ANN AND STACK BASED APPROACH TO EMPLOYEE ATTRITION PREDICTION

STUDENT: TEJAS DINKERRAI DESAI

Barcode: 21111276339116 Person Num: 890526

THESIS SUPERVISOR: MR. JANPREET SINGH

Final Thesis Report

A thesis submitted in partial fulfilment of the requirements of Liverpool John Moores University for the degree of Masters in Data Science

FEBRUARY 2020

ABSTRACT

Attrition means loss of workforce ('Voluntary' or 'Involuntary'). Scope of this study is limited to the *prediction of voluntary attrition*. The attrition rate in major economies is double than the global average of 7%, *making it* a *global problem*. 77% of voluntary attrition can be prevented if predicted in advance, saving millions of dollars. So, predicting attrition is a *huge* business problem. Changing demographics and the advent of gig-economy are making this *timely and relevant* topic to study. Inclusion of human resource analytics as a part of companies' core business process has made plenty of Human Resource data available. This, combined with the availability of powerful analytical platforms and maturity of analytics to predictive level have made this a *feasible* problem *to solve*.

Very few studies have used neural-network (NN) on this problem and, so far, only two studies applied stacks to it. No research has examined this problem *from all angles by using ANN, Stack and visual-analytics*. Further, *the key issues of variable-significance and explainability and interpretability of results, inadequately addressed by earlier studies, are addressed in current study using the latest tools and techniques. This is its novelty-value.*

Two models each were built using neural-network and stack to predict the probability of attrition. The metrics obtained by 'neuralnet' based neural network were: Accuracy=0.8183, Sensitivity=0.8811, F1_Score=0.9002, AUC=0.739. Those by 'nnet' based neural network were: Accuracy=0.8485, Sensitivity=0.9584, F1_Score=0.9117, AUC=0.7620.

For caretStack (Random Forest as meta-classifier and C5.0, NB, GLM, KNN and SVM as base-classifiers), metrics were: Accuracy=0.9199, Sensitivity =0.9757, F1_Score=0.9534, AUC=0.7363, Kappa=0.6696. For H2o based stack (with deeplearning as meta-classifier and GLM, GBM, RF, deeplearning as base classifiers) metrics were: AUC=0.8358. Accuracy=0.8868, Sensitivity/Recall=0.9340, F1_Score=0.9340, Kappa=0.5377

Key Words:

Employee attrition; voluntary turnover; machine learning; Artificial Neural Network (ANN); Stack;[1], explainable AI, interpretable AI

[1] IN THESIS, AFTER CLICKING A HYPERLINK, TO COME BACK TO ORIGINAL PLACE, USE SHORTCUT "Alt + <-" i.e. [ALTER + LEFT ARROW]

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

SR NO	ABBREVIATION	EXPANSION
1	ANN	ARTIFICIAL NEURAL NETWORK
2	ACC	ACCURACY
3	AI	ARTIFICIAL INTELLIGENCE
4	ANFI	ADAPTIVE NEURO FUZZY INTERFACE
5	AUC	AREA UNDER (ROC) CURVE
6	BFGS	BROYDEN–FLETCHER–GOLDFARB-SHANNO
7	BNN	BIOLOGICAL NEURAL NETWORK
8	BPA	BACK PROPAGATION ALGORITHM
9	BPN	BACK PROPAGATION NETWORK
10	BUPA	British United Provident Association (now a multi-insurance group)
12	CRISP-DM	CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING.
13	CV	CROSS VALIDATION
14	DALEX	MODEL AGNOSTIC LANGUAGE FOR EXPLORATION AND EXPLANATION
15	DNN	DEEP NEURAL NETWORK
16	DT	DECISION TREE
17	EDA	EXPLORATORY DATA ANALYSIS
17-B	EOR	EMPLOYEE ORGANISATION RELATIONSHIP
18-A	FPR	FALSE POSITIVE RATE
18-B	FNR	FALSE NEGATIVE RATE
19	GA	GENETIC ALGORITHM
20	GB	GRADIENT BOOSTING
21	GLM	GENERALISED LINEAR MODEL
23	HR	HUMAN RESOURCE
24	IBM	INTERNATIONAL BUSINESS MACHINE
25	K-MEANS	AN ALGORITHM FOR CLUSTERING
26	KNN	K NEAREST NEIGHBOURS
27	LDA	LINEAR DISCRIMINANT ANALYSIS
28	LR	LOGISTIC REGRESSION
29	LVQ	LINEAR VECTOR QUANTIZATION
30	ML	MACHINE LEARNING
31	MLP	MULTI LAYER PERCEPTRON
32	MLR	MULTIPLE LINEAR REGRESSION
33	MSE	MEAN SQUARE ERROR
34	NB	NAÏVE BAYES
35	NN	NEURAL NETWORK

36	NNSOA	NEURALNETWORKSIMULTANEOUSALGORITHM
37	ODbl	OPEN DATA COMMONS OPEN DATABASE LICENSE
38	PD	PROBABILISTIC DISTRIBUTION
39	PDP	PARTIAL DEPENDENCE PLOT
40	PERT	PROGRAM EVALUATION AND REVIEW TECHNIQUE
41	R&D	RESEARCH AND DEVELOPMENT
42	RF	RANDOM FOREST
42-B	RMSE	ROOT MEAN SQUARE ERROR
43	ROC	RECEIVER OPERATING CHARACTERISTICS
44	SMOTE	SYNTHATIC MINOIRITY OVERSAMPLING TECHNIQUE
45	SOM	SELF ORGANIZING MAPS
46	SVM	SUPPORT VECTOR MACHINE
47	SVMRADIAL	SUPPORT VECTOR MACHINE (with) RADIAL KERNEL
48	TPR	TRUE POSITIVE RATE
49	UCI	UNIVERSITY OF CALIFORNIA, IRVINE
50	XGB	EXTREME GRADIENT BOOSTING

ACKNOWLEDGEMENT

I deeply acknowledge LJMU and IIIT-BENGALURU for creating this unique opportunity to learn and acquire a renowned degree.

Heartfelt gratitude to Prof. Dr Dhia, Prof. Dr Manoj and Prof. Thiru and Prof. Jagadeesh Anmala for giving invaluable guidance through webinars and viva.

My Special thanks to Prof. Paulo (HOD, Applied Mathematics, LJMU), Prof. Ahmed, (Dean, Faculty of Engineering, LJMU), Prof. Atif, Prof. Gabriela, Dr Mohit. Dr Rubem, Dr Silvester and Dr Ivo for their enlightening lectures.

Very special and deep gratitude to my Thesis-Supervisor Mr Janpreet Singh for his precious and timely guidance and patient and calm way of handling student.

Sincere gratitude to Team Upgrad, especially two student mentors Ishrat and Anmol for providing constant administrative and procedural support.

Finally, I thank my family for the sacrifices they have given for accommodating my busy schedule without complaints and, in fact, encouraging me.

Tejas Desai. February 15, 2020.

TABLE OF CONTENTS

THIS PAGE IS INTENTIONALLY LEFT BLANK

CHAPTER 1 INTRODUCTION

1.1 Background of the study:

Employee-attrition (here onwards referred to only as attrition) is a reduction of workforce. Quantitatively, for a given period, attrition is:

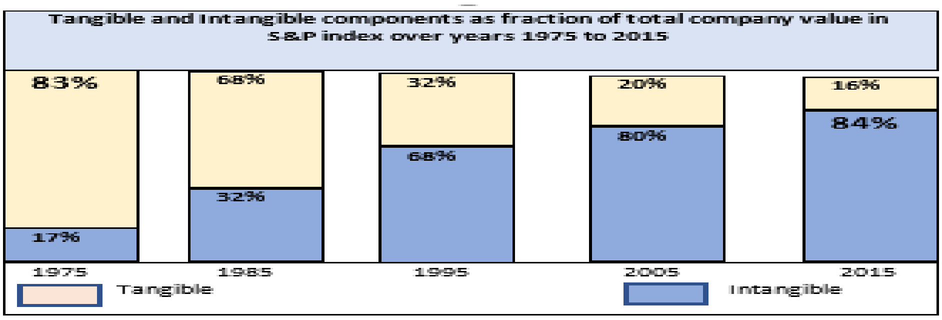
$$\frac{Total\ no\ of\ quitters\ over\ a\ period}{Average\ Total\ no{:}\ of\ Employees\ over\ the\ period} \times 100$$

In literature, attrition is classified into 'Voluntary' and 'Involuntary' attrition. In this research, the *focus is only on the prediction of 'Voluntary' attrition* as 'Involuntary' attrition is already under management's control and knowledge.

1.1.1 How attrition is global, worsening problem?

a) The advent of the knowledge economy and the subsequent rise in the intangible assets as a fraction of total S&P value to 84%. In China, this proportion is 85% [c] of stock market value. *Today $21 trillion in the US is intangible assets (patents, copyrights, code, data, trademark, brand, etc)*. [Figure- 1.1].

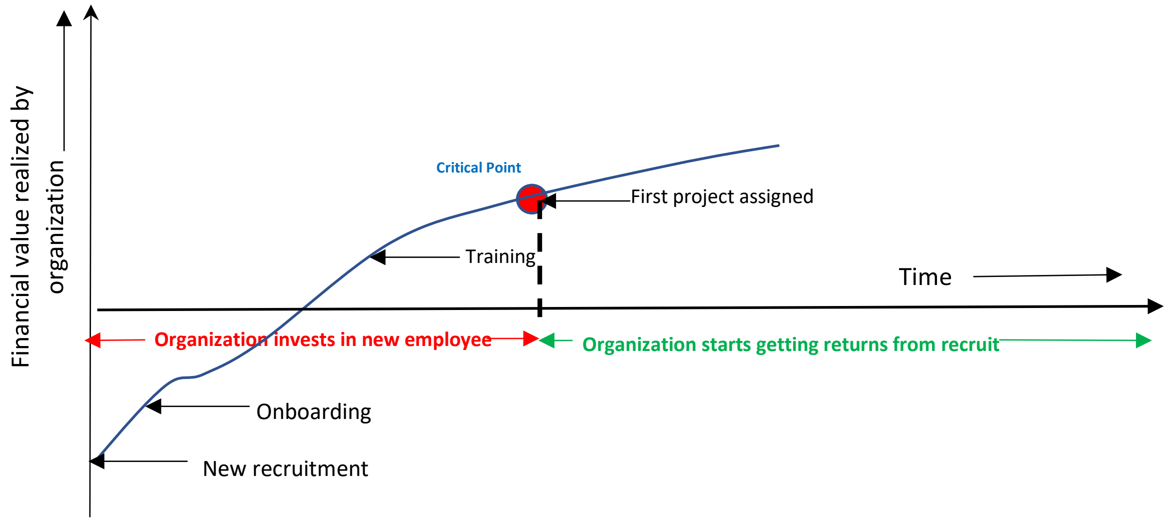
FIGURE-1.1 Proportion of intangible asset in total company value



Intangible assets are created by employees of companies. Thus, in a way, Figure 1.1 shows that employees themselves have become an asset for the company. So, companies want to retain their top talent (creators of intangible assets) and predict their intention to quit.

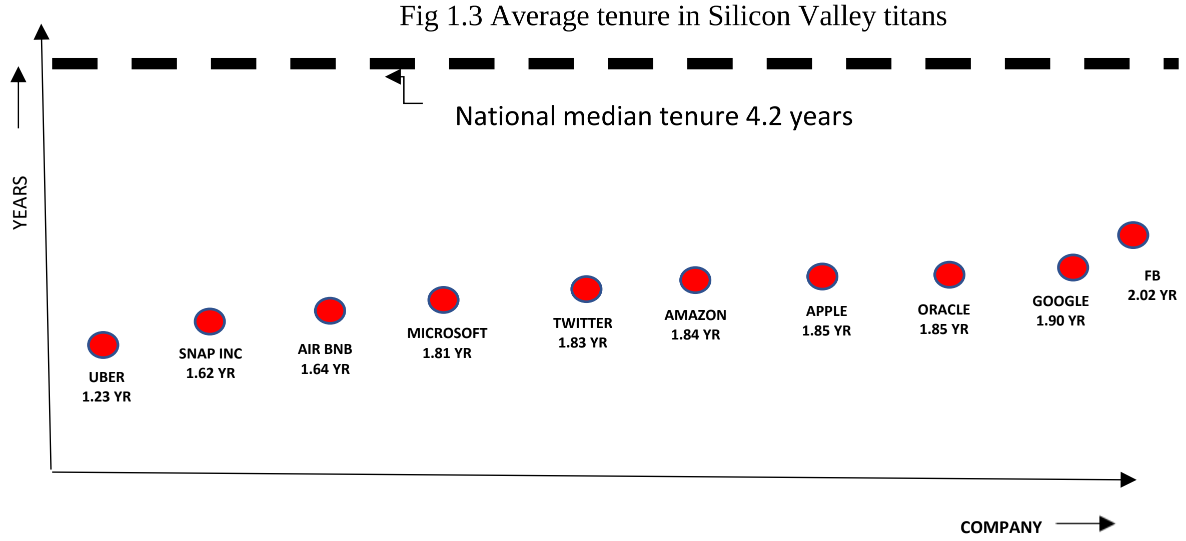
b) Another reason is: when a company hires an employee. it starts deriving value from employee after a time lag. If an employee quits before a critical point, then the company actually makes a loss. [Figure 1.2]

FIGURE-1.2: Cost-value for a new hire



c) Third reason is the advent of the gig economy and the arrival of gen-x and gen-z in the workforce. E.g. in 2018, millennials in silicon-valley changed jobs twice as often compared to the national median job-tenure of 4.2 years, even in top 10 tech titans.

Fig 1.3 Average tenure in Silicon Valley titans



d) Frequent attrition is not just IT-industry-specific[b]. Tenures for other industries are also decreasing. *The roles with the highest attrition both globally and in Europe are finance, sales, and HR with percentages 12.7%, 12.6% and 10.9% globally and* 11%, 9.9% and 9.7% in Europe.

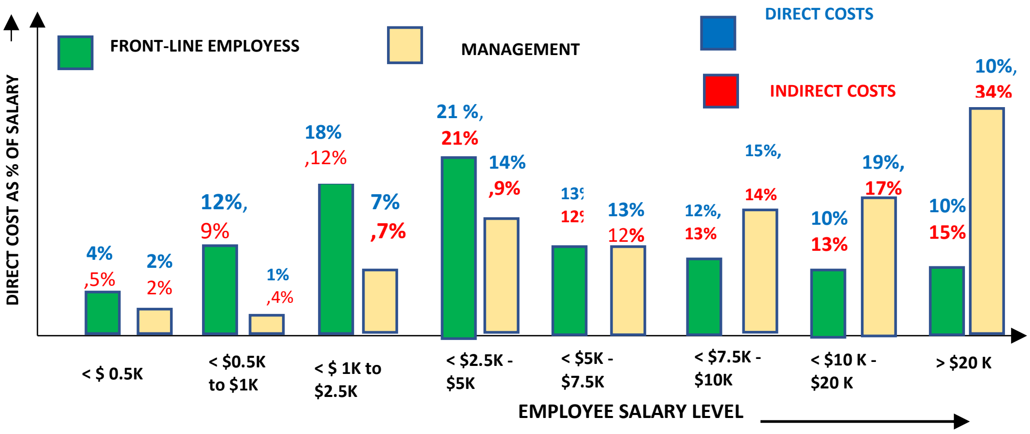
e) The problem is not country-specific either. In Australia and UK, the attrition-rate was 15% in 2018, much above the European average of ≈7%. Employers in Hong Kong, Japan, Romania, Taiwan, and Turkey are also finding skilled-labour shortage.

f) Further, *the problem is only going to worsen with time*. By 2030, talent-shortages will greatly affect the services-sector with estimated shortfalls in all major economies [c]. By 2030, the demand *for skilled workers* will exceed the supply creating a global shortage of 85.2 million skilled workers, resulting in unrealised output to jump from $2.1T in 2020 to $8.5T in 2030. This is *equal to combined current GDP of Germany and Japan.* [c]. So, a *research like this that helps companies handle attrition, hiring and retention in an orderly manner will become increasingly relevant.*

1.1.2 Direct and indirect costs of attrition

Total costs associated with attrition has doubled from $331B in 2010 to $680B in 2018 in the US alone and is likely to reach $930B in 2030. The direct cost of replacing one employee is, on average, $15000 [4], and it increases with skill-level of quitting employee. [d]

FIG 1. 4 Direct and indirect costs of attrition as% of salary, based on salary/skill (BLS, US)



Indirect costs of attrition are *loss of knowledge and trade secrets, reduced productivity, customer unsatisfaction, reduced morale of stayers, lost team balance* and so on. Considering both direct and indirect costs, each attrition can cost *one to two times an employee's salary*.

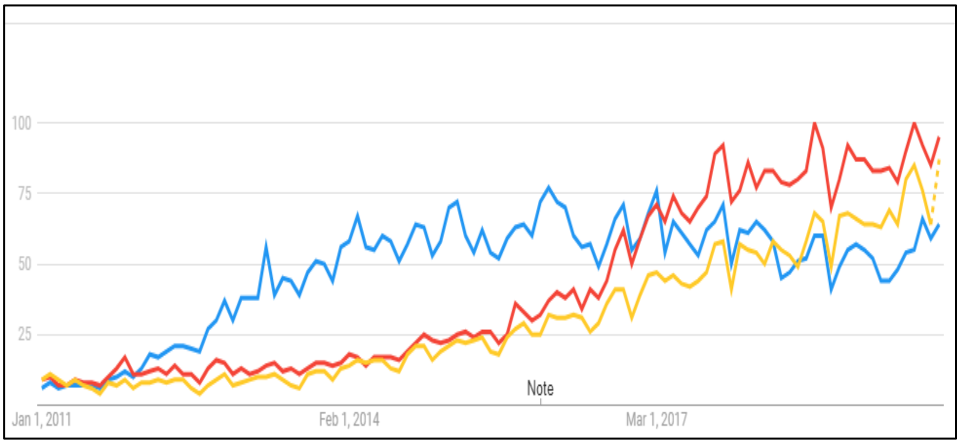
Since 77% of attrition is voluntary, the controllable cost for U.S. alone was around $524B in 2018. Reducing this by just 20% would have saved humongous $105B in 2018 (By 2030 estimates, around $150B). This is possible as nearly *77% of quitters can be retained* as per reliable studies[c], *if the employer can predict their intentions to quit.*

In short, attrition prediction is *timely*, *globally relevant*, and *big* problem which is fit for a research topic. Recent advances in data-analytics *from 'report' to 'analyse' to 'monitor' to 'predict' to 'simulate' has made it feasible* to predict attrition. With companies embedding HR analytics in their core systems and powerful analytical platforms available as a service, *data or hardware is no longer a constrain. Actually, the lack of robust models is slowing down the adoption of HR Analytics.* So, this is an apt problem for thesis.

