National College of Ireland

Evaluation of Employee Attrition by Effective Feature Selection using Hybrid Model of Ensemble Methods

MSc Research Project
Data Analytics

Divyang Jain
x16110323

School of Computing
National College of Ireland

Supervisor: Mr.Brian Buckley

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

Student Name:	Divyang Jain
Student ID:	x16110323
Programme:	Data Analytics
Year:	2017
Module:	MSc Research Project
Lecturer:	Mr.Brian Buckley
Submission Due Date:	11/12/2017
Project Title:	Evaluation of Employee Attrition by Effective Feature Selection using Hybrid Model of Ensemble Methods
Word Count:	6680

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:	11th December 2017

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):	

Evaluation of Employee Attrition by Effective Feature Selection using Hybrid Model of Ensemble Methods

Divyang Jain
x16110323
MSc Research Project in Data Analytics

11th December 2017

Abstract

Employees are leaving organisation's pre-maturely that results in high losses for the organisation which cannot be predicted by HR.This paper contributes to HR predictive analytics (HRPA) that helps in a predicting the employees who will leave the organisation in a certain period of time by using a hybrid model of machine learning techniques.Employee attrition is a major concern today which is related to customer attrition prediction and much research has been done on customer churn by using ensemble methods.This research explains how predicted accuracy,sensitvity and specificity can be enhanced by the use of ensemble methods in determining employee attrition with the feature selection method using efficient feature engineering, data wrangling, visualizing and analyzing results from previous models to increase the accuracy.This work has the potential for greater accuracy to improve employee retention and reducing Human Resource costs.The CRISP-DM method using three different ensemble methods was used (1-stacking) GLM, SVM, Decision Trees, KNN, (2-bagging) Random Forest and (3-boosting) GBM, Adaptive boosting (ADA).This achieved **88.85**% accuracy by these techniques from which HR can place a sound strategy to raise employee retention.

Keywords -*Employee attrition, ensemble model, ADA, data mining, HRPA, employee retention and machine learning.*

1 Introduction

Human Resource Management act as an important role in identifying the key decisions in an organisation and losing high skilled employees can result in negative impact on the working of the organisation.This issue is growing drastically, however the majority of the companies did not have a steady and general perspective of employee attrition thus HRPA is crucial to create explanatory abilities which can be proficient to deliver more Return on Investment (Mishra et al.; 2016).(H) Human (R) Resource (P) Predictive (A) Analytics is the future of the organisation which helps to find out the business insights in the field of data analytics by judging the past factors and making machine learning models to predict the attrition, absences and other risks to improve the employee retention.There are three types of attrition:-Voluntary (employee resignation), Involuntary (induced by the company) and retirements (Ribes et al.; 2017). This paper majorly focuses on voluntary

attrition and rest of the two are out of the scope.

The Advent of machine learning models and HR systems have capabilities in making the organisation achieve their targets of human capital management (HCM). It is very important for an organisation to check what are the factors affecting employee attrition so that managers can motivate them or can find a replacement for the smooth working of the organisation.There are many techniques like predictive retention modelling that can help managers decide pre-emptively. It is evident from (Mishra et al.; 2016) that companies which lack human resource predictive analysis techniques cannot operate in the long run.Big companies like Google, Facebook and Microsoft are adapting HRPA which helps to retain employee and will overcome the traditional analytics in the near future.Prediction can be done using statistic methods, survival analysis expert systems and machine learning techniques based on historical facts.This paper contributes to data analytics by renovating Human Resource Management by predicting voluntary employee attrition using machine learning ensemble methods.

This research is motivated by prior research (Sharma; 2017) and (Mulla et al.; 2013) which were focused on customer churn prediction using survival analysis, statistical and machine learning approaches.Employee attrition is a similar problem.According to (Saradhi and Palshikar; 2011),the indexes of employee attrition for IT service organizations is 12-15% and in other industries it is 10-12%.The attrition rate is quite high and assuming even a lower attrition rate of 5%, the cost involved in an employee leaving an organisation is approximately 1.5 times the annual salary of an employee (Saradhi and Palshikar; 2011, p. 1).Figure 1[1] can explain employee attrition more effectively as the man in the red colour is leaving the organisation and rest of the other employees are working in the same company.

Figure 1: Employee leaving the organisation

So far,most of the related researchers use different models with different techniques and to my best of the knowledge, this research will be a unique first study in evaluating employee attrition through hybrid ensemble methods in machine learning using ADA boost, GBM, Random forest and comparing accuracies with the classification models -SVM, GLM, SVM, decision trees and KNN to get the best accuracy in finding the right deciding factors.Many researchers have taken different datasets and their strongest factors are age, satisfaction, tenure, pay and employees perception of fairness (Ribes et al.; 2017).This work contains many other variables which are discovered through effective feature engineering and data wrangling resulting in high accuracy.The 6 strongest predictors were job

[1] https://www.linkedin.com/pulse/20140606134242-83812109-why-can-t-we-diminish-attrition/

position,stock option level, relationship and job satisfaction, overtime and job involvement to predict the binary main class label Attrition-yes (H1-1) or no (H0-0) which will be the hypothesis on the IBM employee attrition dataset.This research paper aims to bridge the hiatus between machine learning predictive modelling and HR management based on business data mining model-CRISP-DM which finally induces to research on this question which can be measured by finding accuracy,sensitivity and specificity: -

An investigation on analysing employee attrition using machine learning ensemble methods by doing effective feature selection is effective which can be used in the HRM field?

This paper covers 7 sections covering the related work/literature review, methodology,design specification,implementation/solution development, evaluation,conclusion and future work. The last section contains references which motivated this research.

2 Related Work

2.1 Assessment of Employee Attrition using traditional methods-Survival Analysis

An organisations main aim is to gain profits from the customer by building reputation and goodwill of the company and establishing the industry to higher scale but if there is a problem of customer churn then it will be very difficult to acquire new customers as customers feedback and reviews are the crucial points.According to (Saradhi and Palshikar; 2011),it is very crucial to judge customer churn in advance to save the losses and helps in gaining potential profits.In the future work, they suggested to use survival analysis which was carried by (Griffeth and Hom; 2001) in which survival analyses was done on employee attrition over a specified period of time to gain employee retention resulting in valuable workforce for the survival of an organisation and the major parameter was Overtime.They got effective results at that time and suggested that for the proper working of an industry, there should be an appropriate strategy for evaluating attrition. Companies should focus on voluntary-avoidable attrition to improve the staff retention by making sound strategies.

