Empirical Evaluation of Employee Attrition using hybrid model of machine learning techniques

DIVYANG JAIN
16110323
MSc in Data Analytics

4th August 2017

Abstract

Human Resource management works as the backbone of an organisation and taking a good decision is really difficult to reshape and optimize the performance. Losing the high skilled employees is the major concern which negatively impacts the company and cannot be predicted by HR personnel manually. Therefore, HR predictive analytics is an evolving process which can tell whether an employee will leave the company in certain period of time i.e. employee attrition prediction based on data collected to save the new issue of overtime costs or hiring costs of replacements and low productivity. Hence, this paper extends the previous researches in HRPA to get the most accurate results by using hybrid machine learning applications (decision trees-gradient boosting, KNN, SVM, GLM, association rules and genetic algorithms) with CRISP-DM and deploying the best technique so that HR can place a sound strategy to increase the employee retention.

Contents

1 Introduction

In an organisation, HR plays an important role in handling the organizational performance by making people-related decisions but most of the organizations lacked a constant and general view of the workforce thus HR analytics is fundamental to develop analytical skills which should be capable to produce more Return on Investment. Human Resource Predictive Analytics (HRPA) is an art in an organization which pops out the business insights in the field of data analysis which is done by analysing the past factors and predicting the voluntary termination, absences and other risks to fully realize the business profits. It can be done by statistical techniques, data mining models and machine learning methods that can extract the historical facts and can convert it into information. Hence this study will give a new term to data analytics that is HRML-human resource machine learning

Advent of HR systems and machine learning models has ability to achieve the organizational goals of human capital management (HCM). It is crucial to check which employee has the higher chances to leave the firm such that managers can motivate that employee or can find other employee who is capable of filling the vacant position in the advance. It can be achieved by predictive retention modelling, risk modelling in which a manager develops a profile of employees that have a higher risk of leaving prematurely. According to (Mishra et al.; 2016), it is evident that companies cannot continue in the long run if they do not adapt the predictive analytics. HRPA is adapting by many large companies like Facebook, Google, Microsoft and will take over the traditional analytics in upcoming future. Hence, this research is a major step towards HRPA which can judge voluntary attrition, involuntary attrition that includes good attrition, bad attrition which can contribute to data analytics field in renovating Human Resource Management.

So far, most of the related predictions has been done in customer churn and employee churn through the statistical methods or survival analysis which make use of detailed employee information Sharma (2017) and Mulla et al. (2013). To the best of my knowledge, this paper will be the first study to predict employee attrition through hybrid machine learning methods of decision trees-gradient boosting, KNN, SVM, GLM and association rules which will be looking at employees age, education, environment and job satisfaction, job involvement, performance rating, relationship satisfaction, work life balance, turnover and marital status to predict the binary value attrition- yes or no in the IBM employee attrition dataset. This study aims to bridge the hiatus between the HR management framework and prediction modelling using CRISP-DM and motivates me to research on this question. An investigation on machine learning techniques used in analysing employee attrition are efficient?

There are some keywords used in the paper - Employee attrition, Employee retention, data mining, machine learning, HRPA, HRML. This paper will be covering 4 sections, first section describes about the related research work happened on employee attrition that is literature review and what deviations has been done from these papers which is subdivided into other parts. Second section tells about the research methodology and project time plan through the means of Gantt chart. Third section conclude the work happened in this paper and last section includes the references which motivated me to research on this topic.

2 Literature Review

2.1 Evaluation of Employee Turnover using Survival Analysis

Customer churn is the major problem in many companies as loss of customer affects the profits and goodwill of an organisation which makes it difficult to acquire new customers in the market. According to (Saradhi and Palshikar; 2011) it is vital to build predictive models for customer churn as well employee churn and hence demonstrated that Support Vector Machine-a machine learning technique should be used to detect which employee is going to leave the company, in their future work they suggested that survival analysis techniques should be used to build employee turnover prediction models to get the more accuracy for the retention of valuable models. This previous research was carried and explained by (Griffeth and Hom; 2001) in which survival analysis was focused over a specified period of time in which the event of employee retention occurs which retained the valuable workforce for the survival of an industry. They suggested that to remain the quality of an employee there should be appropriate strategy by judging the voluntary and involuntary turnovers from which avoidable turnover should be taken into consideration so that management can improve the sound strategy to improve the staff retention.