1.2 Why machine learning based methods?

After describing the importance of attrition as a research topic, next, it must be described why machine-learning (ML) based methods over statistical methods were chosen. Study of google trends *from 2011 to 2020* extracted by researcher shows an increasing interest in data-science (DS), machine learning (ML), Artificial Intelligence (AI) and Bigdata. This is shown in Fig 1.5 This is one reason why ML-based methods were chosen. Further reasons for choosing ANN and Stack over other ML algorithms shall be described in section-3.4.1 .

Fig- 1.5 Google search trends using keywords 'Big Data (Blue)', 'Machine Learning (Red)' 'Data Science (Yellow)'



1.3 Problem-Statement (Research-Question):

How to build machine-learning (ML) models based on artificial-neural-network (ANN) and stacking which can accurately predict attrition, based on data of employees.

1.4 Aims and objectives:

1.4.1 Aims

- To create ML-based *explainable and interpretable* models based on ANN and Stacking which can accurately predict attrition**.**

1.4.2 Objectives [or Research Question]

- To create two NN based and two stacked models with at least 90% value of F-1 score and sensitivity, without sacrificing other metrics too much. The researcher is more interested in sensitivity (i.e. the number of attritions correctly predicted as such) than overall accuracy.
- To get the relative significance of predictors on attrition by all models**.**
- To compare two ANN models and two stacked models in terms of result-metrics, training-time, programmer's efforts and hardware-resources required.

1.5 Scope/Limitations of the study:

- The scope is limited to voluntary attritions only.
- The scope is restricted to diagnostic and predictive analytics but not prescriptive analytics.
- Deployment and Tuning of models will not be discussed.

1.6 Novel Features of The Study:

This study will add to the existing body of knowledge in the following manner:

- No research so far has studied attrition using the stacking in so much detail as this work.
- Some other studies based on ANN on this topic got poor sensitivity (≈34%) which is the key metric in this problem. This study corrects this problem.

- The study adds rigorous visual-analytics for extra validation and robustness and measures to what extent results of ML based methods match with results of visual-analytics.
- The study is perhaps first in addressing, in detail, key aspects of modern ML *i.e. Interpretability and Explainability. Regulators want to know 'how' the ML model has arrived at a decision. These aspects are also important for winning people and shareholder trust in ML and AI. Without addressing them, it is difficult to convert 'models' into 'products.'*

1.7 Structure of The Thesis:

Initial pages of the thesis are Title, Abstract, List of Tables, List of Figures, List of Abbreviations, Acknowledgement and Table of Contents. Next, thesis is arranged as below:

- Ch:1-Introduction
- Ch:2-Historical Background and Literature Review
- Ch-3-Research Methodology
- Ch-4 Implementation
- Ch-5 Results and Analysis
- Ch-6 Conclusions, Limitations, contribution and Future directions
- Ch-7 References/Bibliography

1.8 Summary:

To summarise, attrition prediction is today a *globally relevant, timely, urgent* and *big* problem. Further, most modern researchers are favouring ML methods rather than statistical methods. Statistical methods make an *assumption that variables follow Gaussian normality in distribution.* This assumption is almost always violated for practical and finite datasets. This makes such methods invalid for a *finite sample*. Further, statistical-methods often assume *that underlying data has linear distribution.* Contradictory to this, ANNs *do not assume any prior distribution of data and they are robust to noisy data.*

Next chapter:2 discusses Historical-Background and Literature-Review of past attrition studies.

CHAPTER 2 BACKGROUND AND LITERATURE REVIEW

2.1 Historical Background:

- Attrition is under study from 1917 and around 1500 academic studies are available on the subject. Their timeline (1917-2017) is shown in figure-2.1. As most pre-2008 studies were in the domain of psychology and organisational-management, (Appendix-I), *review of studies from 2008 onwards only will be made below.*

FIGURE-2.1 Timeline of 100 years of attrition research
Important landmarks are shown in Italics with a star *

1917: Attrition research begins

1) *Journal of Applied Psychology (JAP) begins**.
2) JAP publishes first study on attrition

1920-1960

1) 1930 Reports describing operational-attrition emerge
2) 1940: *Study of correlation of attrition with demographic and psychological factors begins. **
3) *1958: first formal attrition model by March and Simons. **
4) 1959-60: research correlating selection-score and attrition published.
5) 1959-60: Exit-interviews to find reasons for attrition introduced.
6) 1959-60: Work on antecedents of attrition continues.

1970-1980

1) *1973: Theory of Met Expectations proposed by Porters and Steers **
2) *1977: Intermediate Linkage Model proposed by Mobley **
3) 1978-80: A Taxonomy of Determinants of Attrition developed by Price.
4) 1978-80 Mobley et al publish overview of attrition models published till date. *

1980-1990

1) 1986: Review of attrition models till 1986: Cotton and Tuttle. *
2) Complex causal model: Price and Muller. *
3) 'investment-model': Rasbult and Farrell.
4) Improvement and extension of Price-Muller model: Steers and Mowdy.
5) Hulin et al. addresses the role of job opportunity.

1990-2000

1) 1992: Mobley's 1977 model tested and validated by Hom et al.
1) 1994: *Lee and Mitchel: unfolding model **
2) 1996: Lee et al. test basic tenets of unfolding model.
3) 1998: Maertz and Campion: Review of attrition models up to 1998
4) 1998: Shaw et al.: Theorize and investigate attrition antecedents.

2000-2019

1) 2000-2004: Griffith et al: Meta-analysis of attrition antecedents.
2) *2001: Mitchell et al. proposed and tested a concept of called 'job-embeddedness'. **
3) 2001: Trevor tests March-Simon's model and new construct 'movement-capital'.
4) 2005-2009: Shaw et al test alternative attrition performance and relationships.
5) *2008: Holtom et al publish attrition and retention research **
6) 2010-2014: Hom et al review and expand attrition's conceptual domain.
7) *2017: 100 years of JAP. **

- The research up to 2008 (Appendix-I) showed that attrition is *a complex and dynamic process. The entire work can be summarised as efforts to develop so-called 'component-models', 'process -models' and efforts to synthesise them to derive a 'unified theory of attrition' which could not be obtained.*
- Meanwhile, by 1990s, advances in ML prompted some scientists to use ML based approaches like this study. However, the solid background gained by reviewing works mentioned in Appendix-I has helped this researcher a lot *to understand the context in which attrition-studies evolved. It also becomes handy while interpreting the results from ANN and stacked models which are traditionally considered difficult to interpret*.

2.2 Literature review (for studies from 2008 to 2019)

This literature review is divided into four parts in section 2.2.1.1 to 2.2.1.4 below:

2.2.1 Brief overview of five important studies based on data-science but *not* on ML

2.2.1.1

A Trevor, C. O., & Nyberg, A. J. (2008) [20] *examined hypothesis "whether downsizing predicts voluntary attrition rates"* using data from multiple industries. *Results supported hypothesis.*

2.2.1.2

Hom, P. W., Tsui, A. S., Wu, J. B., Lee, T. W., Zhang, A. Y., Fu, P. P., & Li, L. (2009) [11] hypothesised that EOR affects *"quit-propensity"* and "organizational-commitment" and "social-exchange" and "job-embeddedness" mediates this effect. Two studies were done and both *confirmed the hypothesis*.

2.2.1.3

Felps, W., Mitchell, T. R., Hekman, D. R., Lee, T. W., Holtom, B. C., & Harman, W. S. (2009) [6] did a study describing how attrition can spread like an epidemic in organization and how the co-workers' *"job-embeddedness" and "job-search-behaviour" play key roles in deciding why employees quit.*

2.2.1.4

Nyberg, A. (2010) [15] studied the relationship between *'performance* and *voluntary-attrition'* to test two contradictory views:

a) High-performers, if rewarded well, stay with employer.

b) High performers are more likely to attrit as they get more outside employment opportunity.

This research studied data of 12545 insurance-workers' over a 3-year period and showed that *the relation between performance and attrition is influenced by "interaction between pay-rise and unemployment-rate''*.

2.2.1.5

Chen, G., Ployhart, R., Thomas, H., Anderson, N., & Bliese, P. (2011) [2] proved that *"attrition intentions"* which cannot be explained by considering *"static levels of job-satisfaction"* can be explained if *"job-satisfaction is considered as a dynamic function"*.

2.2.2 A review of studies based on artificial neural network (ANN)

The studies above mainly used statistics and empirical methods. But, with the advent of ML, some researchers attempted ML-based methods to study attrition. Among these, few ANN based studies are as follows:

2.2.2.1

Kent A. Spackman (1992) [19] made one such early study. The study used a single layer feed-forward NN using logistic transformation which is equivalent to a LR model. The process of estimating coefficients in LR is equivalent to process of training neural-weights. Kent used BUPA dataset having 345 cases in class-ratio 200:145. The study used a NN containing 2/3/4/5 nodes in the intermediate layer and 10-fold cross-validation. *28.6% of weights were removed using likelihood. Reduced model had higher accuracy (72.1%) than original (71.5%).* Paper also *showed combining LR and ANN can be used for variable selection and predictive model building*.

2.2.2.2

Mark J Somers (1999) [18] in his study used two neural-network (NN) paradigms—'*MLP*' and 'LVQ' with two aims:

a) To check whether ML techniques were indeed superior to conventional methods for attrition-study.

b) If some new insights about attrition are obtained by the use of ML due to their ability to capture non-linear relationships. *The study argued that the presence of two distinct groups 'quitters' and 'stayers' makes the topic of attrition well suitable for study by NN*. In the study, primary data from 577 hospital employees was collected. The research used MLP and LVQ with 80:20 ratio as a *train: test* ratio. Results obtained by him are in TABLE-2.1, showing MLP giving the best overall result.

TABLE-2.1 Result: Somer's study

MODEL	% CORRECTLY CLASSIFIED		
	OVERALL	STAYERS	LEAVERS
LR	76%	99%	1%
MLP	88%	99%	44%
LVQ	84%	87%	77%

2.2.2.3

Andrew Quinn et al. in 2002 paper [16] applied MLP and LR independently on the same dataset. Data had 12 predictors and 536 cases. After listwise deletion for missing data, finally, 429 cases remained of which 246 were stayers and 97 were leavers. For MLP, Statsoft's 'NN Version 4.0B' was used while LR was run on 'SPSS 9.0'. The results of the study are in TABLE-2.2.

TABLE- 2.2 Result of Quinn's study

MODEL	**% CORRECTLY CLASSIFIED**		
	OVERALL	STAYERS	LEAVERS
MLP (TRAINING DATA N=343)	80%	80%	78%
MLP (TESTING DATA N=42)	60%	56%	70%
LR (N=343)	79%	92%	47%

Above results were *not fully consistent with Somer's study* described above. MLP offered just 1% better overall prediction-rate than LR, while in Somer's study, MLP outperformed LR by 12-14%. Also, in this study, LR could predict 'stayers' more accurately than MLP by 12%, while in Somer's model, MLP could predict stayers with 99% accuracy. However, in predicting ‘leavers’, just as in Somer’s study, MLP beat LR by a huge margin. The study gave following important outcomes.

a) *Rebalancing data or increasing number of neurons gives no better results by MLP*.
b) *NN results are not always superior***.** There can be complex attrition-behaviours and *it is better to use at least two methods*.
c) Though MLP models did no better than LR models in this case, MLP can be improved unlike LR models. In fact, on the same data, Schoech, Quinn, & Rycraft (2000) later proved the superiority of MLP over LR.

2.2.2.4

Mari Maisuradze (2017) [14] presented a Master’s thesis on attrition. In this research, two datasets were taken. One: IBM dataset and other Swedbank data. Several ML algorithms were used in the study of which RF performed the best in terms of accuracy (98.62%). Key takeaways from the thesis are:

a) Discussion on ‘data-driven’ Vs ‘user-driven’ approaches of analysis. The data-driven approach normally uses existing data while in user-driven approach data is collected/selected by domain experts. *One weakness of the user-driven approach is that some hidden patterns may remain undiscovered as only subsets of data are analysed*.
b) Discussion about data-types and which data-type is reliable and why. The study also explains variable-types for IBM dataset.
c) Discussion about the problems like outliers, missing-values, skewness, collinearity, high cardinality fields which affect data-validity and methods to treat them.
d) Description of types of ML models, model-evaluation techniques and an elaborate list of ML algorithms (Page 32, Fig-8) and description of RF, MLP and SVM in detail and also that of various ML tools for predictive analytics and their popularity.

e) Mean square validation score obtained by RF, MLP and SVM are respectively: 98.62%, 79.14%, 71.49%. So, RF gave the best result.

f) For IBM dataset important influencers were 'MonthlyIncome', 'Age', 'hourly rate', 'TotalWorkingYears', 'DistanceFromHome', 'PercentSalaryHike', 'marital status' and 'JobLevel', *all in correlation with 'OverTime'.*

2.2.2.5

Next is a much-cited paper by Randall S. Sexton et al. (2005) [17] , which, described an experiment with a Modified Genetic Algorithm (GA) "neural-network-simultaneous-algorithm (NNSOA)" to train the neural network (NN). NNSOA has two advantages: *One:* It decides an optimal number of hidden nodes automatically. *Two:* It adds an ability to identify relevant predictors to GA. The study used data of a small manufacturing company with 447 cases in which 35 were leavers and NNSOA trained NN in a 10-fold cross-validation experimental design. Multiple methods were used. Metrics obtained by them are in Table-2.3 showing NNSOA gave minimum errors. The study found *'salary', 'tenure' and 'full-time Vs part-time employment'* are most important attrition-predictors.

TABLE-2.3 Result of sexton's study

AVERAGE ERROR PERCENTAGES BY VARIOUS MODELS														
NNSOA			GA			NS			NW			DA		
Overall	Type I	Type II	Overall	Type I	Type II	Overall	Type I	Type II	Overall	Type I	Type II	Overall	Type I	Type II
0.68 %	0.25 %	5.83 %	7.31 %	0.25 %	92.1 %	2.27 %	0.49 %	24 %	7.77 %	0.00 %	100 %	29.2 %	30.1 %	14.8 %

2.2.2.6 Chin-Yuan Fan et al. (2012) [5] presented a novel predictor-model using 'Self-Organizing Maps' (SOM) and 'Back-Propagation-Network' (BPN) (SOM+BPN). Primary data was collected by questionnaires to predict trends in attrition rates of tech-professionals. There were 421 valid responses divided in train: test ratio of 385:36. the K-means clustering method was applied to cluster all data-sets, focussing on 28 variables. The research got clustering accuracy rate of 92.7%. *After obtaining the initial clustering groups from the SOM, method used BPN classification for all data-clustering*. Through two-phase SOM + BPN clustering, this approach successfully clustered all data into four groups *in descending order of 'tendency of attrition'*. Then these four samples were randomly chosen to test the accuracy of the hybrid model. The overall comparative accuracy obtained by the study is shown in TABLE-2.4

TABLE-2.4 Result of study of fan et al.

ACCURACY OF FORECASTING FOR K-MEANS, BPN OVERALL, SOM+BPN			
Method	K-Means	BPN	SOM+BPN
Accuracy	63.5%	87.2%	92%

The study identified *'lack of inner fidelity identification'*, *'leadership'* and '*management'* the main attrition-predictors.

2.2.2.7

ZEHRA ÖZGE KISAOG˘ LU, in 2014 thesis [23] used publicly available profiles of employees from the web and made *job-transition-graphs*. *Using features extracted from these graphs as data*, seven ML classification techniques including ANN were applied to predict attrition. The predictions were evaluated with accuracy, precision, recall and F1-score metrics and all models performed better than baseline-models with SVM being the top performer. Some unique features of this study were:

a) *A large data-set of 14000 employees was used*.

b) *Predictions were made for 1/2/3/4/5-year future windows*.

c) *Overall performance was not good for initial unbalanced data*. *Balancing the data-set improved various metrics*. *This is contradictory to the study of Andrew Quinn et al.* mentioned earlier. *This shows that one must be open to using balancing if needed.*

2.2.2.8

Next is a paper by HMN Yousaf (2016) [10]. Here Data-set of the company's US and Europe branches were taken from which three datasets were made. One for the US, one for Europe and one overall dataset. Data had 39 attributes and 10616 instances, but after pre-processing and balancing 8212 records were kept. Just 6, 12 and 8 attributes were kept for US, Europe and overall dataset respectively. Three ML techniques used were GLM, NN, RF). On original, imbalanced data-sets (96%-98% stayers), all 3 methods performed poorly. But after balancing data-set (75% stayers-25% leavers) by SMOTE, performance greatly improved. Results obtained are as in TABLE-2.5 and TABLE-2.6. They show that *for total data*, though RF had marginally high accuracy, its (sensitivity 61.78%) was less than NN (sensitivity 63.80%).

10-fold cv results TABLE -2.5

US

Model	Performance Measure	
	Accuracy	Kappa
RF	94.07%	83.52%
NNET	87.2%	62.05%
GLM	88.2%	66.17%

EU

Model	Performance Measure	
	Accuracy	Kappa
RF	86.57%	61.26%
NNET	86.52%	61.21%
GLM	84.48%	53.56%

TOTAL

Model	Performance Measure	
	Accuracy	Kappa
RF	86.79%	61.98%
NNET	86.17%	60.52%
GLM	83.76%	54.01%

repeated cv results TABLE-2.6

US

Model	Performance Measure			
	Accuracy	Kappa	Sensitivity	Specificity
RF	93.53%	82.03%	81.17%	97.64%
NNET	87.57%	58.82%	59.94%	96.78%
GLM	88.13%	66.27%	67.04%	95.16%

EU

Model	Performance Measure			
	Accuracy	Kappa	Sensitivity	Specificity
RF	86.64%	61.53%	54.66%	95.36%
NNET	86.36%	61.11%	61.84%	94.72%
GLM	84.76%	54.66%	52.89%	95.61%

TOTAL

Model	Performance Measure			
	Accuracy	Kappa	Sensitivity	Specificity
RF	86.65%	61.76%	61.78%	95.13%
NNET	86.32%	61.56%	63.80%	94.00%
GLM	83.69%	53.70%	57.18%	92.72%

2.2.2.9

Yue Zhao et al. (2019) presented a rigorous study [24] on Attrition-Prediction with ML: study gave an extensive review of earlier work and pointed out some of their limitations. For example:

a) Findings from earlier methods were *difficult to generalize* which was because HR-Data inherently contains *confidentiality*, *noise*, *inconsistency*, *missing values* and *imbalance*.

b) Previous studies focused on a narrow set of metrics (e.g. Accuracy) for measuring the performance of the model. But *accuracy is not much meaningful for imbalanced datasets.*

c) Efforts to improve interpretability by incorporating variable-significance can introduce bias *as such rankings were classifier-dependent*.