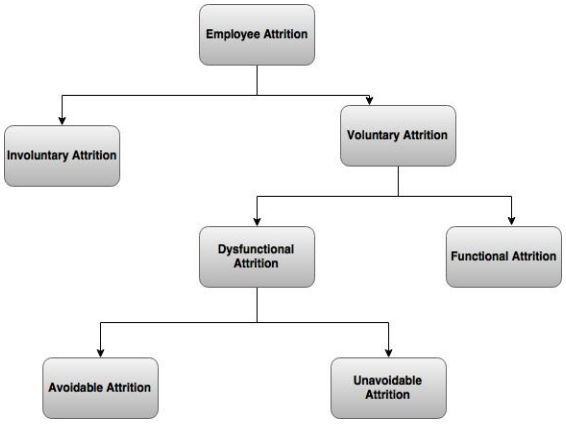

Figure 2: Employee Attrition Flowchart, (Griffeth and Hom; 2001)

According to Figure 2 (www.draw.io),there are two types of attrition and as mentioned in the introduction this paper focuses on voluntary attrition which is further divided into dysfunctional and functional.Lastly,Dysfunctional attrition is divided into further two leaf nodes that are avoidable and unavoidable attrition. (Griffeth and Hom; 2001) opined that HR can evaluate avoidable attrition and if it goes to high attrition which may cause business loses hence, they concluded that survival analysis in evaluating employee attrition is very useful as a benchmark to stop the losses in the near future.

There was other research on employee attrition which was on Bharat Petroleum Corporation Limited(BPCL) by (Mulla et al.; 2013) in 2013 that uses 2140 engineers over thirteen years from 2000-2012.The same survival analyses was used with the help of multiple regression to detect employee attrition taking the time factor of employees worked in the company and got average results. They researched that employees who are unmarried, younger, live far and poor performers leave the company prematurely resulting in company loss (Mulla et al.; 2013) and HR can make strategies to stop the attrition by motivating them or relocating them to their nearest branch if there is that option.

In other research, survival analysis was performed in people analytics based on big data using survival curve in R tool (Isson and Harriott; 2016) with good results too but raised a new problem for machine failure in the prediction using only time-based factors as it is not compulsory to have time attrition in the data and it is impossible to get the survival points without time attribute.

From the above discussion,it can be concluded that traditional methods like survival analysis can be used to avoid voluntary employee churn problem but cannot get the effective accuracy and factors if there is no time data using statistical methods and hence (Isson and Harriott; 2016) recommended in their future work to use expert systems and planned models which help in making a framework for talent management and HRM .This will be covered in the next section.

2.2 Decision making framework for talent and HR management with Expert Systems and Sensitivity Analyses

Decisions are made at the final step when all the models are built, as mentioned in the last section expert system and sensitivity analyses are coined by (Ghosh and Arunava; 2016) and (Ribes et al.; 2017) which made a huge effect in the HR and talent management. Organisations work effectively to identify the best talent in the company and motivate them to keep the employee happy by evaluating the performance and giving rewards to the talented persons.This framework was proposed by (Mishra et al.; 2016) in HR field which helped to find talented people using expert system providing better solutions than other researches like (Varshney et al.; 2014) which used the IBM dataset to find out the talented employees and contributing to employee retention.They evaluated features from various data sources-work products, job title, social tags and HR information using l2-regularized logistic regression and got average accuracy to find talented employees in a department.Hence, these studies were done to increase employee retention by motivating the employees.

Workforce analytics, people analytics and talent analytics are interchangeable terms in HR field which was done using expert systems by (Ghosh and Arunava; 2016) which helped the HR team to predict the employee attrition due to employee job dissatisfaction.The data was collected using employee feedback surveys and counseling.This research gave expert systems to the HR field to find analytics tasks like employee attrition.Other research by (Van Breukelen et al.; 2004) contributed in predicting voluntary employee attrition by merging factors of traditional turnover with the theory of planned behavior which was done to check that variables like job satisfaction, tenure, age and organizational commitment plays an important role or not using statistical methods considering the internal factors.They missed out much important information for the features in the dataset and not up to date results.Hence,in this work more features were included to get better accuracy and efficient results from which effective decisions can be made.

Sensitivity analyses was coined by (Ribes et al.; 2017) in the HR field to find the uncertainty of employees leaving the company and what can be done to improve employee retention by evaluating uncertain parameters. They used a technique called SMOTE to balance the imbalance data which is used in this research work.They included in their future work that if there will be more detailed compensation elements then the models will be effective in the feature mix as they used only the internal environment features and left economic trade-off outside the company in making decisions.

From the above related research, we evaluated that there were features in the dataset which were not taken into consideration and resulted in poor target decisions in making employee attrition a great prediction i.e. not effective decisions.According to (Ghosh and Arunava; 2016),expert systems are not fast and needed a lot of manual work of papers and surveys which can take a lot of time with inaccurate results. In their future work, they included manual labour will be eliminated by computer systems which will be built by Artificial intelligence and making the process automated to evaluate human capital planning and evaluating the performance which will result in a fast and accurate prediction of employee attrition.

This research extends the (Ghosh and Arunava; 2016), (Ribes et al.; 2017) and (Van Breukelen et al.; 2004) researches by considering their future work and removing drawbacks with the help of Artificial Intelligence data mining techniques that will be covered in the next section which will help in making decisions effectively.

