

Improving Employee Retention by Predicting Employee Attrition using Machine Learning Techniques

Dissertation submitted in part fulfilment of the requirements
for the degree of

MSc I.S. with Computing

at Dublin Business School

Tanmay Prakash Salunkhe

Declaration

I, Tanmay Prakash Salunkhe, declare that this research is my original work and that it has never been presented to any institution or university for the award of Degree or Diploma. In addition, I have referenced correctly all literature and sources used in this work and this work is fully compliant with the Dublin Business School's academic honesty policy.

Signature: Tanmay Prakash Salunkhe

Date: 18/08/2018

Acknowledgement

I would first like to thank my Dissertation Supervisor Dr. Shazia Afzal of MSc Information Systems with Computing course at the Dublin Business School. Dr. Shazia was always open to help me whenever I was in trouble or had questions about my research or writing. She consistently ensured that my research paper was unique and my own work. She also corrected me at times when she felt any glitch in my research artefact or writing. As Dr. Shazia was lecturer for the Advanced Databases and Data Analytics subjects and my research was in the same field, I learned a lot from her lectures and followed her advices during the Dissertation.

I would also like to thank the Research Methodology Professor, Dr. Brid Lane, who is also the Program Director for IT and Management Masters at Dublin Business School for sharing the information on how to carry the Research work from start till end and what all factors to be considered while doing the research and writing Dissertation. Also, I would like to thank Dr. Alan Graham professor at Dublin Business School, for giving lectures on Writing for Graduate Studies which helped me to improve my academic writing and Mr. Trevor Haugh, Assistant Librarian at Dublin Business School for providing lectures and guidance on Referencing.

I would also like to acknowledge the HR Professionals from Ireland and India who took part in survey and provided with the valuable inputs which helped me for progressing in Dissertation.

Finally, I wish to express my profound gratitude to my parents, sister, girlfriend and friends for providing me with constant support and continuous encouragement throughout my year of study and process of Dissertation. This accomplishment would not have been possible without them.

Thank you.

Tanmay Salunkhe

Abstract

This Dissertation aims to help the HR and Project Managers in improving the retention rate of the valuable employees in an organization, thereby minimizing the employee turnover cost of the company. The research was carried out in three stages. To improve the retention rate, efforts were made to first, predict the employee attrition; secondly, decide on which employees are valuable and their retention is profitable to the company. Finally, the factors that influence the employee's intention to resign from a company is found out and provided to the HR and Project managers through the HR Analytical application developed using R and Shiny R framework.

Good amount of research has been done while considering the factors for the employee attrition prediction. Also, the survey for these factors has also been carried out among the HR professionals in my network.

Various analysis has been done while selecting the Machine Learning algorithm for training the predictive model. Logistic Regression algorithm is used for building attrition prediction model as it gives the most accurate result. Then, after doing considerable analysis on how to choose the valuable employee and applying methodological assumptions, Decision Model was prepared with the help of conditional logic statements which showcased which employees are valuable and which employees are not. Then the separate employee source file was prepared consisting of the valuable employees and who were also a potential candidate for resignation.

Using the R condition statements, the dashboard was developed which shows all the factors influencing employee attrition so that HR and Project managers can use them accordingly retaining valuable employee.

For developing and testing the application IBM Employee dataset was used (IBM , 2018). This application can be used by the HR Managers to simplify the employee retention decision.

Contents

List of Tables	7
List of Figures	7
Chapter One	10
1.1 Introduction.....	10
1.3 Roadmap to the Dissertation.....	13
Chapter Two.....	15
Literature Review.....	15
2.1 Factors responsible for the employee attrition.....	15
2.2 Study of Data Mining Techniques for predicting Attrition.	18
2.3 Study on factors to be considered for deciding valuable employee.....	21
2.4 Building Decision Tree Model for deciding on valuable employee	22
2.5 Retention Factors and Designing of Result Dashboard displaying the retention factors to users (HR Managers)	24
2.6 Study on IBM Kenexa Talent Management Tool for Attrition Prediction and HRM decisions:.....	26
2.7 Literature Conclusion.....	27
Chapter Three.....	27
Methodology	27
3.1 Introduction.....	27
3.2 Research Philosophy	28
3.3 Research Approach	28
3.4 Research Strategy.....	28
3.5 Primary Data Collection Method	28

3.6	Secondary Data Collection Method	29
3.7	Data Sampling.....	29
3.8	Theoretical Framework	29
3.9	Methodological Assumptions	30
3.10	Technical Framework	31
3.11	Conclusion:	34
3.12	Contextualization:	34
	Chapter Four	35
	Artefact Design and Development.....	35
4.1	Introduction.....	35
4.2	SDLC Methodology	35
4.3	Functional and Non-Functional Requirements	35
4.4	System Design	36
4.4.1	Use Case Diagram.....	36
4.4.2	System Diagram	37
4.4.3	UML Activity Diagram.....	38
4.4.4	Sequence Diagram.....	39
4.4.5	Application framework	40
4.5	Development	42
4.6	Testing.....	48
	Chapter Five.....	49
	Data Analysis and Findings	49
5.1	Introduction.....	49
5.2	Findings.....	49

5.3	Data Analysis	53
5.4	Artifact Results	60
5.5	Conclusion	65
Chapter Six.....		65
Discussions		65
Chapter Seven		66
Conclusion and Recommendation		66
Bibliography		68
Appendices.....		76
Appendix A.....		76
Appendix B		78
Appendix C		78
Appendix D.....		78
Appendix E		78
Appendix F.....		78
Appendix G.....		78
Appendix H.....		78

List of Tables

Table 2.1 1: Correlates of Turnover.....	15
Table 3.7 1: Selected Prediction Models for Attrition prediction	32
Table 5.2 1:Attribute Weighted Multiplier	51

List of Figures

Figure 2.1 1: A model of relationship between quality of work life, satisfaction and retention ..	17
Figure 2.1 2: Model for attrition (Gupta, 2010).....	17
Figure 2.2 1: Decision Tree	19
Figure 2.2 2: Pseudo Code for the Support Vector (Kotsiantis, 2007)	19
Figure 2.2 3: Formalized form of XGBoost.....	20
Figure 2.2 4: Logistic regression (Rohit Punnoose, 2016)	21
Figure 2.2 5: Functional form of the Logistic Regression (Phanish Puranam, 2018)	21
Figure 2.2 6: (Rahul Yedida, 2006)	21
Figure 2.4 1: Entropy (Quinlan, 1985).....	23
Figure 2.4 2: ID3 (Prof. Prashant G. Ahire, 2015) (Quinlan, 1985).....	23
Figure 2.4 3: AF-Association factor	24
Figure 2.4 4: Normalized Form	24
Figure 2.4 5: Gain – Decision Node	24
Figure 2.4 6: Improved ID3 (Prof. Prashant G. Ahire, 2015) (Quinlan, 1985)	24
Figure 2.5 1: Employee retention and Job Satisfaction Model (Bidisha Lahkar Das, 2013)	25
Figure 2.5 2: Basic Model of retention of employees (Gupta, 2010)	26
Figure 2.6 1: IBM Kenexa Talent Management System by IBM Watson (IBM Watson, 2018) .	27

Figure 4.4 1: System Diagram	37
Figure 4.4 2: UML Activity Diagram	39
Figure 4.4 3: Sequence Diagram.....	40
Figure 4.4 4: Use Case Diagram	36
Figure 4.5 1: Application Framework 1	41
Figure 4.5 2: Application Framework 1	42
Figure 4.6 1 : Python Jupyter sample Program for prediction of Employee Salary	42
Figure 4.6 2: Python Jupyter sample output for prediction of Employee Salary	42
Figure 4.6 3: HTML Integration with Python Jupyter Sample Code.....	43
Figure 4.6 4: Before Data Pre-processing UTF Character setting	44
Figure 4.6 5:After Data Pre-processing UTF Character setting.....	44
Figure 4.6 6: With Row Identifier.....	45
Figure 4.6 7: Without Row Identifier	45
Figure 5.2 1: Attrition Attributes selection and weightage.....	50
Figure 5.3 1: KNN Model Accuracy => 85.92%	54
Figure 5.3 2: XgBoost Model Accuracy => 86.97%	54
Figure 5.3 3: Decision Tree Model Accuracy => 86.97%	54
Figure 5.3 4: Random forest Model Accuracy =>86.97%	54
Figure 5.3 5: Logistic Regression Model Accuracy => 91.01%	55
Figure 5.3 6: SVM Model Accuracy => 88.03%	55
Figure 5.3 7: Graph Plot for Age, Business travel, Daily Rate and Department	56
Figure 5.3 8: Plot for Distance, Education, Education level, Education satisfaction and gender. 56	
Figure 5.3 9: Plot for Hourly rate, Job Involvement, Job level and job satisfaction	57
Figure 5.3 10: plot for marital status, monthly income, monthly rate and no. companies worked	58
Figure 5.3 11: Plot for Over time, percent hike, performance rating, relationship satisfaction ...	59

Figure 5.3 12: Plot for stock level, total working years., training times last year, work life balance	59
Figure 5.3 13: Plot for years at company, years in current role, year since last promotion, year with current manager	60
Figure 5.4 1: File Browsing	61
Figure 5.4 2: Load Cleaned data	61
Figure 5.4 3: Model Summary	62
Figure 5.4 4: Prediction output	62
Figure 5.4 5: Prediction output table.....	62
Figure 5.4 6: Valuable employee	63
Figure 5.4 7: Valuable employee table	63
Figure 5.4 8: Retention factors.....	63
Figure 5.4 9: Retention factor table	64
Figure 5.4 10: Data visualization	64

Chapter One

1.1 Introduction

Analytics can be defined as a “logical progression and set of statistical tools”. In a simple way Analytics is a science of analysis (Fitz-enz, 2010). HR analytics is defined as “demonstrating the direct impact of people data on important business outcomes” (Scott Mondore, 2011). “HR Analytics is the systematic identification and quantification of the people drivers of the business outcomes” (IBM Inc., 2017). HR Analytics is the need of the company so that they can spend the money on the right employees rather than spending on wrong ones. With the help HR Analytics, HR Managers can take numerous decisions on investment on employees to get the excellent outcomes that benefits the stakeholders and customers (Scott Mondore, 2011). HR Analytics deals with the Human Resource Management processes and can be used by HR Managers, Project Managers and Line Managers (Fitz-enz, 2010).

Employee Turnover is the most important factor that causes the huge loss to the company. There are many reasons for which the employees leave the company, such as; salary dissatisfaction, stagnant career growth, etc. The loss is not only in terms of the money but also the company sometimes loses the skilled employees who are the most valuable assets to the company (Morrison, 2014). If the company can predict the employee attrition (employees which are going to leave the company) in near future, they can also work on retention beforehand and avoid the loss of valuable employee. The prediction of attrition and retention is the part of the HR Analytics.

Using Machine Learning technique, more specifically Predictive Analytics, we can predict the employee attrition. Machine Learning is a technique used in Artificial Intelligence. “Artificial Intelligence is the science and engineering of making intelligent machines; especially intelligent computer programs” (Mitchell, 1999). “Predictive analytics is the practice of extracting information from existing data sets to determine patterns and predict future outcomes and trends” (Kelleher, et al., 2015). Employee Attrition Prediction helps HR Managers to predict how many employees will resign from the company during a particular period (IBM Inc., 2017). Thus, Managers can work on deciding on the valuable employees and try to retain them. Employee

Attrition Prediction and automation of the retention decision can be done by applying supervised training method and various classification algorithms (Kelleher, 2017). “Supervised machine learning is the search for algorithms that reason from externally supplied instances to produce general hypotheses, which then make predictions about future instances” (Kotsiantis, 2007).

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

1. Direct costs (replacement, recruitment and selection, temporary staff, management time) (Morrell, 2004).
2. Indirect cost (morale, pressure on remaining staff, costs of learning, product/service quality, organizational memory) (Morrell, 2004).

Also, the biggest impact of Employee Turnover includes:

1. Manager faces with lack of employee continuity.
2. The high costs involved in the induction and training of new staff.
3. Issues of organizational productivity (Firth, 2004).

Few of the Currently Used Tools for Employee Attrition Prediction are as follows:

IBM HR Analytics: (IBM Inc., 2017).

IBM Software has one of its products in Talent Management software category called IBM Kenexa HR analytics, powered by IBM Watson Analytics. One of the solutions provided by IBM workforce analytics is “Prediction of employee attrition rate”. It analyses the key drivers for attrition and then the employee attrition is predicted.

SAP Workforce analytics: (SAP Inc., 2017)

SAP SuccessFactors is using predictive analytics to help answer questions related to understanding turnover and identifying and managing flight risk.

Thus, IBM and SAP have found a solution and developed their own application to predict the employee attrition rate, so that company can found a replacement of an employee beforehand.

Research Question - Research Question primarily focuses on how to predict employee attrition, choose valuable employee from them and then find the most effective employee retention factors with the help of machine learning techniques.

Research Objectives - Research objective aims at finding the attributes responsible for employee attrition and then predict the employee attrition with the data mining technique. Once the attrition is found out, study on the factors deciding the valuable employees and after finalizing those factors, build the decision model for valuable employees. And then find the retention factors and display the most effective retention factors for each employee to improve the employee retention.

I am finding the attributes for employee attrition with the help of previous research studies and survey among HR professionals, with the help of questionnaires. Also, with the help of previous study and research on prediction of employee attrition, the most accurate prediction model was developed. Then with the previous case studies, research and methodological assumptions as discussed in methodological chapter, I built the decision model for the valuable employees. Then again with another methodological assumptions I found out the factors most effective for retention and displayed on dashboard.

I have chosen this research topic as I am more interested in Data Analytics and Machine Learning and I have seen the challenges faced by Line Managers, Project Managers and HR Managers in previous companies, when many employees resigned at a same time and project faced various challenges to complete the task in given timelines due to unexpected employee turnover. So, I decided to make an HR Analytical tool so as to help the HR professionals in the employee attrition and retention processes.

Conclusion: Thus, I have explained the research question, research objectives, how I am going to meet the objectives and my interest this research topic.

1.2 Scope and limitations of research

The Scope of the research is divided into five parts - First part contains the study of the factors influencing employee attrition. Second part has study on predictive model analysis along with model accuracy for predicting the employee attrition. Third part has the study on qualities of the valuable employee. Fourth part has study on the Decision tree to be built for deciding on which

Employee is more valuable (like high performance rating) to be Retained. In fifth part I have integrated all the Machine Learning Models and data source and developed the Machine Learning Application displaying the prediction results, model summary, valuable employee and retention result table on dashboard along with graphs and plots. The ML Models and application is developed in R shiny.

The budgeting of the employee turnover to indicate how much the company has saved the employee turnover budget after retaining the valuable employee is not included in in the Dissertation as it gets into the completely other filed of human resource management budget. And hence, I have limited my research work till improvisation of employee retention.

Limitation includes implementation of Prediction System for Employee Attrition due to less training data set, implementation of Decision Making System for Retention due to complex development of the Classification Tree and Algorithm and inconsistent data, accuracy of Decision result, limited access to employee dataset due to GDPR.

My dissertation has a major contribution in the field of HR Analytics in improving the accuracy of the prediction of employee attrition and advancing the application for helping the HR and Project Managers to improve the retention rate of valuable employee by building Decision tree for selecting the valuable employee and finding the factors influencing them to resign, thereby saving the employee turnover budget of the company.

1.3 Roadmap to the Dissertation

The Dissertation is divided into various chapters.

Chapter two consists of the literature review as given below.

Literature theme one shows the study about the factors influencing the employee attrition and the conclusion of the HR survey done for examining these factors.

Literature theme two shows how the machine learning algorithm was chosen for building the predictive model for predicting employee attrition based on the accuracy calculated by confusion matrix.

Literature theme three illustrates the qualities considered during choosing the valuable employees. This also includes the conclusion for qualities chosen for the valuable employee based on research studies.

Literature theme four includes how the decision tree model is built for choosing the valuable employee.

Literature theme five illustrates the study on retention factors and design of result dashboard that was developed for displaying the factors most influencing the valuable employee to resign. So that HR and Project manager can consider those factors while retaining that employee and thereby improve retention rate of the company.

Literature theme Six shows how the IBM Developed the Talent Management System, Kenexa. It also illustrates the features included in it and how the HR Managers utilize it for predicting attrition and taking Human Resource decisions.

Finally, the Literature Review ends with the Conclusion which shows, how the research is different from the previous research and studies done on employee attrition and retention.

Chapter three consists of the research method and methodology wherein I have explained about how I have chosen this research hypothesis, research approach, data collection method, data sampling method and research strategy and its uniqueness along justification for the chosen method.

Chapter four illustrates the artifacts design like application design, UML diagrams of the system design, development method, application testing and process followed during application development and result. It also illustrates how the result was obtained answering the research question on improving the employee retention. It shows all the aspects of the application like data source, data cleaning, data processing, building predictive and decision model for employee attrition and retention respectively.

Chapter five emphasis on data analysis and findings from the research. It also includes how accuracy analysis for the prediction model from the confusion matrix and the analysis on methodological assumptions chosen for finding the valuable employees and retention factors. It

also includes all the parts of the application showing the summary, graphs, tables and analysis results. It shows how the dissertation has brought change in the field of HR Analytics.

Chapter Six includes the in-depth discussion on the research. It shows the objectives of the research and how those objectives were met, and the research was successfully completed. It also shows the limitation of the dissertation and the advancements that can be done to dissertation for obtaining industrial application which can be used in real world industries.

Chapter Seven concludes with the summarization of the findings done during the research work. And the result obtained from the research. It also explains, how this dissertation can further be modified for combining HR and Finance application which shows the budget saved after retaining valuable employee than to hire new employee. Thus, company's overall employee turnover budget can be managed along with Human Resource Analysis.

Chapter Two

Literature Review

2.1 Factors responsible for the employee attrition

There have been considerable studies on Factors responsible for Attrition. One of which is the meta-analysis ("label for a variety of procedures that statistically summarize the information gathered in a literature review") done by John Cotton and Jeffry Tuttle from Purdue University. (John L. Cotton, 1986). The 26 variables which were found out to be responsible for the attrition:

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

Table 2.1 1: Correlates of Turnover (John L. Cotton, 1986)

Organizations invest much in the employee hiring, induction, training, development and retention and hence losing an employee is huge loss to the company. Hence, managers must reduce the employee turnover to reduce the loss of the turnover. “Research estimates indicate that hiring and training a replacement worker for a lost employee costs approximately 50 percent of the worker’s annual salary (Ongori, 2007). Another important factor that plays an important role as a turnover predictor is the job involvement. Job Involvement is the factor that indicates how much an employee is involved in a particular job and carry out the job from start to end with various responsible role. This gives an employee feeling of ownership and motivate him to do that job effectively (Ongori, 2007). One of the factors that help retention, is providing healthy work environment. Also, if reason of attrition is found out, retention can be improved through planning and involvement in discussion with employee (Alex Frye, 2018).

One of an early voluntary study has found that the strongest predictors for voluntary turnover were tenure, overall job satisfaction, job performance, individual demographic characteristics (age/experience, gender, ethnicity, education, marital status), geographical factors, salary, working conditions, job satisfaction, recognition, growth potential etc. (Edouard Ribes, 2017). One of the study reveals about the push factors which includes employee mismatch with job requirement, unbalanced interpersonal relationship with supervisor and peers, work stress, unsatisfactory work-life balance which results into employee unhappiness and influences an employee to resign from an organization (Jessica Sze-Yin Ho, 2010). In a survey of 10,447 business and HR leaders, Deloitte (2017) found that although 71% of companies believe using workforce analytics is important for business performance (Derrick McIver a, 2018).

Another research done on the Indian IT firms reveals the factors responsible for the attrition are Below expectation salary, Low incentives, Relationship with superior, Relationship with subordinates, Job knowledge, Skills utilization, Skills recognition, Acknowledgement of work by superior, Lack of appreciation, Unsatisfied work culture (Archita Banerjee, 2017). The managerial rewards also play an important role in retaining the employee. Thus, manager’s interpersonal skills play a very important role in holding on the evaluable employee (Mitchell Hoffman, 2018).

