ISSN 2278-3091
Volume 9, No.5, September - October 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse91952020.pdf
https://doi.org/10.30534/ijatcse/2020/91952020

Employee Attrition Prediction Using Machine Learning and Sentiment Analysis

Saisanthiya.D[1], Dr.V.M.Gayathri[2], Dr.P.Supraja[3]
[1]Assistant Professor, Department of IT,SRM institute of Science and Technology,Chengalpattu,
saisantd@srmist.edu.in
[2]Assistant Professor, Department of IT,SRM institute of Science and Technology,Chengalpattu,
gayathrm@srmist.edu.in
[3]Assistant Professor, Department of IT,SRM institute of Science and Technology,Chengalpattu,
supraja@srmist.edu.in

ABSTRACT

Consistently a great deal of organizations procures various workers. The organizations put time and cash in preparing those workers, this as well as there are preparing programs inside the organizations for their current representatives too. The point of these projects is to build the viability of their representatives. The point of these projects is to expand the adequacy of their representatives. Human asset investigation (HR examination) is a zone in the field of investigation that alludes to applying expository procedures to the human asset branch of an association in the desire for improving worker execution and hence showing signs of improvement rate of profitability. HR examination doesn't simply manage gathering information on worker productivity. Rather, it plans to allow knowledge into each procedure by social gathering information and afterward utilizing it to decide on important choices about a way to improve these procedures. Whittling down in HR alludes to the progressive loss of workers after your time. within the dataset incorporates highlights like Age, Employee Role, Daily Rate, Job Satisfaction, Years at Company, Years in Current Role and then on we are going to try to break down elements which result in whittling down. By and enormous, generally high steady loss is hazardous for organizations. A significant issue in high representative wearing down is its expense to an association. Occupation postings, contracting procedures, administrative work and new contract preparing are a portion of the normal costs of losing representatives and supplanting them. The point is to arrangement of forecast of representative wearing down and maintenance by perception utilizing business examination standards on enormous information and utilizing assumption investigation to forestall the equivalent.

Key words: Employee Attrition, Random Forrest Classifier, Gradient boosting, Human Resource Predictive Analytics, Support Vector Machine, Machine Learning Models, Heatmap

1. INTRODUCTION

Job hopping is an expanding pattern in India and every one around and this costs the organizations. This issue is developing definitely; anyway, most of the organizations didn't have an equivalent and general point of view of worker steady loss. Along these lines HRPA is critical to make informative capacities which could be capable to convey more Return on Investment [1].(H) Human (R) Resource (P) Predictive (An) Analytics is that the fate of the association which assists with discovering the business bits of data within the sector of knowledge investigation by making a call about the past elements and making AI models to foresee the wearing down, unlucky deficiencies and different dangers to reinforce the representative maintenance. we'll use machine learning to predict turnover rate . Our objective is to seek out out How Attrition affects companies and the way HR Analytics helps in predicting attrition. After analysis we aim at finding factors affecting employee satisfaction and creating environments that promote retention using sentiment analysis of employee E-mail and Feedback dataset.

Multiple statistical analysis is conducted against a detailed of hypothesis to relate which factors affect attrition and which cause retention. Data will be visualized and a model will be created. The investigation was mostly embraced to distinguish the degree of worker's mentality, the disappointment factors they face in the association and for what reason they like to change their activity. When the degrees of Employee's disposition are distinguished, the administration would be able to make essential move to diminish steady loss level. Since they are considered as spine of the association, their movement will prompt the achievement of the association for the since quite a while ago run.

It is important to predict employee attrition and analyze retention decision because according to Morrel (Morrell, 2004), intentional turnover acquires the noteworthy cost, both in terms of:

Direct costs (replacement, recruitment and selection, temporary staff, management time) (Morrell, 2004).

Indirect cost (morale, pressure on remaining staff, costs of learning, product/service quality, organizational memory) (Orrell, 2004).

Applying predictive analytics to the HR department enables them to not just rely on gut feeling but pure historical data and computer prediction models

This serves to:

•Predict which worker will remain or leave the association, in this way helping associations to create and improve the maintenance techniques.

•Predict the hazard scores for a representative, which can enable the enrollment to group to have appropriate substitutions on schedule and forestall income spillage.

•Understand general patterns, examples and indications of whittling down with the goal that particular activities plans can be set up for each pattern, example and side effects.

Our objective is to find out How does Attrition affect companies? and how does HR Analytics help in analyzing attrition? How can we analyze sentiments to take preventive measures and promote retention? we will compare different models and compare reasons causing attrition. We will use the random forest classifier and gradient booster for model building. Outstanding amongst other element Random backwoods model has-it gives the significance of factors/includes in the information/model. For this HR Analytics issue, we are keen on realizing which include/factor contribute the most in the Attrition and RF's one capacity can give us this data. This is simply one more motivation behind why we have utilized RF. dialects to be used are python, R programming and scene programming to make representation and storyboarding of information. The point is to give condition to representative maintenance.