2.3 Prospects of Artificial Intelligence and Data Mining in evaluating Employee Attrition giving rise to HRPA

2.3.1 Artificial Intelligence

Expert systems showed many drawbacks as mentioned in the previous research and to overcome the problem, Artificial Intelligence with the help of data mining techniques came into existence in evaluating employee attrition by building models and automating them. Artificial intelligence (AI) is the intelligence done through data mining methods and displayed by machines like learning or predicting something which makes the work effective and efficient by automating the process, hence this section extends the previous two sections of survival analyses and expert system with the help of data mining

techniques i.e. machine learning.According to (Berson and Smith; 2002), survival of the organisation can happen only if company adapts Artificial intelligence in HR department to predict employee attrition as they can build efficient models which were done using nave Bayes.This research was carried by a Taiwan research on employee attrition which was done by (Hong et al.; 2007) using two comparative test predictive models of logit and probit which were done successfully to solve the classification and regression problems. This research was new to support the flexibility in predicting data and the problems which were faced by classification problems. There was a great research by (Chang and Xi; 2009) which considered the two-mixed approach of KNN-nearest neighbour and Taguchi classification rules. They were applied on the dataset in which attributes like sick leave, salary, gender, seniority and some other attributes were selected showing 78% accuracy.

Feature selection using the Boruta package in this research is motivated by (Dutta et al.; 2010) and (Kursa et al.; 2010) in which the feature selection of attributes was the major focus and was done through artificial intelligence by machine learning techniques to follow the patterns of performance of employees and related attributes.

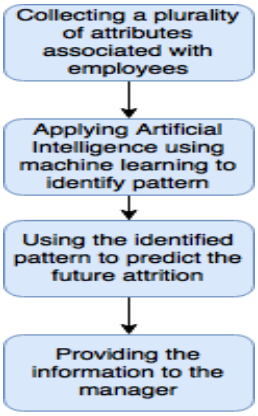

Figure 3: Employee Attributes workflow (Dutta et al.; 2010, p.1)

Banking and finance industry is also a major concern in which employee attrition can be judged so predictive workforce analyses which is done by (Sharma; 2017) using regression and factor analyses to find the employee attrition on which managers can play a sound strategy to retain them. Similarly, this project also considers the previous researches on regression and performs generalized linear model to make the accuracy effective in HR field.

Therefore,this work contributes to these researches by taking KNN and GLM techniques and applying them to Employee attrition to get the effective results which are continued in the next section using data mining approaches.

2.3.2 Data Mining

Expert Systems, sensitivity analyses and survival analyses were replaced by data mining as discussed above and are used to extract data insights to make effective decisions.In todays world, the use of data mining is growing rapidly and much research has been done which aims to support Human Resource management. Selection employees, HR cost plan-

ning, ascertaining competencies of employees, career planning, functions of staffing and predicting employee attrition in HR is ongoing. This topic was covered by (Strohmeier and Piazza; 2013) which was done using machine learning techniques including neural nets,decision trees, cluster analysis, association analysis and SVM from which SVM got the best results and hence this technique is used in this paper to check that this research can upgrade the accuracy from ensemble methods by effective data wrangling.

Employee performance was checked by taking many factors in industrial practices using decision trees by (Gupta et al.; 2014).According to (Thakur et al.; 2015), ensemble models like random forest perform better than normal classification techniques-decision trees if it is a case of sparse data which can overcome the problem of over fitting.Hence random forest is used in this research to overcome the sparse data overfitting problem.

Another research was done by (Aktepe and Ersoz; 2012) and (Zhou et al.; 2016) in the field of HR to check the employee performed where attributes like job satisfaction and strategic plans were clustered into four distinct groups and K-means clustering combined with Artificial Neural Network was used in manufacturing company in Turkey.This study was motivated by the previous research and applied the same techniques using KNN to evaluate which features of employees is the most crucial attribute to the employee attrition and checking the accuracy.

Title	Models implemented	Authors
Data mining to improve human resources in construction company	Decision tree	Chang Youzheng (2008)
Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry	Decision tree (Chi-squared automatic interaction detection)	Chien and Chen (2008)
Employee turnover: a neural network solution	Neural network	Sexton et al. (2005)
Data mining for selection of insurance sales agents	Discriminant analysis, decision tree (C4.5) and neural network (feed-forward)	Cho and Ngai (2003)
Job performance prediction in a call center using a naive Bayes classifier	Naive Bayesian classifier	Valle, Varas et al. (2012)
Using hybrid data mining and machine learning clustering analysis to predict the turnover rate for technology professionals.	Neural network and self-organizing map	Fan, Fan et al. (2012)
Predicting Employee Attrition using Machine learning techniques with Boosting technique (Xgboost)	Support Vector Machine, Naïve Bayes Logistic Regression, Random Forest, LDA, KNN and XGboost	Punnoose and Ajit (2016)

Table 1.Related Work on HRM Employee Attrition (Sikaroudi et al.; 2015, p.3)

Ensemble method- Base Research Data Mining technique

Table 1 shows the information about the previous research and this research project mainly gets an idea from the last row which is done by (Ajit; 2016) and (Chang and Xi; 2009) where all the machine learning which were done previously were checked and new ensemble method XGBoost was used to boost the accuracy with effective feature selection. They got 88% accuracy but failed to check the attributes due to poor feature engineering and data wrangling.

Questions such as Who is at the risk and What can be done on the data were missed and they also told that with the great scalability of network and feature engineering accuracy can be improved with covering all the missing questions and attributes and in the best of the knowledge ADA boosting havent been done on employee attrition IBM dataset.

Hence, this research covers all the drawbacks which were done by previous researches and uses a hybrid model of ensemble methods (ADA, GBM, Random Forest) to upgrade the accuracy and completing all the queries using effective feature engineering, data wrangling, visualizations and variable importance package (Boruta) from R studio which is covered in next sections.

3 Methodology and Design Specification

This research uses (C) Cross (I)Industry (S) Standard (P) Process for (D) Data (M) mining (CRISP-DM) Business model which is motivated by (Sikaroudi et al.; 2015) to use various models of data mining to compare accuracy,senstivity,specificity, area under the curve, calculation time and user- friendliness. It states that employee attrition can be predicted using data mining techniques and can rely on Decision Support Systems for HR recruitment. In this process, CRISP-DM acts as an important model of revision the HR database which can add new employee attributes and delete the attributes which are not useful. Moreover, CRISP-DM provides the ease to understand the model for the business analyst and HR as it goes from Business Understanding of the problem to the evaluation and deployment which covers every part of the problem. Hence, CRISP-DM is used here as a solving strategy to find employee attrition on the IBM dataset step by step which fits better with the hybrid model of data mining contributing to finding out the results of the research question. The process undergoes with six steps:-

3.1 Business Understanding

Understanding the business need is the first step for any company and according to this project understanding the research question and its dataset is important. Dataset has been extracted from IBM website which is among 500 fortune companies of America to judge which employee will stay in the company or which employee will leave the company, hence prediction is done by using data mining ensemble method to find out the decision to take major steps.

