National College of Ireland

Why an Employee Leaves: Predicting using Data Mining Techniques

MSc Research Project
Data Analytics

Tanya Attri
x16146344

School of Computing
National College of Ireland

Supervisor: Dr. Paul Stynes

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	Tanya Attri
Student ID:	x16146344
Programme:	Data Analytics
Year:	2018
Module:	MSc Research Project
Lecturer:	Dr. Paul Stynes
Submission Due Date:	13/08/2018
Project Title:	Why an Employee Leaves: Predicting using Data Mining Techniques
Word Count:	6948

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	13th August 2018

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Contents

Why an Employee Leaves: Predicting using Data Mining Techniques

Tanya Attri
x16146344
MSc Research Project in Data Analytics

13th August 2018

Abstract

HR analyticss is the area of data analyticss helping the organization to understand its employees. Today companies face employee attrition as the major issue effecting the productive functioning of the organization. HR analyticss has enhanced the area of data analytics to an extent that organizations can figure out their employees' characteristics; where inaccuracy leads to incorrect decision making. Data mining is helping the HR department with methods to evaluate the historical data and predict the employee attrition, the baseline for this research. By far, employee attrition is predicted with the suggestions of the domain experts and the use of the classification methods by technical researchers. This research aims to investigate the extent to which ML techniques can help in predicting the employees who might leave, using the optimal hybrid ML models, where oversampling technique (SMOTE) & feature selection technique (SA) are integrated with the classication algorithms such as SVM & LR. The focus is towards the true positive accuracy predicted by the models. A comparison of these results was done between the features selected by the models in this research and the ones listed by the domain expert researchers in the past to see which one has the more reliable outcomes, concluding that the management expert should use the technical methods for their analysis in future to have reliable outcomes. This will help the HR system to adopt the right scenarios in real time & correctly predict the potential employees to leave & know why they do so.

1 Introduction

Human resource is the backbone of any company. Realizing the importance of this, from ages the companies have been investing in selecting & training these resources to compete with the growing market. The hiring process involve the investment of time & money, thus leading to a major loss if any of these resource leaves the organization.

To overcome these loses it is important to not only understand the employees' needs but their behaviour as well. This gives rise to the area of HR Analytics being adopted now by a lot of companies to study the past behaviour of the employees, understand the patterns and thus come up with the possible efficient strategies. HR Analytics is an area of data mining which uses the techniques such as predictive modelling and classification and has helped the companies to avoid the employees attrition (Singh et al.; 2012).

Data mining is "the process for extracting knowledge from data and hence identify and predict the future outcomes based on the past patterns" (Alao and Adeyemo; 2013; Chien and Chen; 2008; Tomassen; 2016). Similarly, this research uses such data mining techniques on the anonymous employee data from IBM to identify the employees which are most likely to leave and determine the factors that influence an employee to do so. The employee attrition has a huge hindrance in the growth & development of an organization. The study of these factor will eventually help the organizations to understand their employee's behaviours and hence take actions to retain them (Singh et al.; 2012; Chien and Chen; 2008).

Various researches have been done in the past where researchers have used data mining techniques to predict the employee attrition using classification methods like random forest & decision trees. Although these past works have focused on the overall accuracy of the models, there has been no major work done in the area of accurately predicting the employees who actually left, giving a motivation to this work. Thus, this research focuses on building the optimal hybrid machine learning models which can accurately predict the employee who might leave, i.e., get a better true positive accuracy. Correct classification of these employees will help the organization to do better decision making for hiring, allocating & managing the future resources, thus it is important to fill the gaps of the research in past to avoid incorrect decision making.

Not only this, it is also very important for an organization to identify the reason for which an employee leaves. Thus, encouraging this research to create a feature selection model which will list out all the possible attributes responsible for the employee attrition. Some of these features were listed out by the management domain experts in the past, thus this research also ensures to have a comparison between the technical & the management researchers to find out the reliable methods.

Nagadevara and Srinivasan (2008) mentioned in their research that there is an improvement in accuracy with the use of hybrid models compared to single algorithms, driving this research to work with hybrid models. Considering this and all of the above, the research question, around which this research revolves is: ***"To what extent can hybrid framework of feature selection and classification algorithms help in efficient prediction of employee attrition?"***

The benchmark for the above is the research conducted by Yiğit and Shourabizadeh (2017) where the authors have used the same dataset of IBM Watson HR Analytics and have come up a good accuracy of the model (highest 85% for SVM), but have a very poor sensitivity value, with the highest recall of 37%. Also, no class imbalance was addressed in this research. This gap is tried to be filled up in this research.

To conduct this research, the IBM Watson - HR analytics data [1] was used on employee attrition, which has 35 attributes with 1471 rows of data, which is the major limitation of this research. Due to the lack of public data on employees to maintain the privacy & confidentiality of the person, it is now very hard to find the suitable data for the analysis. However, methods have been used here to try to overcome this limitation up to an extent.

This report consists of five more sections. The relevant work that was done in the past in the field of data mining as well as employee attrition have been reviewed in the Section 2. The methodologies & implementations done in this research are discussed in the Sections 3 & 4, followed by the evaluation in Section 5. Ultimately, the report is concluded with the mention of the possible future work in Section 6.

[1] `https://www.ibm.com/communities/analytics/watson-analytics-blog/hr-employee-attrition/`

2 Related Work

2.1 HR Analytics & Employee Attrition

HR Analytics includes the study of the employee's behaviours to figure out the major reasons responsible for their actions. This domain is gaining a lot of importance in companies today as it is helping them to maintain the human resource which is the building block for their growth. HR analytics has enhanced a lot, helping organizations to learn the major reasons which lead to the attrition by the employees. The management focuses on the data like employee age, gender, their experience in work environment, and more to predict their behaviors in the future events.