Using sentiment analysis and Machine learning algorithms on the IBM Watson dataset,

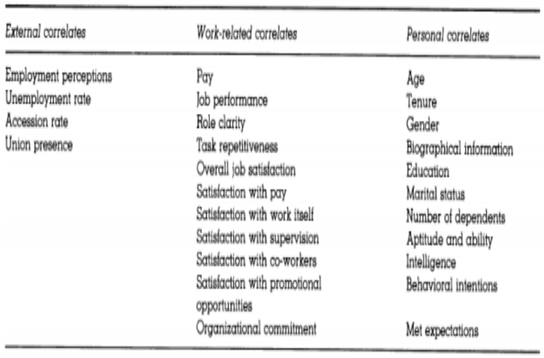
We will try to find the factors which affect attrition of employees the most and aim at providing an in sigh to the companies as to which factors, they need to work on in order t retain employees in this competitive work environment.

So far, most of the related specialists utilize diverse models with various procedures and to our best of the information, this examination will be an interesting first investigation in assessing representative whittling down through mixture outfit strategies in AI utilizing ADA support, GBM, Random woods and contrasting correctness's and the characterization models - SVM, GLM, SVM, choice trees and KNN to get the best exactness in finding the correct choosing factors. Many analysts have taken distinctive datasets and their most grounded elements are age, fulfillment, residency and pay along with other factors.

The Scope of the examination is partitioned into sections - First part contains the investigation of the factors affecting worker attrition. Second part has examination of different predictive models which aim to show precision for anticipating the worker sentiment. In the last part we have incorporated all the Machine Learning Models and information sources and built up the Machine Learning framework showing the expectation results, model outline, important representative and maintenance result table on dashboard alongside diagrams and plots. We have additionally made perception based on Tableau dashboards with graphs and plots.

2. INTRODUCTION

There have been extensive examinations on Factors answerable for Attrition. One of which is the investigation done by John Cotton and Jeffry Tuttle from Purdue University [2]. Figure 1 shows the 26 factors which were discovered to be liable for the whittling down:



External correlates	*Work-related correlates*	*Personal correlates*
Employment perceptions	Pay	Age
Unemployment rate	Job performance	Tenure
Accession rate	Role clarity	Gender
Union presence	Task repetitiveness	Biographical information
	Overall job satisfaction	Education
	Satisfaction with pay	Marital status
	Satisfaction with work itself	Number of dependents
	Satisfaction with supervision	Aptitude and ability
	Satisfaction with co-workers	Intelligence
	Satisfaction with promotional opportunities	Behavioral intentions
	Organizational commitment	Met expectations

Figure 1: Correlates of turnover

Associations put much in the worker enlisting, acceptance, preparing, improvement and maintenance and subsequently losing a representative is colossal misfortune to the organization. Thus, supervisors must decrease the worker turnover to diminish the loss of the turnover. "Research gauges demonstrate that contracting and preparing a swap specialist for a lost representative cost around 50 percent of the laborer's yearly pay [3]. Another significant issue that assumes a fundamental job as a turnover indicator is that the activity association [4].

Occupation Involvement is that the issue that shows what amount a specialist is worried in an exceedingly unequivocal job and do the duty from start to finish with differed responsible job [5]. This offers the specialist sentiment of ownership and motivate him to attempt to that activity successfully (Ongori, 2007). one in everything about components that encourage maintenance, is giving sound work air. Likewise, if reason of whittling down is uncovered, maintenance is improved through planning and association in conversation with specialist [6].

An early willful examination has discovered that the most grounded indicators for intentional turnover were residency, in general employment fulfillment, work execution, singular segment attributes (age/understanding, sex, ethnicity, instruction, marital status), geological components, pay, working conditions, work fulfillment, acknowledgment, development potential and so on [7]. Another investigation done on the Indian IT organizations uncovers the elements at risk for the wearing down are Below desire pay, Low motivators, Relationship with predominant, Relationship with subordinates, Job data, Skills use, Skills acknowledgment, Acknowledgment of work by prevalent, Lack of thankfulness, despondent work culture [8]. The social control remunerates conjointly assume a vital job in holding the specialist. In this way, administrator's social abilities assume an outrageously imperative job in hanging on the evaluable laborer[9].

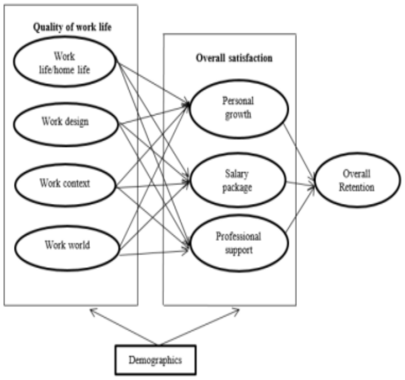


Figure 2: A model of relationship between quality of work life, satisfaction and retention

From Figure 2 shows an associations primary point is to pick up benefits from the client by structure notoriety and generosity of the organization and building up the business to higher scale however in the event that there is an issue of client beat, at that point it will be exceptionally hard to get new clients as clients criticism and audits are the pivotal points [10].According to [11][12][13],it is essential to pass judgment on client agitate ahead of time to spare the misfortunes and aides in increasing potential profits.In the future work, they recommended to utilize survival investigation which was conveyed by [14][15][16] in which survival examinations was done on representative whittling down over a predetermined timeframe to pick up worker maintenance bringing about important workforce for the survival of an association and the real parameter was Overtime. They got viable outcomes around then and proposed that for the best possible working of an industry, there ought to be a fitting technique for assessing steady loss. Organizations should concentrate on deliberate avoidable wearing down to improve the staff maintenance by making sound systems [17].