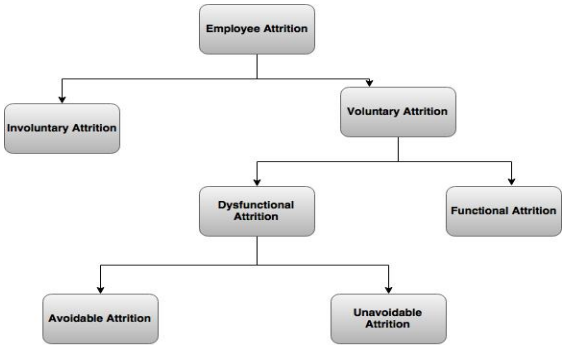


Figure 1: Employee Attrition Flowchart, (Griffeth and Hom; 2001)

Fig 1 (www.draw.io) explains that there are two types of attrition in which voluntary attrition which are dysfunctional and avoidable can be taken as a very crucial loss so that managers can make strategies to avoid employee attrition. (Griffeth and Hom; 2001)opined that high attrition can result in business loss and therefore survival analysis is necessary to act as a benchmark to stop in in near future. The other research was coined by Mulla et al. (2013) which was motivated by the previous researchers and applied survival analysis in 2013 on the data which was collected from Bharat Petroleum Corporation Limited (BPCL) that uses 2141 engineers over 13 years from the year 2000-2012.They performed multiple regression that is a part of survival analysis to detect the drivers of the turnovers and researched that employees who were younger, poor performers, unmarried and lived far from the company leave earlier than other employees Mulla et al. (2013).In the later time survival analysis was done in people analytics research which states that survival analysis can be great help to identity the employee attrition using survival curve in R tool (Isson and Harriott; 2016) but sometimes can be defined as a machine failure for the prediction due to the complications as no one knows when a current employee will terminate like 2 weeks 6 weeks or maybe next month. From the above discussion, we investigated that survival analysis can be used in some fields to find the voluntary attrition which is avoidable but the accuracy was lower and hence

(Isson and Harriott; 2016) guided to use data mining techniques in near future to evaluate employee performance which is covered in next section.

2.2 Data mining role in evaluating Employee performance by HRM

Survival analysis was replaced by data mining as we covered in the previous section and this section contrasts the data mining role in Employee performance as data mining is globally used for extracting insights from data to make efficient decisions. Employee performance is the most important thing in an organisation to validate in well-established subdomains like manufacturing management, customer management and financial management which comes under human resource management. In todays world, there is growing number of research contributions which aims at supporting HRM data mining. Selecting employees, ascertaining competencies of employees, functions of staffing, career planning, HR costs planning and predicting the employee attrition by evaluating the employee performance in HRM is ongoing. A similar research has been done by (Strohmeier and Piazza; 2013) with the help of domain driven data mining techniques including cluster analysis, association analysis, decision trees, neural nets and Support Vector Machines (SVM) which provided the relevant results in the real world and evaluated that SVM is the most accurate method to check the employee performance but there was lack of suitable HR data. Hence SVM is used in this paper to check the employee attrition as an elite method to find the great accuracy. Employee performance was also checked by (Gupta et al.; 2014) in industrial practices from a software project team to check the correlation between the software quality and the project success by using decision trees which was over fitted and this research was extended by (Thakur et al.; 2015) using the same dataset to improve the results. According to (Thakur et al.; 2015) random forest-a bagging based approach showed the better results of employee performance by overcoming through overfitting which have the capability of storing sparse data and hence outperformed decision trees and linear regression.

Performance evaluation is a crucial part HRM which includes many attributes for enhancing the performance of an employee which can be checked by attendance, seminars attended and paper presented to predict the performance of the employee at the end of the year which results in evaluating the employee data for giving promotion, career advancement and yearly increment or not. According to (Waran et al.; 2014) this can be checked by applying rule mining, clustering and decision tree J48 algorithm by judging the root node whether employee will get promotion or not. Extreme gradient boosting (XGBoost) was implemented by (Ajit; n.d.) which showed the great efficiency. Therefore, this paper takes these previous papers into consideration and uses gradient boosting technique which is extended version of decision tree to perform on employee attrition data as similar to employee performance data to get the more efficient results.