Employee attrition has a huge impact on the functioning of the organization. As argued by Singh et al. (2012), both voluntary as well as the involuntary employee attrition, is a major loss to the organization, affecting the company's productivity, causing delays in project deadlines, increase in poor services, eventually leading to disappointed clients. This ultimately leads to the money losses as these resources need to be replaced by hiring new ones and training them again. IT sector has experienced 12-15% of employee attrition rate, causing a money loss of roughly up to "1.5 times the annual salary of the employee who left" (Saradhi and Palshikar; 2011). Even a lot of time is misspend in looking for right replacements, organizing interviews and hire & train them. Even after all this, these replacement do need their own time to settle into the new organization's structure & culture before they can actually produce fruitful results for the organization. These attrition are more troublesome when someone on a higher level leaves, as it is hard to replace someone with such experience both in work as well as to understand the organization in & out. With the loss of money, time & the even resources, the attrition leads to a "Snowball Effect", influencing the other employees to leave as well (Rashid and Jabar; 2016).

The reasons for an employee to leave can be, either because he is getting a better offer or a better career growth, the positive reasons, or can be the unpleasant reasons like environment, manager issues, and many more, and the other negative reasons (Saradhi and Palshikar; 2011). It is suggested by the management experts to retain these employees by providing counter offers like better salary, promotions, etc (Zhou et al.; 2016).

All these losses and issues generated due to the employee attrition, makes it very important to have a solution based system to predict employee attrition so as to avoid it, which can then be used by higher officials as a back up plans to avoid sudden business road blocks (Rashid and Jabar; 2016). This is where we bring technology into the picture.

2.2 Data Mining Methodologies

Data mining is the "process to draw out knowledge from the data available from past events" (Alao and Adeyemo; 2013). Data mining helps in gathering information from the patterns shown in the data. To do so, many researchers have used various machine learning algorithms to run the analysis on the basis of clustering, association, classification and prediction. Methods such as neural networks, statistics, visualization techniques, decision trees, and many more, are all that fall under the category of machine learning algorithms (Chien and Chen; 2008).

Inspired from the various prediction methods that data mining has, a lot of them were used in this research work such as correlation to test the relationship between the features & PCA to test if there are any clusters in the data. Prior to applying PCA, it is

important to check if the dataset is suitable for factor analysis for which the Bartlett's test of Sphericity, which checks the correlation coefficient in the data, and the Kaiser-Meyer-Olkin's test, which checks the "sampling adequacy" of the dataset, should be performed (Marsham et al.; 2007; Tabachnick and Fidell; 2007).

The attributes in a dataset also plays a vital role in building good models for prediction. It is not accurate to use all the features while building up the model as this increases inaccuracy, whereas using each attributes separately is not practical due to the "computational volume", hence the feature selection algorithms comes into picture (Meiri and Zahavi; 2006). Simulated Annealing (SA), used in this research, is inspired by the work done by Debuse and Rayward-Smith (1997); Lin et al. (2012) where both have used prediction models with SA and concluded a better accuracy with the subset of the features identified.

To test the models created for the analysis, the ideal approach of data mining is to distribute the data into groups of test and train data and then run the analysis on the test data based on the observations from the trained data (Nagadevara and Srinivasan; 2008). A similar approach is followed in this research as well as some of the other research done in the past in the area of HR Analytics. Also, there are some tuning parameters in the classification algorithms that need to be tuned to give best results, which are known as the hyperparameters [2], which are hard coded, hence by tuning them, they can give better results. Bayesian optimization, used in this research helps in finding the best hyperparamter value with least error, thus a powerful technique to give significant improvement in the results (Shahriari et al.; 2016).

Before moving into more detail about the research done, it is good to have a look at some of the past work done in the field of HR analytics, not only for the employee attrition, but for other areas as well, with the use of the data mining.

2.3 Data Mining applied to HR Analytics

With the advancement of technology and the science of getting knowledge from data, organizations today has started to gather similar data for their employees as well, and use them with data mining techniques to understand the employee's characteristics. Many of these data mining techniques have been used in past by the researchers to gain the understanding of the behaviours recorded for the employees in an organization. HR analytics is a growing domain, provoking many researchers to explore this area and implement the machine learning algorithms such as decision trees, ANN, and so on, to generate required models and get relevant results.

HR analytics is now enhanced with the use of the prediction techniques of data mining to help organizations for building team of employees who are satisfied with the place & its structure (Mishra et al.; 2016). Employee attrition is indeed the most important area to study in HR Analytics. A predictive analytics approach is mentioned by Mishra et al. (2016) for prediction of the employee attrition.

Regardless of the need to study employee attrition, data mining has also played a major role in the hiring of the candidates. The research of Chien and Chen (2008) has also used data mining in HR analytics to create a model for selecting the future candidates in the organization, managing employee career & job roles focusing on their growth. Rabcan et al. (2017)'s research also focused on employee selection model with the use of C5.0 decision tree and gave an accuracy of 97.27%. Snyder (2016) researched

[2] `http://busigence.com/blog/hyperparameter-optimization-and-why-is-it-important`

on a "Talent Management tool" to check the suitability of the candidate by keeping the candidate's application records & gathering data from web crawling, and then to see whether the person deserves a promotion. Another research to help the hiring systems in HRM was done by Kumar et al. (2017), where the machine learning models were used to rank the resumes of the candidates. The categorization of the jobs posted online by the companies into the big data domain was done in the research of De Mauro et al. (2018) into technical & managerial profiles. Similar work was done in the research of Boucher and Renault (2015) to categorise the jobs, based on the LinkedIn job summary by using NLP & prediction. Not just this, Ramamurthy et al. (2015) built a predictive model to understand their employees in a way that ultimately the companies know which one among them needs to be trained on which skill.

Singh et al. (2012) has used employee attrition with techniques such as decision tree and C5.0 algorithm. On the other hand, Alao and Adeyemo (2013) used the techniques of Artificial Neural Network (ANN), Memory Based-Reasoning Models, Regression analysis, Decision trees, Rule induction, Case Based-Reasoning Models, Clustering and Association rule & correlation on the data of the South-West Nigeria staff of 309 staff members of past 28 years to predict the employee turnover. Nagadevara and Srinivasan (2008) created a model to predict employee attrition using methods such as ANN, C5.0 decision tree, logistic regression and discriminant analysis, where DA outperformed all with the accuracy of 86.84%. On the other hand, the unsupervised machine learning methods - k-means was used by Zhou et al. (2016) where Rombaut and Guerry (2017) used decision tree & LR for the prediction of the employee turnover.