3. PROPOSED METHODOLOGY

This segment gives the hypothetical and specialized walkthrough of the exploration technique utilized for building an investigative application utilizing python for anticipating attrition and how to perceive the significant worker and hold them, along these lines sparing the organization HR spending plan on enlisting new worker [18]. This section depicts the arrangement of strategies utilized for doing the examination and building a product application.

There are four main functional requirements.
Those are as follows:

a.Predict Attrition – User must almost certainly anticipate the wearing down of things to come representative dependent on the authentic dataset of past workers.
b. Find valuable employees – Allow client to order the representative into profitable and normal ones.
c. Discover and rundown down the components influencing the maintenance choice – Display the maintenance factors on a dashboard, for improving the maintenance of the significant representatives.
d. Information Exploration and Data Visualization – User Interactive determination of characteristics for plotting the charts against the Attrition

3.1. System Architecture

Arrangement outline exhibits the succession of activity performed amid the run time of an application. The beneath outline appears, how the client sends the solicitation to each square of code and gets the yield consequently.

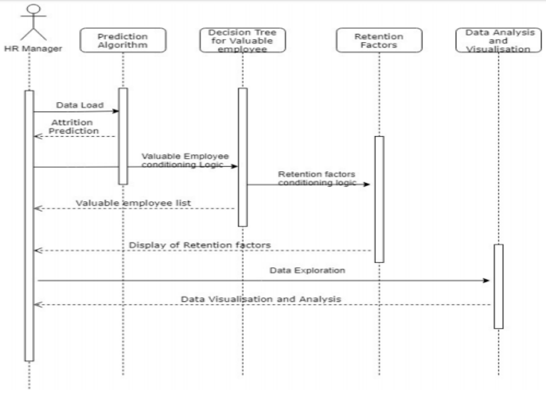

HR Manager
Prediction Algorithm
Decision Tree for Valuable employee
Retention Factors
Data Analysis and Visualisation
Data Load
Attrition Prediction
Valuable Employee conditioning Logic
Retention factors conditioning logic
Valuable employee list
Display of Retention factors
Data Exploration
Data Visualisation and Analysis

Figure 3: Activity Diagram

Figure 3 and 4 shows the framework configuration can be clarified with the assistance of the accompanying UML Diagram which speak to the working of the Analytical Application. As indicated by underneath UML Diagram, we have gathered the information from IBM. At that point we have picked the properties for the forecast and further investigation dependent on research. At that point, we preprocessed the information and put away in csv group in a nearby index. Then, the test information is gone through the prepared prescient model which characterizes the information into positive and negative whittling down and the outcome is appeared as "YES" or "NO" for the steady loss. This choice proclamation is to group the profitable representative from the standard one. When the significant representatives are classified, that rundown is put away in another article and every one of the variables influencing the maintenance are discovered and showed in the last segment of the informational collection. In this manner, result displays features ranked.

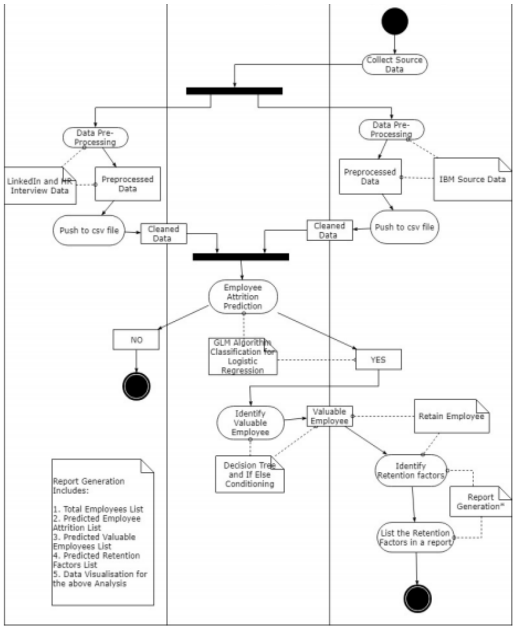

Collect Source Data
Data Pre-Processing
Data Pre-Processing
LinkedIn and HR Interview Data
Preprocessed Data
Preprocessed Data
IBM Source Data
Push to csv file
Cleaned Data
Cleaned Data
Push to csv file
Employee Attrition Prediction
NO
GLM Algorithm Classification for Logistic Regression
YES
Identify Valuable Employee
Valuable Employee
Retain Employee
Decision Tree and If Else Conditioning
Identify Retention factors
Report Generation*
List the Retention Factors in a report
Report Generation Includes:
1. Total Employees List
2. Predicted Employee Attrition List
3. Predicted Valuable Employees List
4. Predicted Retention Factors List
5. Data Visualisation for the above Analysis

Figure 4: UML Diagram

Figure 5 shows the Use case configuration is as per the following which shows how the client can interface with the application and take choices concerning significant representatives.