So, this study aimed at providing a comprehensive guideline for applying ML methods to attrition-problem. In research, two datasets were taken: One IBM dataset and another Bank dataset. For, increased rigour, 10 different datasets of *small*, *medium* and *large* size from these two were created. On each dataset, 10 supervised learning methods were applied like DT, RF, GB, XGB, LR, SVM, NN, LDA, NB and KNN and performance of each method was compared. ACC, Precision, Recall, AUC, ROC and F1-score were chosen as accuracy metrics**.** *Label-encoding was applied while using NN. No feature-selection or dimensionality - reduction was done initially.* The datasets and results are shown in Table-2.7

TABLE–2.7 Result of study of YUE ZHAO et al.

DATASET	GROUP	POPULATION SIZE	FEATURES	ATTRITION RATE	BEST ALGORITHM FOR METRICS SHOWN BELOW				
					ACC	PRC	RCL	F1	ROC
50_BANK	Small	50	19	0.2800	DT	SVM	DT	DT	RF
50_IBM	Small	50	31	0.1600	NN	NN	NN	NN	----
100_BANK	Small	100	19	0.2800	DT	XGB	GBT	XGB	XGB
100_IBM	Small	100	31	0.1600	LR	SVM	NB	NB	LR
500_BANK	Medium	500	19	0.2820	XGB	RF	XGB	XGB	GBT
500_IBM	Medium	500	31	0.1600	NN	LDA	NN	NN	NN
1000_BANK	Medium	1000	19	0.2830	XGB	RF	XGB	XGB	XGB
1500_IBM	Medium	1500	31	0.1612	LR	LDA	NN	NN	GBT
5000_BANK	Large	5000	19	0.2834	GBT	GBT	GBT	GBT	GBT
9000_BANK	Large	9000	19	0.2834	GBT	GBT	GBT	GBT	GBT

a) For small datasets, no algorithm could consistently outperform on all metrics.

b) For medium datasets, XGB performed best on Bank-Data and *NN performed best on IBM-Data*, respectively.

c) For both large datasets, XGB ranked highest.

d) The study also proved that *data-source or data-size do not affect classifier performance* and *grouping data into 10 subsets is valid approach*. Further, the study found *discrepancy in classifier-performance on small data-set by graphical and analytical methods* but the reason for it was explained by researchers. [Ref: Section 5.4 of research paper]. *Details of this discussion are avoided here for brevity.*

TABLE-2.8 Data-size does not affect ROC

Group	Median ROC	Mean ROC	STD DEV IN ROC
Small	0.8295	0.8052	0.1129
Medium	0.8052	0.7940	0.0883
Large	0.8629	0.8606	0.1077

e) The feature importance of predictors was obtained in the study using the best XGB model on 1000_Bank dataset. *GB, XGB and RF found the same three features as most significant*: *'last pay raise'*, *'job tenure'*, *'age'*.

2.2.2.10

Soni Umang, Singh Navjot, Swami Yashish, Deshwal Pankaj [21] did a study. The Train-Data and Test-Data had 5000 and 10 records, respectively. ANN architecture used was two hidden layers with 25 and 50 neurons, respectively. Back-propagation-algorithm (BPA) was used. A logistic sigmoid function was the activation function. ANFI is an algorithm based on neural network theory and it is used to train the fuzzy system. It mixes the least square method with BPA. Results in Table 2.9 show that both RMSE and MSE for both train and test data is less for ANN than ANFI. Also, ANN gave better sensitivity for train and test data.

TABLE-2.9 Comparison of ANN and ANFI models' results

DATASET	RMSE (less for ANN)		MSE (Less for ANN)		SENSITIVITY
	ANN	ANFI	ANN	ANFI	The sensitivity of ANN was also more than ANFI on both train-data and test-data.
TRAIN	0.10	0.19	0.01	0.03	
TEST	0.4	0.50	0.23	0.25	

2.2.2.11

Next document reviewed was, "*Journal of Statistical Software*" [1], which explains package '*NeuralNetTools*'. Neural-network models are often called '*black-box models*' due to their perceived poor interpretability. This package provides that interpretability. It's an *olden ()* function for variable-significance, *plotnet ()* for plotting neural-network itself and *lekprofile ()* for sensitivity analysis.

2.2.2.12

To learn more about '*neuralnet*' package a research paper by Fluke Gunther and Stephan Fritsch [7] [9] were reviewed. In addition, Standard CRAN-documentation on packages used in the modelling were reviewed.

2.2.3 A review of studies based on stacking

Next, a review of studies based on Stacking was made. This part is divided into two categories:

- Studies Describing the fundamentals of stacking. (2.2.3.1)
- Studies describing stacking actually applied to attrition Problem. (2.2.3.2)

2.2.3.1 Studies describing the fundamentals of stacking.

2.2.3.1.1

First is the research paper by Wolpert (1992) [22] , which for the first time introduced ‘stacked generalisation’ as a method for:

a) Minimizing the generalisation-error-rate of one or more generalisers and
b) A more sophisticated version of cross-validation.

The paper also presented two numerical experiments demonstrating how stacked generalisation improves the performance of a single generaliser and also proposes that “*for almost any real-world generalization problem, one should use some form of stacked generalisation to minimise the generalisation error-rate”*. It also gave other experimental evidence in the literature supporting stacking. In the end, the paper discussed some variations of stacked generalisation.

2.2.3.1.2

Next paper on fundamentals of stacking was published in 1999 by Kai Ming Ting and Ian H. Witten [13]. This paper addressed some issues in stacking as proposed by Wolpert. In Wolpert’s paper, two things were mentioned like voodoo magic or an ‘art’. E.g.

a) *“Which type of generalisers is suitable to derive the meta-classifier*?”.
b) “*What kind of attributes should be used as input to this meta-classifier?*”.

This study took two artificial datasets (‘Led 24’ and ‘Waveform’) and eight real-world datasets from UCI repositories. For two artificial datasets (training set) average error-rate of ten repetitions were considered. For eight real-world datasets, k-fold cross-validation was performed. The result was expressed as average error-rate of k-fold cross-validation. At base level C4.5, a decision tree learning algorithm; NB, a re-implementation of a Naive Bayesian classifier; and IB1, a variant of a lazy learning algorithm was used. As meta-classifier C4.5, IB1(using p = 21 nearest neighbours),1 NB, and a multi-response linear regression algorithm (MLR) were used one by one. *The stack with MLR as meta-classifier gave the best result*. The paper finds the following two conditions for stacked generalisations to work:

a) Meta-learner should not use class predictions but class probabilities from base-learners.

b) Among the four algorithms used as meta-classifier in research, only MLR was suitable.

2.2.3.1.3

Next work was by Saso Džeroski et al. [4].Researchers evaluated several state-of-the-art methods (*of that time*) for making a stack of diverse classifiers and showed that "stack performed (at best) equal to selecting the best classifier from the ensemble". They suggested an improvement in prevailing method [*stacking with PD and MLR*)] and advocated use of *'multi-response model trees'* to learn at the meta-level and proved that *it performs better than stacking with MLR and better than selecting the best base classifier too*.

2.2.3.1.4

Funda Güneş, Russ Wolfinger, and Pei-Yi Tan of SAS Institute Inc. presented an excellent paper in 2017 on a stacked ensemble [8]. Following are key take-aways from it:

1. Stacking *not only improves the prediction-accuracy but also improves generalizability by averaging out the noise for different models*.
2. Stacking may have extra overheads *in training many models and in use of cross-validation to avoid overfitting*.
3. The paper quotes Dietterich (2000) that "*The necessary and sufficient condition for a meta-model to be more accurate than any of its base-classifiers is that the base-classifiers are accurate and diverse*".
4. *Overfitting* and *leakage as the two most important problems linked with stacking* and the paper advocates use of *cross-validation*, *regularisation* and *bagging* to tackle them.

5. The approach of combining *weak* learners is outdated and modern research indicates using '*strong but diverse classifiers*' as base-classifiers. *The word 'diverse' here not only means using many algorithms but also multiple subsets of data by bagging or bootstrapping*.
6. The paper suggests 3 techniques for avoiding leakage and finally discusses some advanced methods for stacking.

2.2.3.2 Studies describing the actual attrition problem solved by stacking.,

In this category, *only two studies could be found*.

2.2.3.2.1

First was MSc Thesis by Divyang Jain at the National College of Ireland (2017) [12]. The thesis describes all ensemble methods like stacking, bagging and boosting while using CRISP-DM framework on the problem of attrition. The dataset used was IBM dataset used in this research.

Table 2.10 Results of Divyang Jain Study

ALGORITHM	SENSITIVITY	SPECIFICITY	FPR	FNR	ERROR	ACCURACY
Boosting						
Adaptive Boosting	85.93%	90.68%	9.31%	14.06%	11.1%	88.8%
Gradient Boosting	83.5%	90.19%	9.80%	16.40%	12.3%	87.6%
Bagging						
Random Forest	85.9%	87.25%	12.74%	14.07%	13.25%	86.74%
Stacking						
SVM	77.53%	90.1%	9.83%	22.65%	14.75%	85.24%
GLM	77.34%	90%	9.38%	22.66%	14.75%	85.54%
Decision Trees	81.25%	84.80%	15.1%	18.75%	16.56%	83.43%
KNN	70.41%	88.97%	11.03%	29.59%	16.66%	83.34%

The algorithms used with result-metrics obtained are shown in Table 2.10, clearly showing that all ensemble methods give good accuracy, sensitivity and specificity. The study found five most significant variables as: *'JobLevel', 'StockOptionLevel', 'job satisfaction', 'Relationship Satisfaction' and 'OverTime'*.

2.3.3.2.2

Next work studying attrition by stacking was by Deep Sanghavi et al. (2018) [3]. This study also used IBM dataset and Adaboost and SVM as base classifiers and decision tree as meta-classifier to create a stacked model. The results obtained by them once again confirmed that the stacked model outperforms all individual classifiers with an accuracy of 90.65%.

2.3.3.2.3 CRAN-documentations for packages required for stacking were reviewed.

2.3 RESEARCH-GAPS IDENTIFIED. HOW THIS WORK FILLS THEM UP?

Thus, Attrition-prediction is becoming increasingly data-driven and applying ANN this problem is the latest trend. In spite of reviewing such extensive literature, some research-gaps could be identified in attrition related research.

1) This problem has been approached very little by stacking and neural-network.
2) No study so far does comparative analysis of two different ANN algorithms and two different stacking algorithms.
3) The most important question that gives actionable insights for employee-retention i.e. 'Variable-Significance' appears inadequately addressed by past studies.
4) The interpretability and explainability of previous models appear wanting. *E.g. given data of a particular single employee, how to find if employee will attrit or not and how to explain which predictors lead to that decision to others?* This research uses the latest techniques and packages (*from year 2018,2019*) for better interpretability.

This study tries to fill these gaps. Four separate models were made, two were based on NNs: one using *'nnet'* and second using *'neuralnet'* package. Two other stacked models were built one by package '*caretEnsemble*' and second by '*H2o'*. Extensive EDA was also performed to do visual-analytics for finding some patterns in the data about how attrition is affected by each individual variable. For explainability "DALEX" package was used.

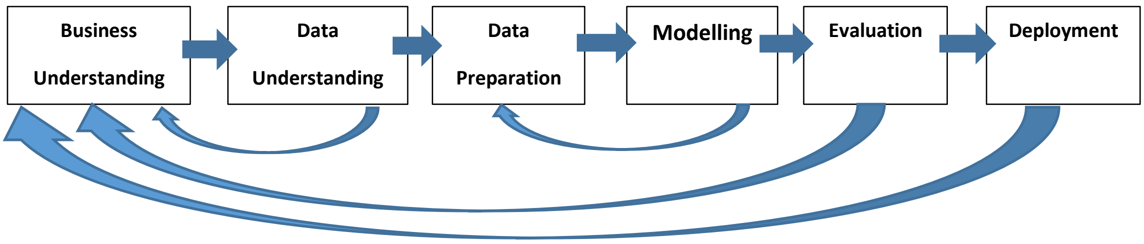
The value of this research lies in the levels of performance achieved, the *novelty of techniques used, depth of analysis of the problem from all angles and better interpretability and explainability not seen so far in earlier studies*.

In the next chapter: 3 'Research Methodology' will be discussed.

CHAPTER 3 RESEARCH METHODOLOGY

The basic methodology used will be the CRISP-DM (Cross Industry Standard Process for Data Mining) framework. It is an iterative process shown generically, in Figure-3.1.

FIGURE-3.1 GENERIC CRISP-DM FRAMEWORK



Below, each stage of CRISP-DM *in the context of current research* is described.

3.1 Business Understanding: [PHASE -1 OF CRISP-DM]

3.1.1 Determining Business Objective:

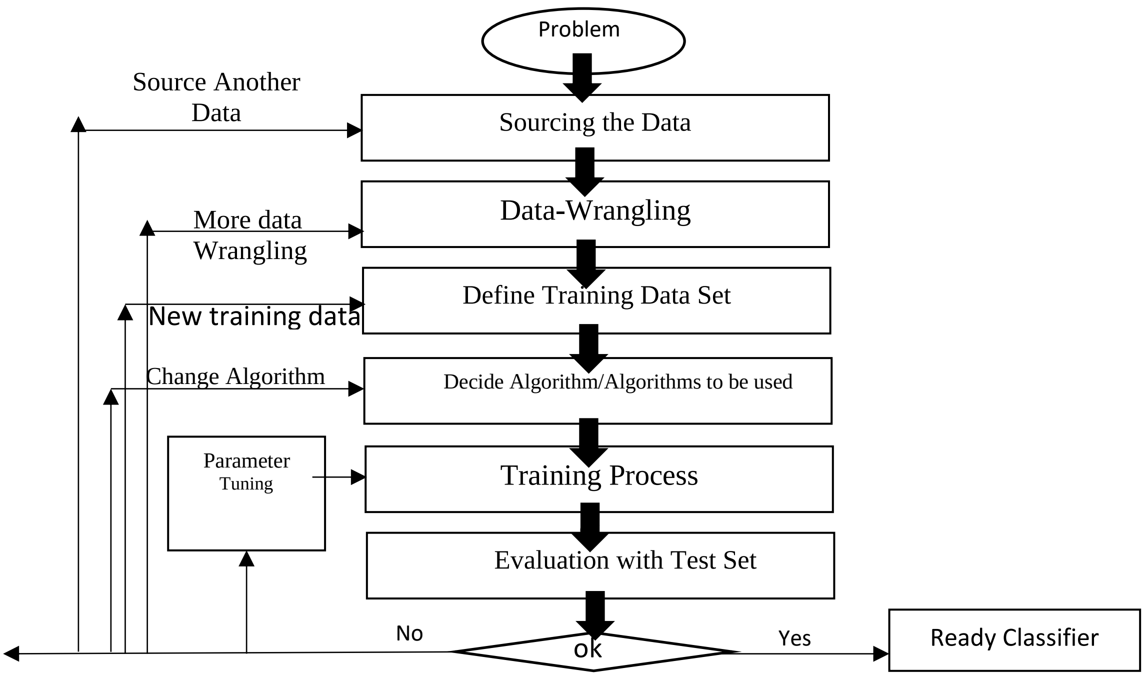
3.1.1.1 Background:

Attrition has become a global, urgent and costly problem which is only expected to worsen in the decade of 2020-2030 as described in chapter 1: Introduction.

3.1.1.2 Business Objective and its conversion to Data science objective:

Business objective is to predict voluntary attrition. The ML objective is to create *ML-based models which can predict the probability of attrition for employees. So, this is a binary classification problem*. A generic iterative binary classifier is shown in Figure-3.2.

FIGURE-3.2 A Generic binary classifier



3.1.1.3 Research success criteria or Research Objectives:

- **One:** Four models, two ANN-based and two stacking based must be created and detailed visual-analytics must be done for extra validation of variable-significance found by ML models.
- **Two:** F1_score and sensitivity of all models should be around 90%.
- **Three:** *the relative significance of predictors* for attrition must be obtained for each model.
- **Four:** At least one model must give explainability meaning that, given data of a single employee how to know his/her probability of quitting and *how to explain, to others, what were contributions of individual predictors to cumulative-probability for that decision.*

3.1.2 Accessing Situation:

3.1.2.1 Inventory of resources:

DELL Laptop, with 300 GB HDD, Intel i5, 64 GB RAM and AMD GPU. Software is 'R 3.6.0' and 'R studio 1.2.5001'. OS is Ubuntu 16.4 LTS and Windows 10 with the latest JAVA installed. The assumption is that this hardware will be able to handle the processing requirement. In terms of time, 28 weeks were allotted for the project and a Gantt-chart is presented in Figure-3.4 on the next page.

3.1.2.2 Risk and contingency plan

TABLE–3.1 Risk and contingency plan

	Risks or ethical/legal issues	Contingency plan
1	Computing resources may be inadequate as stacking and ANN both need high resources.	With an alphanumeric, structured dataset and simple binary classification problem, this analysis can be handled by available hardware. Further, cloud-platform can be used if needed.
2	Predictive algorithms in general and ANN, in particular, do not work well with factors. Same is the case with ordinal variables as ANNs confuse them as ratio variables.	Proper encoding of categorical variables and applying scaling to numeric variables should take care of this. Ordinal variables are already encoded in the given data.
3	Dataset is moderately imbalanced.	ANN and Stacking can handle this moderate imbalance.
4	Using employees' data can create privacy and legal/ethical issues.	IBM data is synthetic and ODbL license data, So, these issues do not arise. *However, anyone who uses these models on others' data must take informed, explicit consent.*
5	Both NN and Stacking take more time in modelling. So, the project may be delayed.	Research is currently as per schedule in Gantt Chart. So, completion should not be an issue.

