Predictive Human Resource Management Model Using Attrition Analysis

Amit Patti, Anisha Ghosh, Anupama Nhavalore, Saicharan Polishetty

Abstract— The key to the survival, growth and development of an organisation stems from the way it structures and manages its human resources. The major pain points seen in Human Resource Management (HRM) is employee commitment, productivity and retention in the long run. Attrition means the gradual reduction in the number of employees through retirement, resignation, lay-offs or death. Machine Learning and Artificial Intelligence have been widely used to provide predictive analytics in various verticals of an organisation. Human Resources is fairly new to this field of predictive decision-making. HRM is a field dominated by the subjective perspective of every individual involved - employee and employer. Our solution is a predictive model that looks to convert this subjective process into a certain degree of objective analysis purely based on data. The dataset we're using to analyse and predict employees' attrition is provided by IBM analytics. This dataset includes 35 features and close to 1500 samples. The paper talks about a comprehensive comparison study of 6 models, independent and ensemble classifiers. The final model is built using an ensemble model and has an accuracy of 92.52%.

Index Terms—*attrition, classification models, comparative study, data analytics, ensemble models, human resource management, machine learning, predictive model*

1 INTRODUCTION

HR Analytics is the systematic identification and quantification of the people drivers of the business outcomes (IBM Inc., 2017). Essentially, it answers the 5 W's (what, why, where, when, how) of Human Resource Management. By capitalising on the insights acquired, HR analytics helps to promote the organisation and stay ahead of the competition.

Most studies analysing a company's growth focus more on the customer acquisition and sales of their company, leaving out the main assets of their company, the employees. Employee attrition has been a prominent issue in most companies nowadays due to many factors related to work-life balance, job satisfaction, and the like. Existing research showed that employee attrition is mainly due to compensation;[5] the main reason for changing jobs is for a higher salary and better benefits. While attrition cannot be attributed to employees alone, the way the industry is projected and the speed at which the companies are expanding has a major part in attrition.

Overall it has been observed that companies lose out on valuable recruits due to time inefficiencies in the recruitment process, as it's a very competitive space. One of the surveys conducted by Randstad Technologies in 2014 showed that 65% of IT managers were being negatively impacted by skill shortages. [1] This study proved that recruitment/talent acquisition is one of the major tasks in HRM and needs to be optimised further.

In this paper, we will be looking at analysing a company's current set of employees - their work ethic, employee sentiment etc and diagnose why attrition is happening in the first place. The paper also looks to address both involuntary and voluntary attrition. Involuntary attrition is when an employee is forced to leave a company while voluntary attrition is when the employee resigns himself. Our solution looks to aid the HRM industry to create a predictive analyzer to speculate attrition given the attributes as input. They can use these analytics to further optimise their in-house diagnostics and hiring process. They can also suggest and bring about changes in company culture and make specific recommendations to promote the retention of valuable employees.

The paper has been organised as per the following sections: II. Related Works; Literature survey corresponding to our problem statement, III Proposed Solution; Dataset description, preprocessing, descriptive analysis, training and testing, details on model building and methodologies, IV Results and Conclusion; Experimental results of models and inferences made along with concluding remarks have been detailed.

2 RELATED WORKS

Most of the papers discussing attrition analysis have taken a case study approach based on convenience sampling i.e. data taken through surveys of people in a company and such. Descriptive analysis has been done quite extensively but data analytics and diagnoses after acquiring the relevant data is a new area, explored by few.

Data acquisition is an important aspect covered by all researchers as data harvested is possibly private information to companies. A notable technique to farm responses is the Likert's scale. [3] Likert's scale is advantageous as it takes into consideration frequency analysis to evaluate various first-hand sources of information. It's useful in the context of gaining personal opinions in a universal scope using which we can further derive data. Further, Rank Order Analysis is used to make comparisons between various elements of attrition and

sequence them in order of priority on basis of frequency analysis.[4]

Describing attrition and its causes is essential to formulating analytics for it. One such term is employee churn rate i.e. the percentage of employees lost over a period usually associated with overworking[5]. This study [6] explores various factors influencing attrition and retention and provides suggestions based on correlation analysis between these variables.