Use Case Scenario: The information is flawlessly taken care of into the framework and wearing down is anticipated with the normal precision. The representatives who will leave the organization are then classified into important and conventional workers utilizing choice tree. The best maintenance factors for significant workers are then shown on the dashboard.

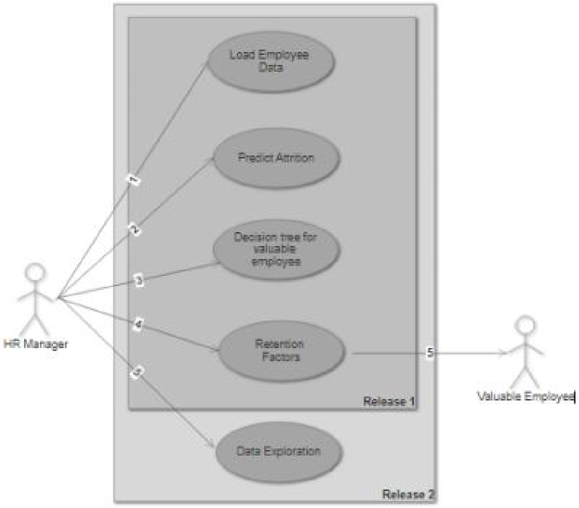
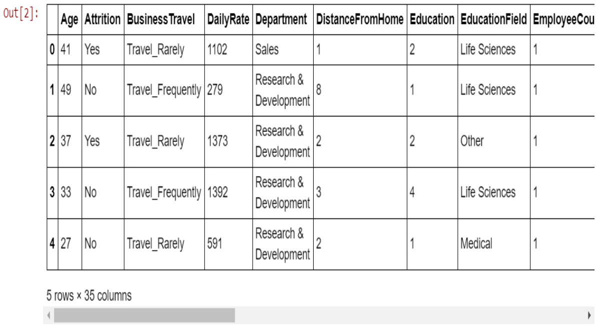


Figure 5: Use case Diagram

4. RESULT AND DISCUSSION

4.1 Exploratory Data Analysis:

Right now, investigate the dataset by investigating the element conveyances, how related one component is to the next and make some Seaborn and Plotly perceptions.


Out[2]:

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCou
0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1
1	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1
3	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1
4	27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1

5 rows × 35 columns

Figure. 6: Dataset

Figure 6 evinced from the dataset, our target column with which we can point our model to train on would be the "Attrition" column. Further more, we see that we have a mix of numerical and categorical data types. For the categorical columns, we shall handle their numerical encoding in the latter chapter. This section will be devoted to data exploration and as a first step, let us quickly carry out some simple data completeness checks to see if there are nulls or infinite values in the data.

4.2. Distribution of Dataset

Generally one of the first few steps in exploring the data would be to have a rough idea of how the features are distributed with one another. To do so, shall invoke the familiar kdeplot function from the Seaborn plotting library and this generates bivariate plots.

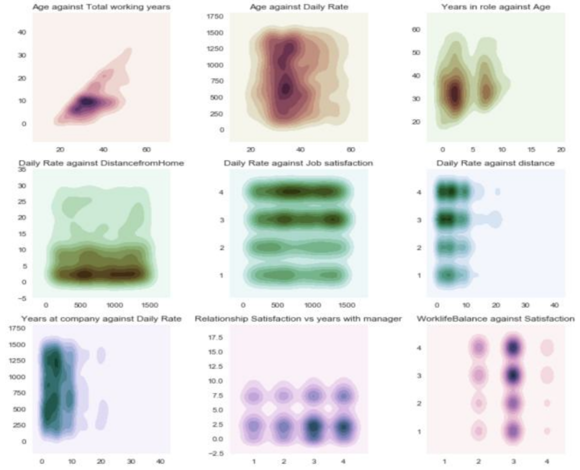
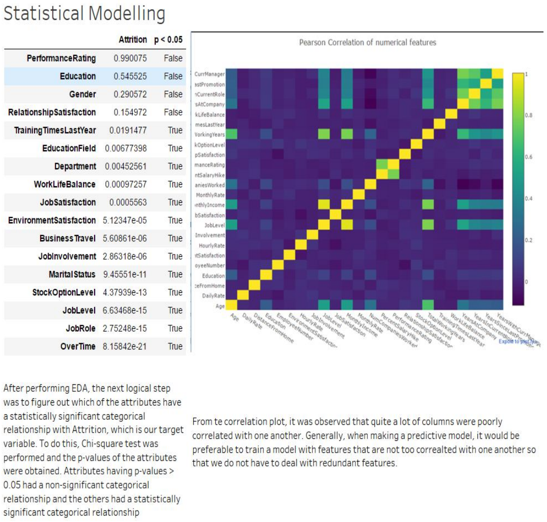


Figure 7: Dataset Distribution

4.2 Correlation of features:

The following instrument in an information adventurer's munitions stockpile is that of a relationship lattice. By plotting a relationship grid, we've an exceptionally pleasant review of how the highlights are associated with one another . For a Pandas data frame, we will helpfully utilize the correlation which in fact gives the Pearson Correlation estimations of the sections pairwise therein data frame Figure 8.