3.1.2.3 Cost-benefit analysis:

Hardware depreciation cost over the lifetime of project and data scientist's charge, based on his hourly rate, are main costs. For benefits' calculation, attriting employees' pay, seniority etc. must be considered for calculating cost saved by retaining them.

3.1.3 Determine Data-Mining Goals:

3.1.3.1: About Dataset:

Data-set is clean, static, structured. Further, both ANN and stack require less data preparation, feature selection, etc. compared to conventional methods. So, the goal of data-mining will be to do just enough steps required by the respective algorithm and feed data into it. *Extensive EDA is done just to understand data in detail and find useful patterns*.

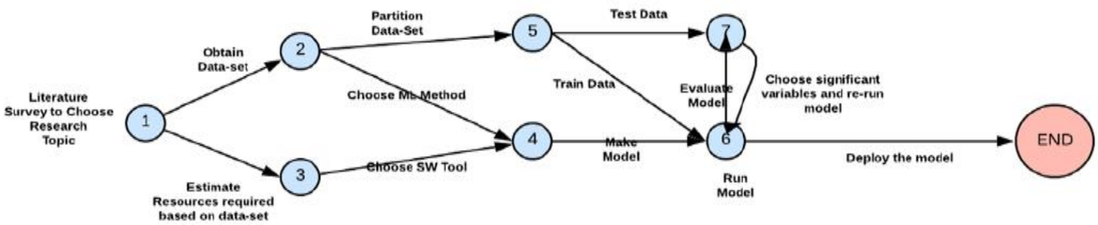
3.1.3.2 Success criteria for data-mining: Success criteria for data mining is that data should be well prepared for the respective algorithm and for visual analytics.

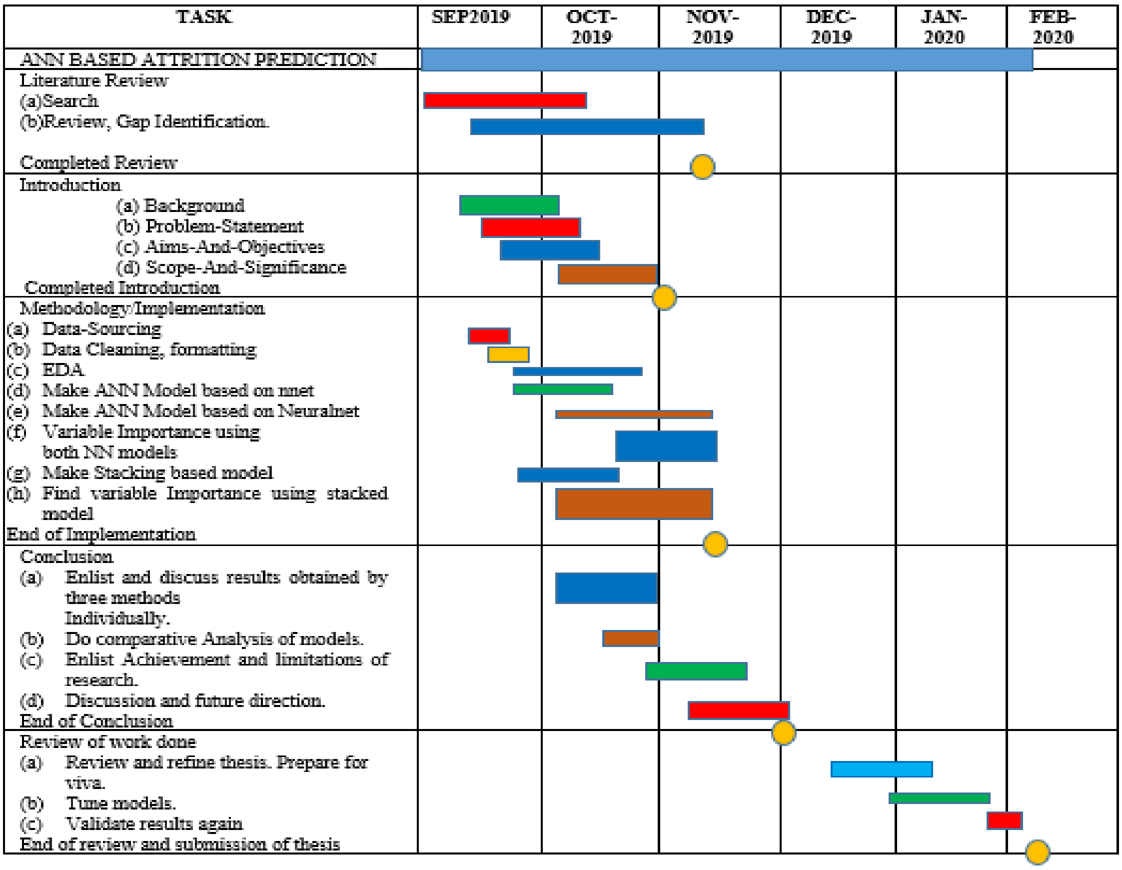
3.1.4 Produce Project Plan:

3.1.4.1 Project Plan:

PERT and GANTT charts for the research project are shown in Figure-3.3 and 3.4

FIGURE- 3.3 Pert chart and FIGURE-3.4 Gantt chart





3.1.4.2 Tools and Techniques

Tools used were MSWORDS for Gantt and PERT chart and R-studio IDE for coding. The technique used was *making two neural network classifiers*: *first based on 'neuralnet' and second based on 'nnet' package*. Then *two stacked classifiers were made one using caretEnsemble () [('C5.0', 'NB', 'GLM', 'KNN', 'SVMRadial')] as base classifiers and GLM and RF, one by one, as meta classifier] and second using H2o package ['GLM', 'GBM', 'RF', 'deeplearning' as four base models and again using 'deeplearning' as meta classifier].*

3.2 Data Understanding: [PHASE -2 OF CRISP-DM]

3.2.1 Data-Sourcing:

Data set is a fictional dataset from IBM with ODbl license from the source [79-A]. Data itself is currently hosted at [79-B].

3.2.2 Data Description, Data Exploration and Verify Data Quality:

the dataset has the following characteristics:

A. Dataset is structured, tabular with 1470 records, 35 attributes of which 9 are 'factor' and 26 are 'integer' data type. 1st column's name was set 'Age' from 'I. Age'.

B. The meaning of attributes is listed in Table 3.2:

TABLE – 3.2 Meaning of names of all variables including ordinal variables like Education

Sr No	Attribute	Explanation
1	Age	Age of Employee
2	Attrition	Whether the employee will quit or not ("Yes", "No")
3	business travel	Frequency of business-related travel that employees will make ('Nontravel', 'Frequently', 'Rarely')
4	daily rate	The daily rate of employee
5	Department	Department to which employee belongs (HR, R&D, Sales)
6	DistanceFromHome	The distance of working place from an employee's home
7	Education	Education level: 1 'Below College' 2 'College' 3 'Bachelor' 4 'Master' 5 'Doctor'
8	EducationField	In which branch has employed been educated (Life Sciences, Medical, Marketing, Technical Degree, Other, Human Resources)
9	EmployeeCount	Count Of employee per that record (row)
10	employee number	A kind of unique employee identifier
11	EnvironmentSatisfaction	Rating of Employee's satisfaction with work environment ('1','2','3','4': 1 'Low' 2 'Medium' 3 'High' 4 'Very High')
12	Gender	Sex (Male or Female)
13	hourly rate	The hourly rate charged by employee (maybe consultants)
14	JobInvolvement	Rating for employee's JobInvolvement ('1','2','3','4': 1 'Low' 2 'Medium' 3 'High' 4 'Very High')
15	JobLevel	Rating for Employee's JobLevel ('1' to '5')
16	job role	Employee's Role in Company

		(‘Healthcare’, ’human resource’,’ LaboratoryTechnician’, ‘Manager’, ‘ManufacturingDirector’, ‘ResearchDirector’, ‘ResearchScientist’, SalesExecutive’, ‘Sales Representative’)
17	job satisfaction	Ranking of how satisfied is an employee with his job (‘1’,’2’,’3’,’4’: 1 'Low' 2 'Medium' 3 'High' 4 'Very High’)
18	marital status	Whether an employee is unmarried, married or divorced
19	monthly income	Monthly income of the employee
20	monthly rate	The monthly rate charged by employee (maybe part-timers)
21	NumCompaniesWorked	Number of companies in which employee has worked prior to joining this one (Varies from ‘0’ to ‘9’)
22	Over18	Whether an employee is over18 years of age. All are ‘Y’ meaning ‘Yes’
23	OverTime	Whether employee did overtime or not (‘Yes’, ‘No’)
24	PercentSalaryHike	Last Percentage hike in employee’s salary.
25	performance rating	Performance rating of employee (1 'Low' 2 'Good' 3 'Excellent' 4 'Outstanding')
26	RelationshipSatisfaction	RelationshipSatisfaction of employee (1 'Low' 2 'Medium' 3 'High' 4 'Very High')
27	StandardHours	Hours worked by the employee (same 80 for all)
28	StockOptionLevel	Rating of how many stocks of the company the employee has? (0,1,2,3)
29	TotalWorkingYears	No of years employee has worked before joining the current company
30	TrainingTimesLastYear	How many trainings did employee undergo last year? (0,1,2,3,4,5,6)
31	work-life balance	Rating of work-life balance for employee (1,2,3,4 for bad, good, better, best)
32	YearsAtCompany	No of years employee has spent at current company
33	YearsInCurrentRole	No of years employee has spent in current role
34	YearsSinceLastPromotion	No of years employee has spent since last promotion
35	YearsWithCurrentManager	No of years employee has spent with his current manager

C. By using ‘*sapply ()’* and ‘*unique ()’* functions on dataset ‘*EmployeeCount*’, ‘*Over18*’, ‘*StandardHours*’ are found to be single-valued with no predictive significance. Similarly, ‘*EmployeeNumber*’ is just ID (as it has all unique values though not in sequence from 1 to 1470). These four columns were dropped giving a 1470 * 31 dataframe. Also, sum (is.na ()) and ‘sum (is. null ())’ are ‘0’. So, no ‘NA’ or ‘NULL’ was present. sum ((duplicated(data)) = 0. So, no duplicate records were found and all rows were kept.

D. Output of table(data$Attrition) command reveals that from 1470 employees, 237 attrited (16.12%) while 1233 did not (83.88%), indicating imbalanced data. *‘Accuracy’ is not a suitable metric for imbalanced data. So, ‘F1-score’ and ‘sensitivity’ were chosen for comparing model-performance.*

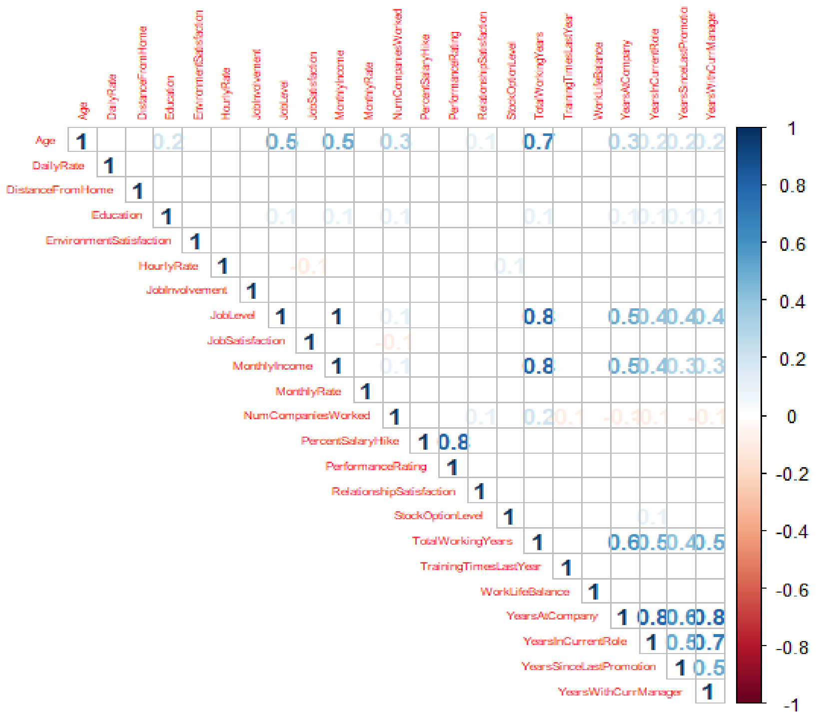
E. Function *dfSummary(data),* of the package ‘*summarytools*’, was used for data exploration. Its output is shown in APPENDIX-III. *This function does almost complete univariate analysis and produces a very presentable and informative output which can be redirected to R-studio’s viewer or even web browser.*

For numeric variables, the output gives data-type, mean and standard deviation, min, max and median values, Inter Quartile Range. number of distinct values, a barplot, number of valid records and % of missing values *in each column*. *Even for continuous variables, the plot shows histogram after doing binning automatically*. I*t can be seen that not many variables follow Gaussian normality*.

For factor variables, the output *gives* levels and the number of distinct values for each level, histogram for each level (category), number of valid records, and % of missing values *in each column*.

F. A Correlation-Matrix was created using "*cor()"* function and its plot was obtained using "*corrplot()"* function.(Fig-3.5), 5 pairs with high correlation (0.8) were seen: '*JobLevel-TotalWorkingYears*', '*MonthlyIncome-TotalWorkingYears*', '*PercentSalaryHike-PerformanceRating*', '*YearsAtCompany-YearsInCurrentRole'*, '*YearsAtCompany-YearsWithCurrManager*'.

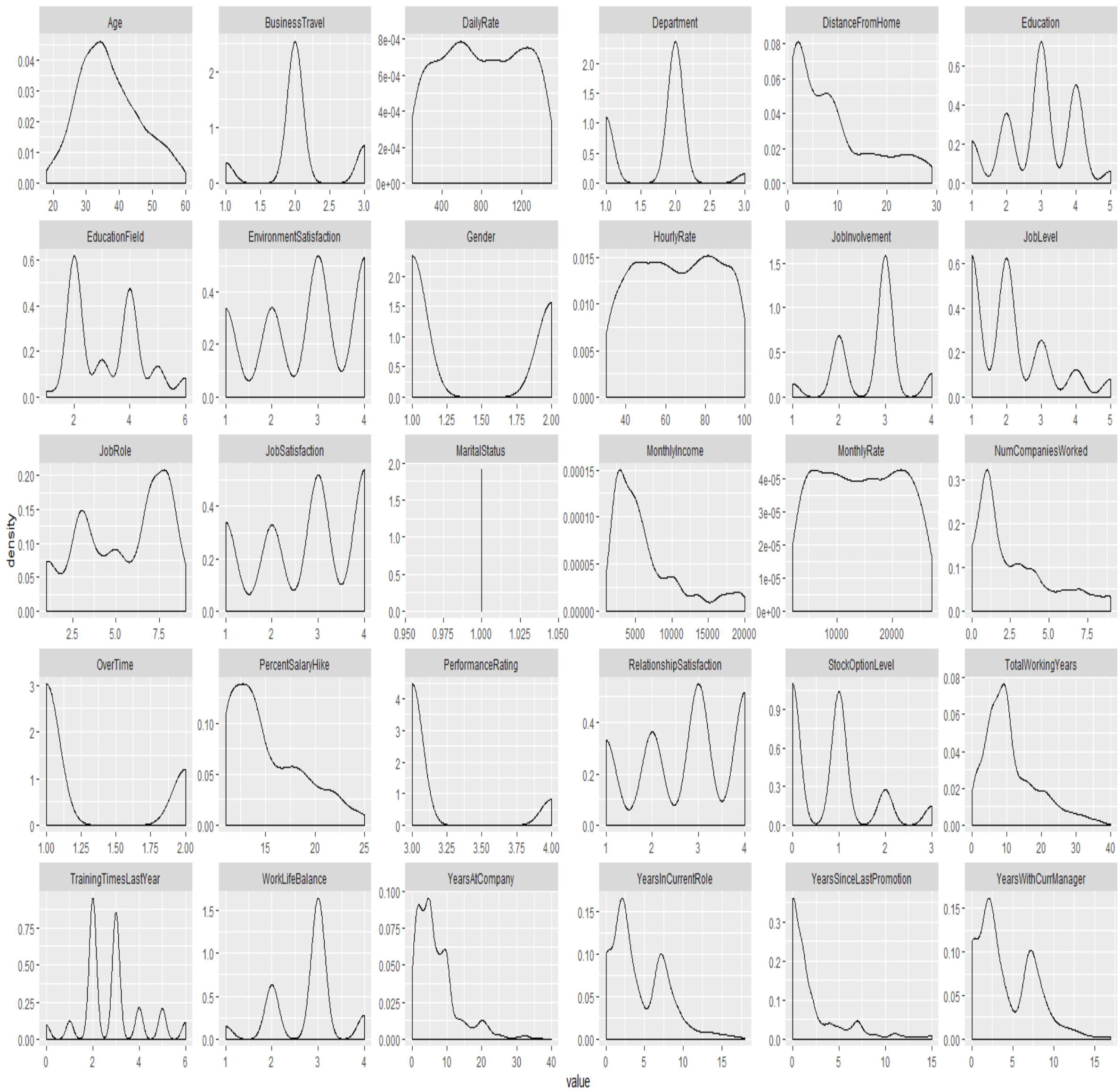
FIGURE-3.5 Correlation-matrix-plot



G. Next, label encoding of categorical variables to convert them into integer/numeric was done. Response variable 'Attrition' was also converted to numeric '1' and '0' for 'Yes' and ‘No’. This gave as a totally numeric data frame, which was scaled (except response variable column).