I have studied the factors influencing the employee attrition in two main industries – Healthcare and BPO. Below are the factors that influence the employee attrition in healthcare industry.

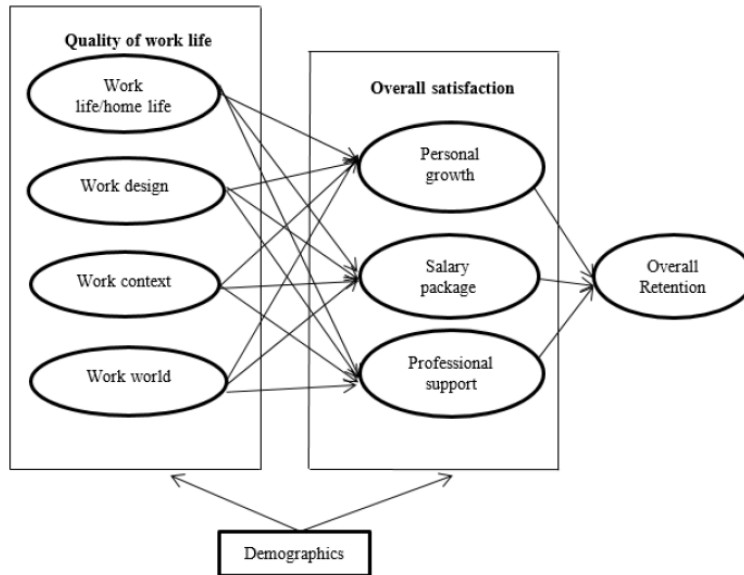


Figure 2.1 1: A model of relationship between quality of work life, satisfaction and retention (Musrrat Parveen, 2016)

The factors influencing the employee attrition in BPO can be illustrated as below:

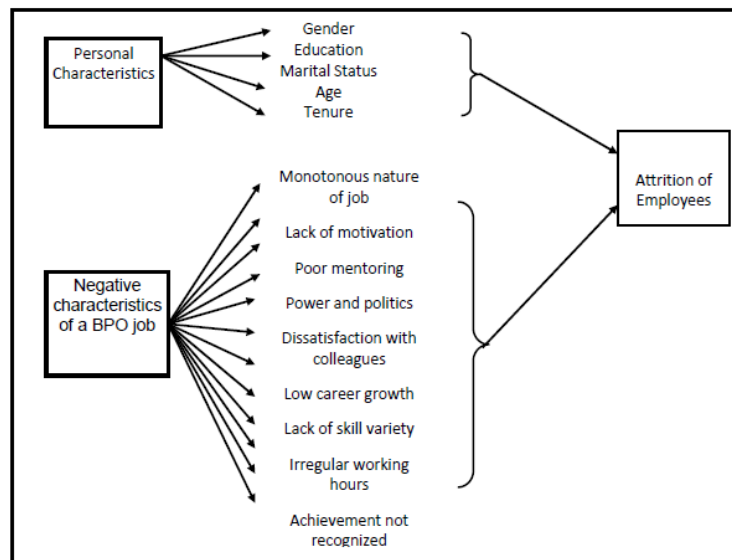


Figure 2.1 2: Model for attrition (Gupta, 2010)

Another mediated and moderated regression analyses states that employee organization misfit and dissatisfaction do not necessarily lead to employee turnover (Anthony R. Wheeler, 2007).

Conclusion: Thus, the factors influencing employee attrition were studied from various above different research studies and case studies.

2.2 Study of Data Mining Techniques for predicting Attrition.

The Predictive Model can be developed by various methods of data mining. Data mining can be helpful to Human Resource Managers in identifying factors influencing employees for high attrition. “Data Mining is a process through which valuable knowledge can be extracted from a large dataset.” One of the researches on the prediction of Employee Attrition was done using Decision Tree as a data mining tool. “Decision Trees are tree-shaped structures that represent decision sets.” It uses Classification algorithms for data mining. It is used to explore the possible outcomes for various inputs (Alao D., 2013). Predictive Analytics is the prediction of future by using past and current data. Predicting future depends on four things which are knowledge of past and present events, cause for the events, understanding of patterns and pattern variations, correct tool to predict the future and accuracy of the same. This can be done by statistical modelling and data mining techniques (S. Chitra, 2018). The model developed for prediction of attrition need to be checked for under-sampling and oversampling. “In under-sampling a random subset of the majority class is used for training, whereas oversampling randomly duplicates instances of the minority class.” (Bernd Bischl, 2016). One of the research studies was carried on using SVM and Random Forest models for employee attrition prediction (Rohit Punnoose, 2016) and another was carried on using KNN algorithm for employee attrition prediction (Raschka, 2017). In one of the researches xgBoost model was used for carrying on the research on employee attrition prediction (Greenwell, 2017) and in one research, GLM i.e. Logistic regression was used (Raschka, 2017).

The ML Algorithms studied are as follows:

Decision Tree – “Decision trees are trees that classify instances by sorting them based on feature values.” Each node in a decision tree signifies a characteristic in an instance to be classified, and each branch denotes a value that the node can assume (Kotsiantis, 2007). The Decision Tree can be depicted as below:

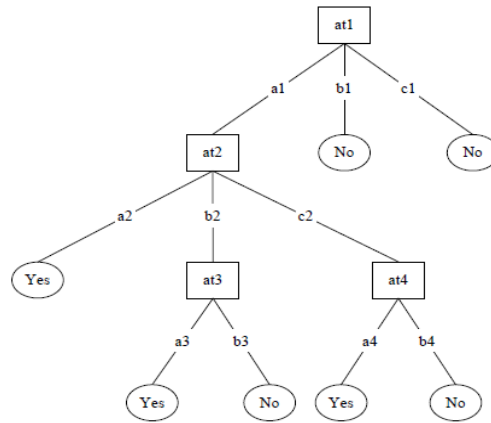


Figure 2.2 1: Decision Tree

The decision tree algorithm studied by Kotsiantis was the C4.5 which is an extension of Quinlan's earlier ID3 algorithm (Kotsiantis, 2007). It is given as J48 in R-Weka. Another study makes the use of C5 decision tree algorithm (Rohit Punnoose, 2016). Bagging is one more type of Decision tree Algorithm which is studied by S. Chitra and Dr. P. Srivaramangai (S. Chitra, 2018). CHi-squared Automatic Interaction Detector (CHAID). Performs multi-level splits when computing classification trees also comes under the umbrella of the Classification and Regression Tree (CART) Analysis (Alao D., 2013).

Support Vector Mechanism – Support vector machines (SVMs) does a class division by finding a division point for categorizing in a hyper plane in a high dimensional space. A decent partition is done by the hyper plane that has the separation between the separate categories as wide as possible. The bigger the edge is, the lower the error and inaccuracy of the classifier (B-Gent, 2006) (S. Chitra, 2018).

```

1) Introduce positive Lagrange
multipliers, one for each of the
inequality constraints (1). This
gives Lagrangian:

$$L_p \equiv \frac{1}{2} \|w\|^2 - \sum_{i=1}^N \alpha_i v_i (x_i \cdot w - b) + \sum_{i=1}^N \alpha_i$$

2) Minimize  $L_p$  with respect to  $w$ ,
 $b$ . This is a convex quadratic
programming problem.
3) In the solution, those points
for which  $\alpha_i > 0$  are called "support
vectors"

```

Figure 2.2 2: Pseudo Code for the Support Vector (Kotsiantis, 2007)

SVM can solve linear as well as nonlinear binary classification problems (Rohit Punnoose, 2016). SVMs are also referred to as maximum margin classifiers (Edouard Ribes, 2017).

K-Nearest Neighbour (KNN) – Neighbors based order is a kind of instance-based learning. Characterization is figured from a basic greater part vote of the k closest neighbors of each point (Raschka, 2017). “The 2 stages for classification using KNN involve determining neighboring data points and then deciding the class based on the classes of these neighbors” (Rohit Punnoose, 2016). “KNN is a non-generalizing method, since the algorithm keeps all of its training data in memory, possibly transformed as a ball tree or a KD tree”. The Manhattan distance is computed using the formula $D = \sum |x_i - y_i|$ (Rohit Punnoose, 2016).

Random Forest – Random Forests (RFs) (Wiener, 2002), result from the combination of tree classifiers. Each tree depends in the estimations of an irregular vector inspected separately. RF has characteristics of correcting decision tree over-fitting for the training dataset.

Random forests are different from standard trees as its each latter node is split using the best category split among all variables. “In a random forest, each node is split using the best among a subset of predictors randomly chosen at that node” (Rohit Punnoose, 2016).

xgBoost (xgbTree) – XGBoost is a boosted tree algorithm. It uses a more regularized model formalization to control over-fitting, which gives it better performance.

$$f_t(x) = w_{q(x)}, w \in \mathbb{R}^T, q: \mathbb{R}^d \rightarrow \{1, 2, \dots, T\}$$

Figure 2.2 3: Formalized form of XGBoost

Where ‘w’ is the vector of scores on leaves, ‘q’ is a function assigning each data point to the corresponding leaf and ‘T’ is the number of leaves (Rohit Punnoose, 2016). xgbTree is the method used for the eXtreme Gradient Boosting (method = ‘xgbTree’) and is used for classification and regression with the help of xgboost and plyr packages (Kuhn, 2018). XGBoost (eXtreme Gradient Boosting) is a general library providing enhanced distributed gradient boosting that is precisely designed to be highly efficient and provide accurate result (Greenwell, 2017).

Logistic Regression (GLM) – Logistic Regression is a machine learning calculation for characterization. In this calculation, the probabilities portraying the possible results of a separate trial are displayed utilizing a Logistic Regression (Raschka, 2017).

$$p(\text{churn}|w) = \frac{1}{1 + e^{-[w_0 + \sum_{i=1}^N w_i x_i]}}$$

Figure 2.2 4: Logistic regression (Rohit Punnoose, 2016)

$$y = \ln\left(\frac{P(Y=1|x)}{1-P(Y=1|x)}\right) = w_0 + \sum_{i=1}^p w_i \cdot x_i$$

(x is a vector of independent variables of dimension p and y is the logit (log odds).
 w_0, w_1, \dots, w_p are model parameters)

Figure 2.2 5: Functional form of the Logistic Regression (Phanish Puranam, 2018)

Logistic regression is a regression model that fits the values to the logistic function. It is useful when the dependent variable is categorical (Rahul Yedida, 2006). The general form of the model is

$$P(Y|\bar{X}, W) = \frac{1}{1 + e^{-(w_0 + \sum w_i x_i)}}$$

Figure 2.2 6: (Rahul Yedida, 2006)

2.3 Study on factors to be considered for deciding valuable employee

Work Engagement describes the extent to which employees are involved in their work. According to the study by Mark Attridge it is revealed that the position and the level of job also decides the employee engagement with work. Employees with Director and Executive level are more involved in the work than the support staff. Also, highly educated and highly skilled workers are also involved more in their work (Attridge, 2009). “Human Resource Information System (HRIS)” plays an important role in the decision-making process for effective Human Resource Management (HRM). “Intelligent Decision Support System (IDSS)” along with “Knowledge Discovery in

Database (KDD)” is applied in HRIS to improve structured, especially semi structured and unstructured HR decision making process. This module will offer a total performance index, considering all criteria, for an employee. This process is very complex as there are many rules for each criterion with different priority. It is one of the factors to be considered during choosing valuable employee. These are the skilled employees and are predicted by Machine Learning and decision-making criteria using historical data (Abdul-Kadar Masum, 2015). Another study reveals that the employee performance is more if the job satisfaction is more and hence the job satisfaction is indirectly responsible for deciding the valuable employee (Alf Crossman, 2003). One of the studies done shows that the work relevance and the employee skillset play a very important role in deciding on the performance of an employee (Blake A. Allan, 2017). Performance of an employee help decide Positive Organizational Behavior based on which valuable employee can be determined (Fred Luthans, 2008). Another thing that impacts the employee performance is the social relationship factor. This needs to be considered while deciding valuable employee (Nina van Loon, 2018). Another major factor to be considered for deciding on employee value is the employee loyalty i.e. how long an employee has been employed with the one organization. This all depends on the job satisfaction (Onsardi, 2017).

2.4 Building Decision Tree Model for deciding on valuable employee

At present in HRM (Human Resource Management), it is necessary to take the timely decision and correct decision for which Intelligent HR Decision Making System is made, one of which is Decision Support System DSS which helps HR to take the decision in shorter period. For this, DSS (Decision Support System) and KDD (Knowledge Discovery in Database) and Data Mining is implemented for Extraction of Knowledge for HRM (Abdul-Kadar Masum, 2015). Valuable Employee can be predicted by the Employee Performance and the employee performance can be found out by K-Mean Clustering Method or Decision tree Method. Decision Tree generates a decision tree from the given training data. “Decision Tree is one of the most used techniques, since it creates the decision tree from the data given using simple equations depending mainly on calculation of the gain ratio, which gives automatically some sort of weights to attributes used, and the researcher can implicitly recognize the most effective attributes on the predicted target.” (Ananya Sarker, 2018). While choosing the valuable employee, various factors were taken into

consideration, hence enhanced decision tree was required. One of the studies revealed about the improved ID3 algorithm which is the enhanced version of the ID3 Decision tree algorithm. ID3 algorithm is a recursive method and there is an evaluation of each subset and decision node is created based on metric known as Information Gain. ID3 uses two major concepts – Entropy and Information Gain. “Entropy is a measure in information theory to measure the impurity of arbitrarily collection of items.” For set S, being p_i the probability of S belonging to class I, formulation is (Algorithm, 2009)

$$\text{Entropy } H(P) = - \sum_{i=1}^n P_i \log_2 (P_i)$$

Figure 2.4 1: Entropy (Quinlan, 1985)

Information Gain = $I(S_1, S_2, S_3, \dots, S_m) - E(A)$ Algorithm for generating a decision tree according to a given data sets (Prof. Prashant G. Ahire, 2015) (Algorithm, 2009).

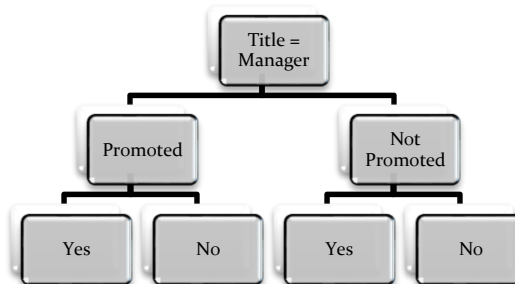


Figure 2.4 2: ID3 (Prof. Prashant G. Ahire, 2015) (Quinlan, 1985)

The shortcoming of the ID3 is only one attribute is tested at a time and it may be over fitted or over classified. To overcome this shortcoming, improved ID3 is used. Complexity is reduced, and time is saved in Improved ID3 algorithm. In Improved ID3, the same gain which is calculated initially in basic ID3 is used which get changed every time when the dataset gets modified as tree grows (Kirandeep, 2018).

$AF(A) = \frac{\sum_{i=1}^n x_{i1} - x_{i2} }{n}$	$V(k) = \frac{AF(k)}{AF(1) + AF(2) + \dots + AF(m)}$	$Gain(A) = (I(s_1, s_2, \dots, s_m) - E(A)) \times V(A)$
Figure 2.4 3: AF-Association factor	Figure 2.4 4: Normalized Form	Figure 2.4 5: Gain – Decision Node

(Prof. Prashant G. Ahire, 2015) (Quinlan, 1985)

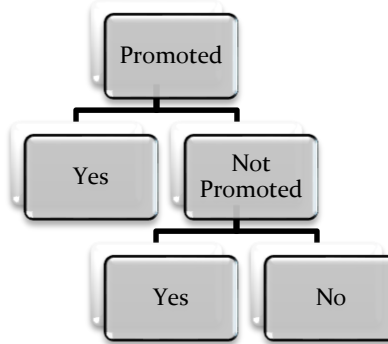


Figure 2.4 6: Improved ID3 (Prof. Prashant G. Ahire, 2015) (Quinlan, 1985)

2.5 Retention Factors and Designing of Result Dashboard displaying the retention factors to users (HR Managers)

Various tools have been developed for improving customer retention by various companies, similarly tools for employee retention needed to be developed for employee retention and the case study done by Edouard Ribes, Karim Touahri and Benoit Perthamez shows how machine learning can help in employee retention by using Employee turnover prediction and classification method (Edouard Ribes, 2017). Employee retention can be defined as an “effort by an employer to keep desirable workers in order to meet the business objectives” by keeping the right people on the right jobs (Jagun, 2015). Bidisha Lahkar Das and Dr. Mukulesh Baruah have found out the various factors affecting the employee retention through their study. These factors are Compensation, Rewards and Recognition, Promotion and Growth opportunity, Participation in Decision making, work-life balance, Work environment, Training and development, leadership and job security (Bidisha Lahkar Das, 2013) (Jagun, 2015).

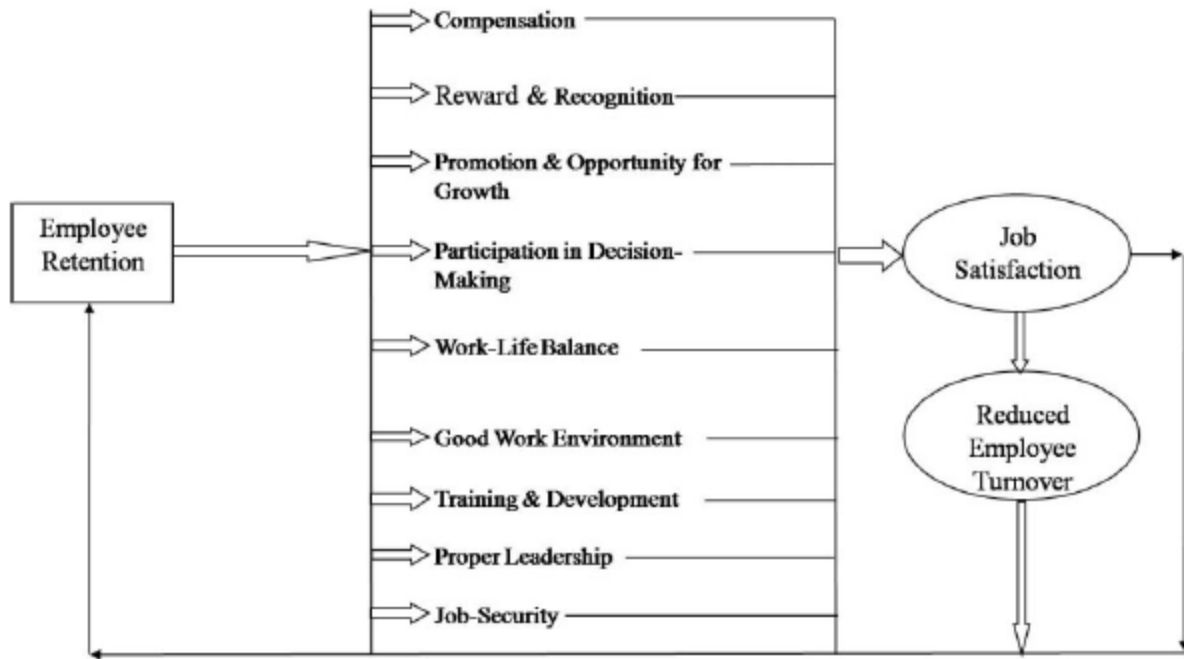


Figure 2.5 1: Employee retention and Job Satisfaction Model (Bidisha Lahkar Das, 2013)

While deciding on the retention, the efforts should be made on holding on the best employee. The best employee can be chosen by his/her loyalty. That is the number of years spend in the company shows the loyalty of the employee (T.MURALIDHARAN, 2017). There was recently one paper published from a Conference in University of Salford, Manchester, which showcased the important factors of employee attrition and retention while comparing between the retention in inhouse offshoring and offshore outsourcing in software development industry (Bass, 2018). Santoshi Sen Gupta has developed a basic retention model after the study done on the attrition and retention of the employee in BPO. The model is as given below: (Gupta, 2010).