Utilize the thePlotly library to deliver an intelligent Pearson correlation network by means of the Heatmap work as follows:



Statistical Modelling

	Attrition	$p < 0.05$
PerformanceRating	0.990075	False
Education	0.545525	False
Gender	0.290572	False
RelationshipSatisfaction	0.154972	False
TrainingTimesLastYear	0.0191477	True
EducationField	0.00677398	True
Department	0.00452561	True
WorkLifeBalance	0.00097257	True
JobSatisfaction	0.0005563	True
EnvironmentSatisfaction	5.12347e-05	True
BusinessTravel	5.60861e-06	True
JobInvolvement	2.86318e-06	True
MaritalStatus	9.45551e-11	True
StockOptionLevel	4.37939e-13	True
JobLevel	6.63468e-15	True
JobRole	2.75248e-15	True
OverTime	8.15842e-21	True

After performing EDA, the next logical step was to figure out which of the attributes have a statistically significant categorical relationship with Attrition, which is our target variable. To do this, Chi-square test was performed and the p-values of the attributes were obtained. Attributes having p-values > 0.05 had a non-significant categorical relationship and the others had a statistically significant categorical relationship

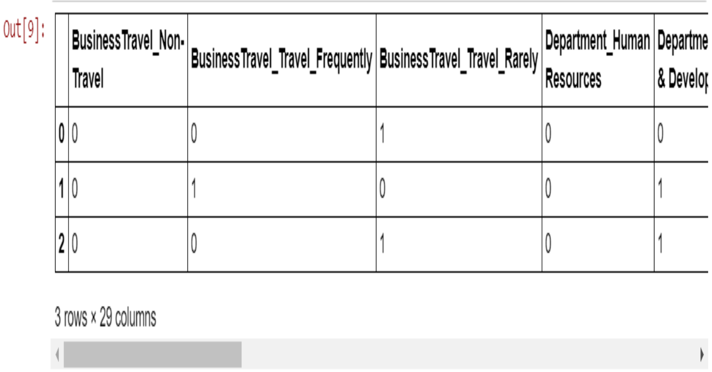
From te correlation plot, it was observed that quite a lot of columns were poorly correlated with one another. Generally, when making a predictive model, it would be preferable to train a model with features that are not too correalted with one another so that we do not have to deal with redundant features.

Figure 8: Correlation Features

From the correlation plots, we are ready to see that quite lot of our columns seem to be poorly correlated with one another . Generally when making a predictive model, it would be preferable to teach a model with features that are not too correlated with one another so we do not need to cater to redundant features. within the case that we've quite lot of correlated features one could perhaps apply a way like Principal Component Analysis (PCA) to reduce the feature space.

4.3 Feature Engineering and Categorial Encoding

Conduct some feature engineering similarly as encode all our categorical features into dummy variables. Having administrated a brief exploration into the dataset, allow us to now proceed onto the task of Feature engineering and numerically encoding the specific values in our dataset. Feature engineering in an exceedingly nutshell involves creating new features and relationships from this features that we've,



Out[9]:

	BusinessTravel_Non-Travel	BusinessTravel_Travel_Frequently	BusinessTravel_Travel_Rarely	Department_Human Resources	Departme & Develop
0	0	0	1	0	0
1	0	1	0	0	1
2	0	0	1	0	1

3 rows × 29 columns

Figure 9: Dummy Variable Matrix

4.4.1 Creating new features from numerical data

Having encoded our categorical columns as well as engineering and created some new features from the numerical data, we can now proceed to merging both data frames into a final set with which we will train and test our models on.

4.4.2. Target variable

One final step that that we have to remember is to generate our target variable. The target in this case is given by the column Attrition which contains categorical variables therefore requires numerical encoding. We numerically encode it by creating a dictionary with the mapping given as 1: Yes and 0 : No. Therefore, we have to keep in mind that there is quite a big imbalance in our target variable. Many statistical techniques have been put forth to treat imbalances in data oversampling or undersampling). In this notebook, I will use an oversampling technique known as SMOTE to treat this imbalance. Implementing Machine Learning models : We actualize a Random Forest and a Gradient Boosted Model after which we take a gander at significant highlights from these separate models. Having played out some exploratory information investigation and straightforward component designing as well as having guaranteed that every single absolute worth are encoded, we are currently prepared to continue onto building our models. As implied in [19][20], we will mean to assess and differentiate the exhibitions of a bunch of various learning models.