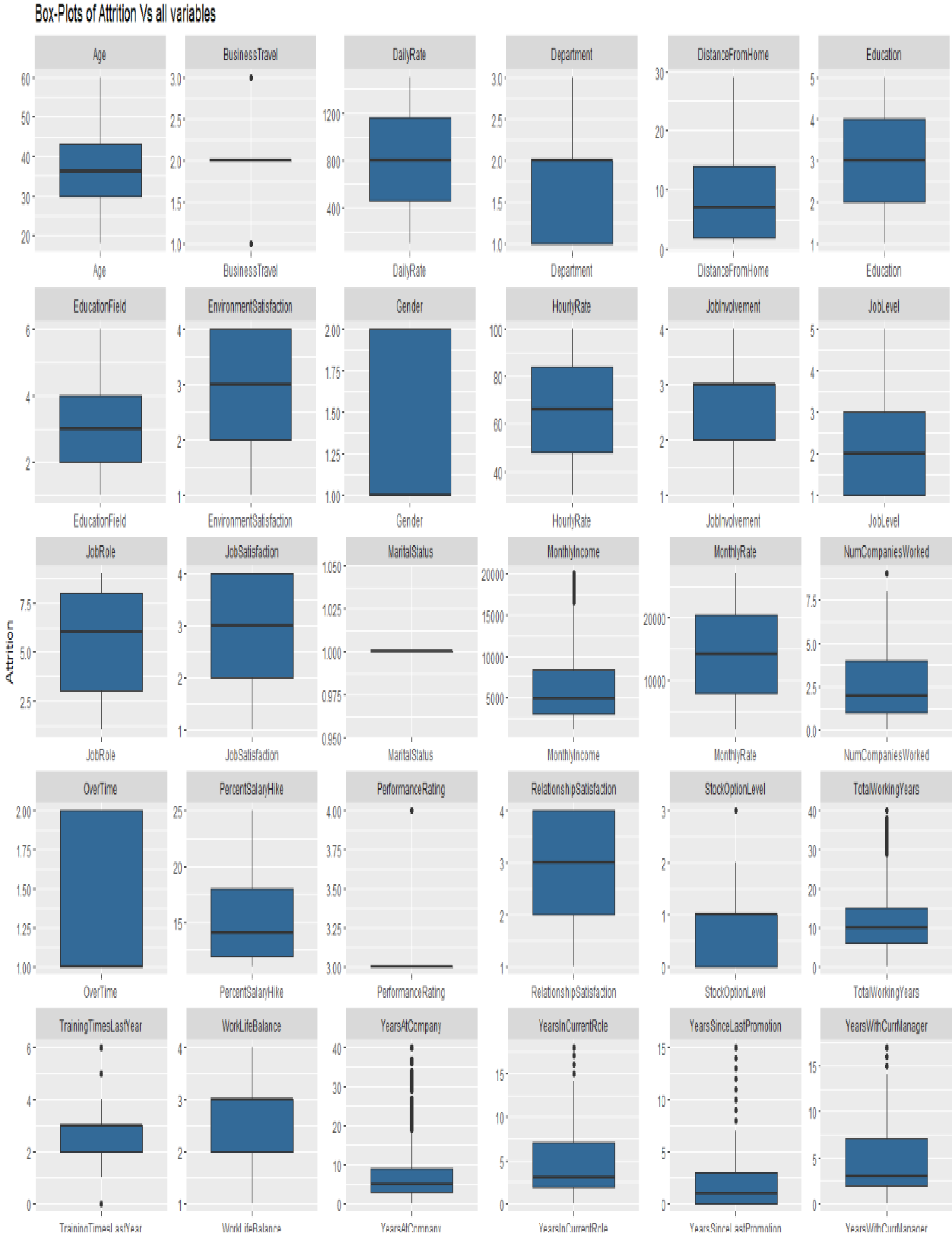
H. Density plots for *all* predictors were plotted as shown below. From this, it is clear that many predictors (e.g. DailyRate, JobRole, PerformanceRating, HourlyRate, MonthlyRate, YearsAtCompany, Gender, MonthlyIncome. *Etc.*) *do not follow Gaussian normality*.

FIGURE-3.6 Density plots for all variables



I. To know the presence of outliers, box-plots of *all* predictors were plotted. These box-plots, shown below, indicate the presence of outliers in some variables (e.g. MonthlyIncome, NumCompaniesWorked, YearsAtCompany, YearsSinceLastPromotion etc.).

FIGURE-3.7 Box-plots for all variables



J. To study the effect of 30 predictors on ‘Attrition’, a subset ‘attrited’ from given data was made consisting of only cases where ‘Attrition’ has a value ‘Yes’. Thus, now for EDA, two datasets *‘attrited’* and *‘data’* were used. *Then graphs of Attrition Vs each predictor were plotted for both original and small dataframe using ggplot ().* Scatter plots or tables (by binning) were obtained for numeric variables. Histograms were obtained for categorical/ordinal variables. Corresponding graphs, tables and how the pattern was identified are described below. *Left side plots* are for ‘attrited’ data frame showing *the effect of the predictor on attrition on an absolute basis*. *Right side plots* show *the effect of the predictor on attrition on a proportionate basis. ‘positive predictor’ means it increases the probability of attrition and vice versa.* As an example, only four plots, two scatterplots for continuous variables ‘Age’ and ‘DistanceFromHome’ (along with their binning table) are shown and also histograms for two categorical variables ‘BusinessTravel’ and ‘Department’ are shown here. However, in APPENDIX-II all variable’s plot can be seen. *Here only patterns identified will be listed for all variables.*

1. Attrition Vs Age:

FIG-3.8

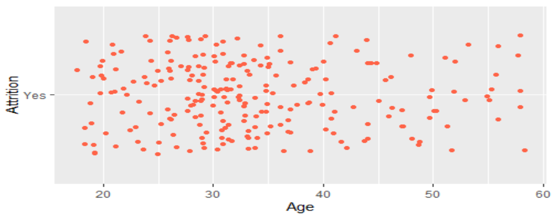


FIG- 3.9

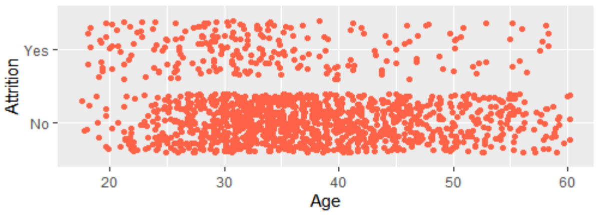


TABLE 3.3 AGE BINS

Age bin	18-24	24-30	30-36	36-42	42-48	48-54	54-60
% in given data	6.05	19.66	28.03	19.93	12.38	8.71	4.69
% in ‘attrited’	14.35	26.16	27.85	11.39	8.02	5.91	4.64

Age is negative predictor meaning that, as the age of employee increases, the probability of quitting decreases. This is clear from table 3.3*;* 68% quitters are below 36 years of age.

2. Attrition Vs BusinessTravel:

FIG-3.10

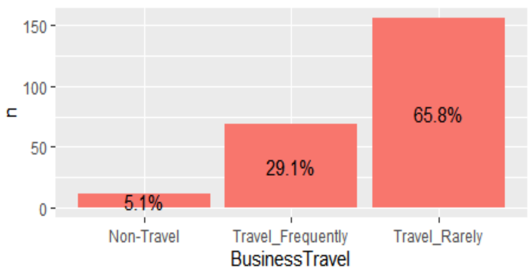
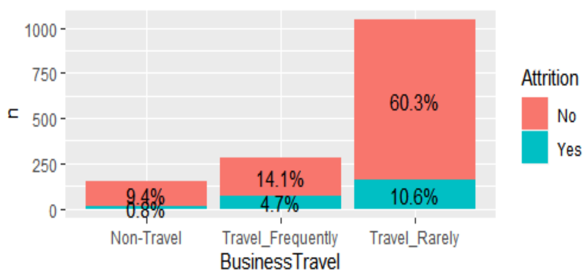


FIG- 3.11



business travel is a positive predictor. Employees who travel more frequently also have more probability of attrition. Proportionately 25% [(4.7/ (4.7+14.1))] of frequent-travellers attrit.

3. Attrition Vs. daily rate:

the daily rate is a negative predictor. DailyRate-less than 900 causes attrition in 64% sample. As DailyRate increases attrition decreases.

4. Attrition Vs Department:

FIG-3.12

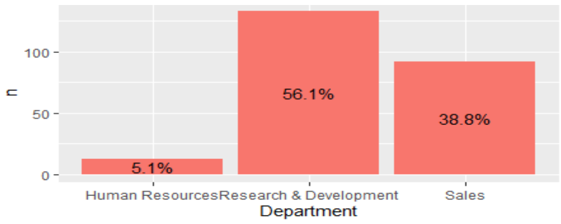
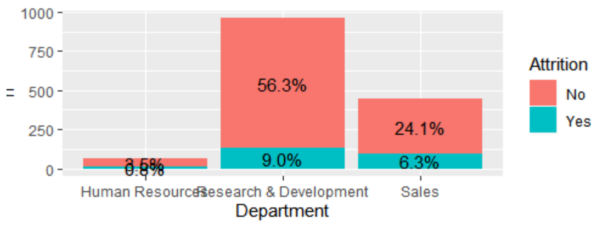


FIG-3.13



Department_Sales is positive predictor followed by department R&D. Proportionately Department_Sales has 20.7% attrition (6.3 / (6.3+24.1)), while R&D has ~14%.*attrition (9/(9+56.3)).*

5. Attrition Vs DistanceFromHome:

TABLE 3.4 DISTANCE BINS

Distance Bin	0-6	6-12	12-18	18-24	24-30
% in given data	47.01	26.12	9.80	9.46	7.62
% in ‘attrited’	39.66	25.32	13.08	13.92	8.02

FIG-3.14

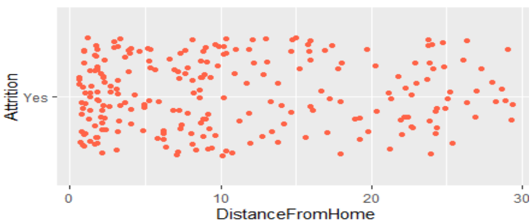
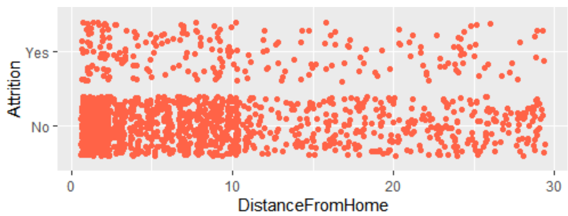


FIG-3.15



DistanceFromHome is a positive predictor. On absolute-basis, 65% attritions are in less than 12 km segment (from the table above). However, on a proportionate basis, attrition tendency is more for employees who live more than 12 km from. As seen from more % in attrited compared to % in given data for rightmost 3 columns.

6. <u>Attrition Vs Education:</u> (*1 'Below College' 2 'College, 3 'Bachelor' 4 'Master' 5 'Doctor')*

Education '5' is a positive predictor. Doctorate employees have the greatest attrition tendency. The overall trend is mixed. On absolute basis 'Bachelors' ['3'] attrit the most while Doctorates ['5'] attrit the least. But proportionately, % attritions for Education '1','2','3','4' and *'5(doctorates)'* are respectively 18%, 15.6%, 17.2%, 14.4%, *34.4%*.

7. <u>Attrition Vs EducationField:</u>

('HumanResources', 'Life-Sciences', 'Marketing', 'Medical', 'Other', 'Technical Degree')

HR, Marketing, TechnicalDegree are positive predictors. Although on an absolute basis, the highest quitters were from EducationField 'LifeSciences', 'Medical' and 'TechnicalDegree'. Proportionately the % quitters from 'HR', 'LifeScience', 'marketing', 'medical', 'other' and 'Technical degree' are respectively 29.4%, 14.7%, 22.22%, 13.6%, 12.72%, 24.44%.

8. <u>Attrition on Vs EnvironmentSatisfaction:</u> *(1 'Low' 2 'Medium' 3 'High' 4 'Very High')*

EnvironmentSatisfaction is a negative predictor. The highest number of quitters had low ('1') environment-satisfaction on an absolute and proportionate basis. Proportionately, *25.4%, 14.9%, 13.6% and 13.5% quitters* are seen in *Satisfaction levels '1','2','3','4'.*

9. <u>Attrition Vs Gender</u>

Gender male is a weak positive predictor. Higher attrition in males compared to female on an absolute basis. Proportionately too, 17% of males quit while 14.75% of females quit.

10. <u>Attrition Vs HourlyRate:</u>

hourly rate is a mild negative predictor. Employees with HourlyRate <70 (58% quitters) have higher attrition on an absolute basis. Proportionately too, *the* trend is the same as shown by table in APPENDIX-II.

11. <u>Attrition Vs JobInvolvement</u> *('1' Low '2 'Medium' 3 'High' 4 'Very High')*

JobInvolvement is a negative predictor. On a proportionate basis, 'Low JobInvolvement' had maximum quitters and 'Very High JobInvolvement' had minimum quitters, ranks

‘1’,’2’,’3’,’4’ had 33.9%, 18.85%, 14.4% and 9.1% quitters. On absolute-basis maximum quitters had high ‘3’ job involvement followed by medium ‘2’.

12. Attrition Vs JobLevel:

JobLevel is a negative predictor. Proportionately for job-levels ‘1’,’2’,’3’,’4’ and ‘5’ the percentage of quitters are 26.3%, 9.6%, 14.7%, 4.1% and 6.4%. So, junior (level ‘1’) and middle (level ‘3’) had maximum attrition. On an absolute basis, junior most (‘1’,’2’) employees have 82% quitters.

13. Attrition Vs JobRole:

(‘Healthcare’, ’HumanResources’, ’LaboratoryTechnician’, ‘Manager’, ‘Manufacturing-Director’, ‘Research-Director’, ‘ResearchScientist’, ‘SalesExecutive’, ‘Sales Representative’)

Proportionately maximum quitters have JobRole ‘SalesRepresentatives’, ‘LabTechnicians’, and ‘HR’. While ‘ResearchDirector’ has the least attrition followed by ‘manager’ and ‘ResearchScientist’. On an absolute basis, maximum quitters are ‘LabTechnicians’ followed by ‘SalesExecutives’ followed by ‘ResearchScientist’.

14. Attrition Vs JobSatisfaction: (1 'Low' 2 'Medium' 3 'High' 4 'Very High')

job satisfaction is a negative predictor. Proportionately for ‘1’,’2’,’3’,’4’ quitter % are 22.8%, 16.3%, 16.6% and 11.2%.

15. Attrition Vs marital status

‘MaritalStatus Single’ is a strong predictor to attrition. On an absolute basis, maximum quitters are ‘Single’ followed by ‘Married’ followed by ‘Divorced’. On the proportionate basis too, the % quitters for the same categories are respectively 25.6%, 12.4% and 9.9%.

16. Attrition Vs MonthlyIncome:

Monthly income is a negative predictor. Maximum leavers had a monthly income less than 5000. 86% of quitters had less than 9000 income. Proportionately in bins 1000-5000, 5000-9000, 9000-13000, 13000-17000, the percentage quitters are 57.4%, 39.01%, 51.1%, 22%.

17. Attrition Vs MonthlyRate:

With MonthlyRate 'attrition' mildly increases. the monthly rate is not a strong attrition-predictor on an absolute basis.

18. *AttritionVsNumCompaniesWorked:* ('0', '1', '2', '3', '4', '5', '6', '7', '8', '9')

NumCompaniesWorked has a mixed effect. On an absolute basis, maximum no: of quitters leave from their very first company. Proportionately % quitters in various groups are 11.9%, 18.9%, 11.1%, 10.9%, 10.1%, 23.8%, 21.3%, 17.28%, 9%,17.8%. *Thus, those in their 6th or 7th company have maximum quit-propensity (maybe retirees).*

19. *Attrition Vs OverTime:*

'*OverTime-Yes' is a strong positive predictor on both absolute and proportionate basis*. % quitters with 'OverTime-Yes' (30.4%) are thrice those with 'Overtime-No' (10.4%).

20. *Attrition Vs PercentSalaryHike*

Trend is mixed on proportionate basis and percent salary hike is a negative predictor on absolute basis. On absolute basis, as PercentSalaryHike increases, count of quitters decreases. Proportionately the % of quitters are 19.6%, 16.4%, 16.2%, 11.8%, 17.4%, 18.5%, 17.8%, 14.75%, 11.5%, 13.1%, 9.3%, 21%, 21%, 28.5%, 7.6% in 15 bars on the right graph.

21. *Attrition Vs Performance Rating:* 1 'Low' 2 'Good' 3 'Excellent' 4 'Outstanding'

PerformanceRating is not strong attrition-predictor. Proportionately 16%, quitters have rating '3' while 16.2% have rating '4'. Rating '3' employees have quit in far larger number on an absolute basis. Data has records for only for ratings '3' and '4'.

22. *Attrition Vs RelationshipSatisfaction:* (1 'Low', 2 'Medium', 3 'High', 4 'Very High')

Low RelationshipSatisfaction ('1') is positive attrition-predictor. Proportionately, in four categories, % quitters are respectively 20.7%, 14.9%, 15.4%, 14.9%.

23. *Attrition Vs StockOptionLevel:* ('0','1','2','3')

StockOptionLevel is negative attrition-predictor. Employees with stock options level ‘0’ or ‘1’ make ≈89% of quitters on an absolute basis. Proportionately, those with StockOptionLevel ‘0’, ‘1’,’2’,’3’ have 24.5%, 9.4%, 7.47%, 17.2% quitters.

24. *Attrition Vs TotalWorkingYears:*

TotalWorkingYears is a negative predictor. Attrition tendency is strongest on an absolute basis in first 8 working-years, moderates in 8-20 years and minimal attrition is seen after 19 years.

25. Attrition Vs TrainingTimesLastYear: (0,1,2,3,4,5,6)

No training (‘0’) is positive predictor. Proportionately those having ‘0’,’1’,’2’,’3’,’4’,’5’,’6’ trainings had percentage-attrition respectively 27%, 12.5%, 18%, 14.07%,21.4%,12.3%, 9%. On absolute basis maximum quitters were those trained 2- or 3-times last year.

26. *Attrition Vs WorkLifeBalance:* (1,2,3,4 for bad, good, better, best)

Bad WorkLifeBalance is a positive predictor. In absolute terms, maximum quitters had good ‘2’ or better ‘3’ WorkLifeBalance. Proportionately, the percentage of quitters in class ‘1’,’2’,’3’ and ‘4’ are respectively 31.48%, 16.66%, 14.16%,17.30%.

27. *Attrition Vs YearsAtCompany*:

YearsAtComapny is a negative predictor. Among quitters, maximum spent less than 4 years at the company. After 10-12 years at the company, attrition is negligible. From the binning table, more than half of the attrition was before completing 4 years at the company.

28. *Attrition Vs YearsInCurrentRole:*

YearsInCurrentRole is a negative predictor. On absolute-basis, attritions are maximum in first ‘4’ years in current role. Proportionately, they are maximum in the first two years.

29. *Attrition Vs YearsSinceLastPromotion:*

YearsSinceLastPromotion is a negative predictor. In absolute terms, 75% attritions occur within 2 years of promotion. 85% leave within 5 years of promotion.