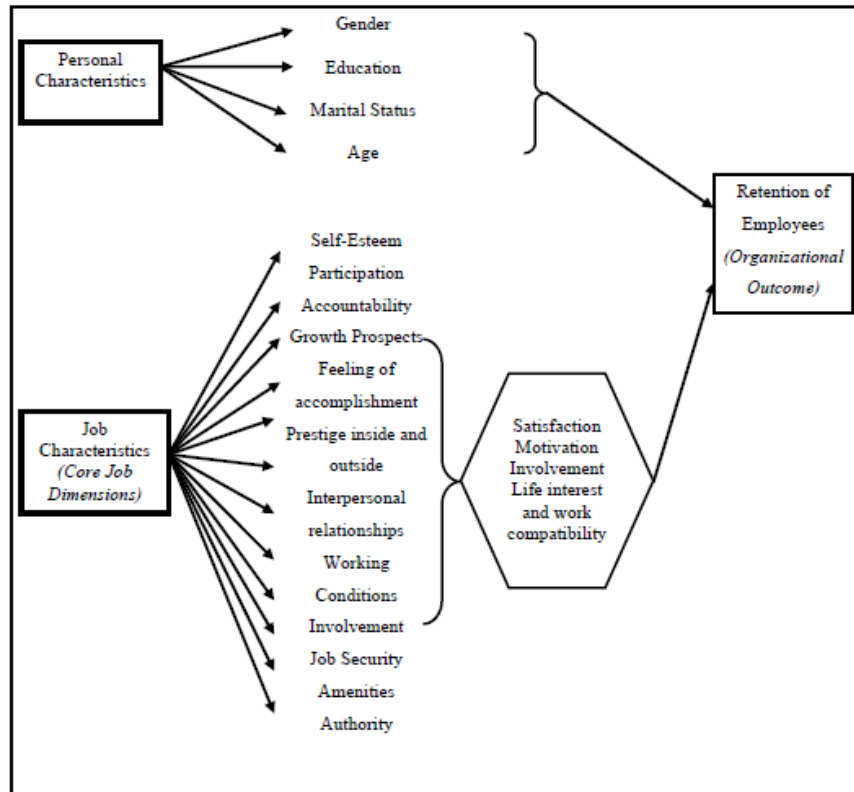


Figure 2.5 2: Basic Model of retention of employees (Gupta, 2010)

2.6 Study on IBM Kenexa Talent Management Tool for Attrition Prediction and HRM decisions:

IBM has developed a Talent Management tool called “Kenexa” and is used to predict the employee attrition with the employee data which is input manually by HR Managers (IBM , 2018). The working and UI Application of the IBM Watson Talen Management tool, Kenexa is as described below (IBM Watson, 2018). IBM Watson has prepared its own UI Application for HR and Hiring Managers. Below is the preview of it and the development and working of an application are explained in the “Appendix A”.



Figure 2.6 1: IBM Kenexa Talent Management System by IBM Watson (IBM Watson, 2018)

2.7 Literature Conclusion

Thus, we can see that till date there has been various theoretical and technical research and studies have been carried out to find the attrition prediction. But there has been no significant study or research on development of tool which can take automated decisions on categorizing valuable employees and ordinary employees. And there is no application that shows the final dashboard that shows the retention factors which HR Managers must consider while retaining the valuable employee; so that the Human Resource Management budget can be reduced significantly if retention rate is increased.

Chapter Three

Methodology

3.1 Introduction

This section provides the theoretical and technical walkthrough of the research method used for building an analytical application using R for predicting employee attrition and how to recognize the valuable employee and retain them, thereby saving the company HRM budget on hiring new

employee. This chapter describes the set of methods used for carrying out the research and building a software application. It also covers the data collection and data analysis methods.

3.2 Research Philosophy

The purpose of this research is to predict the employee attrition and improve retention of valuable employee, thereby saving the HRM cost. For the research study, both the qualitative and quantitative methodologies are used. Qualitative methodology is used for gathering the employee attributes and demographics for predicting attrition as well as finding the retention factors for retaining a valuable employee, whereas I used the quantitative methodology for weighting the factors influencing the attrition and analyzing the accuracy of the prediction model build for predicting attrition.

3.3 Research Approach

For completing my research, I followed the inductive approach. Inductive approach is the one in which there is a systematic study on observation and previous research for proposing the theories and finding the patterns with the use of various assumptions and hypotheses. The assumptions and the approach for research changes with the progress in the research.

3.4 Research Strategy

For successfully completing the research I took the help of Survey which was carried among the HR Professionals to collect the factors and attributes of employee for predicting employee attrition and improving employee retention. Also, Experiments were carried on processing and application of machine learning algorithms on the dataset and increasing the accuracy of the research result.

3.5 Primary Data Collection Method

The primary data source is the most important aspect of any research as it is the initial point of starting the dissertation. Hence it requires to be genuine and must provide accurate data with evidence. In this research, the primary data was collected by the study on various research papers and case studies on employee attrition, predicting employee attrition and Factors influencing retention. The weighting for each factor influencing the attrition was given after the survey was carried among the HR Managers and Professionals. The detailed information about the survey is given in “Appendix B”.

3.6 Secondary Data Collection Method

Secondary data was used in my research for implementing the research technically. Thus, whatever was studied theoretically, needed to be implemented practically wherein I referred to Technical thesis and other research Journals for transforming business logic to technical implementation. Here, I studied about the Machine Learning Algorithms and how they can be implemented for predicting employee attrition and implementing decision tree for categorizing the valuable employee from ordinary one. I also studied about the previous implementations done by IBM Watson for workforce analytics and Talent Management tool (IBM Watson, 2018) and how I can improvise those tools and make considerable advancements in those.

3.7 Data Sampling

For data sampling, I have used the Convenience sampling technique. Convenience sampling is a type of non-probability sampling that includes the sample gathered from that part of the population that can be easily approached. In this case it was collecting the employee attrition and retention factors from the HR professionals within my social connections with the help of Survey questionnaires' form, as it was easy to connect with them and get the survey done.

3.8 Theoretical Framework

Attrition attributes – With the study on various research papers and surveys from HR managers located in India and Ireland, the important attrition factors were found out and those were used as predictors for predicting employee attrition. Those factors are listed below:

Percent Salary Hike, Monthly Income, Years Since Last Promotion, Distance from Home, Job Role, Performance Rating, Job Level, Environment Satisfaction, Years in Current Role, Relationship Satisfaction, Years with Current Manager, Job Satisfaction, Work Life Balance, Number of Companies Worked, Years at Company, Over Time, Total Working Years, Marital Status, Age and Gender.

Valuable Employee Attributes – Similar to the attrition factors, valuable employee factors were found out by study on various research papers. And with the help of methodological assumptions and conditional logic, the valuable employees were categorized from the ordinary ones.

Retention Attributes – Likewise, after doing research on the case studies on retention, the important factors were found out for improving retention. And with the help of conditional statements and logic the respective factors were displayed on the dashboard of the application.

3.9 Methodological Assumptions

While developing an employee analytical application, I made use of two theoretical and two technical assumptions.

Methodological Assumption 1: The first assumption was theoretical and made while choosing the attributes for deciding on the valuable employee. The assumption was made based on the study about past researches and case studies mentioned in Literature. So, the attributes I considered for deciding valuable employees are as follows: (John L. Cotton, 1986) (Jessica Sze-Yin Ho, 2010) (Gupta, 2010) (Archita Banerjee, 2017) (Edouard Ribes, 2017) (Alex Frye, 2018).

Employee Performance, Peer-Manager Relationship, Number of years at current company, Education level, Job Level, Job Involvement, Job Title, Job Satisfaction.

Methodological Assumption 2: Similarly based on past research and case studies, I had considered following attributes as effective retention factors: (Edouard Ribes, 2017) (Bidisha Lahkar Das, 2013) (Jagun, 2015) (T.MURALIDHARAN, 2017) (Bass, 2018) (Gupta, 2010).

Promotion, Percent Salary Hike, Work Life balance, Work environment, Job satisfaction, Job involvement, Peer relationship.

Methodological Assumption 3: For developing decision tree algorithm for choosing valuable employee, decision tree and advanced decision tree algorithm was chosen. But due to data inconsistency and data quality issue, the decision tree was not able to build due to underfitting data. Hence, as a work around, I decided to build the decision tree using if else conditioning statement. For this to work, I grouped the attributes in two parts – one with highest priorities and other with lowest priorities. The Highest priorities attributes would come with the Boolean “OR” logic and lowest priorities attributes would come with Boolean “AND” logic in if else statement. I calculated the threshold value for each column. Based on the value, if any record has highest priority column above the threshold it would come in valuable employee category or if all the lowest priority columns come above the threshold then the employee would be categorized as

valuable. The threshold value was calculated by taking the 1st quartile value (value at approx. 30th percent of the all the records when arranged in ascending order).

Methodological Assumption 4: Similarly, while developing if else conditioning statement, the threshold value was calculated for all the columns considered as retention factors. Here, it was the mean value of the total records from the original dataset. If the value of the attribute is below threshold then it is added in the “retention factors” column or else is “x” character is added. Thus, finally the “retention factors” column will display all the attribute names for all the records whose values are below threshold value. Thus, HR Managers need to pay more attention on those factors while retaining the valuable employees.

3.10 Technical Framework

Data selection and Data processing – As per the above study of Research Journals and various case studies, I have chosen the IBM Watson Data (IBM , 2018) for Building the HR Analytics Attrition and Retention Model and Testing it with the same data.

In the beginning, I tried to develop the prediction model for attrition using Python Jupyter by log in with the help of Anaconda Command Prompt. But then I wanted to develop an application for data visualization too and user interactive application, so I decided to go with Python flask and Django. But the installation of the libraries and environment setup was too time consuming and complex. Hence decided to move with Tableau and HTML for application development, but with Tableau, only data visualization was possible and Predictive Analytics and decision tree logic was not possible. Hence, I required a tool which can do both Machine Learning task as well as Data Visualization and I can use it as user interactive application. Hence, I decided to go with R framework and in R, “R Shiny” provided me the platform to develop the analytical application (Hadley Wckham, 2017).

Before using the data for training in Machine Learning Algorithm and testing the test data, I have preprocessed the data into usable format by performing following steps: (Hadley Wckham, 2017)

1. Removed NULL values
2. Removed Columns with non-unique values
3. Removed unnecessary columns

4. Transformed the unsupported UTF format to supporting UTF format
5. Rearranged the columns according to weightage of each column
6. Saved the cleaned data in a csv format file

Predictive Modeling for Attrition – After studying various algorithms from above research journals and case studies, I have selected following Algorithms to develop a predictive model for attrition, as these algorithms provided the desired result for the prediction. I wanted to build a classification model and these algorithms helped me to develop the same with various ranges of accuracy. (Alao D., 2013) (Bernd Bischl, 2016) (Kotsiantis, 2007) (Rohit Punnoose, 2016) (Edouard Ribes, 2017) (Raschka, 2017) (Wiener, 2002) (Kuhn, 2018) (Greenwell, 2017) (Phanish Puranam, 2018) (Rahul Yedida, 2006).

- | | |
|----------------------------------|--|
| • KNN (k-Nearest Neighbor) | • Decision Tree |
| • SVM (Support Vector Machine) | • Random Forest |
| • GLM (Generalized linear model) | • XGB Tree (Extreme Gradient Boosting) |

Table 3.7 1: Selected Prediction Models for Attrition prediction

After developing the prediction models with above algorithms and training the models with training data set, I ran the test data for each model and validated the confusion matrix. Confusion matrix shows the various accuracy of the model based on prediction result. After comparing the confusion matrix of all, I found out that GLM had the most overall accuracy from all the other models. Hence, I chosen the GLM algorithm to build the model in the final Analytical application (Peter Bruce, 2017).

Decision Tree modeling for Valuable Employee – Then with the help of decision tree Algorithm (Peter Bruce, 2017) and “if else” conditional statement, I found out the valuable employee from the given data. In the Decision model the following attributes were given the most preferences. Performance Rating, Job Involvement, Years at Company, Job Satisfaction and the least preference is given to Job Level, Relationship Satisfaction and Education. Then for Building Decision Model for Deciding on valuable employee, the methodological assumptions 1 and 3 were considered. With the assumption 1, the threshold value of the all the attributes was decided by calculating the value at 30th percent of each data columns. First, I made each column in ascending order. Then, I calculated the count of record at 30th percent of the total count of records. Then, in each column if

the employee had value less than the value at 30th percent, I considered him as ordinary employee and all the above as valuable employee based on the preferences of the attributes. The Decision tree was to be used for the same but due to data inconsistency and data quality the decision tree was not able to be applied. This part of research failure is explained in “Appendix C”. And hence the manual if else conditioning was needed to be done. The company can change the threshold value and the preferences according to their use for further application to get the desired output. According to assumption 3, I have taken the threshold as the value at 1st quartile of the total record. Then the column was added in front of each record that whether an employee is Valuable or Not in form of “YES” and “NO”. If its valuable, the record has “YES”, else its “NO”. The Yes and No decision can also be validated by HR Managers and Project Managers during real time run.

Application building for displaying retention factors of Valuable Employee – Similarly, while displaying the factors for retention, I used the methodological assumption 2 for finding the attributes which would help HR Managers to take into consideration the most while retaining valuable employee. Here I have considered the assumption 4 for deciding on the threshold value of the retention factors attributes. Thus, according to assumption, the threshold value for the decision is taken as the value at the median of the total record. As did in deciding valuable employees, similarly, the attributes required for retention decision were arranged into ascending order and the value at median of the total record was found out. If an employee has any selected attribute value less than the threshold value, then the attribute is added into the retention factor’s list for that employee. Finally, all the attributes whose value are less than threshold value, are concatenated and displayed in one single column for all the employee record.

In the application I have made the functional button which does the above given tasks. The “Logistic Regression” Button helps to build the prediction model and give the prediction output and model summary.

The “Retention Factor” Button displays the valuable employee list along with the factors responsible for improving the retention of valuable employees.

Finally, I have made the provision for the data exploration (John D. Kelleher, 2015) of the given data, where I have provided the drop-down menu for plotting the two 2-Dimensional plots/graphs with one constant axis of Attrition and one 3-Dimensional plot/graph with one constant axis as Attrition. Thus, HR Managers and Hiring Managers can work on Human Resource Analysis for

Predicting Attrition, working on retention and thereby engaging in saving the HRM budget cost of the company.

3.11 Conclusion:

Thus, I developed an analytical Application for the HR Managers to predict attrition and improve retention of valuable employee. For this, first I gathered the attributes required for attrition prediction, deciding on valuable employees and displaying the retention factors. Then I preprocessed the data and using the predictive model predicted the attrition. After, the attrition was predicted, with the help of four methodological assumptions given above, I was able to categorize the valuable employee from the ordinary one and display the most effective retention factors which HR Managers can pay attention to the most while retaining the valuable employees.

Finally, everything was put in an application to perform these activities of predicting attrition, selecting valuable employees and displaying retention factors. Also, drop-down menu for the attributes helped in data visualization and data exploration. Thus, user interactive data visualization and data analytical application was developed using R framework due to its uncomplicatedness (Hadley Wckham, 2017).

3.12 Contextualization:

Thus, it can be seen from the above researches, case studies and IBM Kenexa talent management tool description, that there has been various research on prediction of employee attrition, but there is no application which includes all the three functionalities of attrition prediction, decision on categorizing valuable employee from ordinary ones, displaying most effective retention factors. Whereas, my HR analytical application for the employee performs these three functionalities and help to provide the most effective retention factors to the HR Managers so that they can pay more attention on those factors while retaining valuable employees, thereby saving the Human Resource Management Budget. Along with it, I have also provided the user interactive space in application for the user to perform data exploration and data visualization for further data analysis and taking customized decisions.

Chapter Four

Artefact Design and Development

4.1 Introduction

This section tells about the Software Development Life Cycle (SDLC) methodology used, Functional and Non-Functional Requirements, System Design including System Diagram, Sequence Diagram, Application framework, UML Activity Diagram, Use-case diagram along with various use case scenarios, Development and Testing processes performed.

4.2 SDLC Methodology

I have chosen the Agile methodology for the development of an application. I have chosen this method for development to complete the work in given timelines. I completed the work in sprints and constructed the application with unit testing in every step after development.

4.3 Functional and Non-Functional Requirements

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

- a. Predict Attrition – User must be able to predict the attrition of the future employee based on the historical dataset of previous employees.
- b. Decide on Valuable employee – Allow user to categorize the employee into valuable and ordinary ones.
- c. Find out and list down the factors affecting the retention decision – Display the retention factors on a dashboard, for improving the retention of the valuable employees.
- d. Data Exploration and Data Visualization – User Interactive selection of attributes for plotting the graphs against the Attrition.

The Non-Functional requirement included the time required by the application for making prediction of attrition, decision of valuable employee and displaying retention factors should be

minimal. I have taken it as 90 seconds on an average (for dataset record of 1470) to perform all the tasks in application except data exploration task.

4.4 System Design

4.4.1 Use Case Diagram

The Use case design is as follows which shows how the user (HR Manager) can interact with the application and take decisions with regards to valuable employees.

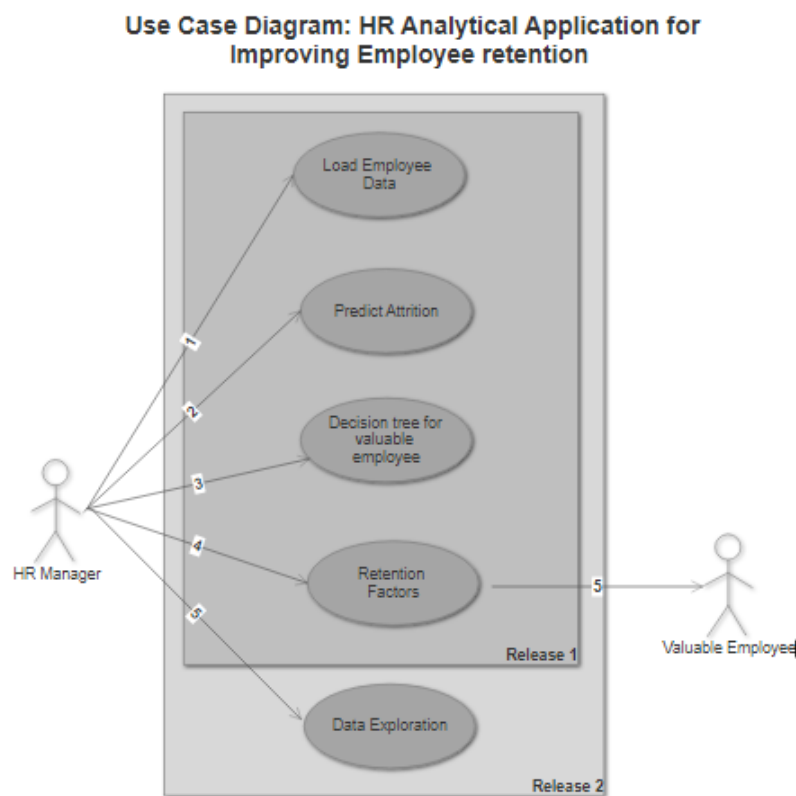


Figure 4.4 1: Use Case Diagram

Below are the various scenarios and exceptions that can take place while user (HR manager) interact with an application.

Use Case Scenario 1: The data is perfectly fed into the system and attrition is predicted with the expected accuracy. The employees who are going to resign from the company are then categorized into valuable and ordinary employees using decision tree. The most effective retention factors for valuable employees are then displayed on the dashboard.

Use Case Scenario 2 (Exception): The data is perfectly fed into the system, but the attrition prediction does not give the expected accuracy. User can manually perform the data analysis with the help of user interactive data visualization tab provided in the application.

Advancement in application to handle these exceptions can be made if the application allows the user to customize the conditioning logic for choosing the valuable employee and finding retention factors of the valuable employees.