Another evaluation of the employees was carried out by (Aktepe and Ersoz; 2012), in which employees were clustered into four distinct groups according to the strategic plans and job satisfaction-performance and k-means clustering approach combined with ANN was followed on the manufacturing company located in Turkey in moving groups A, B, C to group D and keeping the people in group D to find the highly performed employees. This study uses the different clustering method of supervised learning i.e. KNN

by finding the nearest attribute and taking the majority of the labelled class to evaluate which feature of employee contributes the most to the employee attrition. From the topic discussed above, data mining plays a vital role in evaluating the employee performance and similarly different machine learning techniques (SVM, decision trees-gradient boosting and KNN) should be applied, the detailed description of artificial intelligence in the field of HRM is covered in the next section which uses machine learning methods.

2.3 Prospects of Artificial Intelligence in finding Employee Attrition

Artificial intelligence is intelligence displayed by machine such as learning, problem solving or predicting anything, this section extends the previous section of data mining role as predictive analytics is known for state-of-the-art accuracy in decisions which is done by machine learning techniques. Survival of an organisation persists only if it uses Artificial intelligence to predict the churn prediction and management (Berson and Smith; 2002). This research was carried by many other researchers and among one of them (Wei and Chiu; 2002) opined churn prediction in mobile telecommunications industry as the research states that churn prediction and management is the vital in the fast changing way to improve the customer retention, it uses the multi-classifier class-combiner approach to distinguish between the churners and non-churners which showed the improved churn accuracy. This customer churn research was extended by (Hung et al.; 2006) which worked on the same data set of Taiwan mobile telecommunications and evaluated that hybrid machine learning techniques like decision tree with neural nets were effective to predict whether the customer will change the plan and accordingly proactive actions can be taken to valuable churners. Job performance prediction in a call center in Chile was done by (Valle et al.; 2012) using nave Bayes classifier in which objective was to know what were the levels of the attributes that were the indicative of individuals who performed well but the research showed the poor predictive results and hence in their future work they suggested to use different machine learning algorithms that can perform well. The similar research was done on multimedia on demand system by (Tsai and Chen; 2010) using association rules to contain the most important factors of the customer in pre-processing stage which was compared by neural networks and decision trees to check out the customer churn. The results showed that association rules can be used for in data pre-processing state to improve the accuracy of the final outcome. Effective feature selection on employee attrition was done by (Chang and Xi; 2009) in which mixed approach by Taguchi and nearest neighbour classification rules were used on the dataset in which gender, salary, seniority, sick leave and much more were selected and showed 88 percent accuracy which also concludes hybrid methods can show the better results. Therefore, this paper projects hybrid machine learning algorithms with decision tree and association rules to enhance the prediction of employee attrition so that some major steps can be taken to avoid valuable churners.

The new process was given by (Dutta et al.; 2010) in which feature selection of attributes was a main part and applying data mining technique to identify a pattern for future performance and then providing the prediction to a user. This paper follows the guidelines of it.

A predictive workforce analytics was done to find out the employee attrition in banking and financial industry to identify in advance which employee can exit the company using

factor and regression analysis (Sharma; 2017). This paper is motivated by (Sharma; 2017) and uses generalized linear model to check the accuracy of the employee attrition which is similar to previous work. Other related research explains ANN which uses the back-propagation method that can be a good artificial intelligence method as it delivers fast and accurate predictions for a business (Panchal et al.; 2010) but (Panchal and Panchal; 2015) concludes that genetic algorithms based BPN can find accurate and fast results (10 seconds) then ANN and hence genetic algorithm is used in this study to find more accurate employee attrition. This particular section concludes artificial intelligence plays an important role in predicting customer churn and employee churn using genetic algorithm, association rules and generalized linear models, therefore this paper uses these techniques to determine the employee attrition. The data mining extracted patterns and accuracy from artificial intelligence give rise to decisions which is covered in next section.

2.4 Decision-Making framework for HRM and talent management

Decisions are the final and crucial step for an organisation to be made which is done after extracting patterns from data mining and getting accuracy from machine learning methods which is discussed above. Decisions encourages the organisation to work effectively like digging out the big best talents in the organisations, evaluating the performance of employee and motivating employees by giving away rewards and giving promotion so that employee stays in the company. A framework proposed by (Mishra and Lama; 2016)that predicts workforce and talent by using data mining and predictive analytics. It gives contribution to the HRPA field which provided better solutions to the past limitations and similar research happened in 2014 by (Varshney et al.; 2014) which uses the IBM dataset to find out the expertise taxonomy. They evaluated features from four distinct data sources-social tags, HR information, job title, work products which was evaluated by using l2-regularized logistic regression which was used with cross-validation and got 96 percent accuracy if hybrid of three machine learning techniques applied on one single employee. Hence this study uses the hybrid techniques with focusing on the decision making to find out most talented employees to place them as first priority.