30. Attrition Vs YearsWithCurrManager:

YearsWithCurrentManager is a negative predictor. 62% spend only 2 years with the current manager. Nearly 75% spend less than 5 years with the current manager.

3.3 Data Preparation: [PHASE -3 OF CRISP-DM]

In this study, it was required to make two models based on neural network (NN) and two on stacking. So, data preparation is also described in two parts for each class of algorithm.

3.3.1 Data preparation for two neural-network models:

1. After renaming column '1' to 'Age' (Section 3.2.2.A) and removing four insignificant columns (3.2.2.C), a data frame with 1470 rows and 31 columns were left. As described earlier, 5 correlated predictor pairs were found, the moderate class imbalance was noticed and some outliers were also found. *No treatment for class-balancing, outlier-removal or correlated variables was done initially*. So next data preparation steps were started.
2. As described in section 3.2.2.G, label encoding of 7-factor variables '*Overtime*', '*BusinessTravel*', '*Gender*', '*Department*', '*EducationField*', '*JobRole*' and '*marital status*' was done, converting these into a numeric variable. Response variable '*Attrition*' was converted from 'Yes' and 'No' to numeric '1' and '0' respectively. *The 'nnet' algorithm can even take factor-variables as input, but in that case, its two hyper-parameters 'size' (i.e. a number of neurons in its only hidden layer) and 'decay' need to found by trial-error which often makes designing and testing time consuming and architecture complex.* So, a better approach is first using *radiant. model ()* library and using its *cv. nn ()* function to obtain a very finite number of combinations of (size, decay) for which NN converges and gives excellent metrics. This study followed this approach. *However, the use of radiant. model () requires label encoding. So, label-encoding was done.* A*lgorithm 'neuralnet' in any case needs encoding compulsorily.*
3. The value-ranges of various numerical variables differed greatly. For example, 'Age' varies from '18' to '60', 'DistanceFromHome' varies from '1' to '29' but

'MonthlyIncome' varies from '1009' to '19999'. 'MonthlyRate' varies from '2094' to '26999'. So, numeric variables were scaled to bring them within acceptable limits.

4. Both NN based modes give the outcome of 'Attrition' as probability which is always between '0' and '1'. So, all predictor values were converted between '0' and '1' using a 'preprocess range model' created by using method 'range'.
5. The final step in data preparation is partitioning the encoded, scaled and ranged data into a test set and train set. For this *createDataPartition()* function of '*caret*' library is used. 70% of data was kept as train dataframe and 30% as test dataframe. One advantage of createDataPartition() is that it keeps the proportion of classes same as original data (nearly 84% '0' and 16% '1') in train and test dataframe.

3.3.2 Data preparation for stacked-models:

For the stacked model two approaches used were '*CaretEnsemble*' library-based approach and '*H2o and DALEX based*' approach.

3.3.2.1 For CaretStack

The 'caretEnsemble' based approach required only one step of data-preparation i.e. removing four unnecessary columns and renaming column '1' to 'Age'.

3.3.2.2 For H2oStack

For stacking, based on H2o library and DALEX, the steps were as below:

1. First libraries 'dplyr', 'DALEX' and 'caret' were loaded and data was imported. Three columns 'Over18', 'EmployeeCount', 'StandardHours' were removed as they have a single unique value for all rows. All ordered variables were converted to factors as per DALEX requirement. Further, the response variable 'Attrition was changed from 'Yes' or 'No' to numeric '1' or '0' respectively. Column'1' was renamed to 'Age' from 'i..Age'

2. Latest JAVA version and Latest H2o version with its dependencies “RCurl" and "jsonlite" were downloaded and original data-frame df was converted into H2o object by as.h2o () function.
3. The final data-preparation step is creating *training*, *testing* and *validation* frame, named respectively: ‘train’, ‘valid’, ‘test’, by splitting main data using function h2o.splitFrame*()*. All three splits are already H2o objects. ‘train’ has 866 observation, valid has ‘246’ observations and ‘test’ has 358 observation.

3.4 Model-Designing: [PHASE -4 OF CRISP-DM]

3.4.1 Selecting modelling Technique:

Two modelling techniques were chosen; *first*: ‘neural-networks’ (one by ‘nnet’ and other by ‘neural-net’) and *second: ‘stacking’* (one by ‘caretEnsemble’ and other by ‘H2o’). *Reasons for using this approach are as below:*

1. As many variables do not follow *Gaussian Normality and* since the dataset is finite with 1470 records, statistical methods were ruled out.
2. The user-driven approach was ruled out as the researcher does not have much domain-expertise in the HR field. The data-driven approach was selected as there is ready data-set, so domain-expertise in collecting details not required. Not much feature engineering was done as selecting subsets of data, creating new features by combining original features etc. also needs domain expertise. So almost entire dataset was used. Two methods are chosen (ANN and stack) also need little prior domain knowledge and feature-engineering.
3. The class-imbalance in data (16.12% quitters Vs 83.88% stayers) can be handled by either algorithmic-approach (i.e. selecting algorithms which are robust to imbalance, correlations, outliers etc. like ANN or Stack) or data-approach (e.g. doing oversampling using methods like SMOTE). The first approach was preferred as it helps maintain the purity of information contained in data.

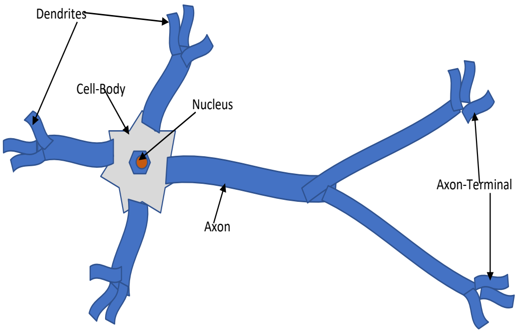
4. ANN does not assume any relationship(distribution) between the 'predictor' and 'response' variable. So it can catch both *'linear'* and *'non-linear'* trends from which new insights may emerge.
5. Attrition-prediction is a binary supervised classification problem with the clear non-overlapping boundary between two classes. ANN is very suitable for such problems. Moreover, compared to the SVM, extreme learning machine, and RF; ANNs are more fault tolerant. (That is, they can handle incomplete data and noise), easily scalable, and can generalize at high speed and make predictions.
6. Andrew Quinn et al. in 2002 paper [13] showed that results obtained by NN are not always superior. *So, for extra validation stacking based method and visual-analytics were added. Further, in both the NN based and stack-based model, two different packages were tried. This approach ensures that employee-attrition problem is examined by multiple algorithms and all angles and it makes the current study more reliable and robust*.
7. Stacking is an ensemble-based method. Recently such methods have always produced better results (in terms of performance, generalizability and averaging out noise) than individual classifiers as diverse models are used. It also shows a paradigm shift in the thinking of data -miners that: "*Rather than trying to find the best possible algorithm for a given problem it is better to try a set of collection of well-performing complementary models and then combining them in a typical manner*".
8. The problem of 'data-leakage' often described as a 'weakness' of stacking can be easily overcome by manyfold cross-validations and many repetitions as is done in this thesis.

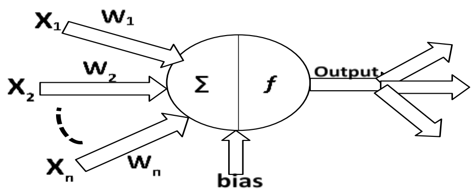
CHAPTER 4 IMPLEMENTATION

4.1 Background of Artificial Neural Networks (ANN)

An ANN is an information-processing system inspired by biological neural network-BNN (i.e. human brain). Although many modern top AI-scientists no longer prefer this analogy, it is useful to learn the basics of comparison between BNN and ANN. (fig 4.1 A below)

FIG-4.1 Comparison of biological neuron with perceptron

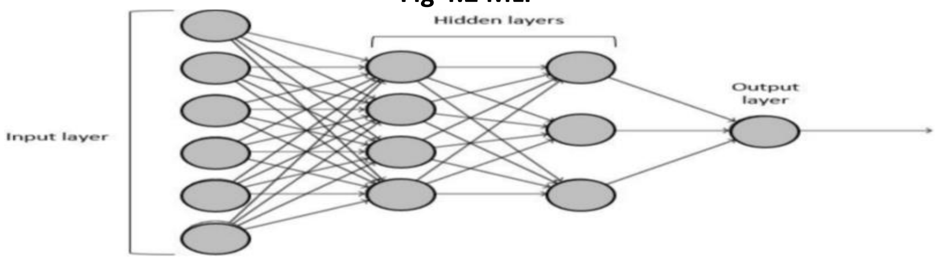




In BNN, neurons are joined with each other through dendrites and axon and there are connecting gaps between axon and dendrites called synapses. *The strengths of synaptic connections change in response to external stimuli*. *This change is how learning takes place in living organisms*. Comparing this with the neuron(perceptron) each perceptron receives inputs from many perceptrons. *Perceptrons are connected by weights which are analogous to synapses*. Each input X_i is scaled by a weight W_i and the weighted sum of inputs i.e. $\varepsilon w_i \, x_i$ is the actual input for a neuron. Often a constant bias 'b' is added then input-signal is. $\varepsilon w_i \, x_i \, + b$ (for i= 1 to n) Then an activation function f is applied to this combined signal to find the output. The difference between actual output $\hat{y}$ and calculated output y is used to find error *using suitable error-function*. This error is back-propagated through ANN *using suitable backpropagation algorithm* (BPA) and *it works analogously to a 'negative feedback' in a biological organism through which learning occurs*. *In BNN learning occurs by adjusting synaptic strength*, but *in ANN, it occurs by adjusting weights*.

The single perceptron described above has very limited computational power. However, when many such neurons are joined, a 'Multi-Layer-Perceptron' (MLP) forms. [Fig 4.2]. Now the flexibility in a varying number of layers, the number of neurons per layer, changing weights, changing error-function and activation-function…..etc. gives MLP ability to approximate almost any functional relationship between 'input' and 'output'. Hence it is called "*universal function approximator*". If there are a very large number of hidden layers, then ANN is called a *deep neural network* (DNN).

Fig 4.2 MLP



However, this biological analogy is considered outdated nowadays. Instead, *an ANN is considered as directed graphs whose vertices are 'neurons' and whose directed edges are synapses. The weights attached to the directed edges (synapses) indicate the effect of originating neurons.* All data travels through NN as signals and as each neuron receives many signals they are combined first by integration function $\sum$ and then by activation function *f*.

For such DNN, the final output is $f = g(Z)$ where:

a) $Z_{0,} Z_{1,} Z_{2,}$……………, Z_K are inputs from all preceding neurons.

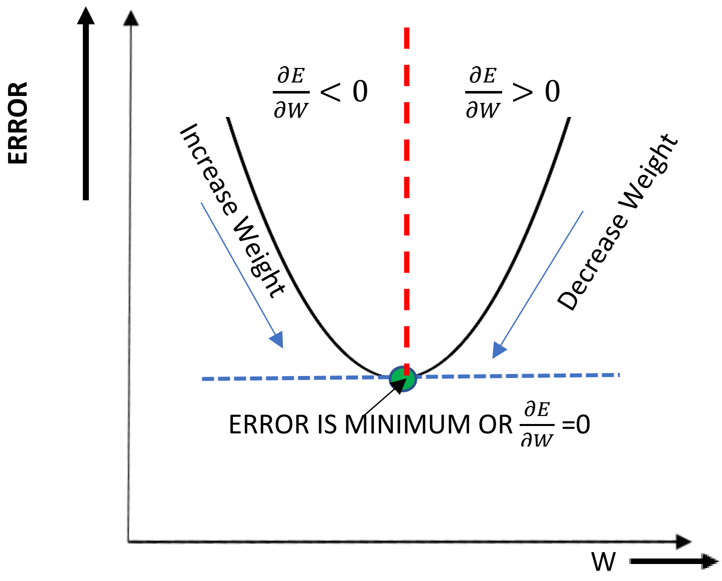
b) g defined over $R^{k+1} \rightarrow R$ (integration function)

c) f defined over $R \rightarrow R$ (activation function) [*bounded, differentiable, non-linear*]

The process of 'training' the 'neural network' depends on backpropagation. Its steps can be summarised as below:

a) The difference between actual output $\hat{y}$ and calculated output y is used to find error E using suitable error-function. *The graph of Error E Vs Weight W* is shown below:

Fig 4.3 Back-propagation



b) Weights are adapted by suitable back-propagation algorithm by finding $\frac{\partial E}{\partial W}$. If $\frac{\partial E}{\partial W} < 0$ then weights are increased and vice versa.

c) This process is iteratively repeated until either $\frac{\partial E}{\partial W}$ becomes equal to selected minimum threshold **OR** it completes designated steps **OR** error function becomes minimum (Shown by the green dot in figure 4.3). With the above general discussion, modelling by specific algorithms '*neuralnet'* and '*nnet'* used in this research can be described.

4.2 Modelling by 'neuralnet' package:

1. Once data-preparation described in section 3.3.1 is done, model-building by '*neuralnet*'. the package can start. This is a flexible package for training NN. It allows several BPA , ability to tweak several parameters, custom choice of *error function* and *activation function*. This package is built in the context of regression analysis (a process of approximating the functional relationship between response variable and predictor variables). *So it can catch both linear and non-linear relationships*. It can also calculate '*generalised weights*'.
2. The generic formula of '*neuralnet*' is shown on left and hyper-parameters used in this research on right.

Fig- 4.3-B Generic-Formula of 'neuralnet'(left) and hyper-parameters used in current work (right)

GENERIC FORMULA- SOME DEFAULT VALUES ARE ALREADY CHOSEN	FORMULA USED IN THIS THESIS. IF NO VALUE IS SPECIFIED, DEFAULT VALUE IS USED.
nn<- neuralnet (formula, data, hidden = 1, threshold = 0.01, stepmax = 1e+05, rep = 1, startweights = NULL, learningrate. limit = NULL, learningrate. factor =list (minus = 0.5, plus = 1.2), learningrate = NULL, lifesign = "none", lifesign. step = 1000, algorithm = "rprop+", err.fct = "sse", act.fct = "logistic", linear. output = TRUE, exclude = NULL, constant. weights = NULL, likelihood = FALSE)	nn<- neuralnet (formula = Attrition ~., data = trainDF, hidden = c (6,4,2), threshold = 0.0017, err.fct = "sse", start weights = NULL, act.fct = "logistic", algorithm = "rprop+", linear. output = F, learningrate. limit = NULL, lifesign = "full", lifesign. step = 100, stepmax = 15000,)

In the above right side, the hyperparameters used in current work have the following meaning: *hidden = c (6,4,2)* means 6,4 and 2 neurons in 1st, 2nd and 3rd hidden layer respectively, *threshold= 0.0017* means when $\left(\frac{\partial E}{\partial w}\right)$ reaches this value, the training stops, *err.fct = "sse"* means '*sum of square*' method is used to find error between actual and expected *output.act.fct = "logistic"* means logistic or sigmoid function is used as activation function, *algorithm=rprop+* means "resilient backpropagation(BP) with weight backtracking", *linear.output = False* means output is mapped by activation function between [0,1] i.e. it is probabilistic, *lifesign= full* produces verbose output, *lifesignstep=100* means output would be seen in the console after every 100 steps and *stepmax=15000 means* training will stop after 15000 steps.

3. The reasons for using *rprop+* are as follows:
    a. It is the fastest algorithm for regression *(Schiffmann et al. ,1994, Rocha et al.,2003, Kumar and Zhang, 2006, Almeida et al. 2010*).
    b. In simple BP, one fix learning rate for the entire training process and whole NN needs to be defined while in 'rprop'(+ or - both) *learning rate can be changed with time.*
    c. It *uses only the signs* (+ve or -ve) of $\left(\frac{\partial E}{\partial W}\right)$ to update the weights. *This guarantees an equal influence of learning rate over the entire network*.
    d. Further, if 'rprop+' is used then still one more benefit of "*weight backtracking*" is added. This is a technique of undoing the last iteration and adding a smaller value to the weight in the next step. It prevents "jumping" of the algorithm over minima several times and hence missing it.
4. The plot of NN obtained by 'neuralnet' package with above hyperparameters is shown in 'Results and Analysis' Fig 5.1.
5. Function gwplot() was used to obtain the plots of generalized weights of six predictors and the response variable Fig_5.2A . Similar plots can be obtained for all predictors.
6. Once neural network was trained using above hyperparameters, it was used to make predictions on test set. As the model gives the probability of Attrition, one needs to select a 'probability-cutoff' while making predictions. Using the ROCR library, the graphs below were obtained from which optimal cutoff can be obtained.

FIGURE – 4.4 Accuracy vs cut-off

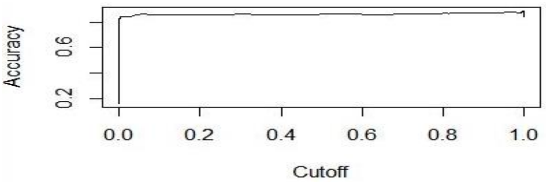


FIGURE- 4.5 F1-score vs cut-off

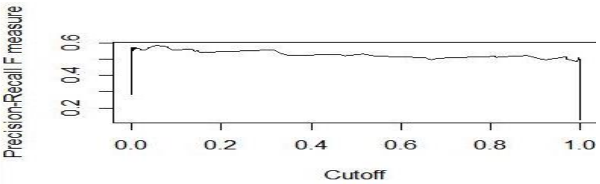


FIGURE- 4.6 SENSITIVITY VS CUT-OFF

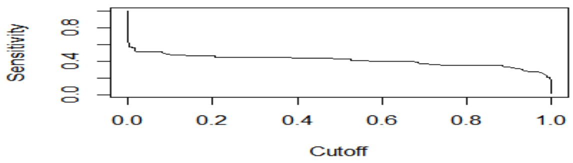
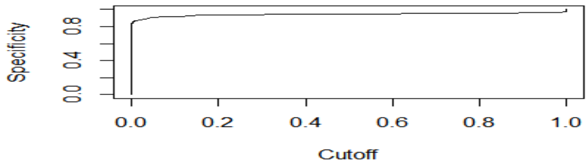


FIGURE- 4.7 SPECIFICITY VS CUT-OFF



7. Figure-4.4 shows that accuracy becomes almost constant for *any* cutoff near ≈0.1. But the two key metrics in this study i.e. *F1_Score* and *Sensitivity* vary with cut-off chosen. (Figure 4.5 and 4.6). Although it is possible to get the highest F1_score and Sensitivity near cut-off ≈0.1, it means that retention measures must be taken even if the employee shows >10% probability of quitting. Such extreme cut-offs do not make business sense. *So a balanced way can be choosing cutoff around 0.35 for 'conservative' view and 0.5-0.6 for 'liberal' view on attrition*. In this study, all *cut-offs 0.1,0.5.0.6 and 0.35* were tried and results are in Table-5.1.
8. The variable-significance plot was obtained by *olden()* function. It is shown in Figure- 5.3. From this top 6 positive predictors and top 6 negative predictors were found (Table 5.2 and Table-5.3) and they were compared with patterns obtained by visual-analytics.
9. More metrics were obtained using the library 'MLmetrics' for extra validation and each neural-weight was calculated by *neuralweights ()* function of library *'NeuralNetTools'*. The metrics are listed in Table 5.4. Lekprofile of top-6 variables was done from which detailed sensitivity analysis can be made as shown in Figure 5.4.