One of the most challenging tasks for a business is managing the retention of employees by keeping engagement high. Various statistical analyses and procedures give us information about the attrition level and related factors. [8] This paper provides methods to decrease the attrition level by extensive analysis of every attribute that they could obtain. The primary sampling techniques used were the Chi-squared test, Correlation-coefficient and the like. Due to limitations concerning data harvesting analysis in this paper was limited to a surface level. On the contrary, the attributes in our dataset give us more freedom to perform an in-depth analysis.

For the comparative study of existing classifiers in the context of attrition, this paper[9] presented the effect of voluntary attrition on organisations, and why predicting it is important. It further outlined various classification algorithms like Naive Bayes, Logistic Regression, Multi-layer Perceptron (MLP) Classifier, K-Nearest Neighbours (KNN) based on supervised learning to solve the prediction problem. The results of the mentioned paper showed the superiority of the KNN classifier in terms of accuracy and predictive effectiveness, using the ROC curve. This paper only focused on model development based on classifiers as opposed to interpreting the data and its analytics.

To date, most of the human resource analytics is done manually leading to delay in analysing the data to come up with significant and useful results. Automating the initial process of analysing potential reasons for attrition is a new field of work, some of the notable methods that have been used to perform predictive analysis in this field have been mentioned in [10]. This paper uses a Gaussian Naive Bayes classifier as it predicts the greatest number of people who could leave the company by minimising the number of false negatives. To validate these models they used two major error estimation techniques: cross-validation and holdout. After multiple iterations and making changes in the model, results obtained by the proposed automatic predictor demonstrate that the main contributors to attrition are monthly income, over time, age, distance from home.

As a result of analysing related work, the need for a good predictive model for real-time analysis and diagnostics is apparent. Our approach to developing such a model is elaborated in the next section.

3 PROPOSED SOLUTION

Data Analytics using Machine Learning can be powerful in the field of HR Management. Abstracting the subjective process of HRM using the analytics on attrition that our solution provides is our goal. Our solution includes the use of various data visualisation tools to find and retrieve relationships between them.

Currently, all existing solutions [reference related study] dealing with this problem statement deal with it in a unidirectional manner i.e. company cost-cutting attrition or reasons for employee attrition and retainment. Our approach is unique as it considers an employee-company relationship in these ways; effects of an employee's job on their personal life, effects of the company culture and environment on employee's work, and effects of an employee's background(field and level of education) on the job quality and appraisals they receive.

3.1 Dataset Description

The dataset [11] used in this analysis is by IBM analytics, it's a real dataset curated by employees in the company. Our dataset has 1470 samples and 35 attributes that describe employee attrition. The dataset has neither inconsistent data nor missing values.

3.2 Processing

Attributes such as EmployeeNumber, Over18 and StandardHours were redundant. As they're the same for all employees, they were dropped from the dataset. Among the remaining relevant attributes, there are 16 categorical variables and 11 numerical variables.

Attributes such as BusinessTravel were removed as most of the employees aren't involved in travelling for work and it doesn't contribute to the attrition as much. To deal with some prominent multi-collinearity, derived attributes like HourlyRate, MonthlyRate, WeeklyRate were removed and they could be derived from MonthlyIncome. Finally, the dataset had 28 attributes, one of which is attrition; the dependent variable. All the categorical variables are encoded using ordinal encoding as all models used need numerical values for analysis.

Outliers are present in individual attributes, but they carry meaning as attrition itself is seen only in a smaller subset of the dataset.

3.3 Descriptive analysis

The first step is to observe the distribution of the target variable - "Attrition" within the dataset. In the sample of 1470 employees, 16% (237 workers) left their jobs, while the

remaining 84% (1233 workers) are still in service with the company.

The breakdown with respect to the target variable vs other attributes and among the independent variables is summarised below.

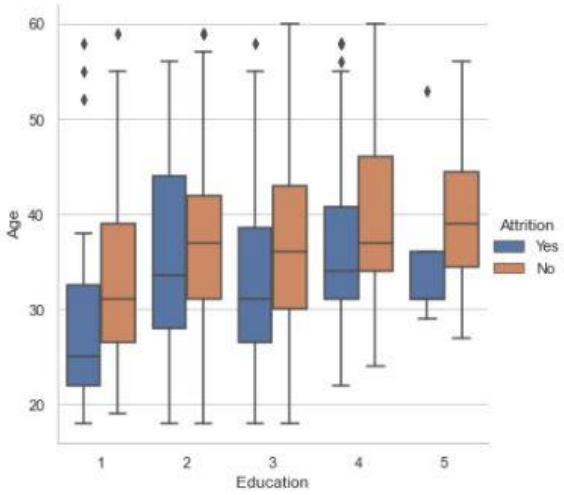

Fig. 1. Education vs Age with Attrition (Box plot)