Correlation has been used in the past to see the relationship between the employee attrition and other attributes like employee satisfaction and demographics attributes, which is a very useful technique to find out the relationships between the predictor and the predicting variable, and also between the different predictor variables in a dataset (Sengupta; 2011; Harter et al.; 2002).

Many feature selection methods were also used to identify the attributes responsible to influence employee's decision for attrition such as age, gender and other demographic characteristics. In the research conducted by Yiğit and Shourabizadeh (2017), which is the benchmark for this research, the author used Recursive Feature Elemination method and compared the various classification models to predict employee attrition, hence concluding that the accuracy is better when models are analyzed with feature selection method. Rashid and Jabar (2016) discussed in his research that it is very important to have a model which is intelligent enough with its predictions, hence used "Fuzzy Rough Set Theory" to identify the features responsible for the attrition. Factor analysis - Principal Component Analysis (PCA) was used by Adhikari (2009) whereas Cluster analysis - Self-organizing map (SOM) was used by Fan et al. (2012) giving a model with accuracy of 92.7%. An "Intelligent Human Resource Management System" was built by Cahyani and Budiharto (2017) which can predict the future status of the employee in regards to the turnover based on the employee data attributes like age, gender, birthdate, and more, with SVM, LR, RF & Adaboost.

The research work done with the techniques used in past for HR analytics & employee attrition which have motivated this work are all listed below in Table 1.

Authors & Year	Methodologies Used
Singh et al. (2012)	Decision Tree, C5.0 algorithm
Alao and Adeyemo (2013)	Memory Based-Reasoning Models Regression Analysis Decision trees, Rule Induction, Correlation Case Based-Reasoning Models Clustering, Association Rule
Nagadevara and Srinivasan (2008)	ANN, C5.0 decision tree, LR Discriminant Analysis
Zhou et al. (2016)	k-means
Rombaut and Guerry (2017)	Decision tree, LR
Yiğit and Shourabizadeh (2017)	Recursive Feature Elimination
Rashid and Jabar (2016)	Fuzzy Rough Set Theory
Adhikari (2009)	Principal Component Analysis
Fan et al. (2012)	Self-Organizing Maps
Sengupta (2011) Harter et al. (2002)	Correlation
Cahyani and Budiharto (2017)	LR, RF, SVM, Adaboost

Table 1: Data mining methodologies used in HR Analytics

2.4 Attributes leading to Employee Attrition

Not just prediction of employees who leave, data mining is also capable of finding out the features which are responsible for attrition of the employees. Some of these features cross over with the ones given by the management domain experts based on their experiences and observations.

To begin first with the technical list of features, the research of Alao and Adeyemo (2013) mentions that the employee demographics data and job-related attributes are the features that affect employee attrition, including the salary and the job length. Demoraphic attributes were also researched by Nagadevara and Srinivasan (2008) including the factors like employee absenteeism and late coming, whereas Rombaut and Guerry (2017) looked into the work specific factors.

From the management point of view, a research by Pande and Chung (2017) hypothesized on the factors like working hours, salary, family & health problems, concluding that employees change jobs quickly and are happy to have monetary benefits. Mihajlović et al. (2008) discussed in their the research the influence of work environment on the job satisfaction of the employees. Some of the major studies done in the past by the management experts to understand the reasons for the employee turnover was reviewed by Allen et al. (2010). More workload and excess business travels leads to work stress which ultimately leads to employee attrition (Avey et al.; 2009).

In Table 2 all the attributes observed both by management experts and technical researchers are listed.

2.5 Background on IBM Employee Attrition Data-set

With the establishment of the data governance policies, it is very important to make sure that we maintain the privacy of the data of the candidates, the customers or even the

	Authors & Year	Attributes Given
Technical Summary	Alao and Adeyemo (2013)	Demographics, Salary, Job-length
	Nagadevara and Srinivasan (2008)	Absenteeism, Late-coming
	Rombaut and Guerry (2017)	Work-specific factors
	Sengupta (2011) Harter et al. (2002)	Employee Satisfaction
	Cahyani and Budiharto (2017)	Age, Tenure, Department, Gender, Health Status, Skills
Management Summary	Pande and Chung (2017)	Working hours, Salary, Employment years, Family & Health problems
	Mihajlović et al. (2008)	Work-environment, Job satisfaction, Job Involvement, Work pressure, Career growth, Education & training, Relationship with manager, Performance rating, Relationships with other team mates, Work-life balance.
	Allen et al. (2010)	Role Clarity & Conflicts, Promotion Opportunities, Job Scope, Tenure, Age, Marital Status, Gender, Organizational Commitment, Race, Work Stress, Job Previews
	Avey et al. (2009)	Work Stress, Work Overoad, Increase business travel
	Batt and Valcour (2003)	Work-Family Balance, Overtime, Bonuses, Job security

Table 2: Attributes responsible for Employee Attrition

employees. Thus, this makes it very difficult to have a real-time data that can be useful for any analysis. To overcome this, the IBM Watson data-set on employee attrition has been used in the past by many researchers to conduct their analysis on employee attrition and for this research as well.

One such research, which is the benchmark for the work here, is conducted by Yiğit and Shourabizadeh (2017), where the authors have shown the importance of introducing the feature selection methods with the classification models to see if there is any boost in the accuracy. Their experiment was done with decision tree, Naive Bayes, LR, SVM, KNN & RF, both with & without the RFE - feature selection method, getting the highest accuracy of 89% with RFE-SVM model. Although, the overall accuracy was improved for almost all of the models, but the recall/sensitivity/true positive accuracy was not good enough. Even the best model RFE-SVM had a recall value of 37%, which means their model is not fully capable of predicting the people who actually left. Also, the problem of class imbalance of this data-set was not referred to anywhere in their work.