4.4.2 System Diagram

The system diagram below, shows the system design of an application which illustrates how the raw data is provided to the system and then retrieved as an output and again fed to system as a second input and get the final report (dashboard result in this case).

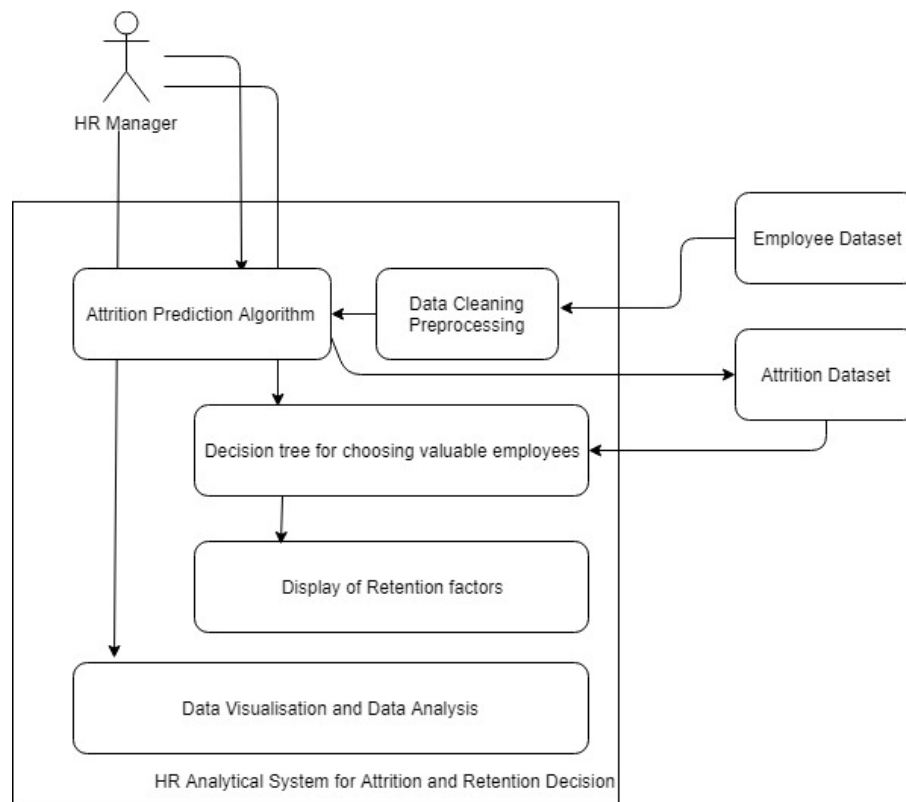


Figure 4.4 2: System Diagram

4.4.3 UML Activity Diagram

The system design can be explained with the help of the following UML Diagram which represent the working of the Analytical Application. According to below UML Diagram, I have collected the data from IBM created by IBM Data Scientist for academic and research work. Then I have chosen the attributes for the prediction and further analysis based on HR feedback form circulated to HR professionals through LinkedIn. Then, I preprocessed the data and stored in csv format in a local directory. Then, the test data is run through the trained predictive model which classifies the data into positive and negative attrition and the result is shown as “YES” or “NO” for the attrition. If the result is YES, all the records with that result are stored in an object and then that data is run through “if else” conditioning statement. This decision statement is to classify the valuable employee from the ordinary one. Once the valuable employees are categorized, that list is stored in another object and all the factors affecting the retention are found out and displayed in the last column of the data set. Thus, in the final stage of application, all the retention factors which are maximal effective to improve retention are displayed on a dashboard. The UML Activity diagram is as follows:

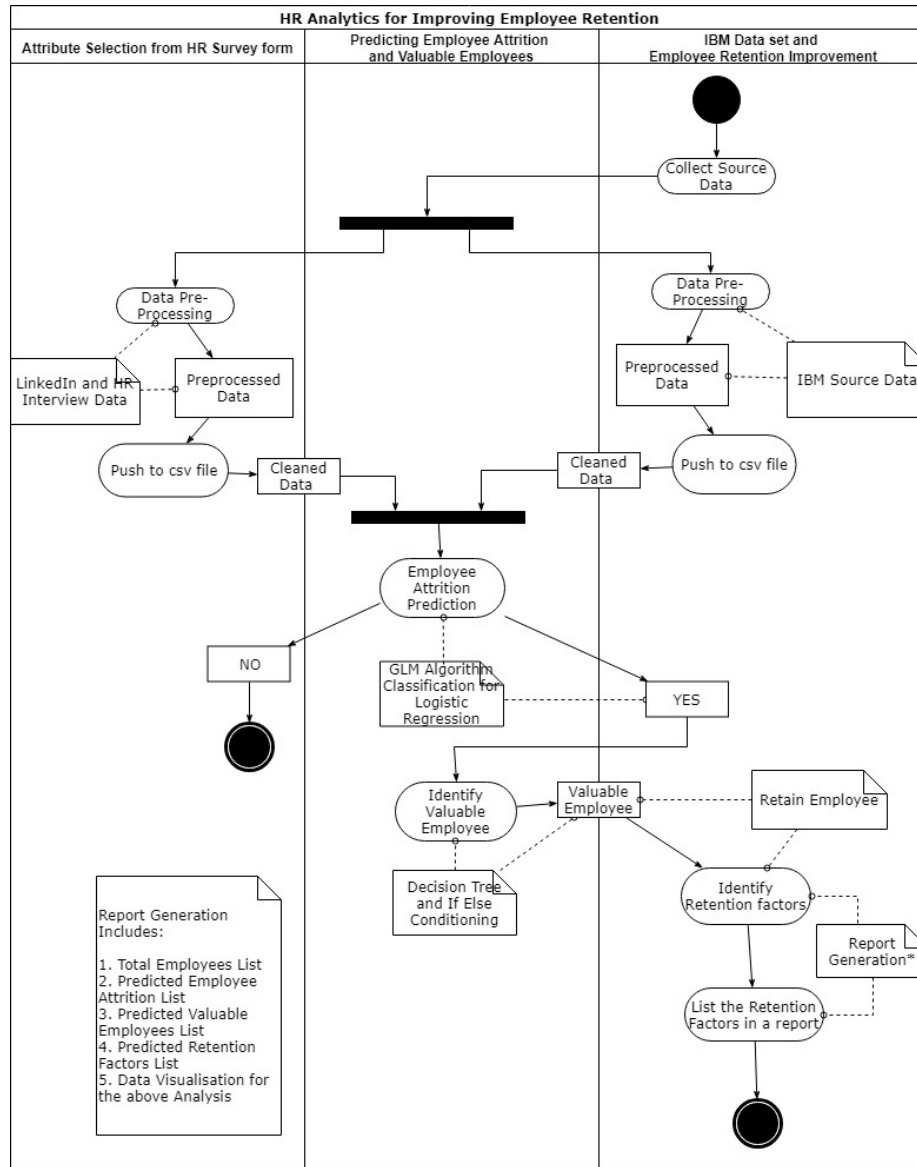


Figure 4.4 3: UML Activity Diagram

4.4.4 Sequence Diagram

Sequence diagram demonstrates the sequence of action performed during the run time of an application. The below diagram shows, how the user sends the request to each block of code and receives the output in return.

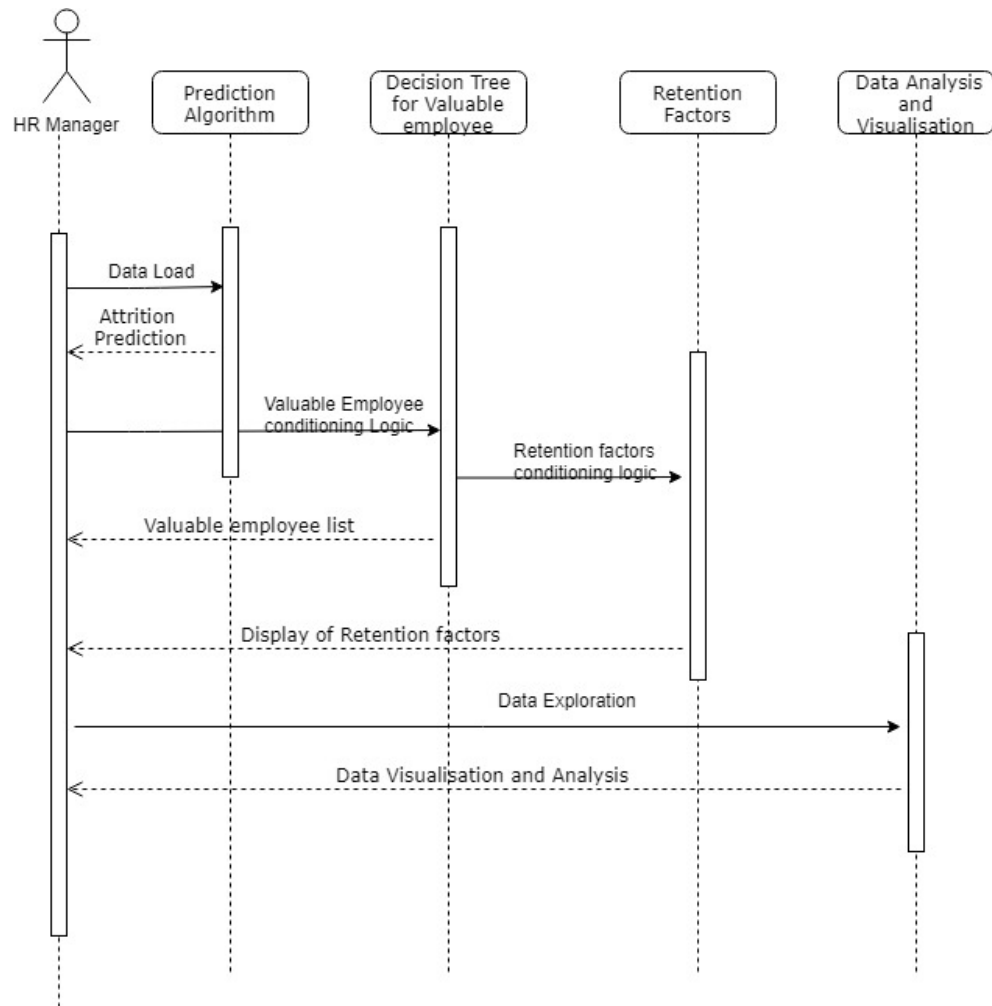


Figure 4.4 4: Sequence Diagram

4.4.5 Application framework

I have chosen R as an application framework as it is less complex, provide both the Machine Learning and data visualization functionality. Also, it's very easy to develop an analytical application with R. The final application representation is as follows:

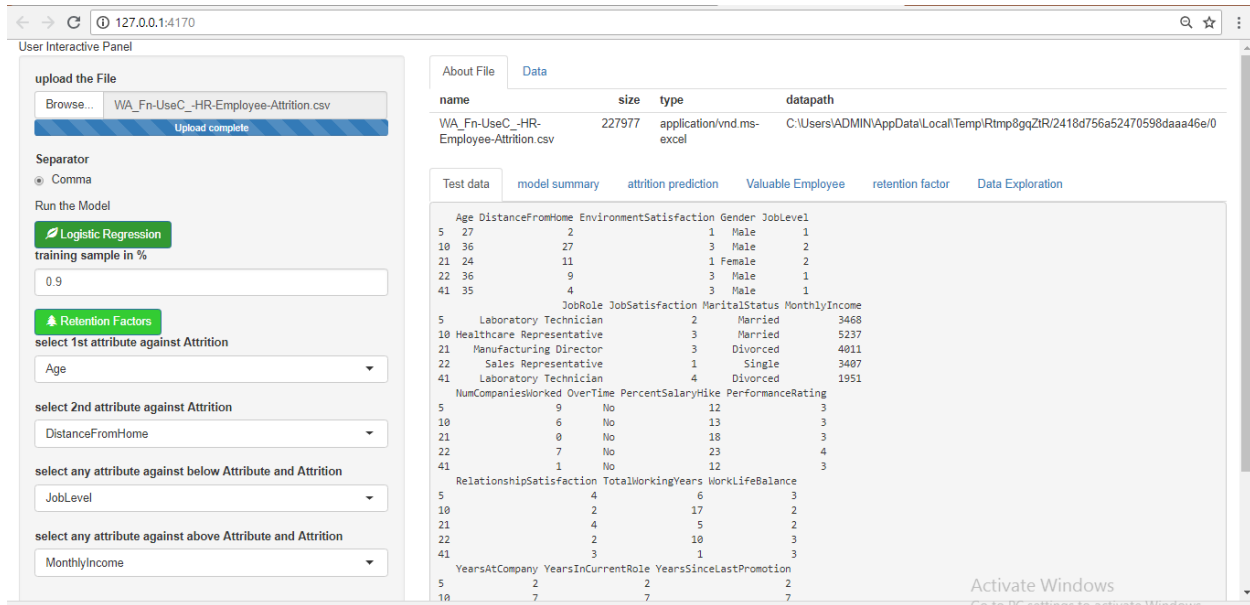


Figure 4.5 1: Application Framework 1

Here, the predefined data set is used for the analysis and prediction testing. I have provided two buttons; one for logistic regression for predicting attrition and other for retention factors publishing. Retention factor button performs two tasks – finds the valuable employee and then lists down the retention factors for each record.

In Addition to that, I have made arrangement where user can manually choose the file and do the further analysis on the chosen dataset. Though, only the prototype is developed for that, but it can be further used in actual application.

Also, there is a provision for data exploration and data visualization, where user can choose one of the attributes as one axis to plot the graph against the Attrition and any two attributes to display the 3-D plot with one axis as Attrition.

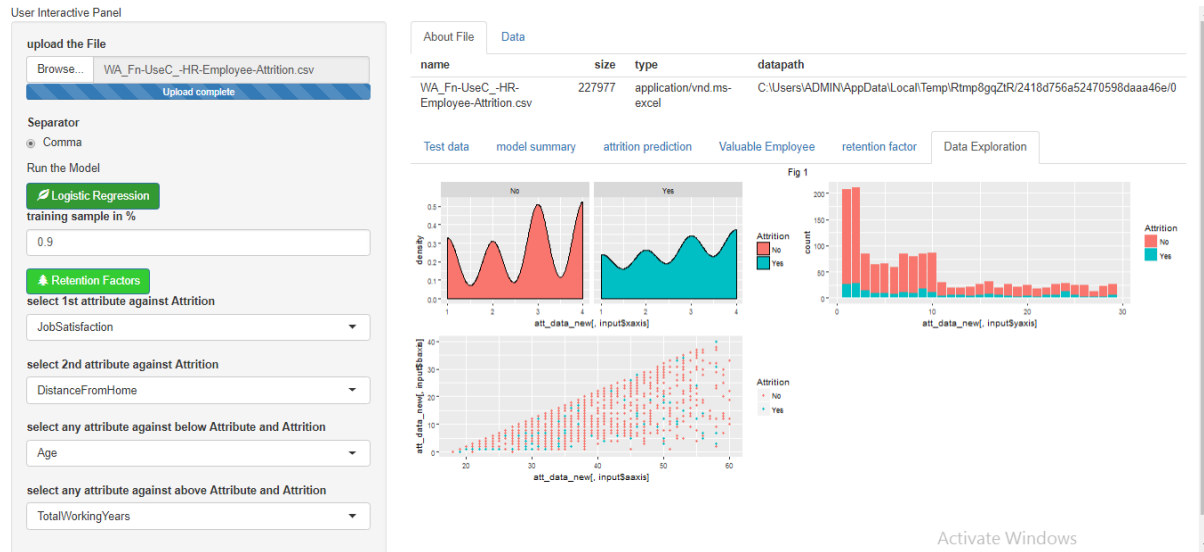


Figure 4.5 2: Application Framework 1

4.5 Development

The development was started with the gathering of data and choosing the attributes for analysis and prediction. I chose the IBM Employee Dataset for building application and testing it. After the data and attributes were finalized, the main task was of choosing the application development framework. In the beginning, I tried to build application by Jupyter Python as below:

```
# This Python 3 environment comes with many helpful analytics libraries installed
# It is defined by the kaggle/python docker image: https://github.com/kaggle/docker-python
# For example, here's several helpful packages to load in

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

# Input data files are available in the "../input/" directory.
# For example, running this (by clicking run or pressing Shift+Enter) will list the files in the
# input directory

from subprocess import check_output
print(check_output(["ls", "../input"]).decode("utf8"))

# Any results you write to the current directory are saved as output.

attrition_file_path = '../input/WA_Fn-UseC_HR-Employee-Attrition.csv'
attrition_file = pd.read_csv(attrition_file_path)
print (attrition_file.describe())

employee_age_data = attrition_file.Age
print (employee_age_data.describe())

employee_interested_columns = ['Age', 'DailyRate']
two_column_data = attrition_file[employee_interested_columns]
print (two_column_data.describe())

y = attrition_file.DailyRate

employee_predictors = ['Age', 'DistanceFromHome', 'Education']

x= attrition_file[employee_predictors]

from sklearn.tree import DecisionTreeRegressor
```

Figure 4.6 1 : Python Jupyter sample Program for prediction of Employee Salary

```
75% 3.000000 3.000000 9.000000
max 6.000000 4.000000 48.000000

YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager
count 1478.000000 1478.000000 1478.000000
mean 4.229252 2.187755 4.123129
std 3.623137 3.224330 3.568136
min 0.000000 0.000000 0.000000
25% 2.000000 0.000000 2.000000
50% 3.000000 1.000000 3.000000
75% 7.000000 3.000000 7.000000
max 18.000000 15.000000 17.000000

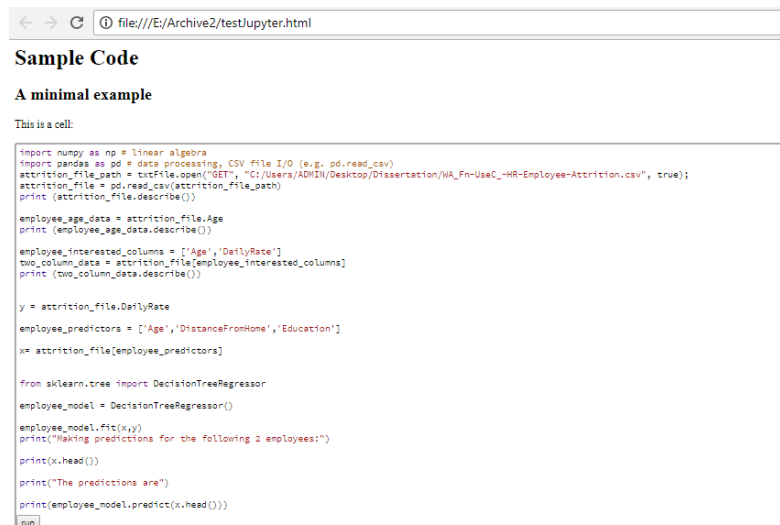
[9 rows x 26 columns]
count 1478.000000
mean 36.923810
std 9.135373
min 18.000000
25% 30.000000
50% 36.000000
75% 43.000000
max 60.000000
Name: Age, dtype: float64
Age DailyRate
count 1478.000000 1478.000000
mean 36.923810 882.485714
std 9.135373 483.509100
min 18.000000 182.000000
25% 30.000000 465.000000
50% 36.000000 882.000000
75% 43.000000 1157.000000
max 60.000000 1499.000000

Making predictions for the following 2 employees:
Age DistanceFromHome Education
0 41 1 2
1 49 8 1
2 37 2 2
3 33 3 4
4 27 2 1

The predictions are
[ 1162. 279. 1206.5 1392. 934.6666667]
```

Figure 4.6 2: Python Jupyter sample output for prediction of Employee Salary

But, as I had to prepare a Web Application, I wanted something that can be integrated with the HTML and then I decided to integrate Jupyter with the HTML as below:



The screenshot shows a web browser window with the address bar displaying 'file:///E:/Archive2/testJupyter.html'. The page title is 'Sample Code'. Below the title, there is a section 'A minimal example' with the text 'This is a cell:'. The main content is a code cell containing the following Python code:

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
attrition_file_path = r"C:\Users\ADRIJ\Desktop\Dissertation\WL_Fn-UseC_HR-Employee-Attrition.csv", True);
attrition_file = pd.read_csv(attrition_file_path)
print(attrition_file.describe())

employee_age_data = attrition_file.Age
print(employee_age_data.describe())

employee_interested_columns = ['Age', 'DailyRate']
two_column_data = attrition_file[employee_interested_columns]
print(two_column_data.describe())

y = attrition_file.DailyRate
employee_predictors = ['Age', 'DistanceFromHome', 'Education']
x = attrition_file[employee_predictors]

from sklearn.tree import DecisionTreeRegressor
employee_model = DecisionTreeRegressor()
employee_model.fit(x,y)
print("Making predictions for the following 2 employees:")
print(x.head())

print("The predictions are")
print(employee_model.predict(x.head()))
```

At the bottom of the code cell, there is a 'run' button.

Figure 4.6 3: HTML Integration with Python Jupyter | Sample Code

But, the result was not obtained when the code was run as it failed to export the csv file from the local desktop. And, then due to limited timeline, I had to go with some other alternative. The other alternative was using the Django or Flask for building the data Visualization application. But again, while loading and installing the libraries of Django and Flask in Ubuntu Virtual Linux System, I faced various difficulties like missing previous library, etc. Hence, I decided to go with R for building the prediction model and R Shiny for building the frontend application which would perform all the tasks including prediction, decision making, display of result of retention attributes or factors on dashboard and even the user interactive data visualization for data exploration.

Once the framework and Data source was selected and finalized, I loaded the data into the R system and then preprocessed and cleaned the data. If the data is not cleaned or preprocessed it does not give the desired output or even sometimes the code fails to run. For cleaning and preprocessing data, I followed below steps:

1. Loading data with UTF-8-BOM compatibility:

Before using the UTF compatibility, I got the unknown characters in the data set like “i»¿Age” instead of “Age” which created problem during recognizing the column name.

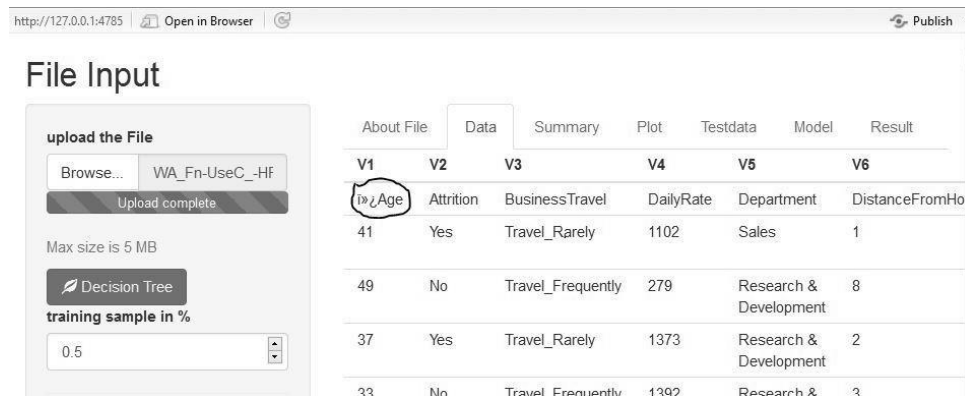


Figure 4.6 4: Before Data Pre-processing | UTF Character setting

After setting the UTF 8 character, I got the required result without any unwanted characters.

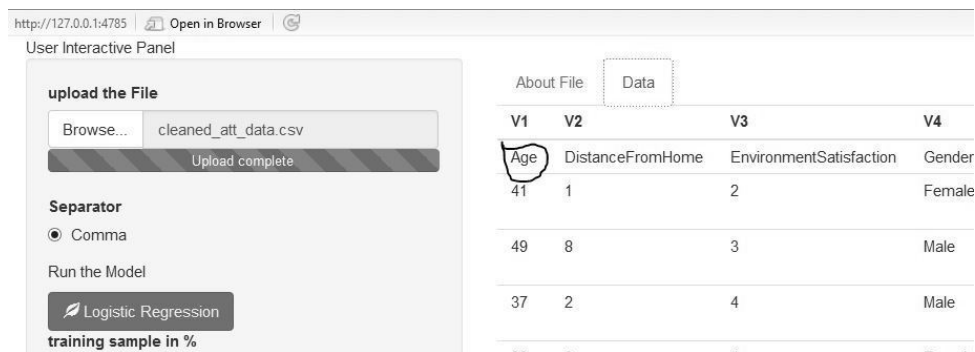


Figure 4.6 5:After Data Pre-processing | UTF Character setting

2. Removing columns with same values

Another task in data cleaning was removing the columns that had no significant role in prediction or analysis. These were the columns with a same value for all the records. Like in the case of IBM Dataset, it had columns like “Employee Count”, “Over 18”, “Standard Hours” which had the same value of “1”, “YES” and “80” respectively. And hence I deleted those rows.

3. Remove all the rows containing NA

From the remaining rows, I deleted the columns with Null value or “NA” value.

4. Move Attrition column to last



As the Attrition column is the one to be predicted, I brought it to the last place from the second place in the dataset for the easily verifying the prediction result.

5. Removing all unwanted columns

As mentioned earlier, I selected few rows for the attrition prediction based on the previous research work, case study and Survey from HR professionals. And deleted all the other remaining unwanted columns.

6. Storing Processed data without the Row identifier

After the data was cleaned and pre-processed, I stored the data in a local directory in csv format. I saved the data in such a way that it does not populated the row identifier as it usually does.

 A screenshot of a CSV file showing data with row identifiers. The first column contains row numbers 1 through 7, which are circled in red. The subsequent columns contain various attributes like Age, BusinessTravel, DailyRate, etc.	 A screenshot of a CSV file showing data without row identifiers. The first column starts directly with the 'Age' attribute, and there are no row numbers.
<code>"", "i..Age", "BusinessTravel", "DailyRate", "Depa: "1", 41, "Travel_Rarely", 1102, "Sales", 1, 2, "Life : "2", 49, "Travel_Frequently", 279, "Research & Dev: "3", 37, "Travel_Rarely", 1373, "Research & Develo "4", 33, "Travel_Frequently", 1392, "Research & De "5", 27, "Travel_Rarely", 591, "Research & Develop "6", 32, "Travel_Frequently", 1005, "Research & De "7", 59, "Travel_Rarely", 1324, "Research & Develo</code>	<code>"Age", "DistanceFromHome", "Er 41, 1, 2, "Female", 2, "Sales Exe 49, 8, 3, "Male", 2, "Research Sc 37, 2, 4, "Male", 1, "Laboratory 33, 3, 4, "Female", 1, "Research 27, 2, 1, "Male", 1, "Laboratory 32, 2, 4, "Male", 1, "Laboratory 59, 3, 3, "Female", 1, "Laborato</code>

Though it does not matter while building application, but it makes the presentation better for the user.

Then, I had to choose the prediction model to predict the attrition. The prediction model in an application is chosen after the analysis and calculating the accuracy of the prediction models. As mentioned earlier, I had chosen six predictive classification algorithms which included KNN, SVM, Decision tree, XgBoost Tree, Random Forest and GLM. Each one gave the different accuracy for the prediction result. I calculated the accuracy with the help of confusion matrix. Confusion matrix compares the prediction result with actual result and calculate the accuracy. After analyzing the selected prediction Algorithm, Logistic Regression, GLM gave the most accurate result. Hence, I used the GLM Algorithm while building a final application. After, it was finalized that I must use GLM model, I included the same in the Application program.

Then I tried to apply the decision tree algorithm to build the decision model for choosing the valuable employee. But due to inconsistent data and data quality, the decision tree was not able to be implemented as it gave incorrect output. Hence, I had to make use of the four Methodological Assumptions, as mentioned in Methodology, to find out the valuable employee and factors of retention. The above four assumptions were logically coded in the program and the output was found out and displayed on the dashboard.

Thus, after executing first and second methodological assumptions, I got the result of valuable employees and then on that result dataset, I executed the second and third methodological assumption, to find out the retention factors and display on the dashboard.

I also developed the tab which displayed the data visualization for data exploration. It displayed two 2-Dimensional graphs and one aesthetic mapping plot. The 2-Dimensional graph has predefined Y axis, which is X-axis variance and X axis can be selected by user from the drop-down menu provided in the application. The drop-down menu contains all the attribute names in the dataset. Thus, it provides the graphical representation of the measure of factors affecting the attrition with Negative attrition in Red color and Positive Attrition in Blue color.

The development was done step by step with each functionality developed in separate sprint. All the missing and required libraries and packages were installed with the “install.packages()” command. The important libraries that are installed are dplyr, shiny, caret, ggplot2 and grid (Peter Bruce, 2017). The related packages are downloaded and installed before using those libraries in application program.

I used the “dplyr” library for using the select function for data manipulation task while removing the unwanted columns (Peter Bruce, 2017). “caret” library was used for implementing the machine learning algorithm for classification and regression training (Hadley Wckham, 2017). “shiny” library is necessary to run the shiny application. “ggplot2” is required for plotting the aesthetic mapping in the data visualization tab. “grid” library is used to add nx by ny rectangular grid to an existing plot after using the ggplot2 (Hadley Wckham, 2017).

Before implementing the glm (logistic regression) machine learning algorithm (Peter Bruce, 2017), I have also made arrangement for reproducing the result for debugging purpose. The set.seed() function does the work of generating the random number so as to generate the reproducible results for validation and debugging (John D. Kelleher, 2015).

I had given the user to select the file format of the data. Such as comma separated, or tab separated (Here, only comma separated is shown in the prototype of the application.) Thus, user has a choice of selecting the file as well as the format of the file. Secondly, I have also provided a space to alter the percentage of training dataset, minimum from 0.5 to maximum 0.9 i.e. from 50% to 90% of the total data can be treated as training dataset and remaining as a test data set. The ideal percentage

of the training dataset is 70-80 % i.e. 80:20 or 70:30 ratio between training and test data set should be there for the better quality and accuracy result. I have also developed two buttons, one of which does the function of running Logistic regression algorithm for predicting the employee attrition. The button includes the functionality of predicting the attrition and displaying the model summary and prediction output.

I have also developed another button which performs two functions of finding the valuable employees and secondly display the most effective retention factors of the valuable employees. This button has the functionality of running the conditional logics aligning to the assumptions requirement and provide the desired output.

Now, talking about the development of the methodological assumptions; the first and third assumption was of the decision of valuable employee wherein I had to calculate the threshold value which was the value at the 1st quartile (approx. 30th percent) record of the total records. So, to calculate the threshold value, I first, arranged the columns in ascending order and then calculated the 1st quartile of record and assigned the value at that location to the threshold variable.

Similarly, I calculated the mean of all the values of the required columns and assigned them as the threshold value while finding the factors for the retention.

There are mainly eight tabs developed in the output panel. First tab shows the file information of the user file selected by browsing the file in the local machine. Second tab displays the file data records with the column names. Third displays the sample test data, forth tab shows the model summary of the prediction model, and fifth tab shows the prediction output. Sixth and seventh tab shows the valuable employees list and retention factors most effective while retaining valuable employee is displayed on the dashboard in the form of table records. Finally, the last tab is the data exploration tab which show the graphs and plots plotted by the user from the attributes selected from the drop-down menu.

Thus, the application is divided into mainly two parts; input which is at left-side panel of the application and output at the right-side panel. The application when run in the R app, it can also be opened in the local host at “<http://127.0.0.1:xxxx/>” in this case (Hadley Wckham, 2017).

Advancement - The application can be further uploaded on the cloud using the Google cloud or AWS cloud and can be used as a web application.

4.6 Testing

As I followed the Agile methodology, I tested the application unit by unit as it got completed. The very first unit was the prediction model for the attrition. So, I tested the prediction model by running the test data and comparing the prediction result and actual result.

Secondly, I tested the working of two assumption models – Valuable employee decision and finding of retention factor.

Next, the testing was done on the front-end application. Whether the three functionalities (Attrition prediction, Decision while categorizing employees into valuable and ordinary employee and finally displaying most effective retention factors). Also, the user interactive data exploration with the help of data visualization was tested by plotting various graphs and plots on the dashboard.

As the prediction functionality was enclosed in the button named as “Logistic Regression”, it was tested by clicking and checking the output panel and the model summary and prediction result tabs in the output panel.

Similarly, the retention factor button performed two tasks of categorizing the valuable employees with ordinary one and finding the most effective retention factors. Those conditioning logics were based on assumptions as mentioned in methodology; hence the result was tested according to the assumptions. By comparing the sample records value and threshold values, I tested whether the decision taken by the if else conditioning logic is proper or not. Like, according to assumption for deciding the valuable employee, if the value of the any of the highest priority is attribute is above threshold value then it is categorized as valuable employee or if all the attributes with least priority are above threshold value then the employee is valuable employee; else it comes under ordinary employee. In the same manner, according to assumption for the most effective retention factors, if the value of any attribute is below the threshold value then it is added in the retention factor list else not. Thus, the unit and functional testing was carried out.

Also, as a future advancement, the option for browsing the data file is provided if user wants to work on any other dataset, other than the predefined dataset, but with same column names. So, the browsing functionality was also tested by browsing the test file and displaying its data in the Data tab as a record table.

After all the test cases were passed, the system application was integrated, and integration and smoke testing were performed when run on R framework as well as on local host. Testing was successful at both the places and finally the application was declared completed. But then too, it is a prototype of a real commercial application. And can be enhanced more further with the same functionalities and making functions more user interactive as discussed in above methodology and development section.

Chapter Five

Data Analysis and Findings

5.1 Introduction

This chapter mainly focuses on the Findings obtained from the research and how they were used for progressing in the research work. It also includes the Data Analysis with covering the both Qualitative and Quantitative aspects of the research. Finally, it highlights the result obtained from the research in the same manner how they would be presented to the user. Also, this includes how the Methodological Assumptions were used to complete the research work and build the final working analytical application.

5.2 Findings

When the, research was started, with the study on the previous research work and case studies, I found that various methods and tools were developed for predicting the attrition, but it was for making the company ready for the new hire as a replacement of the old one. Sometimes, company lost the valuable employee and had to spend a good amount of money on hiring new employee and on new hire's training. There was no tool developed which performed the automated decision of retention. HR Managers needed to take the decision manually by involving in the discussion with

the line managers and employee's reporting managers to know about the capability of the employee and decide whether to retain the employee or not and what are the factors which can help them to retain a valuable employee.

But, with the help of an analytical application, it performs the two main type of analysis – predictive and prescriptive analysis. With the predictive analytics it predicts the employee attrition and the prescriptive analysis helps HR Managers to decide on valuable employees and know the factors most effective in retaining the employees.

The weightage of each of the factor is calculated with the help of survey done for analyzing the attributes. The weightage graph of all factors is as below:

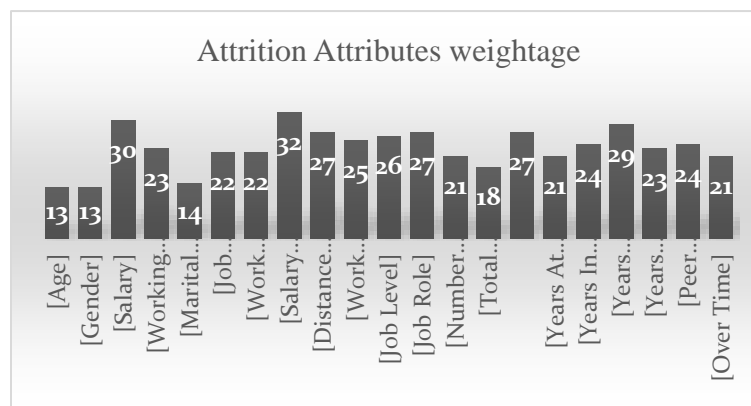


Figure 5.2 1: Attrition Attributes selection and weightage

Thus, with the help of the above attributes and their respective weightage attrition was to be calculated. But due to non-numeric values of few of the attributes, the weighting of the attribute was not possible. It can be done in future advancement while developing application for commercial purpose. Thus, above weightage can be used for calculating the weightage of the attributes by multiplying the actual value with the proportionate value as below:

Attribute	Multiplier
[Age]	0.03
[Distance from Home]	0.06
[Gender]	0.03
[Job Level]	0.06
[Job Role]	0.06
[Job Satisfaction]	0.05
[Marital Status]	0.03
[No. of Companies Worked at]	0.05

[Over Time]	0.05
[Peer Relationship]	0.05
[Performance Rating]	0.06
[Salary Hike]	0.07
[Salary]	0.07
[Total Working Years]	0.04
[Work Env. Satisfaction]	0.05
[Work Life Balance]	0.05
[Years at Company]	0.05
[Years in Current Role]	0.05
[Years Since Last Promotion]	0.06
[Years with Current Manager]	0.05

Table 5.2 1:Attribute Weighted Multiplier

Even though the weighted attributes were not considered, the prediction result was found out with maximum accuracy by studying and implementing the best selected prediction algorithm from the previous research work. After comparing the prediction result from the selected six algorithms, it was found that Logistic Regression, GLM gave the most accurate output. The accuracy was calculated by the confusion matrix.

Suppose, “cleaned_att_data” is a cleaned data file and is saved in “ind” object. Training dataset is stored in training and test dataset is stored in test and “test_glm” has the GLM prediction model stored. “predict_glm” predicts the value for the target in test dataset. Now once the test dataset has the predicted value, then the confusion matrix calculates the accuracy of the model by comparing the actual target value with the predicted target value.

GLM Algorithm Code with ratio of training test data of 0.8:0.2 and “Attrition” as a target.