From the above literature review. Employee attrition can be find out on the IBM dataset using different data mining techniques to extract the pattern, artificial intelligence by machine learning to find the best accuracy and make decisions to find the most talented employees who can leave the company which can be used in the field of HRM.

3 Research Method and Specification

3.1 Methodology

In this modern era, there is much competitive world to build good relations with customers and not allowing them to leave the company i.e. customer churn problem. Similarly, there should be good relations between company and its employees to prevent the employee attrition problem. This research covers employee attrition prediction with the intend to use machine learning techniques like KNN, decision tree-gradient boosting, SVM, association rules, GLM and genetic algorithm to get the best accuracy. This research can be extended to next research in which model can be built through CRM point of

view like SEMMA or KDD with the deep analysis of the attributes using deep learning which extends the artificial neural nets and adding more hidden layers with storing the memory through LSTM and RNN that can help the companies to make good decisions. A case study by (Sikaroudi et al.; 2015) motivated this paper to use different data mining methods to compare accuracy, calculation time and user friendliness with the CRISP-DM methodology-Cross Industry Standard process for data mining. Similarly, CRISP-DM is used here as a solving strategy that can fit well with the data mining process that contains six iterative main sectors:1. Business understanding, data understanding, data preparation, modelling, evaluation and deployment (Sikaroudi et al.; 2015).Figure 3 is made from draw.io which shows the CRISP-DM model.

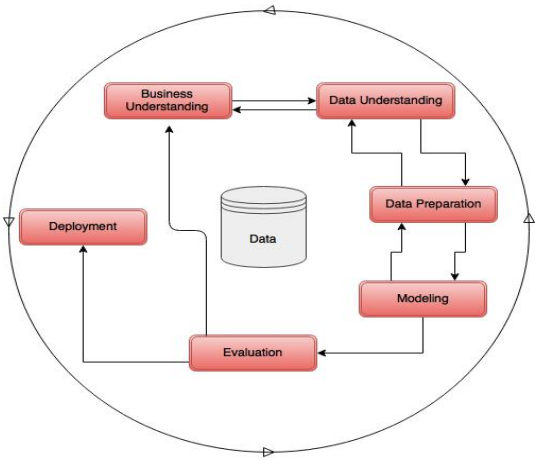


Figure 2: Fig-3 CRISP-DM

3.1.1 Business understanding

This research includes the employees from IBM company that is one of the fortune 500 companies from America, to know which employee can leave the company in the future it is very crucial to know attributes which will be effecting the attrition. DSS which is capable of detecting the employees who are highly probable to leave their jobs can be the first step so that hybrid model of different machine learning techniques can be applied.

3.1.2 Data Understanding

For the evaluation and modelling the first step is to understand the data which has been gathered from IBM site. Here are the attributes :-

Fig 4 shows the information about the attributes and the IBM data information which is marked as blue and yellow shows the class label which is employee attrition that is the class label yes or no. Research is about this label which factor is effecting the most and what is the best model to be used to find this label.

3.1.3 Data Pre-Processing

After understanding the data, we need to clean and transform the data with labelling the information. The data should be appropriate for the models otherwise it will not

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

Figure 3: Fig-4 Data set information from IBM site

match the respective models hence, data will be pre-processed according to the different machine learning techniques. Labelling will be changed like this:Education-Below college -2 College-2 Bachelor-3 Masters-4 Doctor-5.Environment Satisfaction-Low-1 Medium- 2 High-3 Very High-4.Job Involvement-Low-1 Medium-2 High-3 Very High-4.Job Satisfaction-Low-1 Medium-2 High-3 Very High-4.Performance Rating- Low-1 Medium-2 High-3 Very High-4.Relationship Satisfaction- Low-1 Medium-2 High-3 Very High-4.Work Life Balance-Bad-1 Good-2 Better-3 Best-4

Many levels of measurements will be seen like categorical or nominal data, ordinal data, interval and ratio and accordingly data mining techniques will be applied like association rules to pre-process into categories and many more. Data should be free from redundant values,no missing Values,data will be transformed into the consistent form and checking incorrect attributes of data.Outliers, data distributions with the help of histograms and scatter plot will be done in R studio.

