised-model formalization to control over-fitting, giving it a better execution.

IV. RESEARCH FINDINGS

Application of fundamental statistical strategies is utilized to study the employee. A ROC curve was generated in order to show which algorithm was accurate of them all.
After the parameters with tuned, models were evaluated using 10-fold cross-validation. For evaluating model performances we have used the area under the curve (AUC) from prediction scores, the receiver operating characteristics (ROC) curve.

A. EVALUATION CRITERIA

"Classification accuracy is detected by evaluating the area of the region that occurs under the receiver operating characteristic curve (ROC-AUC). The AUC is a general measure of ' predictability' and decouples the classifier assessment from working conditions i.e., class distributions and misclassification costs[16]." Moreover, AUC is favoured over other indicators like error rate because it gauges the probability of a classifier choosing a positive instance in a random manner over choosing a negative instance in a random manner. This is also in line with the Wilcoxon trial of ranks[17]. For the prediction model the confusion matrix will be:

Table 3: Confusion Matrix

	Class-1/true	Class-2/false
Class-1/true	e	l
Class-2/false	$\bar{e}$	$\bar{l}$

Where e is a number of the correctly classified data object (called true positive), in this study case correctly predicted the number of employees. $\bar{e}$ is number of misclassified employees (false negatives). Analogically, $\bar{l}$ is a number of misclassified leavers (false positive) and l is number of correctly predicted leavers (true negative).

$True\ positive\ rate = e /(e + \bar{e})$

$False\ positive\ rate = \bar{l}/(\bar{l} + l)$

ROC curve is plotted using true positive rate and false positive rate coordinates, shown in Figure 13. AUC calculates the area under the ROC curve.

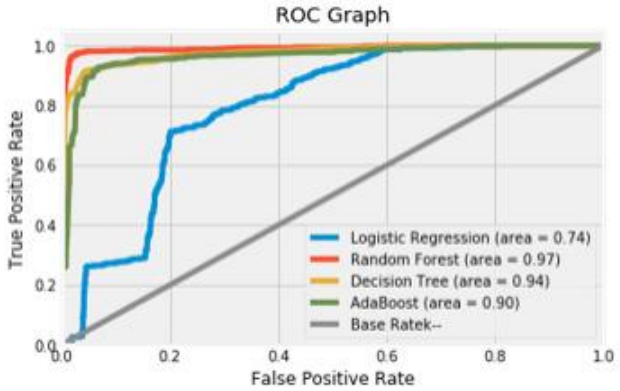


Fig. 11. ROC graph

Furthermore, the model runtime is additionally used to analyse the execution of the classifiers. This measure is vital to be considered, as it fabricates a case from an expert's point of view on figuring out that an algorithm which is worthy to implement for real-life business issues, answering facility and execution.

Table 4: AUC Score

Algorithm	AUC	Runtime(Sec)
Logistic Regression	0.74	0.021
Random Forest	0.97	8.292
Decision Tree	0.94	0.167
AdaBoost	0.90	2.515

The machine learning models performed at varying performance levels for the different input datasets. By considering the average F1 score, model rankings are obtained- i.e. a higher F1 score is indicative of a better performing model.

Table 5: Model Outcome

Model Name	F1 score	Ranking
Base Model	0.66	5
Logistic Model	0.76	4
Adaptive Boosting Model	0.93	3
Decision Tree Model	0.95	2
Random Forest Model	0.98	1

V. CONCLUSION & FUTURE WORK

This paper demonstrates the use of Machine Learning/Data Mining algorithms to build turnover models which foresees employee turnover in companies and discusses the significance of the latter. Based on the experiment to anticipate the turnover, two important criteria have been considered: Accuracy and Runtime. Also, we found that there is a direct relationship between employee satisfaction v/s project served, hours worked and how they were evaluated. Based on different evaluation measures for predicting turnover using different data mining models, our work has shown that the Random Forest classifier has been proved to be the predominant algorithm among the others.