The same IBM dataset was also used in the work of Frye et al. (2018), with two other sources of data, where the authors has predicted employee attrition figuring out

the factors responsible for the same. PCA with k-means, Random forest & LR was applied where LR outperformed with an accuracy of 74%. Although the accuracy seems good, but the work does not have any mention of the sensitivity of the model. Another work on the same data-set was done by Barvey et al. (2018), where the author has used feature engineering to increase the attributes of the data to have reliable business usable.

A class imbalance occurs when one of the category in the data has very less number of observation compared to the other and this, hence, is one of the issues to have a better accuracy in the models (García et al.; 2012). This problem is observed in the dataset used for this research as well and after getting motivated from the work of Han et al. (2005) & El-Sayed et al. (2015), Synthetic Minority Over-sampling Technique (SMOTE) is used to deal with the imbalance in data.

Another limitation of the dataset, which is the less number of observation (1471) is not addressed by any of the research in the past. Whereas, this research uses k-fold cross validation as this technique is ideal to be used with small data by taking a large value of k (Yadav and Shukla; 2016).

3 Methodology

When building up a research work, it is very important to start with a road-map. For building up this plan, the data mining is always divided into various phases with some plan of actions to be achieved at each phase. Also, to have a successful data mining project, it is important to have a "standard methodology with good research outcome" [3]. A similar process model is used in this research, which is CRISP-DM. This approach is adopted by many researchers in the past to get expected results. It is a "hierarchical process model" with small phases, where each phase comprises of list of generic tasks (Wirth and Hipp; 2000).

"Cross-industry process for Data Mining" is ideal for data mining research projects. It gives a detailed understanding of the flow of the tasks. This is because, each phase gives a clear & detailed understanding of each task that is required to be done to achieve the end result. In the Figure 1, the phases of CRISP-DM are shown diagrammatically and also listed below.

- Business Understanding
- Data Understanding
- Data Preparation
- Modelling
- Evaluation
- Deployment

[3] https://www.datasciencecentral.com/profiles/blogs/

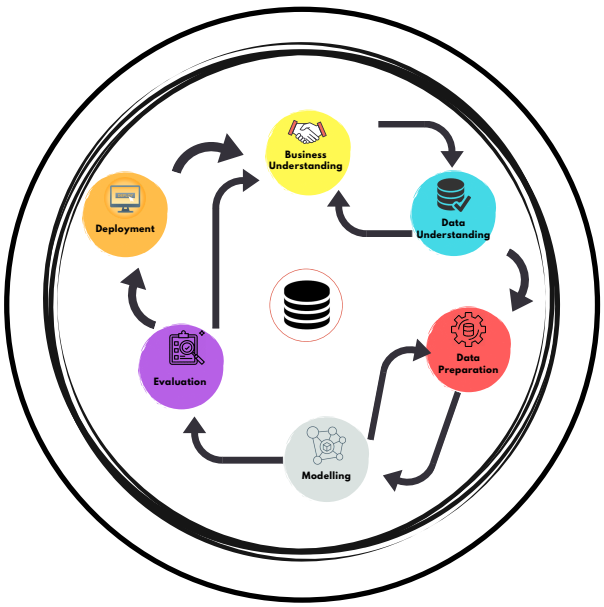

Figure 1: CRISP-DM - Proposed Research Methodology

3.1 Business Understanding

Business Understanding is the first phase of CRISP-DM approach. The major task at this phase is to choose and understand the business objective of the research. This research is motivated with the issue faced by the organizations due to the employee attrition, thus the business objective is to help the organizations to predict the future attrition on the basis of the actions of the employees in the past.

The business objective of this research is described in the first section in this paper. The another task in this phase is to come up with the questions about the results & contribution the research is going to make. This research will help the higher management to take better decisions and create preliminary plan of actions to avoid the sudden attrition of resources. Not just this, but the target is to provide the management with the accurate prediction of employees who have the highest chances of leaving, thus focusing the research towards the true positive accuracy.

3.2 Data Understanding

This second phase of data understanding includes the tasks of gathering and understanding the data that will be used in the analysis. For this research, a dataset of IBM Watson was used with 35 attributes and 1471 observations. This data was carefully selected after going through different publically available datasets on the open sources.

It was also important to maintain the data privacy, the reason why a real-time data was not used. But keeping this in mind, the IBM data was chosen as it has the attributes which can be related to the real-time scenarios. This has such a data which is easily available with the HR department in any organization or can be related to.

3.3 Data Preparation

Data comes from various raw sources, which means it might contain noise as well as irrelevant information. It is important to refine this data so that it can be suitable for

the models and generate better results. Once the data is selected, the third phase is to prepare this data. This phase includes tasks like cleaning, transformation and removing the unwanted data.

Below are the tasks that were performed in this phase for this research:

- For this research, the data for the employee attrition had various attributes which were not relevant, i.e. was not giving any useful information, like Employee Number, Employee count, etc., hence these attributes were removed in the process of data cleaning. The below Table 3 shows the all the final attributes of this dataset which were obtained after removing the unwanted ones.

S.No.	**Attribute**
1.	Age
2.	Attrition
3.	BusinessTravel
4.	DailyRate
5.	DistanceFromHome
6.	EducationField
7.	EmployeeCount
8.	EnvironmentSatisfaction
9.	Gender
10.	HourlyRate
11.	JobInvolvement
12.	JobLevel
13.	JobSatisfaction
14.	MaritalStatus
15.	MonthlyIncome
16.	MonthlyRate
17.	NumCompaniesWorked
18.	OverTime
19.	PercentSalaryHike
20.	PerformanceRating
21.	RelationshipSatisfaction
22.	StockOptionLevel
23.	TotalWorkingYears
24.	TrainingTimesLastYear
25.	WorkLifeBalance
26.	YearsAtCompany
27.	YearsInCurrentRole
28.	YearsSinceLastPromotion
29.	YearsWithCurrManager

Table 3: Final Attributes Used in this Research

- Data transformation was done, i.e. the attributes were correctly type-casted. The categorical data was converted into factors. As this dataset had a lot of categorical variables, like, the variable Attrition has the Yes and No values, and all the rating

attribute had Likert scale varying from 1 to 5, all of these attributes were type-casted to factors.

Once these tasks were done, the dataset was ready to be analyzed for further test.