4.3 Modelling by 'nnet' package:

1. Although 'neuralnet' based model had many positives like above 90% F1_Score, and 88% sensitivity for all four cut-offs chosen and matching with patterns identified for 11 out of top 12 predictors, the parameters (Figure 4.3-B) *were obtained by trial-error*. This took a lot of time. So it was decided to make another model based on *'nnet'* package.

2. 'nnet', at its root, uses BFGS (Broyden–Fletcher–Goldfarb–Shanno) optimization. BFGS belongs to quasi-newton methods [77] which is a type of hill-climbing optimisation technique.

3. The generic formula and values used in this study for '*nnet*' are shown below.

NN<-nnet (formula, data, weights, ..., subset, na. action, contrasts = NULL) *OR*

NN<-nnet(x, y, weights, size, Wts, mask, linout = FALSE, entropy = FALSE, softmax = FALSE, censored = FALSE, skip = FALSE, rang = 0.7, decay = 0, maxit = 100, Hess = FALSE, trace = TRUE, MaxNWts = 1000, abstol = 1.0e$^{-4,}$ reltol = 1.0e^{-8}, ...)

THE HYPERPARAMETERS USED IN THIS STUDY

NN1<-nnet (Attrition~. , trainDF, size= 1, rang=0.07, Hess=FALSE, decay=13e-4,maxit=3000)

4. In above list, meaning of important parameters is as follows: 'x' =train-set, 'y'= train-set$responsevariable, 'weights'= (case) weights for each example. 'Wts'=intial parameter vector, if missing chosen at random. 'mask'= logical vector indicating which parameter should be optimized. If 'linout'=F then logistic output if it is T then linear output. If 'entropy' = F then use least square regression and 'entropy'= T then 'cross-entropy' regression. The 'softmax'= F means 'nnet' will use 'loglinear model' for fitting, if it is T then it will use 'maximum conditional likelihood' for fitting. If 'rang' = 0.3 then all initial random weights are chosen from [-0.3,0.3]. The 'decay'= 0 by default but one must set some value to start training. *The 'decay' is a kind of penalty on larger coefficient*. Ideal 'decay' was found as described in point 6,7,8 below. The 'maxit' = the maximum number of iterations after which training stops.

5. Thus, *'nnet'* also requires specifying many hyper-parameters. *However, For most of these, default values are fine or there are only 2-3 available alternatives*, information about which can be found from CRAN documentation of 'nnet' package.

6. Two parameters which are difficult to decide by trial-error are *'size'* (*no of neurons in the only hidden layer*) and *'decay'* (*a kind of penalty used for larger coefficients while*

tuning the NN). So, this time, instead of guessing them, first several pairs of (*size*, *decay rate*) which give minima of error-function were obtained by package radiant.model, using its function cv.nn().

7. So,once data-preparation described in section 3.3.1 was done, library *'radiant.model'* was loaded and a temporary neural-network(NN) with size = 6 and decay = 0.01 (arbitrary values) was used to train the train-data.
8. Then using this temporary NN, sizes '1 to 29' and decay rates from '$10e^{-4}$ to $20e^{-4}$ in an increment of $1e^{-4}$' were tried in function cv.nn() with fivefold cross-validation and one repeat to optimize AUC. This gave hundreds of combinations of (*size* and *decay*) for which the subsequent *neural-network* which was made by '*nnet*' always converged and gave good metrics.
9. Five top combinations, as described in Table 4.1 below, were used to train *'nnet'* based models. *For each case probability cut-off of 0.365 was used*.

TABLE- 4.1 Various 'nnet' models trained

size	decay	F1_Score	Sensitivity	Accuracy
1	$13e^{-4}$	0.9117	0.9584	0.8485
16	$14e^{-4}$	0.8883	0.9066	0.8254
17	$19e^{-4}$	0.9004	0.9312	0.8299
21	$20e^{-4}$	0.9039	0.9176	0.8390
18	$17e^{-4}$	0.9039	0.9176	0.8390

The model with one hidden neuron and decay 13e-4 was the best. Its diagram, obtained by *plotnet()* function is shown in figure 5.4.

10. The variable importance plot by the best *'nnet'* based model was obtained by the *olden()* function which is shown in figure 5.6. From this plot, Top 6 positive predictors and Top 6 negative predictors were identified and they were compared with the pattern identified by visual-analytics as shown in table 5.5 and table 5.6.
11. LekProfile and ROC plot for sensitivity analysis were obtained by *'nnet'* based model which are shown in figure-5.7 and figure-5.8.

4 Background of 'Stacking-ensemble':

A lot of discussion on fundamentals of stacking was done in literature review (2.2.3.1). So, only brief information will be added here. Stacking is an ensembling technique that uses a diverse set of classifiers to improve predictive-power, generalizability and stability. *As various types of classifiers have different "inductive biases", they learn about data in a different manner***.** This *diversity reduces variance-error without increasing bias-error*. Sometimes, an ensemble can reduce bias-error too. From base-classifier level (also called level '0') a meta-data set is created containing a tuple for each tuple in the original dataset. The meta-classifier is also called level '1' classifier. Instead of using the original input attributes, meta-learner *uses the predicted classification(probabilities) of base-classifiers as the input attributes*. The target attribute remains the same as the original training set.

4.5 Steps in caret-stacking:

stacked-classifier used in the current study is shown in Fig 4.8 in a symbolic manner

FIGURE-4.8 Diagram of stack

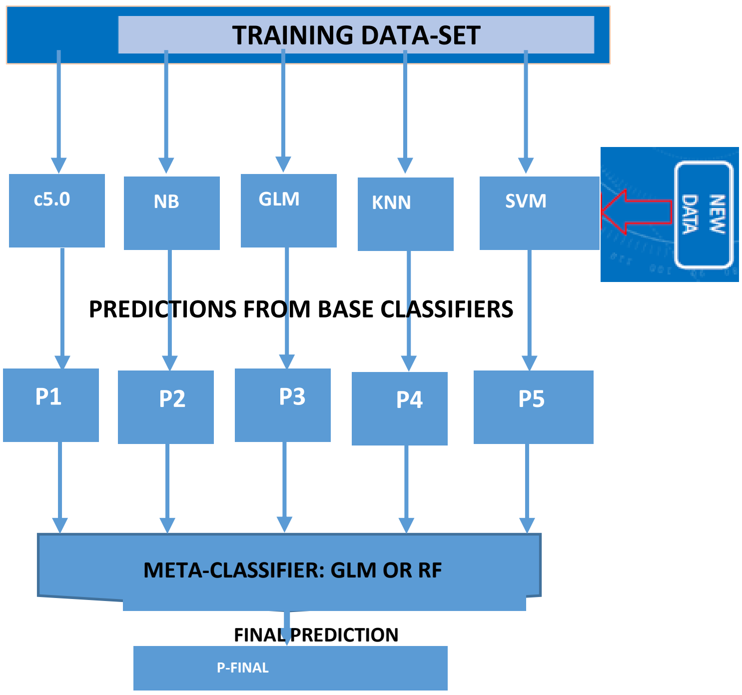
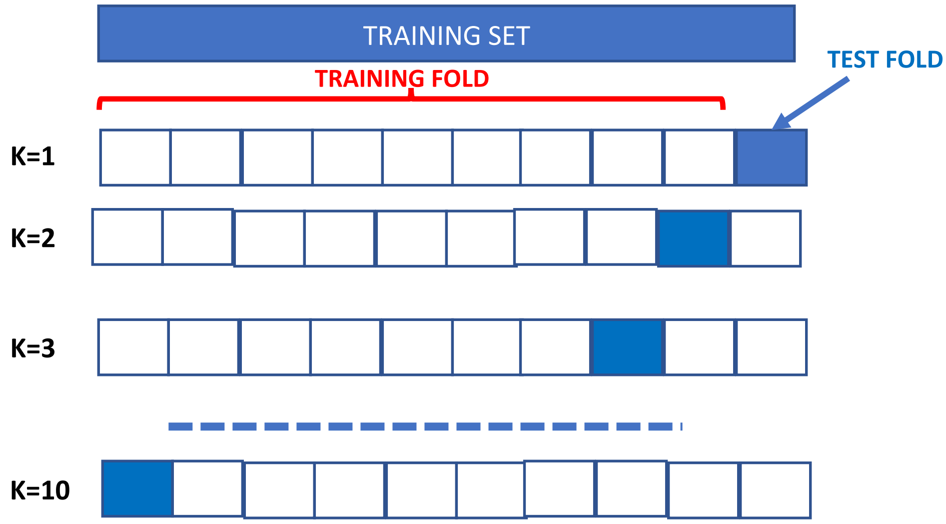


FIG-4.9 K-Fold cross validation



For stacking, *'caret'*, *'caretEnsemble'* and *'doParallel'* libraries were loaded and 3 clusters were registered to make processing parallel. Data was read and only data-preparation was removing four unnecessary columns and changing the name of column-1 to *'Age'*.

1. Then hyper-parameters were set up, in *traincontrol ()*, for obtaining '*control1'* using 10-fold cross-validation using method '*repeatedcv*', with 10 repeats, with option *'summaryFunction = twoClassSummary'* to get *ROC*, *Sensitivity*, *Specificity*.
2. Similarly, *'control2'* was obtained, keeping all hyper-parameters same but removing the option *'summaryFunction = twoClassSummary'* to get *Accuracy and Kappa*.
3. Next, a pair of five base-models were created using function *caretList ()*. '*models_1*' with '*control1*' and '*models_2*' with '*control2*'. The data used was entire dataset (*with cross-validation its subsets will be used as data for various models*, as fig 4.9 above shows, for example of 10-fold validation). *This cross-validation avoids so-called data-leakage problem* [8]. As argument *'methodlist'* the *list* of all base-classifiers *('C5.0', 'nb', 'glm', 'knn', 'svmRadial') was passed. As argument 'metric' (to be optimized) 'ROC' was used.*
4. Two base-models *'models1'* and *'models2'*, one using '*control1*' and other using '*control2*', were created and used to obtain '*results1*' and '*results2*' by *resamples ()*. The metrics for

various base classifiers and the ensemble are in Table- 5.7. Here, just metrics for best RF-based ensemble stack are noted here: *ROC=0.8696, Sensitivity=0.9692, Specificity=0.4761, Accuracy=0.9156 and Kappa=0.6476*.

5. Then *modelCor()* was used with *'results1'* and *'results2'* as arguments to check if the 5 base models have some excessive correlation. This is required as, for *stacks to give increment in metrics, it must be made of diverse models*. No excessive correlation was found, so base models were diverse enough.

6. Then *traincontrol ()* was used a second time, again using 10-fold cross-validation using method '*repeatedcv*', with 10 repeats to obtain '*stackcontrol1*' and '*stackcontrol2*' with and without option '*summaryFunction = twoClassSummary*' respectively.

7 Finally, *caretStack ()* was used to obtain two stacked models (*stack1.glm* and *stack2.glm*) with *GLM* as meta-classifier and other two (*stack3.rf* and *stack4.rf*) with *RF* as *meta-classifier*. Results obtained with them are as below:

GLM as meta-classifier (*stack1.glm* and *stack2.glm*): *0.8838 accuracy, 0.8435 ROC, 0.9781 sensitivity, 0.4885 kappa and 0.4353 specificity*.

RF as meta-classifier (stack3.rf and stack4.rf): *0.9156 accuracy, 0.8696 ROC, 0.9692 sensitivity, 0.6476 kappa and 0.4761specificity*.

As stack4.rf was best, it was saved as deployment-model by using *saveRDS ()*.

8. Now when predictions using this stack are wanted, the user just loads the saved model, with *readRDS ()*, reads input data, removes 4 columns, sets up train-set and test-set, makes predictions and obtains confusion-matrix. For the RF-based stack, metrics obtained are shown in the figure in table 5.7 and table 5.8 on *test-set which were better than the base classifiers.*

9. ROC plot for ensemble stack was plotted as shown in figure-5.8. Variable Importance, in descending order, by the stacked model for top 12 variables is shown in Table 5.9.

10. The stack was also used for actual business-objective i.e. one employee from test-set was isolated as case-employee whose attributes are as tabulated below:

TABLE 4.2 Attributes and prediction for a single employee

Age 41	Business Travel Rarely	daily rate 1102	Department Sales	Distance from Home 1 km	Education '2'	EducationField LifeSciences
Environmental Satisfaction '2'	Gender Female	Hourly Rate 94	Job Involvement 3	JobLevel 2	job role sales executive	Sales Executive 4
marital status Single	monthly income 5993	monthly rate 19479	NumCompaniesWorked 8	OverTime Yes	PercentSalaryHike 11	performance rating 3
Relationship Satisfaction 1	Stock Option Level 0	TotalWorkingYears 8	TrainingTimesLastYear 0	work-life balance 1	YearsAtCompany 6	YearsInCurrentRole 4
YearsSinceLastPromotion 0	YearsWithCurrentManager 5	Using Stack4.rf when the prediction was made for this employee the answer was *"No"* means that the above employee will *not* attrit. His probability was 0.090 while cutoff was 0.365				

4.6 Stacking by H20, Interpretation by DALEX: (Requires JAVA installed)

Although the caretStack gave very good sensitivity, F1_Score, Accuracy and other metrics, variable-significance obtained by it is limited in usefulness as it lacks direction unlike *olden ()* method. So, the model suffers from *poor interpretability* and *poor explainability*. To solve this problem, another H2o library-based stacked model was made. This method needs some extra data preparation as described in 3.3.2.2. Once these steps are done, steps to make base-classifier and ensemble are as below:

1. GLM based base-classifier was created with hyper-parameters family=" binomial', keep_cross_validation_predictions=TRUE, keep_cross_validation_predictions = TRUE, fold_assignment="Modulo", *which gave an AUC 0.8284.*
2. GBM based base-classifier was made with hyperparameters fold assignment="Modulo", keep_cross_validation_predictions = TRUE , nfolds=10, ntrees=100, max_depth = 3, stopping_metric= "AUC", stopping_rounds = 5, stopping_tolerance = 0.005 and seed=100 which gave AUC= 0.8103.

3. RF based base-classifier which gave with parameters nfolds = 10, fold_assignment =" Modulo", keep_cross_validation_predictions = TRUE, ntrees = 1000, stopping_metric = "AUC", stopping_rounds = 10, stopping_tolerance = 0.05 and seed = 123. *It gave AUC=0.7590.*

4. Deeplearning based classifier with hyper-parameters nfolds=10, fold_assignment = "Modulo", keep_cross_validation_predictions = TRUE, stopping_metric = "AUC" and seed= 123. *Which gave AUC = 0.7823*.

5. Finally, an ensemble-stack using above four base-classifiers and again '*deeplearning*' as meta-classifier was made, *which gave AUC of 0.8358*. Thus, once again *ensemble gives better performance than the best of base classifiers*. Although in this case, the increment from 0.8284 to 0.8358 (i.e. ≈0.0074≈0.74%) looks small but later it will be seen that it sometimes makes a big difference in decision making.

6. A further advantage of this H2o based stacking is that it *can be easily combined with DALEX package to obtain much better interpretability and explainability* compared to caret-stacked models as described below.

7. Next performance of ensemble was tested *on the validation set. AUC= 0.8333, MSE =0.09467, maximum accuracy=0.8965* at probability threshold 0.955 are good values. However, maximum sensitivity was 1 at a probability threshold 0.017. These results are consistent with figure 4.2 and 4.4. As such extreme probability values do not make business-sense, the threshold must be decided based on the importance of employee.

8. Next performance of the ensemble was tested *on test-set*. Again, *AUC=0.8262, MSE =0.09191, maximum accuracy=0.8976 at probability threshold 0.4203 are good values*. However, maximum sensitivity was 1 at a very low probability threshold 0.019.

9. For further interpretability use of the package, DALEX was made. However, DALEX does not have native support for H2o objects. So, feature-set was converted into its original form and response-variable- ‘Attrition’- into 0/1 form.
10. One also needs to develop a custom predict-function ‘cust_predict’ which takes base/ensemble classifier and ‘newdata’ as an argument and predicts the probability of ‘response’ and returns it. Internally it uses H2o’s prediction function only but for using DALEX one cannot use H2o’s predict function directly.
11. Next job done was to create *explainer object* for all four base models and the ensemble stack. For all of the arguments passed to function *explain ()* are common: *‘model’, ‘data’, ‘y’, ‘predict_function’ and ‘label’*.
12. These explainer objects were passed as arguments to function *model_performance ()* to find residuals for each model. A comparison plot is shown in Fig 5.10. The residuals can also be used to obtain box-plots of residuals as shown in Fig 5.11.
13. Next variable importance was found using ‘variable_importance ()’ with stacked-explainer object and *loss_function = loss_root_mean_square as arguments*. It is shown in figure 5.12. From that plot, a table of Top -6 positive predictors and Top-6 negative predictors were obtained. It is shown in table 5.10.
14. Next, using ensemble and *h2o.partialplot()* function Partial Dependence Plots (PDP) of *showing effect of each predictor on response-variable (attrition)* was plotted. Here assumption is that all other predictors are constant, which is often violated in practice. Even then PDPs are useful in finding dominant predictors and deciding if a predictor has linear relation with response variable. PDPs and analysis are shown in APPENDIX-IV.
15. Finally, the most important business-question was studied i.e. “*given data of an employee how to predict employee’s probability of quitting, how to know what dominant factors lead to that decision and how to present/explain that information to various stack*

holders?". For this record of an arbitrary employee was taken and the *break_down()* function of *breakdown()* package was used and finally using *ggplot()* package, the graphical view showing the probability of attrition for that employee and predictors responsible for quit decision was obtained by ensemble-model is shown in Fig. 5.13. It shows that the probability of attrition is 0.171 and if this employee quits (*if the cut-off is < 0.171*) then which predictors positively or negatively contribute to decision and what is the contribution of each to cumulative probability.