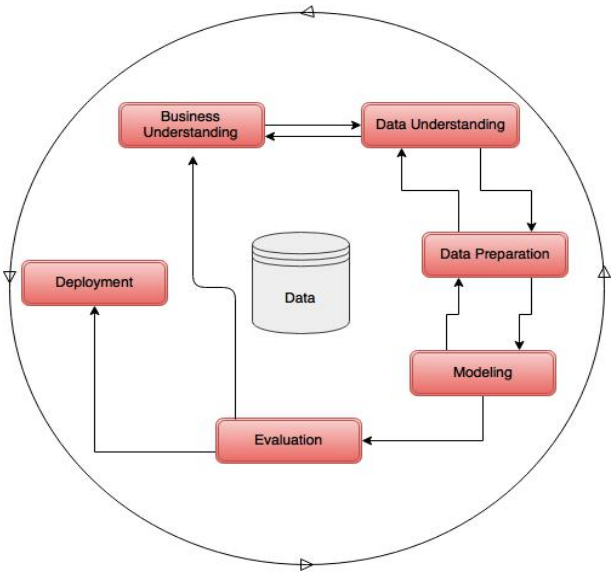

Figure 4: CRISP-DM Process

3.2 Data Understanding

Data understanding is the major step for this project as if someone is not aware about the attributes of the dataset then attributes will be unaware resulting to bad modelling. Here is the snippet of the dataset:-

	A	B
1	Attributes	IBM data information
2	Age	Employee's Age
3	Attrition	employee will leave the company or not
4	Business Travel	Frequency of travel
5	Daily rate	Daily rate of employee
6	Department	To which department it belongs
7	Distance from Home	what is the distance from home
8	Education	rating for college
9	EducationField	education field like medical,life sciences
10	Employee Count	count of the employee
11	Employee Number	numbers of employee
12	Environment Satisfaction	satisfaction rate
13	Gender	gender
14	Hourly Rate	price per hour
15	Job Involvement	how much an employee involved In a company
16	Job Level	level of job
17	Job Role	role of employee doing job in a company
18	Job Satisfaction	how much a employee is satisfied
19	Martial Status	is the employee maried or not
20	Monthly Income	how much an employee is earning per month
21	Monthly Rate	monthly rate which is related to income
22	Num Companies Worked	in how many companies an employee has worked
23	Over18	Employee is over 18 or under 18
24	Percent Salary Hike	has the employee worked for over time
25	Performance Rating	how much percent of the salary is hiked
26	Relationship Satisfaction	relationship satisfaction rate
27	Standard Hours	hour for an employee worked
28	Stock Option Level	stock level for an employee
29	Total Working Years	for how many hours an empoyee has worked
30	Training Times Last Year	training times for the last year
31	Work Life Balance	work life balance rate for an employee
32	Years At Company	how many yours he/she has worked in IBM company
33	YearsInCurrent Role	how many yours he/she has worked as this role
34	Year Since Last Promotion	how many years happened for the last promotion
35	Year With Curr Manager	how many years happened with this manager.

Figure 5: Snippet of the IBM dataset

Figure 5 explains about the IBM dataset attributes where the yellow coloured row is the label class which is **employee attrition-yes or no**.Research finds out the attributes affecting this label by considering major attributes from the dataset.

3.3 Data Pre-Processing

Data Pre-processing is very important to build a data model as with the uncleaned and unorganized values in the data set evaluation and results can be effective.Many levels of measurements are presented in the data set like interval, ordinal, ratio and categorical/nominal data and accordingly data mining techniques using machine learning has been applied to remove the:

1.Missing Values 2.Redundant Values 3.Outliers

Data will be transformed into proper and consistent form and checking the incorrect attributes using the histogram and other visualization plots.

3.4 Modelling

After the data is cleaned and ready modelling takes the 4th step in which all the machine learning techniques are parameterized and checked. This research project uses some techniques which are covered in the next section.

3.5 Evaluation

This phase is linked with the previous modelling step in which it evaluates the best techniques by checking the accuracy, area under the curve and mean squared error through plots which are finally checked and applied in the next deployment phase.

3.6 Deployment

CRISP-DM gives the opportunity to apply the best technique which is evaluated in the evaluation phase and giving the best technique to the market which can be used by the company and that is the main aim for this research to find out the effective technique for the organization which is covered in the next section.

Hence, CRISP-DM plays an important part to find out employee attrition and can result in retention of employees by motivating them or manager can find a new replacement resulting in saving money and Return on investment (ROI). All the codes and steps are written in an attached configuration file which can help the reader to understand the CRISP-DM better.

4 Implementation

Every implementation follows one development cycle and as discussed above in the methodology section this projects design and implementation follow the CRISP-DM methodology for the data analysis. In the related work section, we discussed the data analysis done on employee attrition and how their implementation helped to find employee attrition. Similarly, this project identifies factors which affect employee attrition which is further divided into train and test dataset on which models were built then finally the best technique is chosen which is being deployed.

Dataset was downloaded from IBM website in comma separated value (.CSV) format which was explored in Microsoft Excel that had 1470 rows and 35 columns. Attrition is the main column which is to be predicted whether employee leaves the company or not (yes or no) and other columns are the independent variables which are taken into consideration to build the models.

After the data set is explored, it is imported into R studio and some packages like mlr (smote), boruta (Dutta et al.; 2010) and some others (Kursa et al.; 2010) were installed and imported to give functions to modelling in figure 6.