4 Implementation

Implementation is the fourth phase of the CRISP-DM approach. All the models are built in this phase & executed. Before beginning with applying the models to the data-set, a process flow was planned & followed in this research as shown in the Figure 2. This is the process flow architecture of the presented research.

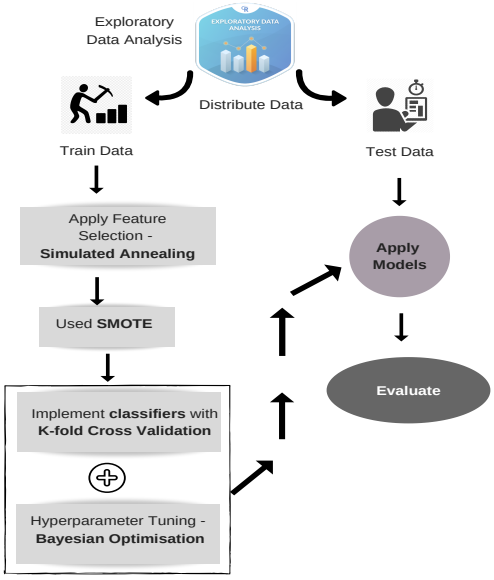

Figure 2: Process Flow Architecture of the Research

The implementation phase was initiated with doing an exploratory data analysis. This is used to summarize the data and its characteristics. This helps in exploration of the data and to see what we have in the data and to get the best out of it.

For this research, several methods were executed to explore the data as are mentioned below:

- **Correlation:** A statistical test to check if the attributes have any correlation among each other was done with the correlation matrix. A data-set with good attributes should not have a correlation among themselves. Correlation matrix was used to find the highly correlated attributes, as these variables can adversely effect the models as there is a possibility of them carrying the same information. Hence, it is important to handle the problem of correlation. After checking the Correlation matrix, it was observed that there are several attributes with high correlation (Figure 3). Under the condition of them carrying the same information, these will be eliminated by feature selection model.

- **Bartlett's test of Sphericity & Kaiser-Meyer-Olkin's Test:** BTS and KMO tests were done to check the homogeneity of variance and sampling adequancy

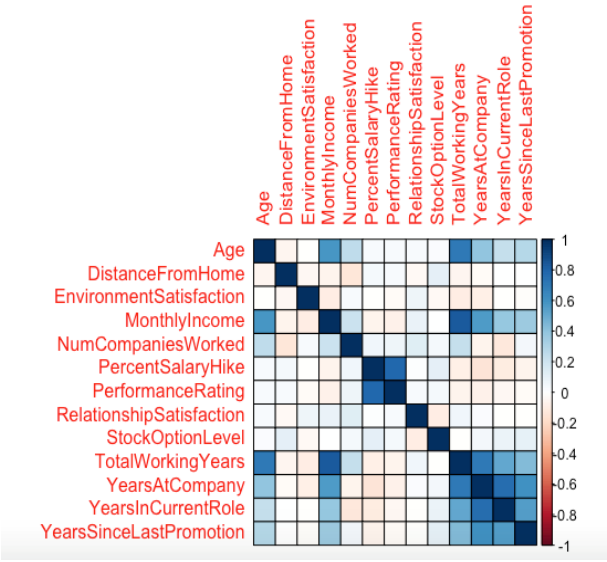

Figure 3: Correlation Matrix

among the data. These two tests are used to check how suitable the data is for factor analysis.

In the BTS test, the null hypothesis was accepted as the p-value of test is 2.22e-16 which is less than 0.01.

The KMO test tells if the data is suited for factor analysis, only when the Measure of Sampling Adequacy (MSA) value lies between 0.6 and 1. In the test for our analysis, the MSA value for KMO is equal to 0.77. This means that factor analysis can be applied on this dataset.

- **Principle Component Analysis:** As the test for BTS and KMO was successful, we proceeded to apply factor analysis - PCA to our data, a statistical procedure to find the linearly uncorrelated variables, which are called principle components. Out of 20, 13 principle components were selected as these were explaining the variance of 90%.

- **k-means Clusters:** After the shortlisting of 13 principle components, they were then used with k-means to see if there are any hidden clusters in the data. A k-means plot was built and it was observed that the clusters were not distinguishable, i.e. no clusters were formed as seen in Figure 4.

- **Class Imbalance:** The class imbalance of the data set was also identified. It was observed that there is approximately 84% of observations that belong to 'No' class in dataset.

After the exploratory data, we then moved towards the implementations of the techniques for this research.

The data was distributed into train and test for the further analysis. The feature selection technique - Simulated Annealing was used. This algorithm gives a global optimal of a function. In regards to feature selection, this algorithm computes the external

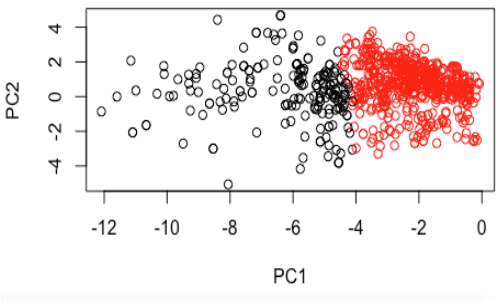

Figure 4: No clusters in the data

performance estimates and focuses on eliminating the overfitting of the features in the subset. The code was run with 50 iterations to reach a level which is good for the models. Out of the 29 attributes that were finalised after the exploratory data analysis, the Simulated annealing gave 15 features which according to the algorithm were influential for the employee attrition prediction listed in Table 4.

Age	BusinessTravel	EnvironmentSatisfaction
Gender	Hourly Rate	JobInvolvement
JobLevel	MonthlyIncome	MonthlyRate
NumCompaniesWorked	OverTime	StockOptionLevel
TrainingTimesLastYear	YearsInCurrentRole	YearsWithCurrManager

Table 4: Attributes finalized by Simulated Annealing

As was observed in the exploratory data analysis, it was seen that there exits a huge class imbalance in the dataset. To overcome this, it is advisable to use either oversampling or undersampling techniques. As the dataset used in this research is very small, the undersampling technique cannot be considered, the reason why the oversampling technique, SMOTE was used. For oversampling, SMOTE uses the k-nearest neighbours value on the sample of dataset. The value of k=5 was observed to be the best value to oversample the minority class in this data.