```
“test_glm <- train(Attrition~.,training, method = 'glm',trControl = trainControl(method = 'repeatedcv',number = 3))
predict_glm <- predict(test_glm, test)
confusionMatrix(predict_glm,test$Attrition)”
```

The output and the interpretation of the Confusion matrix result is as below:

Confusion Matrix and Statistics	The first part is the contents of the matrix. Within each row is the prediction and diagonal
Reference	

Pos Pred Value : 0.9292	With the help of the sensitivity, specificity and prevalence, caret calculates the Positive and Negative prediction value.
Neg Pred Value : 0.7407	
'Positive' Class : No	

The Decision tree and Advanced decision tree algorithm was to be used for deciding on the valuable employee. But, it was found that, due to data inconsistency and low data quality and attribute underfitting, the decision tree was not able to be implemented as it gave the undesired output. So, I made use of two methodological assumptions to make decision of valuable employee and other two for finding the retention factors. The assumptions are as explained in the Methodology. It was also found that, the result obtained from the application of the assumptions gave the desired output. Though, the result can be modified by making the application more user interactive where, they can choose the criteria for deciding on valuable employees and the most effective retention factors; users can get the more desirable output rather on staying on any assumptions. This also increases the efficiency of an application and accuracy of the predictive and prescriptive analysis.

5.3 Data Analysis

There was excessive study done on attributes for predicting employee attrition and various analysis was done on selecting the machine learning algorithm. The accuracy and the confusion matrix were deeply analyzed for selecting the algorithm. The research study for predictive algorithm can be found in “Appendix D”. All the prediction models and their respective accuracy is given below:

```

26 # Knn model
27 fit_knn <- train(Attrition~.,training,method = 'k
28               trControl = trainControl(
29                 method = 'repeatedcv',number =
30 Predictions_knn <- predict(fit_knn,test)
31 summary(fit_knn)
32 confusionMatrix(Predictions_knn,test$Attrition)
33 <

```

26:1 (Top Level) ↕

Console C:/Users/ADMIN/Desktop/Dissertation_Model/ ↗

```

Accuracy : 0.8592
95% CI : (0.8132, 0.8974)
No Information Rate : 0.8592
P-Value [Acc > NIR] : 0.542

```

```

Kappa : 0.1773
McNemar's Test P-Value : 1.963e-05

```

```

Sensitivity : 0.9754
Specificity : 0.1500
Pos Pred Value : 0.8750
Neg Pred Value : 0.5000
Prevalence : 0.8592
Detection Rate : 0.8380
Detection Prevalence : 0.9577
Balanced Accuracy : 0.5627

```

'Positive' Class : No

Figure 5.3 1: KNN Model Accuracy => 85.92%

```

60 #XgbTree
61 fit_xgb <- train(Attrition ~.,training,
62               method = 'xgbTree',tuneGrid = xgbGrid)
63 Predictions_xgb <- predict(fit_xgb,test)
64 summary(fit_xgb)
65 confusionMatrix(Predictions_xgb,test$Attrition)
66 <

```

64:17 (Top Level) ↕

Console C:/Users/ADMIN/Desktop/Dissertation_Model/ ↗

```

Accuracy : 0.8697
95% CI : (0.8249, 0.9066)
No Information Rate : 0.8592
P-Value [Acc > NIR] : 0.34110

```

```

Kappa : 0.3616
McNemar's Test P-Value : 0.02136

```

```

Sensitivity : 0.9549
Specificity : 0.3500
Pos Pred Value : 0.8996
Neg Pred Value : 0.5600
Prevalence : 0.8592
Detection Rate : 0.8204
Detection Prevalence : 0.9120
Balanced Accuracy : 0.6525

```

'Positive' Class : No

Figure 5.3 2: XgBoost Model Accuracy => 86.97%

```

75 #Decision Tree
76 fit_rpart <- train(Attrition ~.,training,
77               method = 'rpart', trControl = 1
78               method = 'cv',number = 3)) #
79 Predictions_rpart <- predict(fit_rpart,test)
80 summary(fit_rpart)
81 confusionMatrix(Predictions_rpart,test$Attrition)
82 <

```

81:50 (Top Level) ↕

Console C:/Users/ADMIN/Desktop/Dissertation_Model/ ↗

```

Accuracy : 0.8697
95% CI : (0.8249, 0.9066)
No Information Rate : 0.8592
P-Value [Acc > NIR] : 0.3411

```

```

Kappa : 0.2713
McNemar's Test P-Value : 7.961e-05

```

```

Sensitivity : 0.9754
Specificity : 0.2250
Pos Pred Value : 0.8848
Neg Pred Value : 0.6000
Prevalence : 0.8592
Detection Rate : 0.8380
Detection Prevalence : 0.9472
Balanced Accuracy : 0.6002

```

'Positive' Class : No

Figure 5.3 3: Decision Tree Model Accuracy => 86.97%

```

83 #random forest
84 set.seed(123)
85 fit_rf <- train(Attrition ~.,training,
86               method = 'rf', trControl = trainContro
87               method = 'repeatedcv',number = 3)) #
88 Predictions_rf <- predict(fit_rf, test)
89 summary(fit_rf)
90 confusionMatrix(Predictions_rf,test$Attrition)
91 <

```

87:19 (Top Level) ↕

Console C:/Users/ADMIN/Desktop/Dissertation_Model/ ↗

```

Accuracy : 0.8697
95% CI : (0.8249, 0.9066)
No Information Rate : 0.8592
P-Value [Acc > NIR] : 0.341097

```

```

Kappa : 0.3103
McNemar's Test P-Value : 0.001009

```

```

Sensitivity : 0.9672
Specificity : 0.2750
Pos Pred Value : 0.8906
Neg Pred Value : 0.5789
Prevalence : 0.8592
Detection Rate : 0.8310
Detection Prevalence : 0.9331
Balanced Accuracy : 0.6211

```

'Positive' Class : No

Figure 5.3 4: Random forest Model Accuracy =>86.97%

```

95 library(caret)
96 #GLM - Logistic regression
97 fit_glm <- train(Attrition~.,training,
98                 method = 'glm',trControl = 1
99                 method = 'repeatedcv',numf
100 Predictions_glm <- predict(fit_glm, test)
101 summary(fit_glm)
102 confusionMatrix(Predictions_glm,test$Attritio
103 <
104 (Top Level)

```

```

Console C:/Users/ADMIN/Desktop/
Accuracy : 0.9101
95% CI : (0.8692, 0.9416)
No Information Rate : 0.8614
P-Value [Acc > NIR] : 0.01031

Kappa : 0.5753
McNemar's Test P-Value : 0.06619

Sensitivity : 0.9696
Specificity : 0.5405
Pos Pred Value : 0.9292
Neg Pred Value : 0.7407
Prevalence : 0.8614
Detection Rate : 0.8352
Detection Prevalence : 0.8989
Balanced Accuracy : 0.7551

'Positive' Class : No

```

Figure 5.3 5: Logistic Regression Model Accuracy => 91.01%

```

130 #Support Vector Mechanism - SVM
131 fit_svm <- train(Attrition~.,training,
132                 method = 'svmRadial',trControl =
133                 method = 'repeatedcv',number =
134 Predictions_svm <- predict(fit_svm,test)
135 summary(fit_svm)
136 confusionMatrix(Predictions_svm,test$Attrition)
137 <
138 (Top Level)

```

```

Console C:/Users/ADMIN/Desktop/Dissertation_Model/
Accuracy : 0.8803
95% CI : (0.8367, 0.9156)
No Information Rate : 0.8592
P-Value [Acc > NIR] : 0.1746

Kappa : 0.2327
McNemar's Test P-Value : 1.519e-08

Sensitivity : 1.0000
Specificity : 0.1500
Pos Pred Value : 0.8777
Neg Pred Value : 1.0000
Prevalence : 0.8592
Detection Rate : 0.8592
Detection Prevalence : 0.9789
Balanced Accuracy : 0.5750