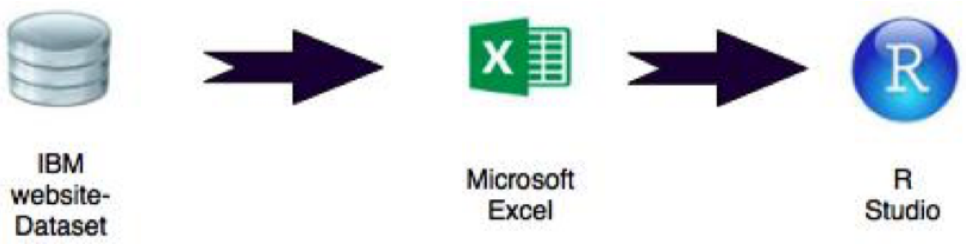

Figure 6: Data Flow Diagram

The Dataset contains factors like age, marital status, gender, job satisfaction, monthly rate and work-life balance, ethnicity, education were mostly used to predict the employee attrition in many researches (Ajit; 2016) and (Kursa et al.; 2010) but this research extends these researches by considering 25 attributes and dropping 10 columns which were of no use and uncleaned. Data was pre-processed where null values, duplicate values/redundant values and also the variables which showed multicollinearity which was checked by correlation plots in R were removed from the dataset making the dataset cleaned.

There was a class imbalance problem in the dataset as there were 1233 values for no and 237 values for yes which could result in bad and ineffective accuracy hence, smote a (synthetic minority oversampling technique) is used in this research to balance the data as there should be at least 70-30 ratio for the particular rows (Ribes et al.; 2017).Some visualizations were done to check the mid-ranges, values and distribution using histogram, bar graph and scatter plot which helps to do feature engineering and data wrapping as 80% of the time is invested in data pre-processing and feature engineering and 20% on modelling.

This project uses 7 machine learning techniques which were motivated by related researchers as mentioned in the literature review and compared to find the best accuracy with effective feature selection and Adaptive boosting which is an ensemble method has not been done on employee attrition as best of our knowledge and trying to compare Ada with other related works with effective selection.

The algorithms which were tested in this research were KNN (Chang and Xi; 2009), Generalized Linear Model (Sharma; 2017), SVM (Strohmeier and Piazza; 2013), Random Forest (Thakur et al.; 2015), decision trees (Gupta et al.; 2014) and Adaptive boosting. The models were implemented by the mlr package in R studio which gives flexibility and ease of use. The dataset was divided into 80:20 ratio between training and testing datasets. All the modelling is done on training dataset and predicted on testing dataset to evaluate the accuracy,sensitivty,specificty and error after feature engineering and data

wrangling.

Feature engineering is a process to make extensive use of data using domain knowledge to create features which result in making effective models and hence done effectively in this project to create best ensemble models.In this research, variables have been dropped which were redundant and showed multicollinearity as discussed above and converted into factor for all models.Attributes like stock option level, training times last year, number of companies worked are converted into factors and all the attributes are converted into integer back again for KNN. Many attributes were converted into factor with data wrangling.

Data wrangling or sometimes called Data munging is the process which is used to transform and map the data from raw to another format to make the data more appropriate and valuable by merging the columns/rows, aggregation or by split the values into more relatable form to make data model more effective.This is done effectively in this project to make bests use of feature selection with systematic modelling. Employee attrition data set has been wrangled to make extensive use of it like age group-has been divided into young, middle-age and adult then using applying feature engineering on it to make it as a factor which is finally added into the dataset.Similarly, many attributes like total satisfaction, distance from home, monthly rate, percent salary hike, total working years, years at the company, total satisfaction, education, environment satisfaction, job involvement, performance rating, relationship satisfaction, work-life balance and job level have been wrangled to make the models effectively.

Variable Selection

After feature engineering and data wrangling the most effective variables are evaluated using R package Boruta (Dutta et al.; 2010) which helps to decide whether a variable is important or not which is represented on the plot below.

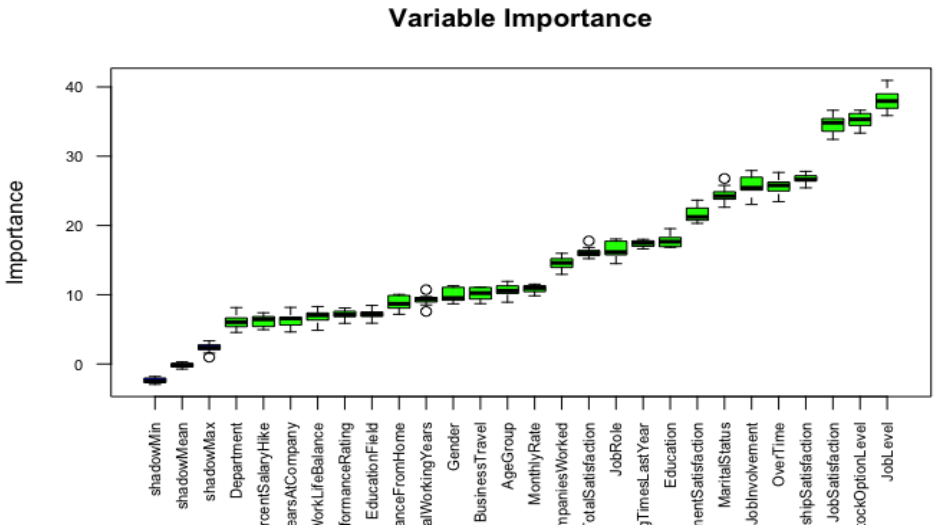

Figure 7: Variable Importance plot for Employee Attrition attributes

5 Evaluation

There are many evaluation measures which can be used in machine learning classification techniques as posted in (Huang and Ling; 2005) like **accuracy(acc)**, area under the curve (auc), mean misclassification error (mmce) but according to (McNee et al.; 2006) accuracy is not enough to be calculated for small or sparse dataset like this hence this research uses **sensitivity/recall,specificity(tpr)**,false positive rate (fall-out) and fnr (false negative rate or miss rate) are used to evaluate the performance of a technique.