Employees with higher education tend to be retained for longer. Ones with lower education levels are subject to attrition within 10 years of their joining the organisation. [Fig. 1]

TABLE 1
FREQUENCY DENSITY OF EMPLOYEES GETTING A SALARY HIKE IN EVERY DEPARTMENT

		Count
PercentSalaryHike	Department	
11	Human Resources	0.076190
	Research & Development	0.609524
	Sales	0.314286
12	Human Resources	0.030303
	Research & Development	0.646465
	Sales	0.323232
23	Human Resources	0.035714
	Research & Development	0.714286
	Sales	0.250000
24	Research & Development	0.761905
	Sales	0.238095
25	Research & Development	0.777778
	Sales	0.222222

From the table, we observe that employees in R&D receive a hike in salary on a very regular basis whereas in HR the proportion of people getting a salary hike is significantly lower. Further, at higher salary hikes(24,25) - Sales and HR are both neglected.[TABLE 1]

3.3 Training and Testing

The given dataset was trained using 6 different models spread across independent and ensemble classifiers for a comparative analysis to build a robust model for attrition analysis.

To deal with the high dimensionality we performed PCA to transform the data into 2 components and use it for training using Logistic Regression, the model had an accuracy of 82.3129% after using PCA. Before performing PCA, training the model with all the attributes gave a significantly lower accuracy, hence this paved the way for us to use a better model and lower the dimension using dimensionality reduction techniques like PCA for getting potentially better results.

Naive Bayes Classifier is based on Bayes Theorem which has two fundamental assumptions: independence of features and that all features contribute equally to the model. The priors have been set to None and the smoothing constant is set to 10^-9 to gain calculation stability. Naive Bayes usually performs well with high dimensional data due to the independence assumption of attributes, but in our model due to indirect correlation between the attributes and the fact that some attributes contributed more than the others (further proved using Random Forests) caused it to give us an accuracy of 79.591%. As all the reduced attributes were important, we decided to explore other models instead.

Based on our research, we found that SVM is very effective for training high dimensional data as it uses a hyperplane to separate the data points which are essentially vectors in the n-dimensional space. Moreover, it uses only the support vectors to construct the decision boundary hence, it's time and space-efficient. The parameters for modelling SVM; the C value was set to 1 while the gamma value was set to scale according to the

dataset. The kernel used here is the RBF kernel. The approach gave us only 83.333% accuracy. As the dataset was imbalanced, SVM cannot construct a good decision boundary to classify the target classes, hence, it yielded an accuracy comparable to the Logistic Regression model even though it was a more regularised model.

The final classification model we implemented was a KNN model. The optimal 'k' value was found using the elbow method where the model was trained across a range of k values between 1-19 and the error rate was measured, the optimal k value found was 9. There was significant work done to reduce the dimensionality and use oversampling techniques but it did not improve the accuracy significantly. The final model had an accuracy of 82.993%.

Among all these models, SVM gave the best results, but these results were still not good enough compared to the related works in this field, this implied that we had to introduce a new perspective to understand the underlying patterns in the data.

Ensemble models for classification are useful as it is simple to implement, very robust, and easy to tune the hyperparameters as it uses simple multiple weak learners. We trained the dataset using the XGBoost model and a Random Forest classifier.

XGBoost is an implementation of gradient boosted decision trees. XGBoost uses gradient descent to minimise the loss function which is predicted using the Taylor expansion, instead of fitting the current hypothesis $h_m(x)$ on the residuals, it fits it on the gradient of the loss function, where m denotes the iteration number. It essentially does not explore all the possible trees, rather it builds a tree using the greedy method, also preventing overfitting by penalising trees that are too deep using regularisation. Every $h_m(x)$ is multiplied with γ, which accounts for the difference in the impact of each branch of the split.

For our dataset, this algorithm is able to deal with the imbalance in our dataset by configuring the class weight in the dataset and configuring the hyperparameters. For our model the regularisation terms alpha and beta were set to 0.1 and 0.05 respectively, the number of estimators used was 80, with a maximum depth of the individual trees limited to 1, and subsample ratio was set as 0.6. The model gave an accuracy of 87.074% [Fig. 2] which was higher than the independent classifiers.