3.1.4 Modelling

After the pre-processing, the data is ready to be processed to check the patterns through the data mining techniques and then doing predictive analysis with the help of machine learning which will be done by R studio, Weka, SPSS and Rapid Miner. Following techniques will be used in the research:

1.Decision trees (Gradient boosting) This is the most usable and compatible machine learning technique which is used to explore the data and showing the attributes to the end user as it the most transparent supervised classification technique in white box which helps in understanding the data from root node to leaf node than the leaf node is tells, in this research the leaf node will tell whether the employee will stay or not. Gradient boosting will be used to generalize the data into more optimized form with minimum cost.

2.Support Vector Machine- This is the black box technique used in many researches as it mostly has the best accuracy which converts the non-linear data into high dimensional plane through the kernel trick and then performs same as the linear data by selecting the hyper-planes hence this research uses this technique to get the more accurate data if the data is in non-linear data.

3.Generalized Linear Model- It is an extension of linear model and general linear model which have the ability to predict the categorical or non- normal dependent variable without having linearly dependency on the response variable. It can take yes/no values and similarly the attributes will be predicting the attrition values. They have the self-assertive distribution which we can be used for this analysis.

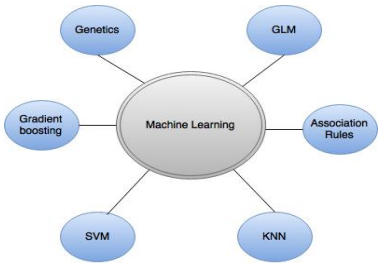


Figure 4: Fig-5 Modelling Techniques

4.Association Rules-It is used which can create rule-based to ascertain the interesting patterns between the attributes and then predicting the values through other data mining techniques, basically it can act as data transforming or pre-processing step for the other data mining techniques. This research will be finding the most interesting patterns then will be applying data mining techniques.

5.KNN- It is a supervised classification techniques which is non-parametric that is used for discovering the patterns by taking the nearest neighbours and taking the majority from them and hence this paper uses this technique to find the attrition by taking another attributes distance and then taking the majority from them. If yes or no is given and the value of K=3 then it will take the three nearest neighbours and take the yes if there is 2 yes values and 1 no variable which can be used in the analysis.

6. Genetics Algorithm This method plays an important role in solving unconstrained and constrained optimization difficulties which is based on a normal natural process that mimics biological evolution. This algorithm plays role like genes in a human body and generates a population of points for each and every iteration and among them best point which is an optimal solution is selected. The main advantage of genetic algorithm is that it uses random number generator which can be very good for this research to have random sampling and can have the trade-off between bias and variance.

3.1.5 Evaluation

This phase evaluates the best machine learning technique from the modelling phase and compares different accuracies according to the performance measures and ranking based, the technique having the maximum accuracy is applied in the next deployment phase and sometimes it gets back to the business understanding phase.