Tuning of parameters has been done while training the model and predicted the test data set to get tuned evaluation measures. Here is the table from which we can easily judge the best technique that can be used in HRM to evaluate employee attrition:-

Algorithm	Sensitivity	Specificity	FPR	FNR	Error	Accuracy
Boosting						
Adaptive Boosting	85.93%	90.68%	9.31%	14.06%	11.1%	**88.8%**
Gradient Boosting	83.5%	90.19%	9.80%	16.40%	12.3%	87.6%
Bagging						
Random Forest	85.9%	87.25%	12.74%	14.07%	13.25%	86.74%
Stacking						
SVM	77.53%	90.1%	9.83%	22.65%	14.75%	85.24%
GLM	77.34%	90%	9.38%	22.66%	14.45%	85.54%
Decision Trees	81.25%	84.80%	15.1%	18.75%	16.56%	83.43%
K-nearest Neighbour	70.41%	88.97%	11.03%	29.59%	16.66%	83.34%

Table 2: Ensemble Models Scores Comparison

As we can see from Table 2 that boosting algorithms performed the best among ensemble models and Adaptive Boosting (ADA) exceed the other models accuracy,sensitivity and specificity by getting **88.8%** accuracy which is done taking the average of weak learners and boosting them to get higher accuracy.

Sensitivity measures the proportion of positive class which actually predicted correctly also known as true positive and here ADA boosting got **85.93%** which is the maximum among other techniques.Specificity measures the proportion of negative class which actually negative which is known as true negative rate.Here ADA got the maximum specificity that is **90.68%**.As mentioned above,accuracy is not the only measure to find if there is a sparse dataset hence by looking at sensitivity and specificity we can conclude that Ada performed the best.Attrition positive class was No and negative class was yes hence it was perfectly done by Ada boosting measuring sensitivity and specificity. Bagging algorithm also performed good in this research but Boosting algorithm takes the crown.However, but stacking algorithm does perform that well and gave low results.

Ensemble methods gave better results as all the algorithm got more than 80% which can be called as a good model.

After evaluating the best technique this research undergoes with some case studies that are shown below: -

5.1 Experiment / Case Study 1

First case study compares the 3 ensemble techniques i.e. boosting, bagging and stacking which is further divided into 6 techniques on the bases of area under the curve by showing and comparing the accuracy on (R) Receiver (O) Operating (C) Characteristic curve and judging the best technique from the graph which can be seen in Figure 8,9 and 10.

5.2 Experiment / Case Study 2

Second case study shows information about the most important variable which we got from variable importance graph figure 7 and hence most important variable is Job level followed by Stock Option Level, Job Satisfaction, Relationship Satisfaction, Overtime and etc. This variable is further explored via scatter plot which can be seen in figure 11 using R studio.

5.3 Experiment / Case Study 3

Third case study is related with figure 7 which showed the most important variables affecting the employee attrition which showed stock option level is the second most important variable which affects the employee attrition hence it is been explained via bar graph in figure 12 using R studio.

5.4 Discussion

After judging the case studies the research shows the following results which is discussed further:-

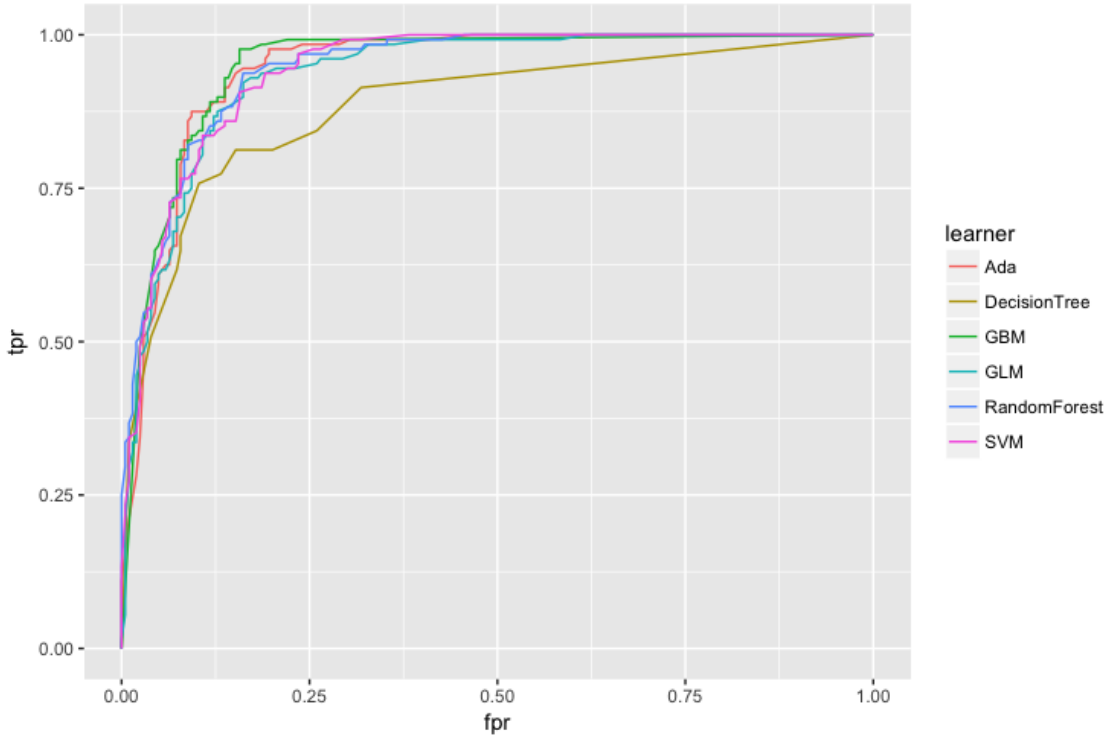

Figure 8: ROC curve for Comparison

Figure 8 shows that red curve i.e. is Ada boosting has the maximum ROC comparing with other techniques and decision trees show the lowest. Hence, by judging from table 2 and figure 8, we can say that Ada performed the best among other techniques taking all the measures. This can be seen in figure 9 as well and figure 10 shows the individually area under the curve:-