a. Splitting Data into Train and Test sets

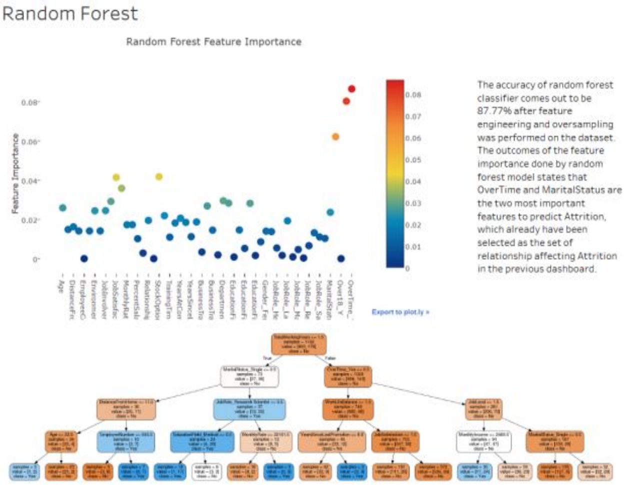
But before we even start training a model, we will have to separate the generated dataset into training and testing data. Since we have just noticed the extreme awkwardness in the qualities inside the objective variable, let us execute the SMOTE strategy in the managing this slanted worth by means of the imblearn Python bundle.

b. SMOTE to oversample due to the skewness in target.

A. Random Forest Classifier

Feature Ranking via the Random Forest

The Random Forest classifier in Sklearn additionally contains an advantageous and most useful trait highlight significances which discloses to us which includes inside our dataset has been given most significance through the Random Forest calculation. Appeared underneath is an Interactive Plotly outline of the different component significances. Visualising tree Diagram with Graphviz, Figure 10 visualise how a single decision tree traverses the features in our data as the DecisionTreeClassifier object of sklearn comes with a very convenient export_graphviz method that exports the tree diagram into a .png format which you can view from the output of this kernel.

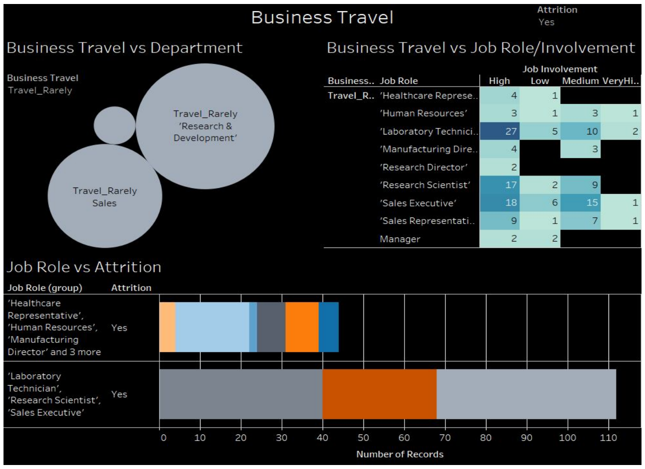


The accuracy of random forest classifier comes out to be 87.77% after feature engineering and oversampling was performed on the dataset. The outcomes of the feature importance done by random forest model states that OverTime and MaritalStatus are the two most important features to predict Attrition, which already have been selected as the set of relationship affecting Attrition in the previous dashboard.

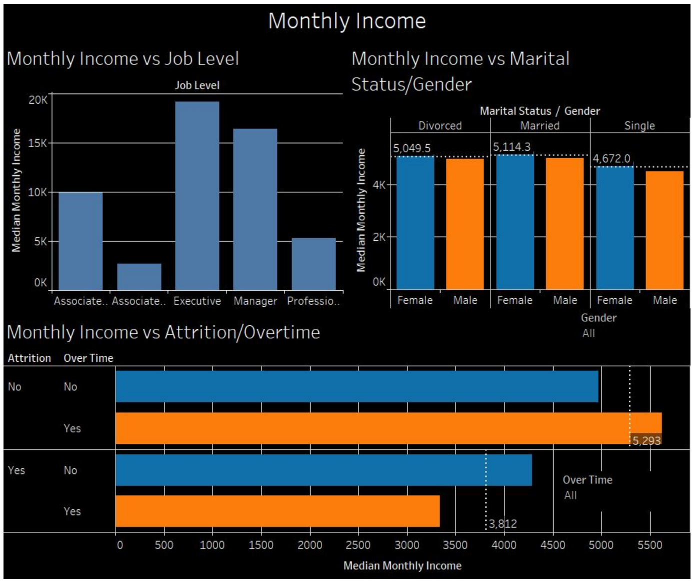
Figure 10: Random Forrest Classifier Outputs