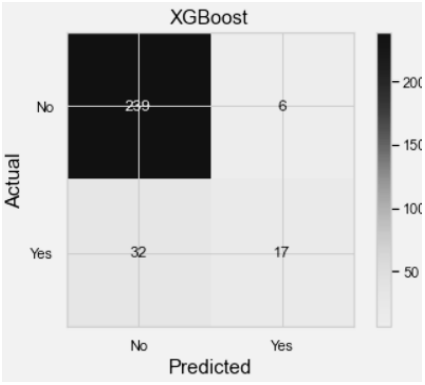

Fig. 2. Confusion matrix for XGBoost

The final ensemble model that was implemented was the Random Forest classifier. Random Forest classifiers work on aggregating independent, uncorrelated diverse decision trees and ensembling them to compensate for errors in each other to give optimal classification [Fig. 3].

It uses bagging and randomness to build each tree and selects features based on Gini impurity or Information Gain. The impurity decrease across features can be averaged across trees to find the final importance of features.[TABLE 3]

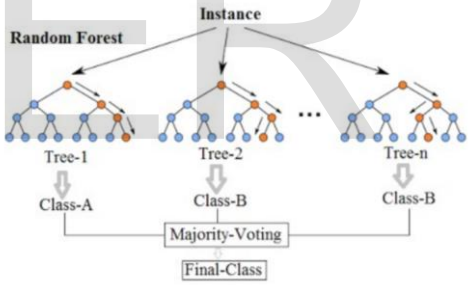

Fig. 3. Random Forest classifier [12]

Random Forest= DT(base learner)+ Bagging(row sampling with replacement)+ feature bagging(column sampling) + aggregation(mean/median, majority vote) (1)

TABLE 3

IMPORTANCE OF FEATURES AS INFERRED FROM RANDOM FOREST

feature	importance
MonthlyIncome	0.088
Age	0.071
DistanceFromHome	0.062
YearsWithCurrManager	0.053
PercentSalaryHike	0.052
YearsAtCompany	0.048
JobRole	0.044
TotalWorkingYears	0.042
OverTime	0.040
JobSatisfaction	0.039
YearsSinceLastPromotion	0.038
YearsInCurrentRole	0.038
EnvironmentSatisfaction	0.037
NumCompaniesWorked	0.037
Education	0.036
TrainingTimesLastYear	0.034
WorkLifeBalance	0.032
StockOptionLevel	0.031
JobInvolvement	0.031
RelationshipSatisfaction	0.030
EducationField	0.030
MaritalStatus	0.028
JobLevel	0.028
Gender	0.014
Department	0.011
PerformanceRating	0.005

In our model, we used 10 estimators given the size of our dataset and inferred from the importance levels to decide which features had to be given more weight. We also changed the hyperparameter random_state to 0 to make sure the outcome for the classifier is consistent across runs i.e. it gives outputs in the same league of values, given that it would be working with a different set of decision trees every time it's run. This model was one of the best models with an accuracy of 92.52% [Fig. 4]. Given our dataset is unbalanced with respect to the target variable, Random Forests being an ensemble of decision trees was able to deal with this imbalance and produce the best model to predict attrition.

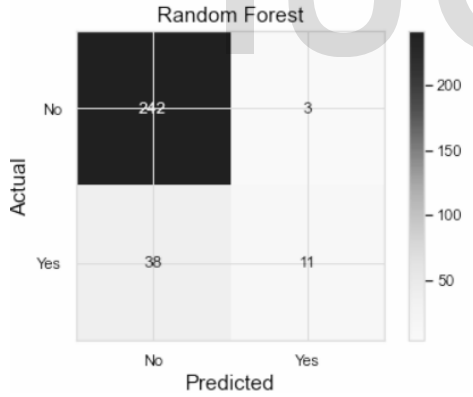

Fig. 4. Confusion matrix for Random Forest classifier

4 RESULTS AND CONCLUSION

In this paper, our proposed solution looked to make a robust predictive model which would serve as the underlying model for analysis in the HRM industry. In the process, we compared 6 models to build the best predictive model.