'Positive' Class : No

```

Figure 5.3 6: SVM Model Accuracy => 88.03%

Thus, from the above all confusion matrices, it can be seen that; GLM Algorithm provides the most accurate result. The accuracy provided by the GLM is 91.01%. Hence, I have chosen the GLM Algorithm for building the predictive model in application.

Further the data analysis was done for finding the factors responsible for attrition and help HR Managers during retention of the valuable employees. Various graphs and plots were plotted for analyzing the effect of the attributes on attrition.

Below is the detailed analysis done while selecting the attributes for predicting the employee attrition. The program and coding for the below data analysis can be found in “Appendix E”.

As per the below, graphs (Figure 5.3.7), it can be seen that, most of the employees between 25 and 35 leave the company. People who travel most for Business, mostly leave the company. Daily rate and Department did not help to distinguish between attrited and non attrited employees.

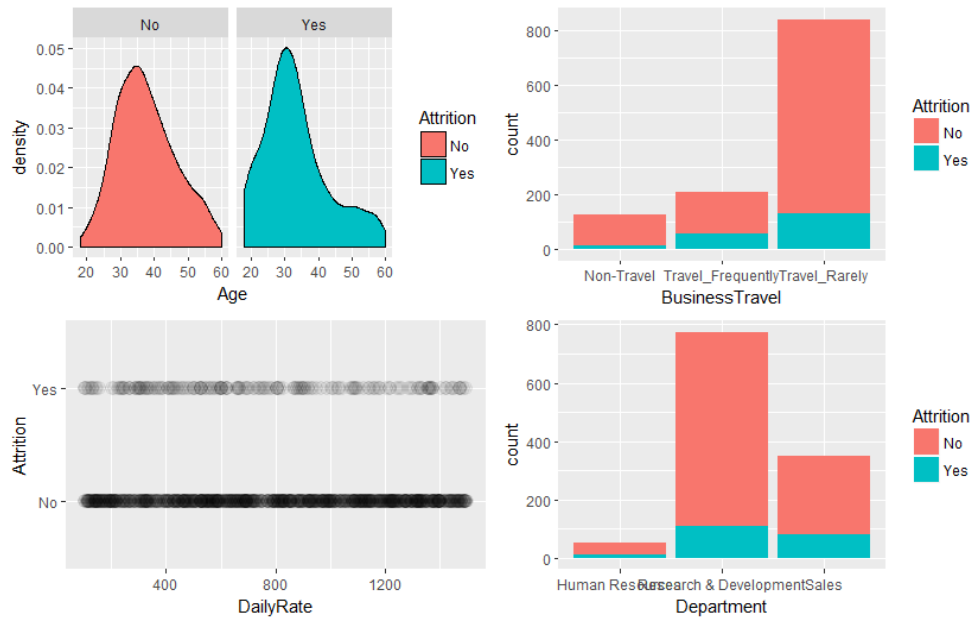


Figure 5.3 7: Graph Plot for Age, Business travel, Daily Rate and Department

It is also seen in another graph (Figure 5.3.8) that, employees staying closest to the office have left the company. Education, Education field, Employee count, Employee number are all insignificant for the attrition analysis. Environment satisfaction does not play important role in analyzing the attrition. Though the most people attrited with the least environment satisfaction i.e. “1”. It is found that 11.76% of total female employees attrited, 19.82% of total male employees in dataset attrited.

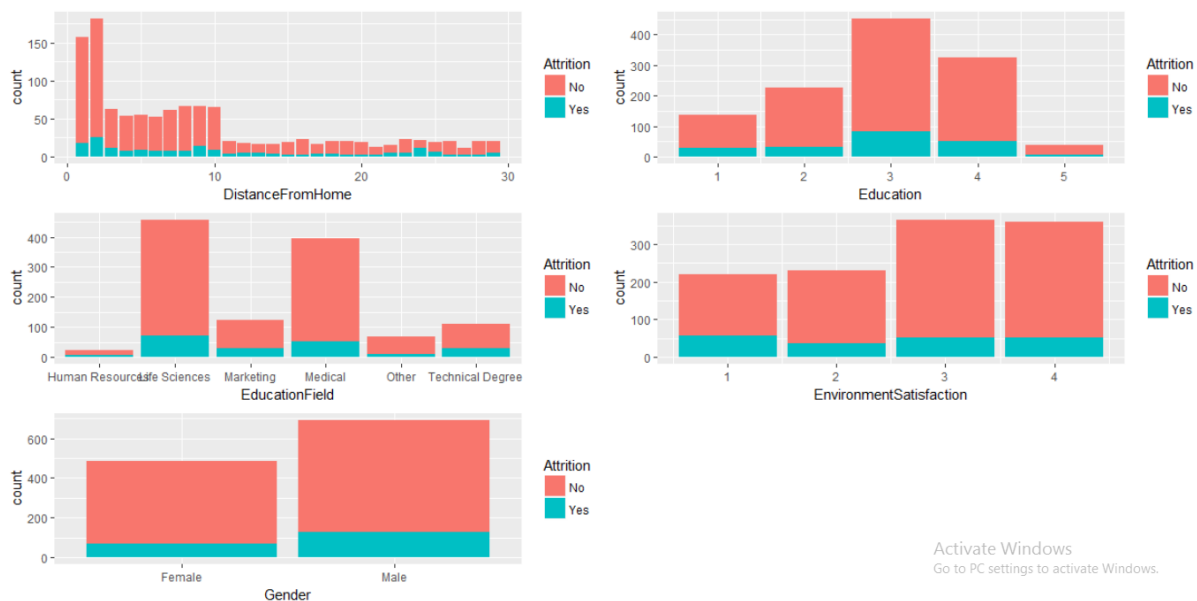


Figure 5.3 8: Plot for Distance, Education, Education level, Education satisfaction and gender

From the next graphs (Figure 5.3.9), it can be seen that nothing can be interfered from the Hourly Rate. Also, the employees with Job Involvement rating “3” attrit the most while with job involvement “4” attrit the least. From the Job Level graph, it can be seen that employees with Job level “1” i.e. Entry level have the highest attrition. Job Satisfaction does not infer any significance about attrition.

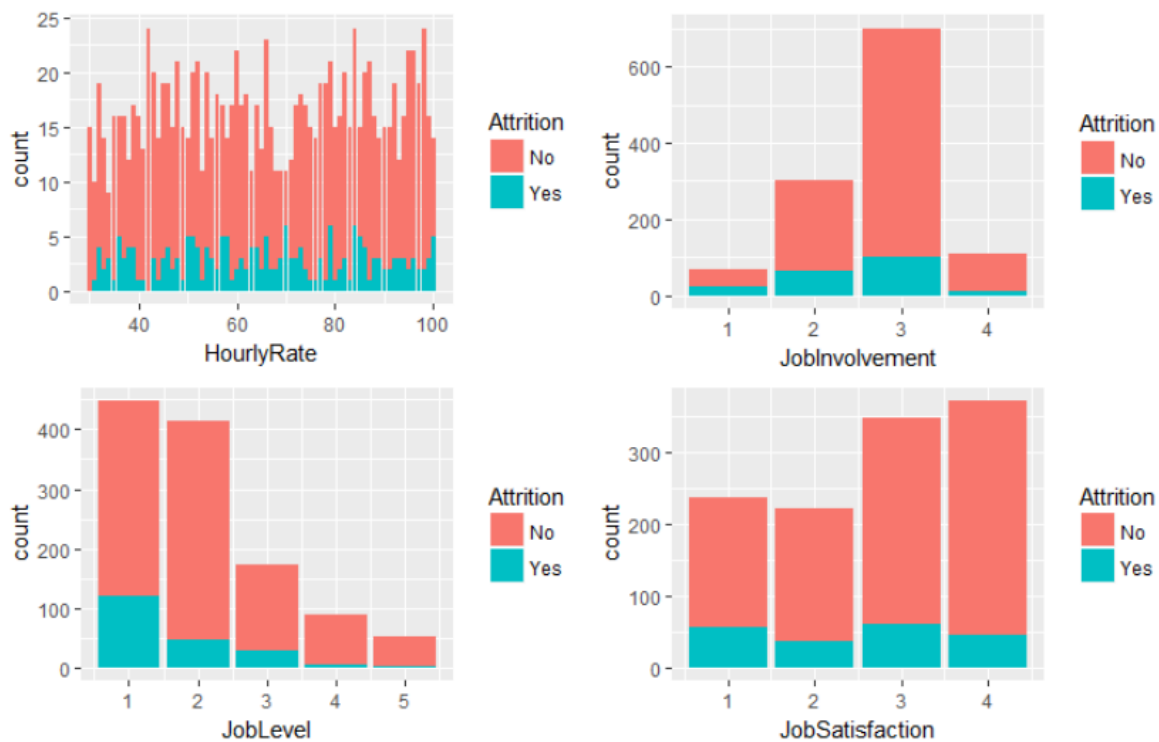


Figure 5.3.9: Plot for Hourly rate, Job Involvement, Job level and job satisfaction

Now, let’s look into the other given below graph (Figure 5.3.10); it shows that employees with “Single” marital status have most attrition. Also, higher attrition is seen at lower segment of Monthly Income. Monthly rate does not show any trend for attrition while employees who have worked in 1 or 2 companies before the current company quit a lot.

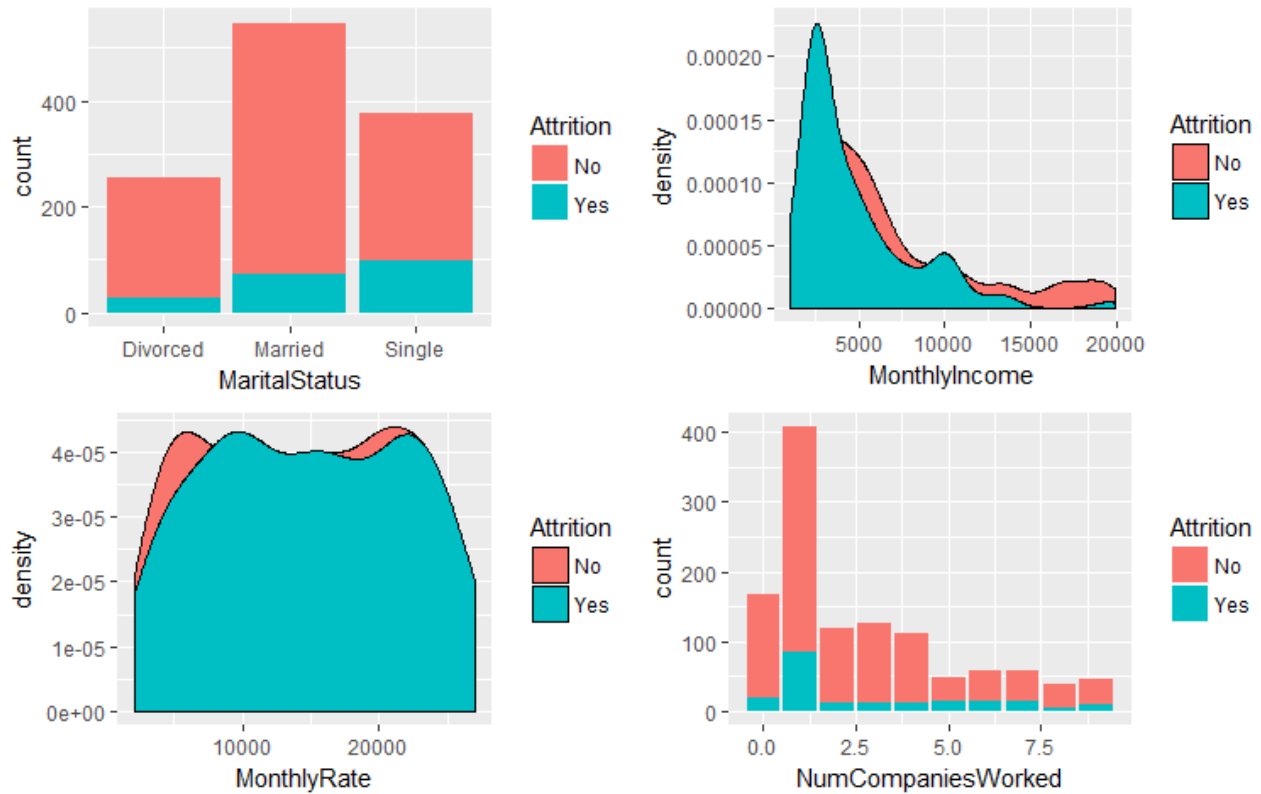


Figure 5.3 10: plot for marital status, monthly income, monthly rate and no. companies worked

Another graph (Figure 5.3.11) indicates that Over18 and Over time attribute does not signify anything about attrition, but Percent Salary hike, performance rating and relationship satisfaction does. Employee with less than 15% salary hike have more chances of quitting and employees with lower rating, here we have only “3” and “4”, hence employee with rating 3 attrit the most. It is seen that people with higher rating in peer relationship satisfaction quit company the most.

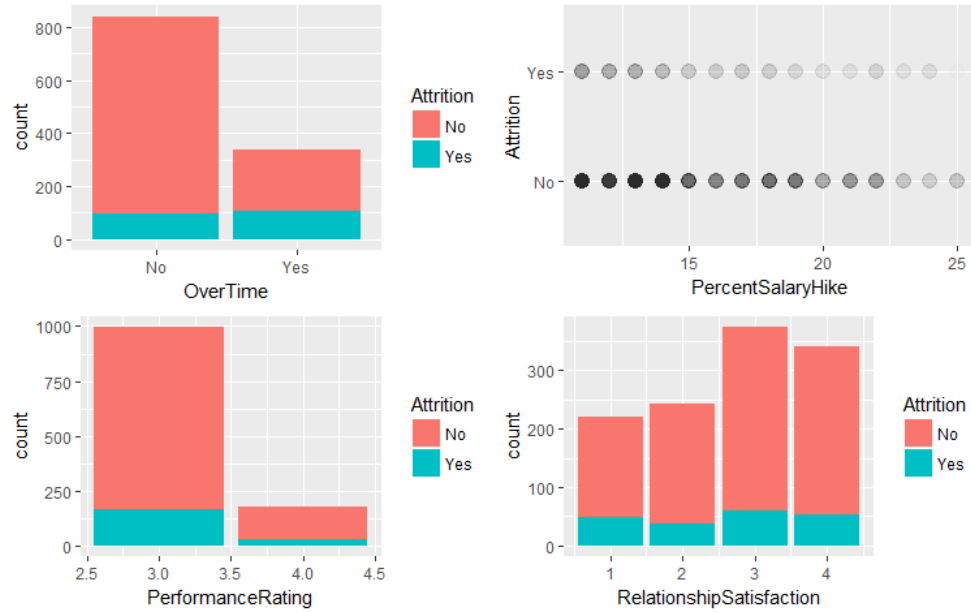


Figure 5.3 11: Plot for Over time, percent hike, performance rating, relationship satisfaction

From the below graph (Figure 5.3.12), it is seen that standard hours, stock option level, training times last year does not interpret anything about attrition. Also, it can be seen from the total working years graph, that employees with 1 to 10 years have larger proportion of attrition than the senior employees. Also, employees with the least rating for work life balance quit the most.



Figure 5.3 12: Plot for stock level, total working years., training times last year, work life balance

The Final graph (Figure 5.3.13) below shows the few important factors that can be used for interpreting the attrition. Years at company, years in current role, years since last promotion and years with current manager. It is also evident from the graph that most new joiners quit the job with least number of years at company. Graph shows the employee with the 0 years in current role quit the most. Also, it is seen that the employees who have been promoted recently have quit the organization. Also, it is found that employee with new managers quit he most.

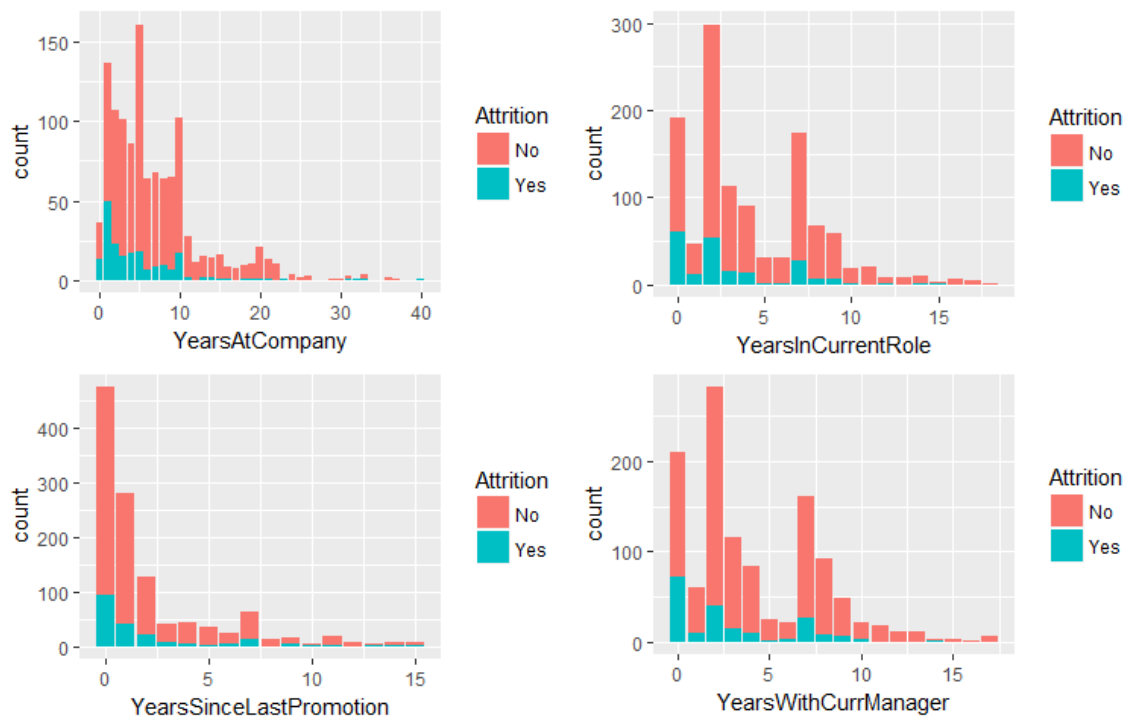


Figure 5.3 13: Plot for years at company, years in current role, year since last promotion, year with current manager

Thus, the data analysis of the data set which was considered was made and similarly user can make the further analysis on various datasets and interpret the conclusion.

5.4 Artifact Results

After the application was finally completed and tested, the results of the working of an application were captured and are explained as follows: (Refer “Appendix F” for the source code of an application).

File Browsing and Data Loading: Below figure shows the file browsing functionality and display of the file information and data on the dashboard.

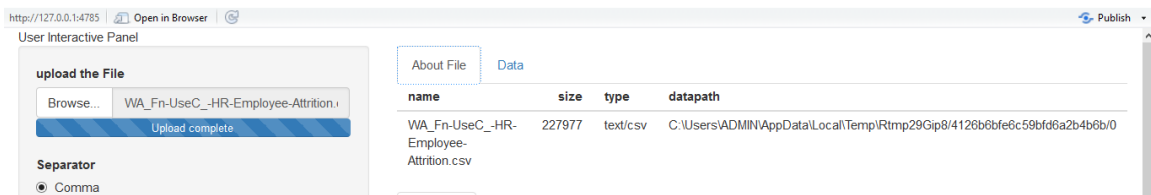


Figure 5.4 1: File Browsing

Cleaned data Loading: This part shows the clean data loaded into the application with the sample test data displayed on the dashboard along with the selection of the training data percentage.

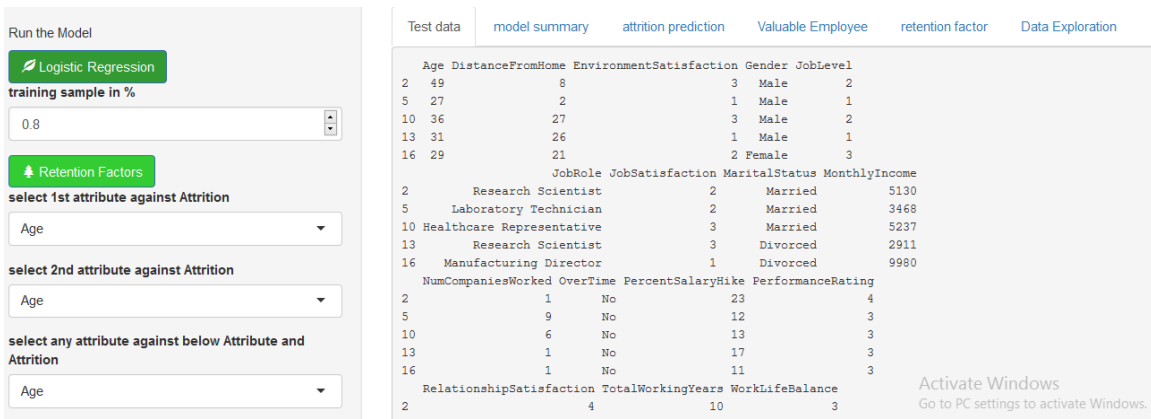


Figure 5.4 2: Load Cleaned data

Prediction Model Summary after the Prediction Model is run: After the Logistic regression button is clicked, it runs the prediction algorithm and shows the summary of the model which can be seen below which shows the Deviance residuals. “The deviance residual is the measure of deviance contributed from each observation”. It also shows the min, max, median, 1st and 3rd quartiles from each observation. Also, it shows the estimated standard error, standard error and predicted value for each attribute. Also, the Prediction tab shows the prediction result with the actual and predicted attrition column at the last of the table grid in the output dashboard.

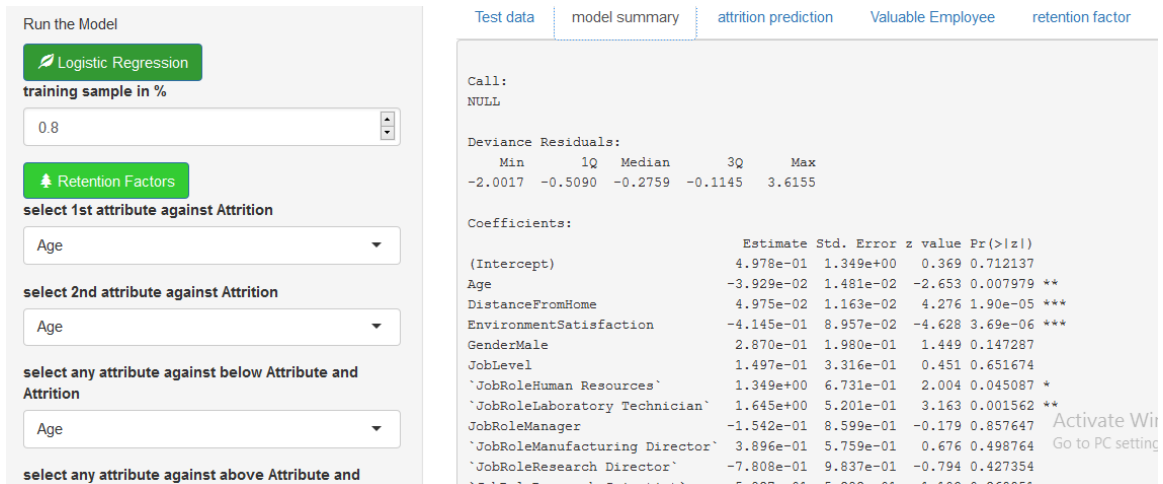


Figure 5.4 3: Model Summary

Prediction Model output: This is the predicted output as displayed in the dashboard.

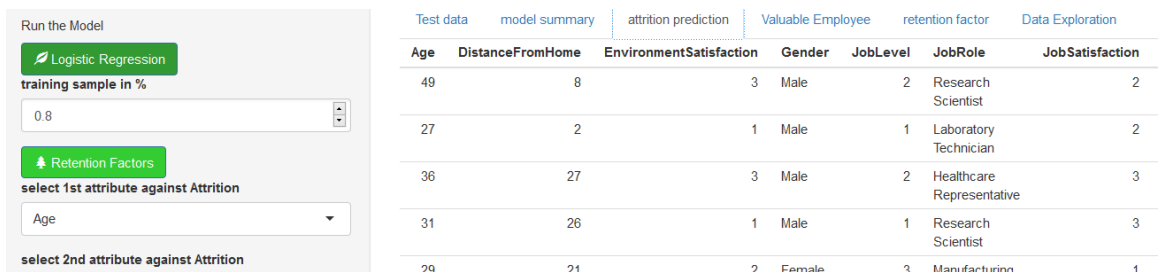


Figure 5.4 4: Prediction output

http://127.0.0.1:4785									
1	3	3	0	0	0	0	No	No	
1	11	3	11	10	10	8	No	No	
4	22	2	22	3	11	11	No	No	
2	11	2	11	8	2	7	No	No	
1	4	2	3	2	0	2	No	No	
4	7	3	3	2	0	2	Yes	No	
3	1	3	0	0	0	0	No	No	
4	14	3	11	10	5	8	No	No	
2	12	2	7	7	7	7	Yes	Yes	
3	5	2	5	4	4	3	No	No	
4	9	4	8	7	0	7	No	No	
4	8	3	3	2	2	Activate W2nd No's			

Figure 5.4 5: Prediction output table

Valuable employee output after clicking retention factors button: After clicking the Retention factors, it executes the conditioning statements which run the conditioning statements satisfying the methodological assumptions and decide on valuable employees and retention factors as below.

separator

☒ Comma

Run the Model

Logistic Regression

training sample in %

Retention Factors

select 1st attribute against Attrition

select 2nd attribute against Attrition

Test data model summary attrition prediction **Valuable Employee** retention factor Data Exploration

Age	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	Employee
41	Travel_Rarely	1102	Sales	1	2	Life Sciences	
37	Travel_Rarely	1373	Research & Development	2	2	Other	
28	Travel_Rarely	103	Research & Development	24	3	Life Sciences	
36	Travel_Rarely	1218	Sales	9	4	Life Sciences	
34	Travel_Rarely	699	Research & Development	6	1	Medical	

Figure 5.4 6: Valuable employee

http://127.0.0.1:4785 Open in Browser Publish

23	2	3	1	0	0	0	Yes	NO
2	3	2	2	2	2	2	Yes	NO
2	0	2	1	0	0	0	Yes	NO
9	3	3	9	8	4	7	Yes	NO
7	3	3	3	2	0	2	Yes	YES
1	5	3	1	0	1	0	Yes	NO
6	1	3	6	4	0	3	Yes	NO
9	3	3	9	7	0	6	Yes	NO
7	2	3	5	4	4	3	Yes	NO

Figure 5.4 7: Valuable employee table

Retention factors displayed on dashboard after clicking retention factors button:

Separator

☒ Comma

Run the Model

Logistic Regression

training sample in %

Retention Factors

select 1st attribute against Attrition

select 2nd attribute against Attrition

Employee-Attrition.csv

Test data model summary attrition prediction Valuable Employee **retention factor** Data Exploration

Age	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeNu
37	Travel_Rarely	807	Human Resources	6	4	Human Resources	
32	Travel_Rarely	1033	Research & Development	9	3	Medical	
41	Travel_Rarely	1356	Sales	20	2	Marketing	

Figure 5.4 8: Retention factors

iningTimesLastYear	WorkLifeBalance	YearsAtCompany	YearsInCurrentRole	YearsSinceLastPromotion	YearsWithCurrManager	Attrition	Isvaluable	Factors
3	3	3	2	0	2	Yes	YES	x x JobLevel JobSatisfaction x x x TrainingTimesLastYear x YearsSinceLastPromotion
2	4	5	4	0	4	Yes	YES	EnvironmentSatisfaction x JobLevel JobSatisfaction x x x TrainingTimesLastYear x YearsSinceLastPromotion
5	2	4	3	0	2	Yes	YES	EnvironmentSatisfaction x JobLevel JobSatisfaction x x x x WorkLifeBalance YearsSinceLastPromotion
3	2	10	4	1	9	Yes	YES	EnvironmentSatisfaction x JobLevel x x x x TrainingTimesLastYear WorkLifeBalance YearsSinceLastPromotion
2	3	6	4	0	4	Yes	YES	x x JobLevel JobSatisfaction x x x TrainingTimesLastYear x YearsSinceLastPromotion
2	1	2	2	2	2	Yes	YES	x x JobLevel JobSatisfaction x x x TrainingTimesLastYear WorkLifeBalance YearsSinceLastPromotion

Figure 5.4 9: Retention factor table

Data Visualization: Data Exploration and Data Visualization with the graphs and plots plotted with the help of attributes selected by the user.

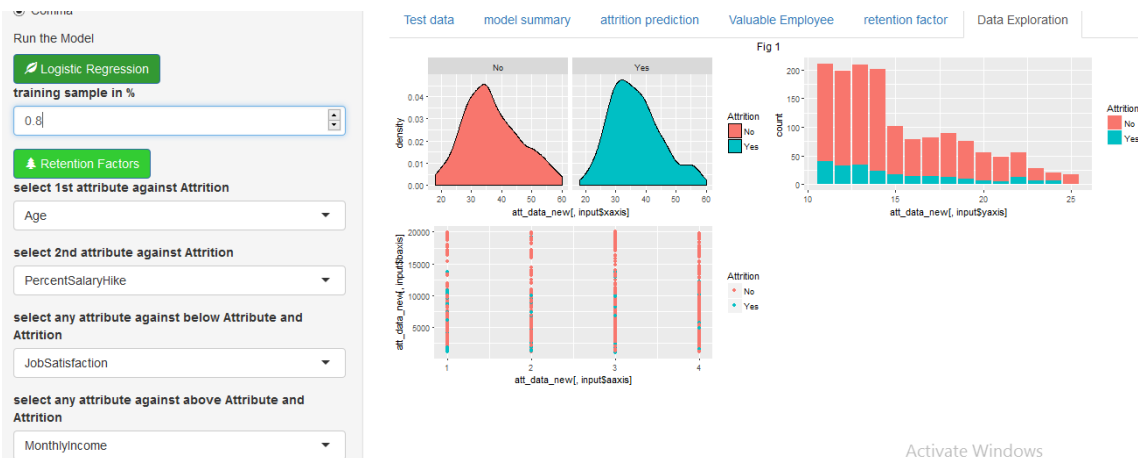


Figure 5.4 10: Data visualization

Thus, the results of the application can be illustrated. This application can be used by HR Managers and Project Managers while retaining the valuable employee so as to reduce the hiring cost of the new employee as well as preserve the quality employees.

5.5 Conclusion

Thus, the Artifact findings included how the employee attrition retention were found out including the factors for choosing valuable employees. Also, data analysis was done included data exploration for selecting the data for predicting attrition and deciding on retention factors. Artifact results were also added in this chapter to show the working of the application.

Chapter Six

Discussions

This chapter shows the objectives of the research and how those objectives were met with the help of qualitative and quantitative methodologies. It also explains various aspects of advancements for overcoming the limitations while developing the HR Analytics application for commercial use.

The primary objective of the research was to improve the retention by finding the most effective factor to be considered while retaining the valuable employee. But, before this, it was necessary to predict the employee attrition so as to know which employees would be leaving in near future and work on their retention. Also, to manage the employee management budget it is necessary to know on which employee to make investment on. Hence, decision model was to be developed for choosing valuable employees. For carrying the research, it was required to gather the data and get the factors responsible for employee attrition and retention and factors for deciding on valuable employees. Also, the machine learning algorithm was to be selected which provided the most accurate output.

With the help of previous research studies and survey conducted among the HR Professionals, the most important factors influencing employee attrition was found out. In the similar way, factors for deciding valuable employees and retention factors were found out with the help of previous case studies and research work. The machine learning algorithms which were used in previous researches were studied to select one of them which provided the maximum accuracy.

With the dataset collected form the IBM and selected six data mining algorithms (Decision Tree, SVM, Random Forest, xgBoost Tree, KNN and Logistic Regression), various experiments were

carried out to predict the employee attrition and the algorithm (Logistic Regression, GLM) which provided the most accurate output, was selected for developing an application.

Finally, with the methodological assumptions, valuable employees and their retentions factors were found out and displayed on dashboard in table format.

The application framework for developing analytical application to carry on the research was selected. From the Python and R, R was selected due to its uncomplicatedness.

The application also requires few advancements to improve its efficiency. The very first advancement that can be done is allowing user to browse and load the dataset of his interest and make the predictions of attrition on the same and find the valuable employees and most effective retention factors used while retaining valuable employees. Secondly while predicting the attrition, user can be given the option of choosing the prediction algorithm and not depend only on Logistic regression as in this case. Another advancement is using the advanced decision tree instead of conditioning logic for deciding on valuable employees. Also, if decision tree is not be used, one can even make the use of the methodological assumptions and provide the user the option to select the threshold values for the attributes while deciding on valuable employees and even retention factors. As, in the application prototype, I am taking the threshold value as 1st quartile of the values while selecting valuable employees and mean value as a threshold value of the all the attributes while displaying the retention factors. This would help the HR Managers to get the desired output from the application.

Chapter Seven

Conclusion and Recommendation

This chapter concludes the research work and recommends on the future advancements of the application and research.

Thus, it can be seen from the above Literature reviews how the study on previous research work and case studies helped to find the attributes for predicting attrition, know the valuable employee and select the retention factors. From the study it is found that the factors responsible for attrition

and retention are - Attrition, Percent Salary Hike, Monthly Income, Years Since Last Promotion, Distance From Home, Job Role, Performance Rating, Job Level, Environment Satisfaction, Years In Current Role, Relationship Satisfaction, Years With Current Manager, Job Satisfaction, Work Life Balance, Number of Companies Worked, Years At Company, Over-Time, Total Working Years, Marital Status, Age and Gender. Methodology demonstrated the processes followed while carrying on the research and developing an analytical application prototype. Finally, an Artifact design illustrated the development and working of system and application architecture. The advancement for improving the result accuracy is also mentioned so that it can be used for developing the application for commercial purpose. In this research, it is found that with the above attributes and Logistic regression algorithm, most accurate prediction result is obtained if the training dataset is 80% of the total data.

As a recommendation, the above analytical application can be integrated with the Human resource management finance budgeting application and there by predict the overall profit or savings in Human Resource Management process which include attrition, retention, hiring of new employee, amount spent on the training and development of new employee and loss to the project due to loss of valuable employee and prescribe on the further actions to be taken. Thus, company can keep track of the amount of budget it had spent on Human resource management and budget to be spent on future and take necessary actions. Also, the system can be put on cloud and the data can directly be taken from the cloud storage through server connections by using FTP and SFTP commands in Unix environment.

Bibliography

1. A. Amin, F. R. I. A. C. K. a. S. A., 2015. A Comparison of Two Oversampling Techniques (SMOTE vs MTDf) for Handling Class Imbalance Problem: A Case Study of Customer Churn Prediction. *Cham: Springer International Publishing*, p. 215–225.
2. Abdul-Kadar Masum, L.-S. B. A.-K. A. a. K. H., 2015. *Intelligent Human Resource Information System (i-HRIS): A Holistic Decision Support Framework for HR Excellence*, Malaysia: IAJIT.
3. Adhikari, A., 2009. Factors Affecting Employee Attrition: A Multiple Regression Approach. *The IUP Journal of Management Research*, VIII(5).
4. Alao D., A. A. B., 2013. ANALYZING EMPLOYEE ATTRITION USING DECISION TREE ALGORITHMS. *Computing, Information Systems & Development Informatics*, 4(1), pp. 17-28.
5. Alex Frye, C. B. M. S., 2018. Employee Attrition: What Makes an Employee Quit?. *SMU Data Science Review*, 1(1).
6. Alf Crossman, B. A., 2003. Job satisfaction and employee performance of Lebanese banking staff. *Journal of Managerial Psychology*, 18(4), pp. 368-376.
7. Algorithm, A. I. I. D. T., 2009. *An Improved ID3 Decision Tree Algorithm*. Tsingtao, China, Proceeding of 2009 4th International Conference on Computer Science & Education.
8. Amir Ahmad, L. D., 2007. A k-mean clustering algorithm for mixed numeric and categorical data. *Science Direct*.
9. Ananya Sarker, S. S. D. M. S. Z. M. M. R., 2018. Employee's Performance Analysis and Prediction using K-Means Clustering & Decision Tree Algorithm. *Global Journal of Computer Science and Technology*, 18(1), pp. 1-5.