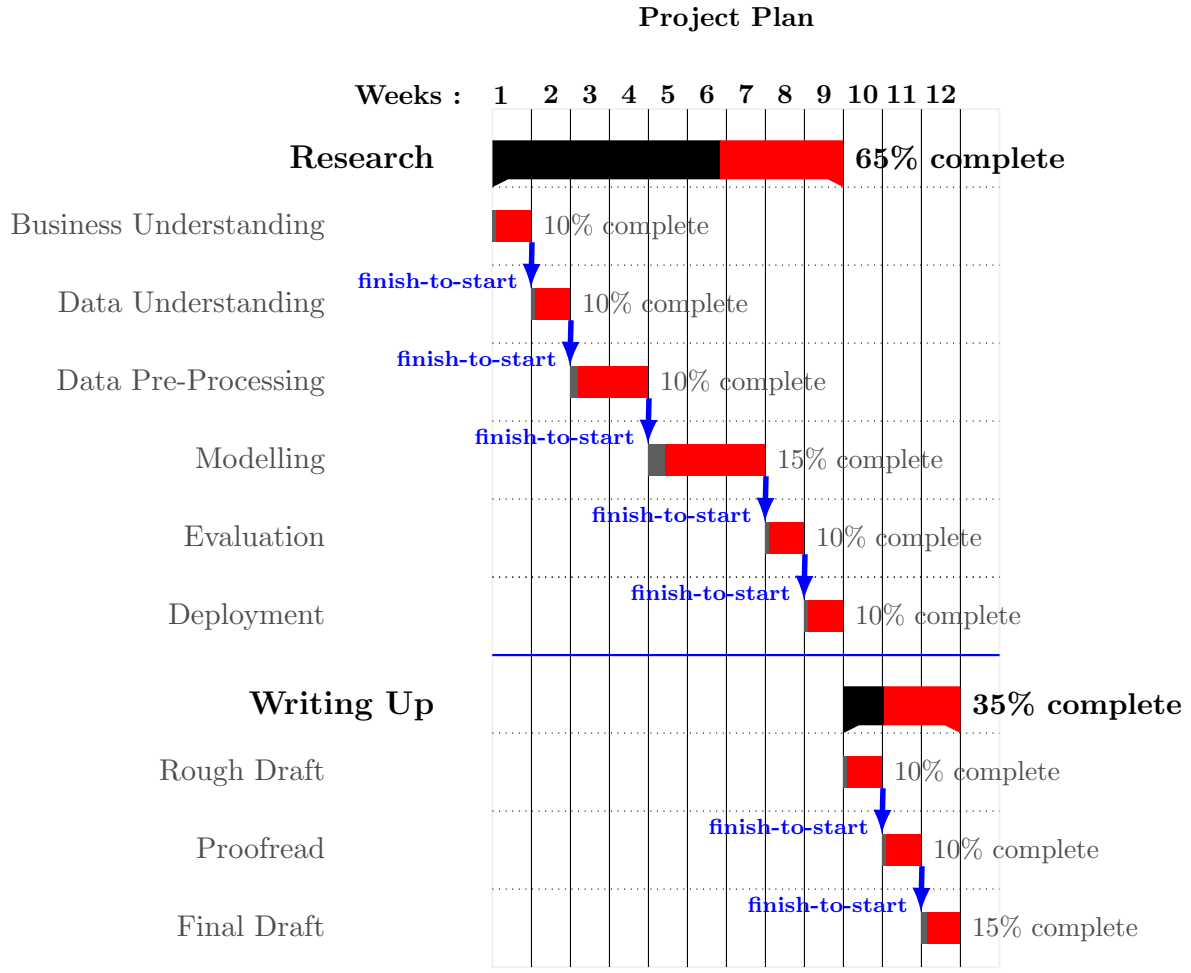
3.1.6 Deployment

The last phase of the CRISP-DM is deployment which is connected to the evaluation phase having a single sided connection. The technique which is the best evaluated is then deployed on the data which is the main purpose of this research to find out.

From the above research methodology, we can see CRISP-DM plays an important part in which process goes around the data from business understanding to the deployment phase and the technique which is optimistic is then applied to the IBM dataset from which managers can take some measurable steps to find out which employee can leave the company in the future and can motivate them or can find a new replacement to save the money and ROI.

3.2 Project Plan

For the time period of 12 weeks, the research is divided among two phases that is Research and Writing part which further divided into the different parts. Gantt chart explains the proper division of work into 12 weeks in which 65 percent work is divided into research part and writing part is divided into 35 percent. The research will be followed according to the Gantt chart time stamp :-



References

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

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

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

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

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

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

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

Hung, S.-Y., Yen, D. C. and Wang, H.-Y. (2006). Applying data mining to telecom churn management, *Expert Systems with Applications* **31**(3): 515–524.

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

Mishra, S. N. and Lama, D. R. (2016). A decision making model for human resource management in organizations using data mining and predictive analytics, *International Journal of Computer Science and Information Security* **14**(5): 217.

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

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

Panchal, G., Ganatra, A., Kosta, Y. and Panchal, D. (2010). Forecasting employee retention probability using back propagation neural network algorithm, *Machine Learning and Computing (ICMLC), 2010 Second International Conference on*, IEEE, pp. 248–251.

Panchal, G. and Panchal, D. (2015). Hybridization of genetic algorithm and neural network for optimization problem, *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)* **4**(4).

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

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

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

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

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

Tsai, C.-F. and Chen, M.-Y. (2010). Variable selection by association rules for customer churn prediction of multimedia on demand, *Expert Systems with Applications* **37**(3): 2006–2015.

Valle, M. A., Varas, S. and Ruz, G. A. (2012). Job performance prediction in a call center using a naive bayes classifier, *Expert Systems with Applications* **39**(11): 9939–9945.

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

Waran, M. et al. (2014). Employee performance evaluation using machine learning algorithm, *International Journal of Computer Communications and Networks (IJCCN)* **4**(2).

Wei, C.-P. and Chiu, I.-T. (2002). Turning telecommunications call details to churn prediction: a data mining approach, *Expert systems with applications* **23**(2): 103–112.