As this dataset has limited observations, there was a high chance of overfitting in the data. To eliminate this, k-fold repeated cross validation was used. With this technique, the sample of data was randomly divided into equally sized samples by the algorithm, giving out a single best sample for the testing.

The hyperparameter tuning was done by implementing Bayesian Optimisation to get the best fit tuning paramters of the classification models. These parameters are usually hard-coded in the models, and vary depending on the model used, such as SVMRadial has the sigma and C value as its parameters. Thus, these parameters were tuned with respect to the models.

- **Model 1 - SA & Random Forest with Bayesian Optimisation**

  Random Forest was applied as the first model to classify the employee attrition. The Bayesian Optimisation was used for hyperparamater tuning. In Random forest, the tuning parameter are mtry, no. of trees and node size, hence these were used

for the tuning using Bayesian Optimisation. The below values were obtained after tuning:

mtry = 12.41822; min_node_size = 13; ntree = 500

These values were further used to evaluate the model.

- **Model 2 - SA & Logistic Regression**

  Logistic regression second model was applied for this research on the selected attributes of SA. The best probability cut-off value was identified as 0.5461403 with the accuracy 0.7787958 as shown in the graph below in Figure 5

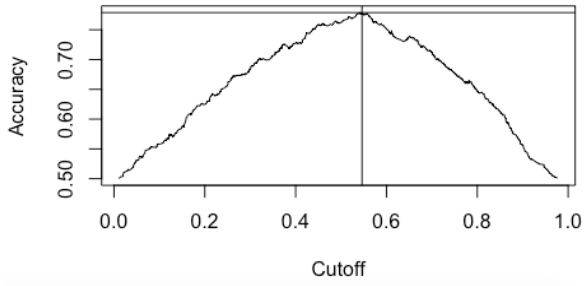

Figure 5: Logistic Regression Plot

- **Model 3 - SA & Support Vector Machine (Radial kernel) with Bayesian Optimisation**

  The SVMRadial model was applying on attributes selected in SA. SVM has sigma and C as its hyperparameters, which were tuned using Bayesian Optimisation to get better results for sensitivity. Below are the values obtained:

  sigma = 4.726128e-07; C = 100

- **Model 4 - Gradient Boosting Machine with Bayesian Optimisation**

  The GBM model was applied. The tuning parameters were optimised by Bayesian Optimisation, giving the below best values:

  interaction.depth = 2.922158379; ntrees = 67.833530238;
  Shrinkage = 0.001995287; n. minobsinnode = 14.988470203

5 Evaluation

Evaluation is the fifth stage of CRISP-DM process. This is the phase where all the models built and implemented are evaluated to see if the relevant results are generated or not. Each model in this research was built and then compared with each other to see which one had the best outputs.

5.1 Evaluation of Models with features selected by Simulated Annealing

SA-Random Forest with Bayesian Optimisation

The attributes given by Simulated Annealing were used with Random forest, and after doing the hyperparameter tuning, the model was evaluated. Here, the accuracy of 78.17% was achieved with the sensitivity & specificity values of 58.69% & 81.93% respectively.

SA-Logistic Regression

The features selected by Simulated Annealing were then used with Logistic Regression and an overall accuracy of 75.35% was obtained with sensitivity & specificity values as 67.39 & 76.89, which was more than the Random forest.

SA-Support Vector Machine with Bayesian Optimisation

The model of Support vector machine on the features finalised by Simulated annealing gave the following results after hyperparamter tuning with Bayesian Optimisation. The overall accuracy of 53.17% was observed, which was quiet low than the other two models, but there was a significant increase in the sensitivity value which was 80.43%, best so far.

Gradiant Boosting Machine with Bayesian Optimisation

Ultimately, the GBM model was tested after Applying Bayesian optimisation for hyperparameter tuning. An overall accuracy of 64.44% was obtained, with sensitivity & specificity values as 58.69% & 65.54%.

Model	**Attribute Observed**
SA-Random Forest	JobLevel, OverTime, StockOptionLevel
SA-Logistic Regression	OverTime, EnvironmentSatisfaction, Age
SA-Support Vector Machine	Monthly Income, Job Level, Age
Gradient Boosting Machine	Job Level, OverTime, StockOptionLevel

Table 5: Comparison of Top 3 Important Attributes By all the models in this Research

5.2 Evaluation of Models with Features Given by Domain Experts

Some features were exclusively selected from the dataset based on the research of management domain experts. The selected features are mentioned in the below Table 6.To see how relevant the management observations are with respect to the technical implementations, all the models built were used with these features.

Age	BusinessTravel	DailyRate
Department	Education	EnvironmentSatisfaction
Gender	Hourly Rate	JobInvolvement
JobLevel	JobSatisfaction	MonthlyIncome
MonthlyRate	OverTime	PerformanceRating
RelationshipSatisfaction	StandardHours	TotalWorkinYears
WorkLifeBalance	YearsWithCurrManager	

Table 6: Attributes finalized by Manual Feature Selection

Random Forest with Bayesian Optimisation

Random forest with Bayesian optimisation was implemented and below best values of hyperparamater tuning were observed: mtry_opt = 5.365047; min_node_size = 17. Although the accuracy was very good with these features, i.e. 86.27%, the sensitivity value was poor, which was equal to 19.56%.

Logistic Regression

Logistic Regression model was used with the new set of features. The accuracy of 85.92% achieved, but the sensitivity was 26.08% lesser than the one selected by Simulated Annealing.

Support Vector Machine with Bayesian Optimisation

SVM was used and after hyperparamater tuning below values were obtained: Sigma = 5.581795e-07; C = 66.21233. Again the accuracy of the model was good with the value of 85.56%, but the sensitivity value dropped down to 19.56%.

Gradiant Boosting Machine with Bayesian Optimisation

After tuning the paramters of GBM, just like the other models, an overall accuracy was increased to 86.97% but the sensitivity fell down to 23.91%.