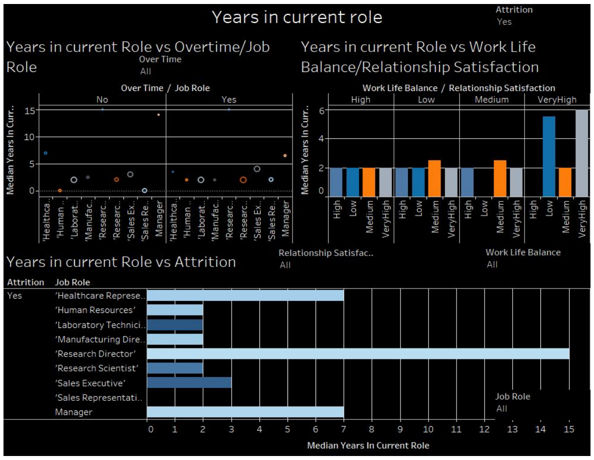
B. Gradient Boosted Classifier

Gradient Boosting is also an ensemble technique much like the Random Forest where a combination of weak Tree learners is brought together to form a relatively stronger learner. The technique[22] involves defining some sort of function (loss function) that you want minimised and a method/algorithm to minimise this. Therefore, as the name suggests, the algorithm used to minimise the loss function is that of a gradient descent method which adds decision trees which "point" in the direction that reduces our loss function (downward gradient). Initialising Gradient Boosting Parameters Feature Ranking via the Gradient Boosting Model[22][23], Much like the Random Forest, we can summon the element significances property of the gradient boosting model and dump it in an intuitive Plotly graph[24].



		Job Involvement			
Business..	Job Role	High	Low	Medium	VeryHi..
Travel_R..	'Healthcare Represe..	4	1		
	'Human Resources'	3	1	3	1
	'Laboratory Technici..	27	5	10	2
	'Manufacturing Dire..	4		3	
	'Research Director'	2			
	'Research Scientist'	17	2	9	
	'Sales Executive'	18	6	15	1
	'Sales Representati..	9	1	7	1
	Manager	2	2		





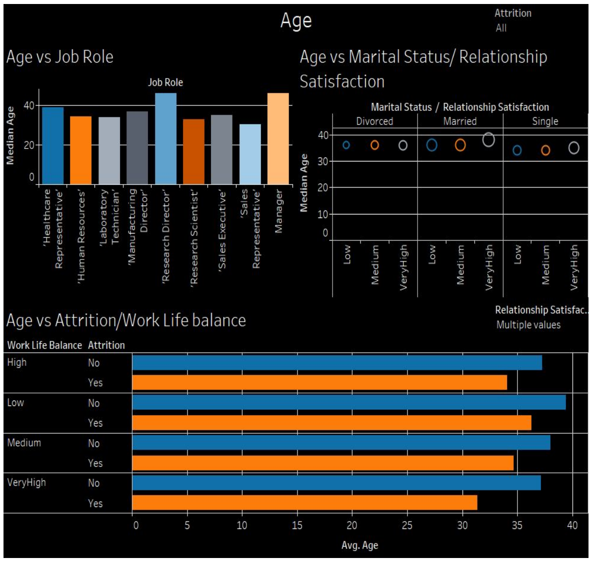


Figure 11: Sample Analysis

5.CONCLUSION

This venture executed prescient examinations on representative steady loss by powerful component choice utilizing a relapse model put together expectation framework with respect to IBM Watson dataset and an organization surveys dataset for opinion investigation utilizing the Vader bundle in python which is a worldwide innovation organization. The Random forest classifier and versatile inclination boosting gave the best assessment scores with finding a good pace exactness [21].

We can outline from the above Literature surveys how the investigation past research work assisted with finding the properties for foreseeing whittling down, know the important worker and select the maintenance factors. Along these lines, we can see that till date there has been diverse speculative and particular research and studies have been done to find the consistent misfortune gauge.

In any case, there has been no basic examination or look at on progress of equipment which can take robotized decisions on requesting significant agents and ordinary specialists. Additionally, there is no application that shows the last dashboard that exhibits the support factors which HR Managers must consider while holding the productive laborer; so the Human Resource Management spending plan can be reduced through and through if standard for steadfastness is extended.

From the investigation it is discovered that the elements liable for turnover of workers are :

•Paying not exactly the business principles.

•Cutting pays. There may come when the organization isn't making benefit.

•Impractical and unfeasible desires.

•Lack of acknowledgment and development openings.

A few factors considered were Attrition, p.c wage Hike,Monthly monetary benefit, Years Since Last Promotion,Distance From Home, Job Role, Performance Rating,Job Level, setting Satisfaction, Years In Current Role, Relationship Satisfaction, Years With Current Manager, Job Satisfaction, WorkLife Balance, scope of partnerships Worked, Years At Company, Over-Time, Total operatingYears, legitimate status, Age and Gender.

As a proposition, the above symptomatic application can be fused with the Human resource the officials cash arranging application and there by envision the general advantage or speculation assets in Human Resource Management process which fuse consistent misfortune, support, enrolling of new delegate, entirety spent on the planning and progression of new laborer and hardship to the assignment due to setback of huge specialist and underwrite on the further moves to be made Thus, association can monitor the proportion of spending it had spent on Human resource the board and spending intend to be spent on future and take imperative exercises.