10. Andrew, N., 2017. *Andrew Ng*. [Online] Available at: <http://www.andrewng.org/portfolio/an-experimental-and-theoretical->

comparison-of-model-selection-methods/

[Accessed March 2018].

11. Anthony R. Wheeler, V. C. G. R. L. B. C. J. S., 2007. When person-organization (mis)fit and (dis)satisfaction lead to turnover. *Journal of Managerial Psychology*, 22(10.1108/02683940710726447), pp. 203-219.
12. Archita Banerjee, R. K. G. M. G., 2017. A Study on the Factors Influencing the Rate of Attrition in IT Sector: Based on Indian Scenario. *Pacific Business Review International*, 9(7), p. 01.
13. Attridge, M., 2009. Measuring and Managing Employee Work Engagement: A Review of the Research and Business Literature. *Journal of Workplace Behavioral Health*, 24(10.1080/15555240903188398), pp. 383-398.
14. Bass, J. B. S. R. M. a. N. J., 2018. *Employee retention and turnover in global software development : comparing inhouse offshoring and offshore outsourcing*. Manchester, University of Salford.
15. Bernd Bischl, G. S., 2016. *On Class Imbalance Correction for Classification Algorithms in Credit Scoring*. Munich, ResearchGate.
16. Beynon, M. J. P. P. D. a. P. G., 2015. Investigating the impact of training influence on employee retention in small and medium enterprises: a regression-type classification and ranking believe simplex analysis on sparse data.. *Expert Systems* , Issue 32(1), pp. 141-154.
17. B-Gent, K. C. a. D. V. D. P., 2006. Churn prediction in subscription services: an application of support vector machines while comparing two parameter-selection techniques.
18. Bidisha Lahkar Das, D. M. B., 2013. Employee Retention: A Review of Literature. *IOSR Journal of Business and Management*, 14(2), pp. 08-16.
19. Blake A. Allan, R. D. D. a. B. C., 2017. Task Significance and Performance: Meaningfulness as a Mediator. *Journal of Career Assessment*, 1(10.1177/1069072716680047), pp. 1-11.

20. Derrick McIver a, M. L. L.-H. b. C. A. L.-H., 2018. A strategic approach to workforce analytics: Integrating science and agility. *ScienceDirect*, 1(BUSHOR-1458;).
21. Dessler, G., 2016. Pay For Performance and Employee Benefits. In: *Fundamentals of human resource management*. 4th ISBN: 9781292098470 ed. Bay 10A 658.3: Harlow Pearson Education, p. 362.
22. Edouard Ribes, K. T. B. P., 2017. *Employee turnover prediction and retention policies design: a case study*, US: Cornell university.
23. Fabian Pedregosa, G. V. A. G., 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 8/11(10/11).
24. Firth, L. M. D. M. K. A. a. L. C., 2004. How can managers reduce employee intention to quit?. *Journal of managerial psychology*, 19(2), pp. 170-187.
25. Fitz-enz, J., 2010. *The New HR Analytics, Predicting the Economic Value of*. 3 ed. United States of America: Library of Congress Cataloging-in-Publication.
26. Fred Luthans, S. M. N. B. J. A. a. J. B. A., 2008. The Mediating Role of Psychological Capital in the Supportive Organizational Climate:Employee Performance Relationship. *Journal of Organizational Behavior*, 29(2), pp. 219-238.
27. Goddard, M., 2017. The EU General Data Protection Regulation (GDPR): European regulation that has a global impact.. *International Journal of Market Research*, Issue 59(6), pp. 703-705.
28. Greenwell, B. M., 2017. pdp: An R Package for Constructing Partial Dependence Plots. *The R Journal*, 9(1), pp. 421-436.
29. Gupta, S. S., 2010. *Employee Attrition and Retention: Exploring the Dimensions in the urban centric BPO Industry*, A-10, SECTOR 62, NOIDA, INDIA: JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY, NOIDA.
30. Hadley Wckham, G. G., 2017. Data Visualization with ggplot2. In: M. L. Marie Beaugureau, ed. *R for Data Science*. United States of America: O'Reilly Media, Inc, p. 7.

31. Hausknecht, J. P. R. J. M. & H. M. J., 2008. Targeted Employee Retention: Performance-Based and Job-Related Differences in Reported Reasons for Staying.. *Working Paper Series, Cornell University, School of Industrial and Labour Relations, Centre for Advanced Human Resource Studies, Working Paper, Issue 08 – 06..*
32. IBM , 2018. *Employee Attrition and Performance - IBM Analytics*. [Online] Available at: <https://www.ibm.com/communities/analytics/watson-analytics-blog/hr-employee-attrition/> [Accessed June 2018].
33. IBM Inc., 2017. *HR Analytics*. [Online] Available at: www-01.ibm.com/software/analytics/solutions/operational-analytics/hr-analytics/ [Accessed 2017 Dec].
34. IBM Watson, 2018. *Kenexa IBM Talent management*. [Online] Available at: <https://www.ibm.com/talent-management/case-studies/ibm-talent-management-case-studies> [Accessed June 2018].
35. Jagun, V., 2015. *An Investigation into the High Turnover of Employees within the Irish Hospitality Sector, Identifying What Methods of Retention Should Be Adopted*, Dublin, Ireland: NCI College.
36. Jessica Sze-Yin Ho, A. G. D. S.-P. L., 2010. Employee Attrition in the Malaysian Service Industry: Push and Pull Factors. *The IUP Journal of Organizational Behavior*, 9(1,2).
37. Jin, G.-e. X. a. W.-d., 2008. Model of customer churn prediction on support vector machine. *Systems Engineering-Theory & Practice*, 28(1), p. 71–77.
38. John D. Kelleher, B. M. N. A. D. A., 2015. Visualization between features. In: *Machine Learning for Predictive Data Analytics*. London: MIT Press, p. 77.
39. John Hausknecht, J. M. R., 2008. Targeted Employee Retention: Performance-Based. *Center for Advanced Human Resource Studies, Volume CAHRS WP08-06*, pp. 2-34.

40. John L. Cotton, J. M. T., 1986. Employee Turnover: A Meta-Analysis and Review with Implications for Research. *Academy of Management*, 11(1), pp. 55-70.
41. K. W. Bowyer, N. V. C. L. O. H. a. W. P. K., 2011. "SMOTE: synthetic minority over-sampling technique,". *CoRR*, Issue abs/1106.1813.
42. Kelleher, J. D. M. B. & D. A., 2017. *Supervised Learning, Fundamentals of machine learning for predictive data analytics : algorithms, worked examples, and case studies*. Bay 1B 006.312 KEL: s.n.
43. Kirandeep, P. N. M., 2018. Analysis of Improved ID3 Algorithm using Havrda & Charvat Entropy. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 3(3), pp. 8-13.
44. Kotsiantis, S. B., 2007. Supervised Machine Learning: A Review of Classification Techniques. *Informatica* 31, 1(1), pp. 249-268.
45. Kuhn, M., 2018. *GitHub*. [Online] Available at: <https://github.com/topepo/caret/> [Accessed 13 July 2018].
46. Kuvaas, B. & D. A., 2009. Perceived investment in employee development, intrinsic. *Human Resource Management Journal*, Issue 19, pp. 217-236.
47. L. A. . B. L. Chen, C., 2004. "Using random forests to learn imbalanced data,". *Technical Report 666 Statistics Department of University of California at Berkeley*.
48. M. Hanke, Y. H. P. S., 2009. PyMVPA: A Python toolbox for multivariate pattern analysis of fMRI data. *Neuroinformatics*, Issue 7(1), p. 37–53.
49. Mitchell Hoffman, S. T., 2018. *People Management Skills, Employee Attrition, and Manager Rewards: An Empirical Analysis*. s.l.:s.n.
50. Mitchell, T. M., 1999. *Machine learning*. International Student Edition ed. Maidenhead, U.K.: McGraw-Hill.

51. Morrell, K. M., 2004. Organisational change and employee turnover.. *Emerald Group Publishing Limited*, 33(DOI 10.1108/00483480410518022).
52. Morrison, P. S., 2014. *Who Cares about Job Security?*. 2 ed. s.l.:Australian Journal of Labour Economics.
53. Musrrat Parveen, K. M. N. M. K., 2016. QUALITY OF WORK LIFE: THE DETERMINANTS OF JOB SATISFACTION AND JOB RETENTION AMONG RNs AND OHPs. *International Journal for Quality Research*, 11(1)(ISSN 1800-6450), p. 173–194.
54. N. V. Chawla, K. W. B. L. O. H. a. W. P. K., 2002. “Smote: Synthetic minority over-sampling technique,”. *Journal of Artificial Intelligence Research*, 16(321–357).
55. Nina van Loon, A. M. K. B. A. W. V. P. L., 2018. Only When the Societal Impact Potential Is High? A Panel Study of the Relationship Between Public Service Motivation and Perceived Performance. *Review of Public Personnel Administration*, 38(10.1177/0734371X16639111), p. 139–166.
56. Ongori, H., 2007. A review of the literature on employee turnover. *African Journal of Business Management*, 1(ISSN 1993-8233), pp. 049-054,.
57. Onsardi, M. A. T. A., 2017. The Effect Of Compensation, Empowerment, And Job Satisfaction On Employee Loyalty. *International Journal of Scientific Research and Management*, 5(2), pp. 7590-7599.
58. Peter Bruce, A. B., 2017. Classification and Statistical Machine Learning. In: S. Cutt, ed. *Practical Statistics for Data Scientists*. United States of America: O'Reilly Media, Inc., pp. 183,210,219,230,237.
59. Phanish Puranam, Y. R. S. V. F. H. G. v. K., 2018. *ALGORITHMIC INDUCTION THROUGH MACHINE LEARNING: OPPORTUNITIES AHEAD FOR ORGANIZATION SCIENCE*, Zurich: s.n.

60. Poon, J. M., 2004. Effects of performance appraisal politics on job satisfaction and turnover intention. *Personnel Review : Emerald Group Publishing Limited*, 33(10.1108/00483480410528850), pp. 322-334.
61. Prof. Prashant G. Ahire, S. K. K. K. H. K. A. B., 2015. Implementation of Improved ID3 Algorithm to Obtain more Optimal Decision Tree. *International Journal of Engineering Research and Development*, 11(2), pp. 44-47.
62. Quinlan, J., 1985. *Induction of Decision Trees*, Boston: Kluwer Academic Publishers.
63. R.E. Fan, K. C. C. H., 2008. LIBLINEAR: a library for large linear classification. *The Journal of Machine Learning Research*, Issue 9, p. 1871–1874.
64. Rahul Yedida, R. R. R. V. R. J. A. D. K., 2006. *Employee Attrition Prediction*, s.l.: s.n.
65. Raschka, S. a. M. V., 2017. *Python Machine Learning*. s.l.:Packt Publishing Ltd.
66. Rohit Punnoose, P. A., 2016. Prediction of Employee Turnover in Organizations using Machine Learning Algorithms. (*IJARAI*) *International Journal of Advanced Research in Artificial Intelligence*, 5(9), pp. 22-26.
67. S. Chitra, D. P. S., 2018. A Study on Analytics of Human Resource Management in Big Data. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 3(3 ISSN : 2456-3307), pp. 58-68.
68. S. Chitra, D. P. S., 2018. A Study on Analytics of Human Resource Management in Big Data. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 3(3 ISSN : 2456-3307), pp. 58-68.
69. SAP Inc., 2017. *Applying predictive analytics to manage employee turnover*. [Online] Available at: <https://blogs.sap.com/2017/07/20/applying-predictive-analytics-to-manage-employee-turnover/> [Accessed Dec 2017].

70. Scott Mondore, S. D. a. M. C., 2011. Maximizing the Impact and Effectiveness of HR Analytics to Drive Business Outcomes. *Strategic Management Decisions*, 34(2), pp. 20 - 28.
71. Stovel, N. B. M., 2002. Voluntary Turnover: Knowledge Management Friend or Foe. *Journal of Intellectual Capital*, 1(10.1108/14691930210435633), p. 303.
72. T. Schaul, J. B. D. W. Y. S., 2010. PyBrain. *The Journal of Machine Learning Research*, Volume 11, p. 743–746.
73. T.MURALIDHARAN., 2017, Feb. *Holding On the Best - humancapital*. [Online] Available at: www.humancapitalonline.com [Accessed Dec 2017].
74. Tom De Smedt, W. D., 2012. Pattern for Python. *Journal of Machine Learning Research*, Volume 6/12(13), pp. 2063-2067.
75. Torelli, G. M. a. N., 2014. “Training and assessing classification rules with imbalanced data,”. *Data Mining and Knowledge Discovery*, 28(1), p. 92–122.
76. TUTTLE, J. L. C. J. M., 1986. Employee Turnover: A Meta-Analysis and Review with Implications for Research. *Academy of Management Review*, Volume 11(1).
77. Wiener, A. L. a. M., 2002. Classification and regression by random forest. *R News*, 2(3), p. 18–22.
78. Yin, P. P. M. O. A. a. O. S., 2017. Stochastic Backward Euler: An Implicit Gradient Descent Algorithm for \$ k \$-means Clustering. *arXiv preprint arXiv:1710.07746*..
79. Zito, N. W. L. W. a. P. B., 2008. Modular toolkit for data processing (MDP): A Python. *Frontiers in Neuroinformatics*, Volume 2.

Appendices

Appendix A

Appendix A explains the development and working of the IBM HR Analytical tool Kenexa developed by IBM Watson.

IBM has prepared the Predictive Model by following the below steps:

Data Loading and Pre-processing

1. Data Preparation: IBM Data Scientist have prepared dataset which is of score of 68. This score indicates that the dataset is of Medium Quality. Hence before using the data they have cleaned the data before uploading to IBM Watson. To clean the data, they have followed below steps:
 - a) Remove summary rows and columns from your data file.
 - b) Eliminate nested column headings and nested row headings.
2. Data Quality and Usage Optimization: After cleaning the data also, it cannot be directly used for prediction. It must be of better quality. Hence to improve the data quality IBM follows the following steps:
 - a) To improve the data quality, first decide on the attributes to be used for prediction. IBM has provided with option to select and deselect the attributes for prediction.
 - b) Set the proper Fields as Input Field and Output fields.
 - c) Consider the missing values of the required attributes and see how it affects the result.
 - d) Make use of more attributes for better results.
 - e) Make use of domain knowledge to check whether the result makes any sense.
3. Using Aggregate Calculation Functions: Using functions like ABS, COUNT, SUM, MIN, MAX, MEAN, MEDIAN, etc. for exploring the dataset.
4. Changing the Data Format: Date, Time, Currency, etc.
5. Creating Hierarchies of the fields to be used for prediction: IBM has provided an option of refine to choosing the hierarchy for the Attributes and to add the additional field for prediction purpose.
6. IBM provides the option of grouping of the data columns: It also provides the option to group the non-group columns in another group. Thus, the new column containing individual column set is prepared and is used for prediction.

7. Filtering the Columns: IBM provides filter option to remove unwanted columns.
8. Replace Dataset after use: Use Specific dataset when needed.
9. Combining Data Sets option: IBM has combining datasets option by Joins method.

Prediction

1. Opening Prediction: The prediction opens for viewing. Use the Top Predictors, Main Insight and Details pages to view the prediction's insights. Steps to be followed:
 - a) Go to the IBM® Watson™ Analytics Welcome page.
 - b) Tap the name of the prediction that you want to view. Use the filter box to view a subset of predictions that match the entered text.
2. Exploring Prediction: Explore the fields for the prediction. AS soon as the Prediction page is open it shows the most Top Predictor. We can then increase the predictor fields based on the requirements.
3. Editing Prediction: Can change the predictor name and field properties.
Insights in Prediction:
 - a) Text insights - Text insights describe the results.
 - b) Visual insights - Visual insights are visualizations that support and visually demonstrate the text insights
 - c) Dynamic visual insights - Dynamic visual insights are dynamic changes to the visualization that result from expanding a text insight. It uses animation for visualization.
4. Setting aside Visualization from Predict use it Assemble: IBM provides option to set aside interesting or important visualizations from Predict. You then add the visualizations to the dashboards and stories that you create in Assemble. It can be used for reporting purpose.

Visualization - Visualization contains following features:

Interactive Tile, Changing Columns in Visualization, Adding More columns to a visualization, show multiple measures in one Visualization, Navigate your data, Sorting in Numerical order, Sorting in alphabetical order. Visualization Type: Bar, Bubble, Packed Bubble, Line, Area, Pie, Tree Map, Heat Map, Map, World Cloud, Grid, Summary, Network.

Appendix B

The survey result obtained from the seven HR professionals is provided in the attached password protected excel sheet named “MSc Information Systems and Computing Research Project (Responses)”. Due to GDPR, the responses from the HR professionals is password protected to maintain the privacy of the survey and related result. It can be found in Primary Data Source folder. The questionnaire used for doing the survey can be found at [“https://goo.gl/forms/N1JIS9gfpUGsBAzH2”](https://goo.gl/forms/N1JIS9gfpUGsBAzH2). The dataset used as a primary dataset can be found in csv file named “WA_Fn-UseC_-HR-Employee-Attrition.csv” in Primary Data Source folder.

Appendix C

The source code for the research part failure for the implementation of decision tree for deciding on valuable employees can be found in the R files “Failed Tree 1.R” and “Failed Tree 2.R”. It can be found in Artifact folder.

Appendix D

The source code for the research on predictive algorithm for attrition prediction is given in R file named “PredictionModel.R”. It can be found in Artifact folder.

Appendix E

The source for the data exploration and data visualization can be found in file “Attrition Data Analysis Graphs.R”. It can be found in Artifact folder.

Appendix F

The final application source code can be found in the file “10364706_Dissertation_FinalApplication.R”. It can be found in Artifact folder.

Appendix G

Dissertation Timeline Grid is given in Information sheets folder.

Appendix H

Dissertation supervisor meeting report is given in folder Information sheets folder named as “Dissertation meeting and progress monitoring report”.