5.3 Discussion

After the implementation of all the four models with the features selected by Simulated Annealing and the ones selected by the domain experts, it was observed that the models with manual feature selection are very helpful in regards to getting a better accuracy, but they have poor sensitivity accuracy. Figure 6 gives all the values in detail of all the models with and without feature selection, which clearly shows that though the manual feature selection gives better accuracy, it is not efficient in real time to help the organizations to identify the potential employee who might actually leave in near future.

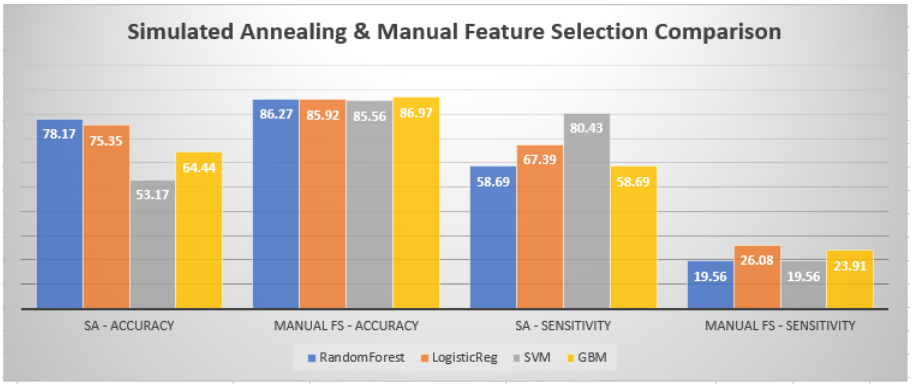

Figure 6: Comparison of SA & Manual Feature Selection

From the chart shown, it is pretty evident that there is a significant difference between the results of delivered by classifier by using manual feature selection and Simulated Annealing. But just to achieve more accuracy is not the objective of this research. In organization, it is very important that the higher management have the right information about the employees and their potential future actions. According to this research, it is important that the organization has the right employees' details of those who have a

significant chance to leave. This information is not successfully delivered by the manual FS as the maximum sensitivity is 26.08% (achieved by Logistic Regression, with accuracy of 85.92%). Whereas, the highest sensitivity observed by the Simulated Annealing with SVM is 80.48%, which means that this model is capable to identify approximately 80% of the employees who have high chances to leave. This can be deployed by the organizations in real time scenarios so that they can focus only on the weak-links and invest their time & resources to motivate them to stay.

As all the manually selected features were the ones given by the management domain researchers, it is interesting to see that there is a difference in what they assume as the features responsible for an employee to leave. Thus, it can be seen that with the advancement of technology, these management assumptions can be modified and implemented in real-time scenario.

6 Conclusion and Future Work

This research study has set out to be a real time application in the organizations where the management can predict the future actions of the employees based on the there records and observations. The main focus of this paper was to build a model which can efficiently predict the employees that might attrit in future, and considering the real scenario, the higher management will be more interested to know the potential employees who might actually leave so that they can set their attention on them to stop them to do so. With the use of Simulated Annealing feature selection technique, this research is able to figure out the major reasons for the turnover. The list of these features were also compared with the ones given by the management researchers in the past.

After applying various models, it was observed that SA-SVM tuned with Bayesian Optimisation, although gives an accuracy lower than other models, gives the best sensitivity of 80.43%. On the other hand, when these models were executed on the features selected manually based on the research of domain experts, it was observed that all the models have very good accuracy but the best sensitivity goes down to 26.08%. Hence, it can be concluded that the technical methods of feature selection are more reliable than the one given by the management domain expert for employee attrition.

It was found that there were some more features given by the management experts like work pressure, job security, job previews, which they mentioned as the leading factors for employee attrition. For future scope of this research, these attributes should be used to analyzed whether there is any technical validation for these. Also, as mentioned in this research, some factors like, StockOptionLevel and TrainingTime have an influence on attrition, but were not considered by the domain experts. Not just this, but some data mining researchers have given factors like absenteeism and leaves taken by the employees as important features. The management should incorporate these attributes in their research and show if this has any practical implementations.

Apart from these, as this dataset was limited to the small observations, it is advised in future to conduct a research with a larger dataset and more attributes so as to have more clarity towards the employee attrition and see if there is any difference in the results depending on the size of dataset.

Acknowledgment

I would like to thank Dr. Pramod Pathank, Dr. Paul Stynes and Dr. Dympna O'Sullivan to guide me throughout with my work. I would appreciate IBM for providing the public data that could be used for my thesis.

References

Adhikari, A. (2009). Factors affecting employee attrition: a multiple regression approach, *IUP Journal of Management Research* **8**(5): 38.

Alao, D. and Adeyemo, A. (2013). Analyzing employee attrition using decision tree algorithms, *Computing, Information Systems, Development Informatics and Allied Research Journal* **4**.

Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2010). Retaining talent: Replacing misconceptions with evidence-based strategies, *Academy of management Perspectives* **24**(2): 48–64.

Avey, J. B., Luthans, F. and Jensen, S. M. (2009). Psychological capital: A positive resource for combating employee stress and turnover, *Human resource management* **48**(5): 677–693.

Barvey, A., Kapila, J. and Pathak, K. (2018). Proactive intervention to downtrend employee attrition using artificial intelligence techniques, *arXiv preprint arXiv:1807.04081* .

Batt, R. and Valcour, P. M. (2003). Human resources practices as predictors of work-family outcomes and employee turnover, *Industrial Relations: A Journal of Economy and Society* **42**(2): 189–220.

Boucher, E. and Renault, C. (2015). Job classification based on linkedin summaries, *CS 224D, Stanford* .

Cahyani, A. D. and Budiharto, W. (2017). Modeling intelligent human resources systems (irhs) using big data and support vector machine (svm), *Proceedings of the 9th International Conference on Machine Learning and Computing*, ACM, pp. 137–140.

Chien, C.-F. and Chen, L.-F. (2008). Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry, *Expert Systems with applications* **34**(1): 280–290.