In any case, this endeavor has a couple of controls. This investigation is limited to a little dataset which needs to set up the model well that may give low results and getting agents data from an affiliation is mystery consequently this assessment is confined to IBM dataset which is the fundamental open dataset on the web. The subsequent drawback is with the model is obliged to simply oversaw AI that requires a lot of count time, sometimes decision cutoff might be over arranged that and customer input is required each time when new features must be incorporated. This endeavor can be loosened up in future as it has a huge number of conceivable outcomes to improve by applying significant learning frameworks with an inside and out arranged arrangement of satisfactory covered layers on tremendous educational assortment which can disguise the limitations of this undertaking. There can be time course of action what's more, design examination which may improve the conjecture execution if the data is in date group.

REFERENCES

1. J. L. Cotton and J. M. Tuttle, "**Employee turnover: A meta-analysis and review with implications for research**", *Academy of management Review,* 11(1), 55-70, 1986.
2. RohitPunnoose **"Prediction of Employee Turnover in Organizations using Machine Learning Algorithms A case for Extreme Gradient Boosting"**, *PhD dissertation Ph.d candidate XLRI – Xavier School of Management Jamshedpur, India Pankaj Ajit ,2*010
3. B. Holtom, T. Mitchell, T. Lee, and M. Eberly, "**Turnover and retention research: A glance at the past, a closer review of the present, and a venture into the future",** *Academy of Management Annals,* 2: 231-274, 2008.
4. N.Silpa , "**Study on Reasons of Attrition and Strategies for Employee Retention**", *Annamacharya P.G college of Management Studies, Rajampet, Andhra Pradesh, India, December 2*015
5. S. L. Peterson, "**Toward a theoretical model of employee turnover: A human resource development perspective**", *Human Resource Development Review*, 3(3), 209-227, 2004.
6. H. Jantan, A. R. Hamdan, and Z. A. Othman, "**Towards Applying Data Mining Techniques for Talent Managements**", *2009 International Conference on Computer Engineering and Applications, IPCSIT vol.2, Singapore, IACSIT Press, 2011*.
7. KarmenVerle and MirkoMarkic," **Process Organization and Employee Satisfaction**" Ph.D dissertation 2010.
8. W. C. Hong, S. Y. Wei, and Y. F. Chen, "**A comparative test of two employee turnover prediction models**", *International Journal of Management*, 24(4), 808, 2007.
9. L. K. Marjorie, "**Predictive Models of Employee Voluntary Turnover in a North American Professional Sales Force using Data-Mining Analysis**", *Texas, A&M University College of Education,* 2007.
10. D.Alao and A. B. Adeyemo, "**Analyzing employee attrition using decision tree algorithms**", *Computing, Information Systems, Development Informatics and Allied Research Journal*, 4, 2013.
11. D. Michie, D. J. Spiegelhalter, and C. C. Taylor, "**Machine Learning, Neural and Statistical Classification"** *Ellis Horwood Limited*, 1994.
12. G. King and L. Zeng, "**Logistic regression in rare events data**", *Political Analy*sis, 9(2), 137–163, 2001.
13. A. Liaw and M. Wiener, "**Classification and regression by randomForest**", *R news, 2(3*), 18-22, 2002.
14. P. Cunningham and S. J. Delany, "**k-Nearest neighbour classifiers**", *Multiple Classifier Systems*, 1-17, 2007.
15. Y. Freund and R. E. Schapire, "**A decision-theoretic generalization of on-line learning and an application to boosting**", *Journal of computer and system sciences,* 55(1), 119-139, 1997.
16. J. H. Friedman, "**Greedy function approximation: a gradient boosting machine**", *Annals of statistics,* 1189-1232, 2001.
17. T. Chen and C. Guestrin, "**XGBoost: Reliable Large-scale Tree Boosting System, 2015**", Retrieved from http://learningsys.org/papers/LearningSys_2015_paper_32.pdf. Accessed 12 December 2015.
18. Apoorv Agarwal BoyiXie Ilia Vovsha Owen Rambow Rebecca Passonneau "**Sentiment Analysis of Twitter Data"** *Department of Computer Science Columbia University,2010*.
19. Rudy Prabowo, Mike Thelwall , "**Sentiment Analysis: A Combined Approach**", *School of Computing and Information Technology University of Wolverhampton* (2009).
20. Ribes, E., Touahri, K. and Perthame, B."**Employee turnover prediction and retention policies design: a case study**", arXiv preprint arXiv:1707.01377 .
21. Quinlan, J.,"**Induction of Decision Trees**", *Boston: Kluwer Academic Publishers 1995.*
22. Morrell, K. M., 2004.,"**Organisational change and employee turnover**" *Emerald Group Publishing Limited.* https://doi.org/10.1108/00483480410518022
23. Lakshmana Phaneendra Maguluri et al. **A New sentiment score based improved Bayesian networks for real-time intraday stock trend classification**. *International Journal of Advanced Trends in Computer Science and Engineering* vol.8 no.4, pp.1045-1055, July-Aug2019. https://doi.org/10.30534/ijatcse/2019/10842019
24. J. C. Mogi, N. P. T. Rahmanto, C. Wiranto, and M. Tuga, **Lending Club Default Prediction using Naïve Bayes and Decision Tree**, *Int. J. Adv. Trends Comput. Sci. Eng., vol. 8,* pp. 2528– 2534, Oct. 2019.