TABLE 3

PERFORMANCE METRICS

Performance Metrics/Model	Logistic Regression	Naive Bayes	SVM	KNN	XGBoost	Random Forest Classifier
Accuracy (%)	82.3129	79.5918	83.3334	82.9932	87.0748	92.5200
Precision	0.8339	0.9186	0.8334	0.8328	0.8892	0.8643
Recall	0.9837	0.8286	1.0	0.9959	0.9755	0.9878
F1 score	0.9026	0.8712	0.9091	0.9071	0.9264	0.9219
Time taken for execution(s)	0.2322	0.01501	0.1681	0.0070	90.6854	0.1661

As seen from [TABLE 3], Random Forest Classifier and XGBoost gave us the best accuracy values. These being ensemble models also show that they're not overfitting the data and account for the reproducibility of results for any test dataset. However, we see that XGBoost has significantly more time to train as it's a boosting model that builds decision trees sequentially.

Naive Bayes yields the lowest accuracy due to independence and equal contribution assumption of attributes which is not valid with regards to our dataset as seen from the importance features table given by the Random Forest Classifier. [TABLE 2]

Logistic Regression yielded good accuracy as it worked on the principal components; it, however, failed due to the imbalance in the dataset and some hidden correlation amongst the features. SVM also fell short due to the imbalance in the dataset as seen from the high recall and f1 score values and caused some misclassifications. SVM in a better-balanced dataset however is bound to give much better accuracy values.

Even though KNN yields a high accuracy, due to the curse of dimensionality it does tend to have a bias towards the majority class as for KNN to work the point needs to be very close to the instance class in every dimension. We also note that KNN takes the shortest time to train as it doesn't actually train and directly plots the values with classes and tests immediately.

Overall, we see that the ensemble models performed much better than the classification models due to the nature of the dataset being imbalanced in terms of attrition (yes: no) which the ensemble models were able to deal with as a function of using multiple decision trees in combination.

In conclusion, given the different models we have analysed some of the other models like SVM, and XGBoost can work better with a more balanced dataset. The attributes that contribute the most to attrition [TABLE 2] are monthly income,

age, distance from home, years with current manager, percent salary hike, years at the current company, and Job role.

The Random Forest Classifier being the most accurate model, this can be used for making relevant tools in the HRM industry for attrition analysis. The model is easy to train and it can be scaled in the future with more data as data can change with time. The model doesn't need a lot of parameter tuning, it's simpler to visualise and improve with changing times. Given the comprehensive analysis that is used to predict attrition, we can positively say that we can create some improvement in the HRM industry by automating the process of attrition analysis.

ACKNOWLEDGMENT

The authors wish to thank Dr. Gowri Srinivas for her unending support in this paper.

REFERENCES

[1] https://www.randstadusa.com/staffing-and-solutions/salary-guide-2014/confirmation/it-salary-guide-confirmation/technologies-sg-2014.pdf

[2] https://www.concentra.co.uk/resources/research-report/strategic-workforce-analytics-research-report/

[3] Karan Hiren Bhalgat,(2012) "An exploration of how Artificial Intelligence is impacting Recruitment and Selection process"

[4] Augustin, Prince & Mohanty, R.. (2012). "A diagnostic study of employee attrition in an Indian automotive company. Int. J. of Indian Culture and Business Management." 5. 593 - 612. 10.1504/IJICBM.2012.048773.

[5] Jaya Sharma (2015), "Employee Attrition And Retention In A Cut-Throat Competitive Environment In India: A Holistic Approach"

[6] M.S.Kamalaveni , S.Ramesh , T.Vetrivel.,2019, "A Review Of Literature On Employee Retention"

[7] Mike Johnson (2004) The new rules of engagement.

[8] N.Silpa "Int. Journal of Engineering Research and Applications" www.ijera.com ISSN: 2248-9622, Vol. 5, Issue 12, (Part - 1) December 2015, pp.57-60

[9] Rahul Yedia, Rahul Reddy, Rakshit Vahi, Rahul J, Abhilash, Deepti Kulkarni, "Employee Attrition Prediction"

[10] Fallucchi, Francesca, Marco Coladangelo, Romeo Giuliano, and Ernesto William De Luca. 2020. "Predicting Employee Attrition Using Machine Learning Techniques" Computers 9, no. 4: 86. https://doi.org/10.3390/computers9040086

[11] https://www.kaggle.com/ghoshanisha/ibm-employee-attrition

[12] https://en.wikipedia.org/wiki/Random_forest