For future examinations, the author proposes the capture of information about the organization's interventions for employees at risk and their result. This will turn the model into a normative one, not only addressing "Who is at risk?" but also "What can we do?" Studying the application of deep learning models for anticipating turnover is also recommended. A very much outlined system with adequate hidden layers may enhance the precision, be that as it may, the scalability and viable implementation perspective must be considered also.

REFERENCES

1. https://trends.google.com/trends/explore?date=2004-01-01%202018-02-01&q=Data%20Science,Big%20Data,Machine%20Learning.
2. Ajit, Pankaj. "Prediction of employee turnover in organizations using machine learning algorithms." algorithms 4.5 (2016): C5.
3. Tamizharasi, K., and U. Rani. "Employee Turnover Analysis with Application of Data Mining Methods." International Journal of Computer Science and Information Technologies 5.1 (2014): 562-566.
4. simpli learn. [Online]. https://www.simplilearn.com/data-science-vs-big-data-vsdata-analytics-article.
5. http://www.tud.ttu.ee/im/Vladimir.Viies/materials/L%C3%95PETAMINE/17MAGhannesele/Thesis_Mari_Maisuradze
6. J M. Stoval and N. Bontis, "Voluntary turnover: Knowledge management – Friend or foe?", Journal of Intellectual Capital, 3(3), 303-322, 2002.
7. K. W. Bowyer, N. V. Chawla, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," CoRR, vol. abs/1106.1813, 2011.
8. A. Amin, F. Rahim, I. Ali, C. Khan, and S. Anwar, "A Comparison of Two Oversampling Techniques (SMOTE vs MTDF) for Handling Class Imbalance Problem: A Case Study of Customer Churn Prediction", pp. 215–225. Cham: Springer International Publishing, 2015.
9. Desai, Ankit B., and Jalpesh Vasa. "Cost-Sensitive Decision Tree Induction for Feature Selection and Sequential Minimal Optimisation for Classification: CSattribSelectorC4. 5 ()." International Journal Of Data Mining And Emerging Technologies 3.2 (2013): 58-62.

10. Alao, D. A. B. A., and A. B. Adeyemo. "Analyzing employee attrition using decision tree algorithms." Computing, Information Systems, Development Informatics and Allied Research Journal 4 (2013).
11. Freund, Yoav, and Robert E. Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting." Journal of computer and system sciences 55.1 (1997): 119-139.
12. Friedman, Jerome H. "Greedy function approximation: a gradient boosting machine." Annals of statistics (2001): 1189-1232.
13. Goel, Eesha, et al. "Random forest: A review." International Journal of Advanced Research in Computer Science and Software Engineering 7.1 (2017).
14. Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." R news 2.3 (2002): 18-22.
15. Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.
16. Lessmann, Stefan, and Stefan Voß. "A reference model for customer-centric data mining with support vector machines." European Journal of Operational Research 199.2 (2009): 520-530.
17. Fawcett, Tom. "An introduction to ROC analysis." Pattern recognition letters 27.8 (2006): 861-874.
18. Grobler, P., Wärnich, S., Carrell, M.R., Elbert, N.F., Hatfield, R.D. (2006) Human Resource Management in South Africa, 3rd edition. [Internet].

AUTHORS PROFILE

Jalpesh Vasa is working as an Assistant Professor at Department of Information Technology in Chandubhai S Patel Institute of Technology, CHARUSAT since July 2012.He has completed his B.E.(CE) and M.Tech.(IT) from VNSGU and Ganapt University respectively. His research interest includes Data Mining, Machine Learning, Privacy Preserving Data Mining/Data Publishing.He has published various articles in international journal/conference.He is currently working on Privacy preserving Big data.

Kanksha Masrani is an Information Technology Graduate from CHARUSAT, India, currently working towards a Masters Degree in Computer Science with a specialization in Big Data at Simon Fraser University, Vancouver. Through her coursework, she is fine-tuning her ability to analyze data and appreciate the complexity of businesses. Upon completion, I will be equipped to solve the most demanding and pressing problems of the industry. Through her experience, she has become adept at translating the language of data into actionable business decisions.