Figure 9: Bar Chart showing the accuracies

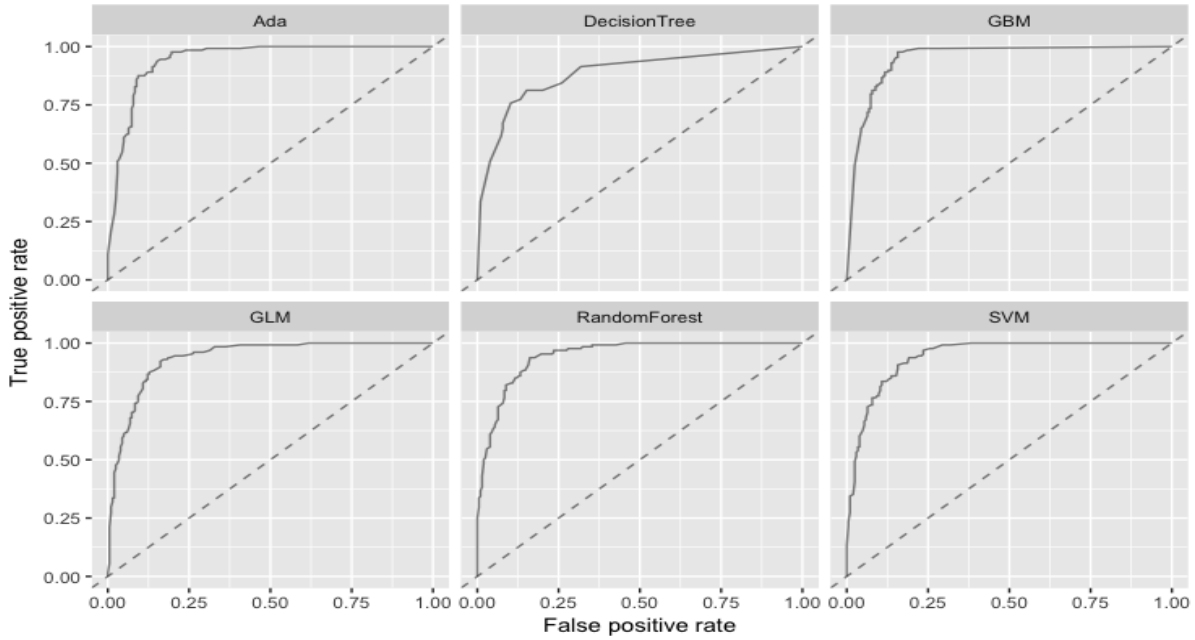

Figure 10: ROC Graph

Figure 11 shows that employees who are at entry level leaves the company more frequently then senior level and we can see the pattern from the graph as the employee level is growing to more seniority the level of attrition is going down. Hence, job level plays an important role in determining employee attrition.

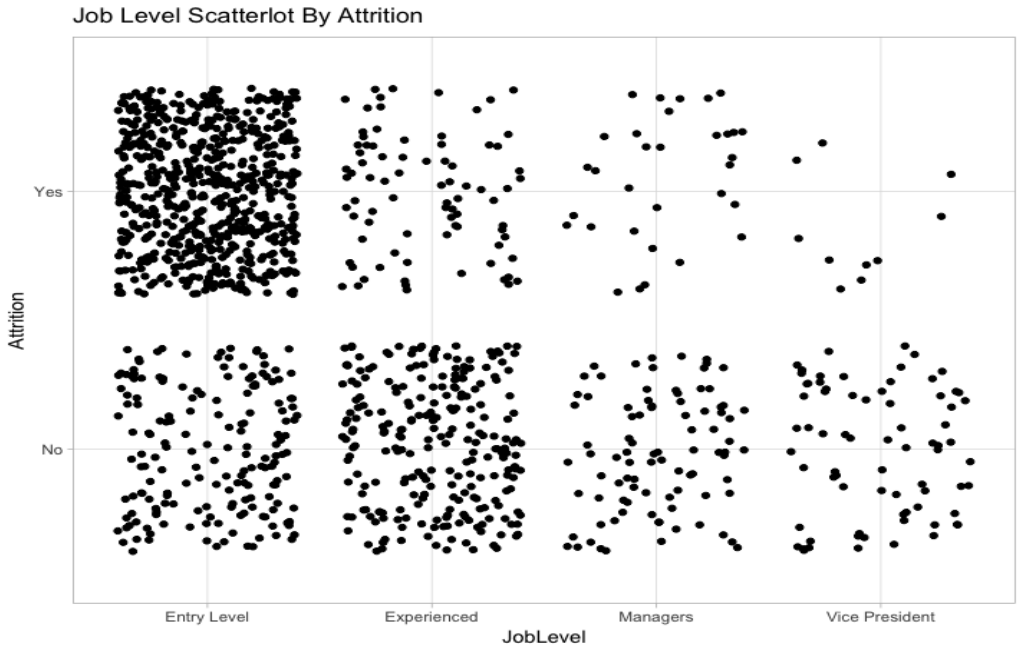

Figure 11: Exploration of Most Important Variable- Job Level

Figure 12 explains that employees are motivated towards stock option as an employee who has more number of stocks level leaves less comparatively to employees who have less number of stocks resulting stock option level as a great contributor to employee attrition.