De Mauro, A., Greco, M., Grimaldi, M. and Ritala, P. (2018). Human resources for big data professions: A systematic classification of job roles and required skill sets, *Information Processing & Management* **54**(5): 807–817.

Debuse, J. C. and Rayward-Smith, V. J. (1997). Feature subset selection within a simulated annealing data mining algorithm, *Journal of Intelligent Information Systems* **9**(1): 57–81.

El-Sayed, A. A., Mahmood, M. A. M., Meguid, N. A. and Hefny, H. A. (2015). Handling autism imbalanced data using synthetic minority over-sampling technique (smote), *Complex Systems (WCCS), 2015 Third World Conference on*, IEEE, pp. 1–5.

Fan, C.-Y., Fan, P.-S., Chan, T.-Y. and Chang, S.-H. (2012). Using hybrid data mining and machine learning clustering analysis to predict the turnover rate for technology professionals, *Expert Systems with Applications* **39**(10): 8844–8851.

Frye, A., Boomhower, C., Smith, M., Vitovsky, L. and Fabricant, S. (2018). Employee attrition: What makes an employee quit?, *SMU Data Science Review* **1**(1): 9.

García, V., Sánchez, J. S. and Mollineda, R. A. (2012). On the effectiveness of preprocessing methods when dealing with different levels of class imbalance, *Knowledge-Based Systems* **25**(1): 13–21.

Han, H., Wang, W.-Y. and Mao, B.-H. (2005). Borderline-smote: a new over-sampling method in imbalanced data sets learning, *International Conference on Intelligent Computing*, Springer, pp. 878–887.

Harter, J. K., Schmidt, F. L. and Hayes, T. L. (2002). Business-unit-level relationship between employee satisfaction, employee engagement, and business outcomes: a meta-analysis., *Journal of applied psychology* **87**(2): 268.

Kumar, A., Pandey, A. and Kaushik, S. (2017). Machine learning methods for solving complex ranking and sorting issues in human resourcing, *2017 IEEE 7th International Advance Computing Conference (IACC)*, IEEE, pp. 43–47.

Lin, S.-W., Ying, K.-C., Lee, C.-Y. and Lee, Z.-J. (2012). An intelligent algorithm with feature selection and decision rules applied to anomaly intrusion detection, *Applied Soft Computing* **12**(10): 3285–3290.

Marsham, S., Scott, G. W. and Tobin, M. L. (2007). Comparison of nutritive chemistry of a range of temperate seaweeds, *Food chemistry* **100**(4): 1331–1336.

Meiri, R. and Zahavi, J. (2006). Using simulated annealing to optimize the feature selection problem in marketing applications, *European Journal of Operational Research* **171**(3): 842–858.

Mihajlović, I., Živković, Ž., Prvulović, S., Štrbac, N. and Živković, D. (2008). Factors influencing job satisfaction in transitional economies, *journal of General Management* **34**(2): 71–87.

Mishra, S. N., Lama, D. R. and Pal, Y. (2016). Human resource predictive analytics (hrpa) for hr management in organizations, *International Journal of Scientific & Technology Research* **5**(5): 33–35.

Nagadevara, V. and Srinivasan, V. (2008). Early prediction of employee attrition in software companies–application of data mining techniques, *Research and Practice in Human Resource Management* **16**: 2020–2032.

Pande, G. and Chung, L. (2017). A descriptive study on reasons for employee attrition behavior in hotels and restaurants of lucknow city: Owners/managers perspective., *CLEAR International Journal of Research in Commerce & Management* **8**(8).

Rabcan, J., Vaclavkova, M. and Blasko, R. (2017). Selection of appropriate candidates for a type position using c4. 5 decision tree, *Information and Digital Technologies (IDT), 2017 International Conference on*, IEEE, pp. 332–338.

Ramamurthy, K. N., Singh, M., Davis, M., Kevern, J. A., Klein, U. and Peran, M. (2015). Identifying employees for re-skilling using an analytics-based approach, *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, IEEE, pp. 345–354.

Rashid, T. A. and Jabar, A. L. (2016). Improvement on predicting employee behaviour through intelligent techniques, *IET Networks* **5**(5): 136–142.

Rombaut, E. and Guerry, M.-A. (2017). Predicting voluntary turnover through human resources database analysis, *Management Research Review* (just-accepted): 00–00.

Saradhi, V. V. and Palshikar, G. K. (2011). Employee churn prediction, *Expert Systems with Applications* **38**(3): 1999–2006.

Sengupta, S. (2011). An exploratory study on job and demographic attributes affecting employee satisfaction in the indian bpo industry, *Strategic Outsourcing: An International Journal* **4**(3): 248–273.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P. and De Freitas, N. (2016). Taking the human out of the loop: A review of bayesian optimization, *Proceedings of the IEEE* **104**(1): 148–175.

Singh, M., Varshney, K. R., Wang, J., Mojsilovic, A., Gill, A. R., Faur, P. I. and Ezry, R. (2012). An analytics approach for proactively combating voluntary attrition of employees, *Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on*, IEEE, pp. 317–323.

Snyder, T. M. (2016). You're fired: A case for agency moderation of machine data in the employment context, *Geo. Mason L. Rev.* **24**: 243.

Tabachnick, B. G. and Fidell, L. S. (2007). *Using multivariate statistics*, Allyn & Bacon/Pearson Education.

Tomassen, M. (2016). *Exploring the black box of machine learning in human resource management: An hr perspective on the consequences for hr professionals*, Master's thesis, University of Twente.

Wirth, R. and Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining, *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, Citeseer, pp. 29–39.

Yadav, S. and Shukla, S. (2016). Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification, *Advanced Computing (IACC), 2016 IEEE 6th International Conference on*, IEEE, pp. 78–83.

Yiğit, İ. O. and Shourabizadeh, H. (2017). An approach for predicting employee churn by using data mining, *Artificial Intelligence and Data Processing Symposium (IDAP), 2017 International*, IEEE, pp. 1–4.

Zhou, N., Gifford, W. M., Yan, J. and Li, H. (2016). End-to-end solution with clustering method for attrition analysis, *Services Computing (SCC), 2016 IEEE International Conference on*, IEEE, pp. 363–370.