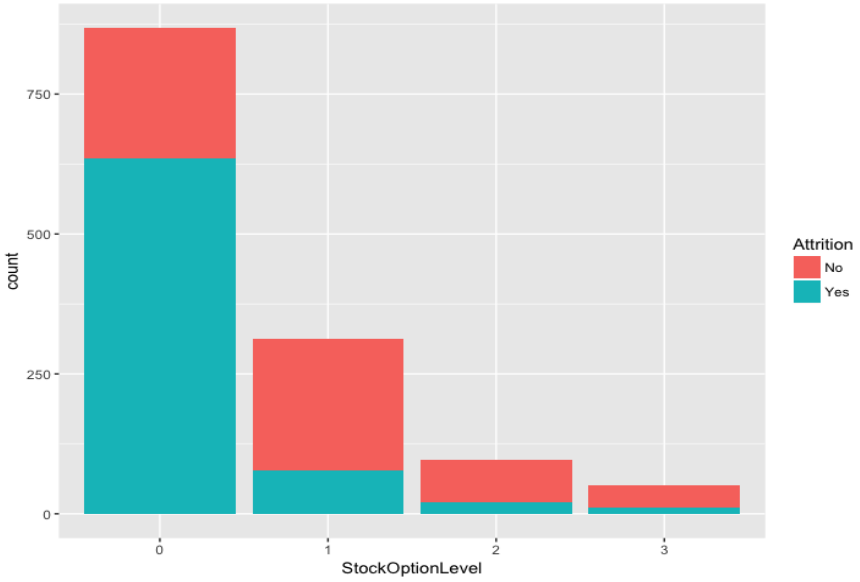

Figure 12: Stock Option Level Exploration

6 Conclusion and Future Work

This project implemented predictive analyses on employee attrition by effective feature selection using a hybrid model of ensemble methods based on the CRISP-DM business model for IBM which is a multinational technology company. Different types of ensemble methods like stacking, boosting and bagging were tested for classification, however boosting algorithms performed the best but Adaptive boosting gave the best evaluation scores with getting up to 88.8% accuracy.Due to sparse data set two other important evaluation scores were evaluated which showed 85.93% sensitivity and 90.68% specificity by ADA that can be applied to any companys dataset.

This project had 3 main case studies which showed transparency in HR field. In this first case study, it showed that area under the curve is higher in Adaptive boosting with higher accuracy and second case study showed that employees who are working at entry-level have higher chance to leave the organisation. Similarly, third case study explains that stock option level is an important variable factor in determining employee attrition.

These case studies further evaluated the research question by showing the result that ensemble method with effective feature selection are effective in predicting employee attrition contributing to Human resource management sector which can be seen by visualizations and accuracies by different models thus managers should look into the top needs of the employee by motivating entry level employees, giving more stock option level, increasing job satisfaction, relationship satisfaction and avoiding overtime by employees.

However, this project has some limitations. This research is limited to a small dataset which lacks to train the model well that might give low results and getting employees data from an organisation is confidential hence this research is limited to IBM dataset which is the only available dataset online. The second drawback is with the model is limited to only supervised machine learning that requires a lot of computation time, sometimes decision boundary might be over trained that and user input is required every time when new features have to be added.

This project can be extended in future as it has a lot of potentials to improve by applying deep learning techniques with a well-designed network of sufficient hidden layers on big data set which can cover up the limitations of this project. There can be time series and trend analysis which might improve the prediction performance if the data is in date format.

Acknowledgement

I would like to thank for the endless guidance and help provided by Mr. Brian Buckley, Mr. Jason Roche and Mr. Noel for the divine powers that be and to my family for their constant support and vote of confidence.

References

Ajit, P. (2016). Prediction of employee turnover in organizations using machine learning algorithms, *algorithms* **4**(5): C5.

Aktepe, A. and Ersoz, S. (2012). A quantitative performance evaluation model based on a job satisfaction-performance matrix and application in a manufacturing company., *International Journal of Industrial Engineering* **19**(6).

Berson, A. and Smith, S. J. (2002). *Building data mining applications for CRM*, McGraw-Hill, Inc.

Chang, H.-Y. and Xi, L. (2009). Employee turnover: a novel prediction solution with effective feature selection, *WSEAS International Conference. Proceedings. Mathematics and Computers in Science and Engineering*, number 3, World Scientific and Engineering Academy and Society.

Dutta, D., Gaspar, B., Challenger, J. and Arora, D. (2010). Determining employee characteristics using predictive analytics. US Patent App. 12/814,756.

Ghosh and Arunava (2016). Use of expert systems to predict employee turnover in organizations, *AGU International Journal of Management Studies and Research* .

Griffeth, R. W. and Hom, P. W. (2001). *Retaining valued employees*, Sage Publications.

Gupta, S. et al. (2014). Empirical study on selection of team members for software projects-data mining approach, *arXiv preprint arXiv:1402.2377* .

Hong, W.-C., Wei, S.-Y. and Chen, Y.-F. (2007). A comparative test of two employee turnover prediction models, *International Journal of Management* **24**(4): 808.

Huang, J. and Ling, C. X. (2005). Using auc and accuracy in evaluating learning algorithms, *IEEE Transactions on knowledge and Data Engineering* **17**(3): 299–310.

Isson, J. P. and Harriott, J. S. (2016). People analytics in the era of big data.

Kursa, M. B., Rudnicki, W. R. et al. (2010). Feature selection with the boruta package, *J Stat Softw* **36**(11): 1–13.

McNee, S. M., Riedl, J. and Konstan, J. A. (2006). Being accurate is not enough: how accuracy metrics have hurt recommender systems, *CHI'06 extended abstracts on Human factors in computing systems*, ACM, pp. 1097–1101.

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).

Mulla, Z. R., Kelkar, K., Agarwal, M., Singh, S. and Sen, N. E. (2013). Engineers' voluntary turnover: application of survival analysis, *The Indian Journal of Industrial Relations* pp. 328–341.

Ribes, E., Touahri, K. and Perthame, B. (2017). Employee turnover prediction and retention policies design: a case study, *arXiv preprint arXiv:1707.01377* .

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

Sharma, S. (2017). A predictive workforce-analytics model for voluntary employee turnover in the banking/financial-service industry, *Global Journal of Human Resource Management* **5**(1): 47–59.

Sikaroudi, E., Mohammad, A., Ghousi, R. and Sikaroudi, A. (2015). A data mining approach to employee turnover prediction (case study: Arak automotive parts manufacturing), *Journal of Industrial and Systems Engineering* **8**(4): 106–121.

Strohmeier, S. and Piazza, F. (2013). Domain driven data mining in human resource management: A review of current research, *Expert Systems with Applications* **40**(7): 2410–2420.

Thakur, G. S., Gupta, A. and Gupta, S. (2015). Data mining for prediction of human performance capability in the software-industry, *arXiv preprint arXiv:1504.01934* .

Van Breukelen, W., Van der Vlist, R. and Steensma, H. (2004). Voluntary employee turnover: Combining variables from the traditionalturnover literature with the theory of planned behavior, *Journal of Organizational Behavior* **25**(7): 893–914.

Varshney, K. R., Chenthamarakshan, V., Fancher, S. W., Wang, J., Fang, D. and Mojsilović, A. (2014). Predicting employee expertise for talent management in the enterprise, *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 1729–1738.

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.