16. To dig further, the role of all 30 predictors for this employee was checked out. This can be studied in table 5.11.
17. A comparison of predictions made by four base classifiers and the stack and Top_5 variables in each case along with cumulative probability are shown in Fig-5.14.
18. All above analysis, except point-8 above, was done with training_frame = trainDF_H2o, validation frame = validDF_H2o. So, for further validity, predictions were made on testDF_H2o using function h2o.predict(), this h2o prediction result data-frame was converted to ordinary R-dataframe and confusion-matrix was obtained for testDF too. Metrics obtained were *accuracy 0.8868, sensitivity/recall 0.9340, AUC = 0.8331, Kappa: 0.5377, precision 0.9340, F1_Score 0.9340, balanced accuracy 0.7689.*

Next chapter will be results and analysis:

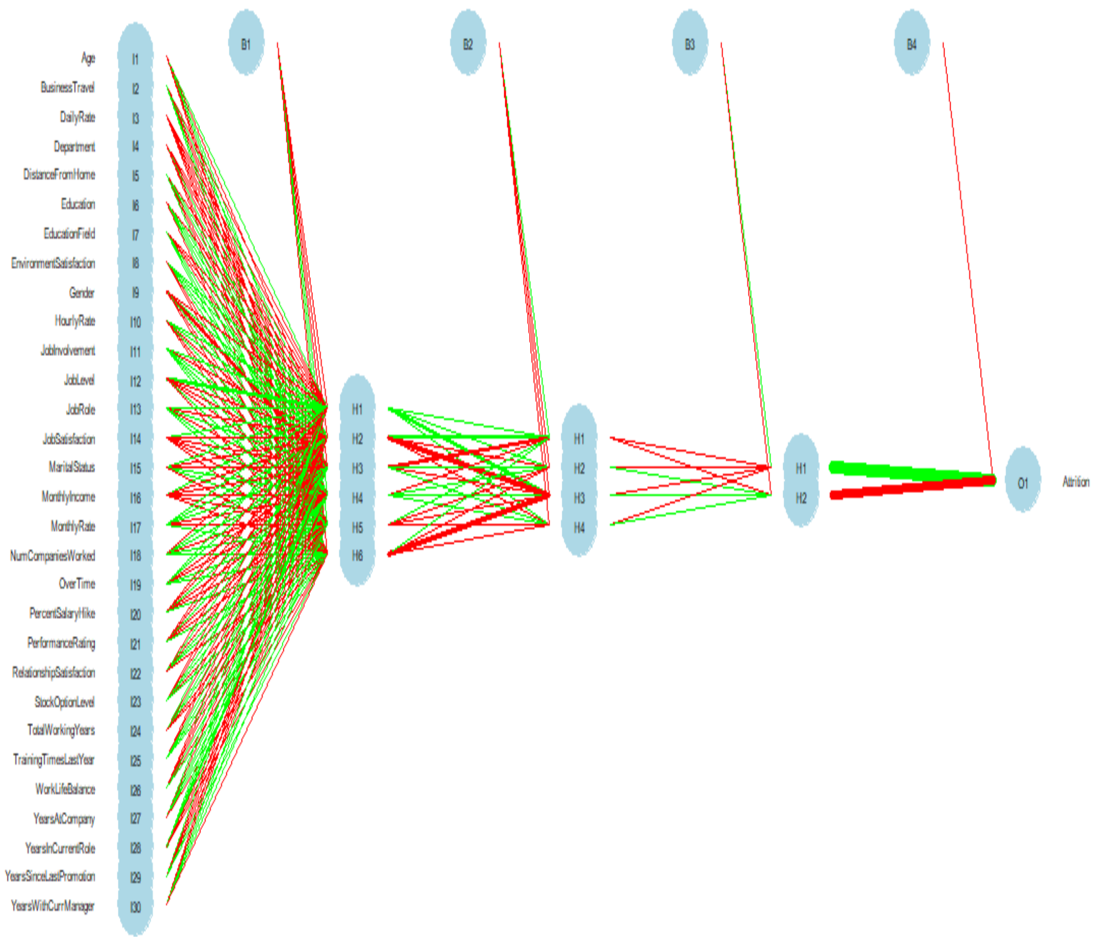
CHAPTER 5 RESULTS AND ANALYSIS

The chapter will discuss results obtained and their analysis for each of the four methods that have been employed on same data, i.e. '*neuralnet*', '*nnet'*, '*caretStack'* and '*h2o+DALEX*'

5.1 Results obtained with 'neuralnet' algorithm

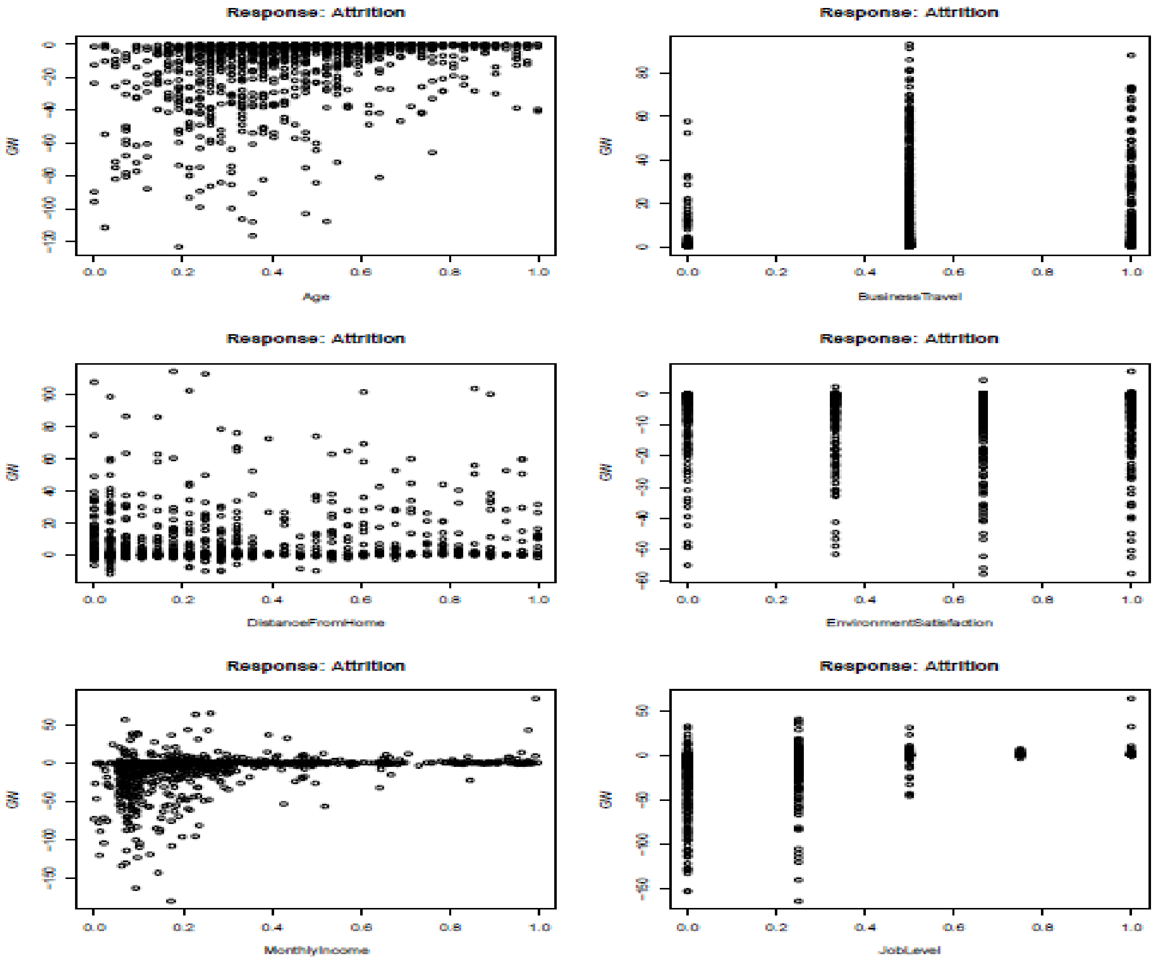
- The plot of the neural network obtained using chosen hyper-parameters is shown in Figure 5.1 below. (red = negative-weight, green = positive-weight, thickness of weight is proportional to its magnitude)

FIGURE 5.1- The neural-network obtained by 'neuralnet'



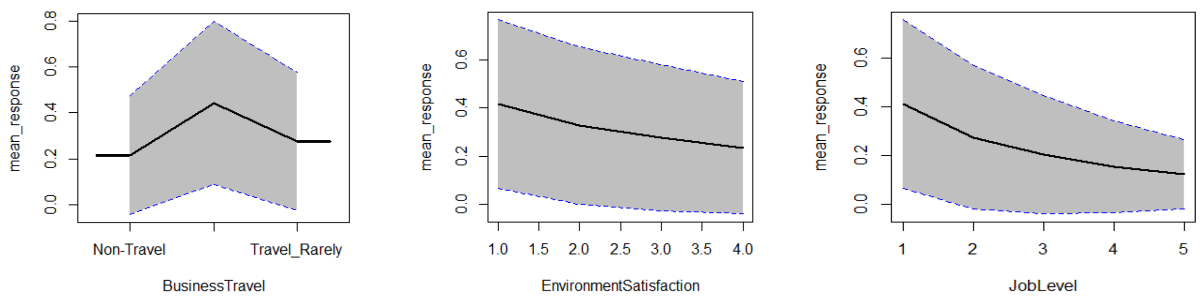
- The complexity of model above shows the perils of using a trial-error approach for finding hyperparameters. As *'nnet'* modelling, later showed, it is better to first obtain a smaller subset of hyperparameters by methods like cv. nn () and use them *to obtain a much simpler architecture, which converges faster and still gives better metrics*.

FIG 5.2-A Generalised neural-weights for six predictors



- Higher variance in generalized-weights of ‘BusinessTravel’, ‘EnvironmentSatisfaction’ and ‘JobLevel’ indicates that *they may have a nonlinear effect on response variable. [ANN](ANN) was chosen in this study due to its ability to catch such non-linear patterns.* Nonlinearity was later confirmed by PDP of three predictors obtained by H2o-Stack. [Figure 5.2-B]. Similarly, neuralweights+ PDP can be used for all predictors to confirm nonlinearity.

FIG 5.2-B PDP OF NON-LINEARLY LINKED PREDICTORS



- As was seen in Fig-4.5 and Fig-4.6, two key metrics F1_Score and Sensitivity were varying with probability threshold chosen while making predictions. In this study, four cut-offs were tried: 0.1,0.5,0.35 and 0.6. The results obtained are shown in Table 5.1.

TABLE-5.1 Metrics with different cut-offs for 'neuralnet' based model on train-set

Cut-Off	Accuracy	Sensitivity/ Recall	F-1 Score	AUC
0.5	0.8526	0.8814	0.9132	0.784
0.1	0.8526	0.8834	0.9130	0.784
0.35	0.8503	0.8811	0.9118	0.784
0.6	0.8503	0.8791	0.9120	0.784

- Choice of threshold is a subjective aspect and variation in metrics with thresholds is small. So, a 'conservative' and a 'liberal' threshold of 0.35 and 0.6 respectively can be used.
- The variable-significance plot obtained by this model is shown in figure 5.3 below from which top-6 positive predictors and top-6 negative predictors were obtained:

FIGURE 5.3- The variable significance obtained by 'neuralnet' model

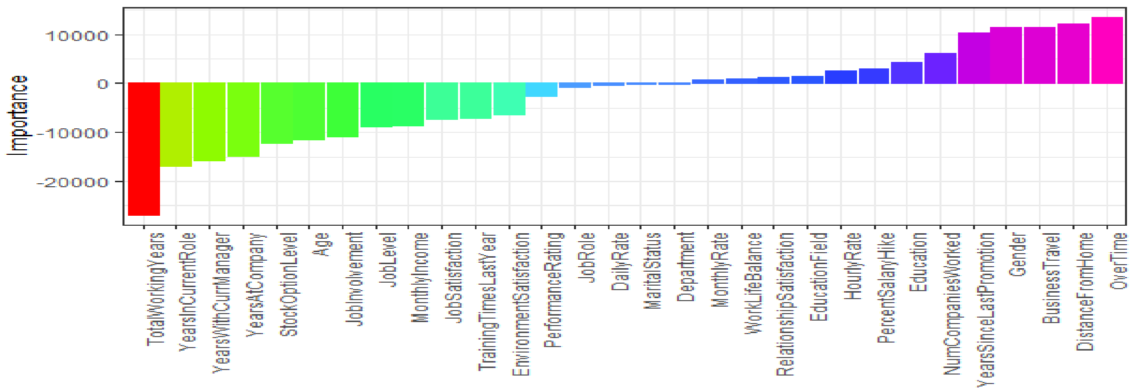


TABLE- 5.2 Top6 positive predictors

Top Positive Predictors i.e. as the value of the predictor increases, attrition tendency also increases	Whether matches with the pattern identified by visual-analytics
OverTime	Yes
DistanceFromHome	Yes
business travel	Yes
Gender	Yes
YearsSinceLastPromotion	No
NumCompaniesWorked	Yes

TABLE- 5.3 Top6 negative predictor

Top Negative Predictors i.e. as the value of the predictor increases, attrition tendency decreases	Whether matches with pattern identified by visual-analytics
TotalWorkingYears	Yes
YearsInCurrentRole	Yes
YearsWithCurrentManager	Yes
YearsAtCompany	Yes
StockOptionLevel	Yes
Age	Yes

11 out of 12 variables match with patterns identified in section 3.2.2. J *by visual-analytics.*

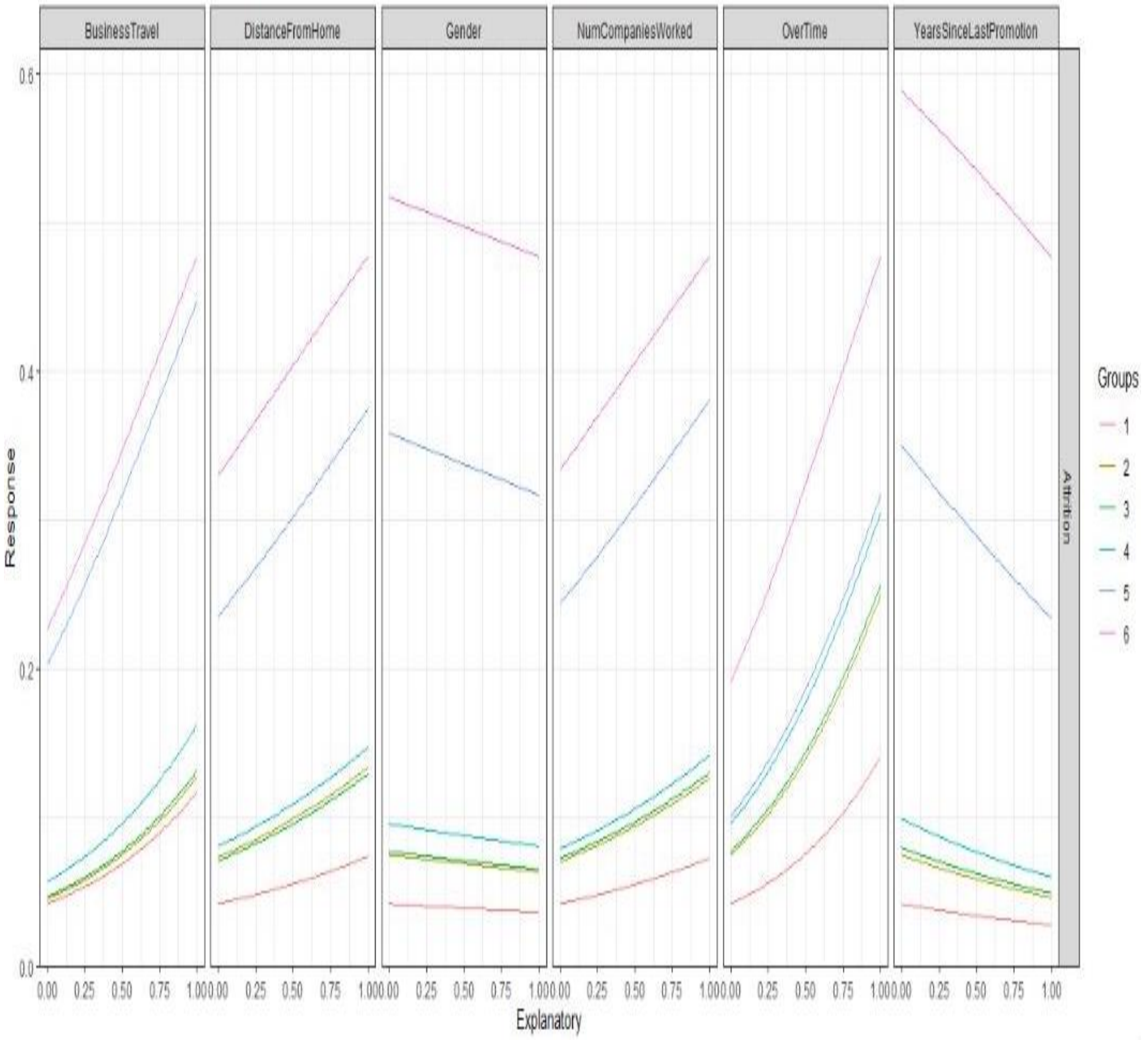
- Following are important metrics obtained by using the neuralnet model *on test-set*, using the library 'MLmetrics'.

TABLE- 5.4 Important metrics for 'neuralnet' model

Accuracy	Sensitivity	F_1 Score	AUC	MSE	R^2 ERROR
0.8183	0.8811	0.9002	0.7391	0.1837	-0.2369

- The lekprofile was also obtained for Top-6 variables using neuralnet model *which can be used for sensitivity analysis*. It is shown in figure 5.4 below. The six groups correspond to keeping other predictors at *minima, 20th, 40th, 60th, 80th quantiles and at their maxima* and *finding a relation between outcome-probability and a single predictor of interest* at a time.

FIGURE – 5.4 Lekprofile top 6 predictors by 'neuralnet' model

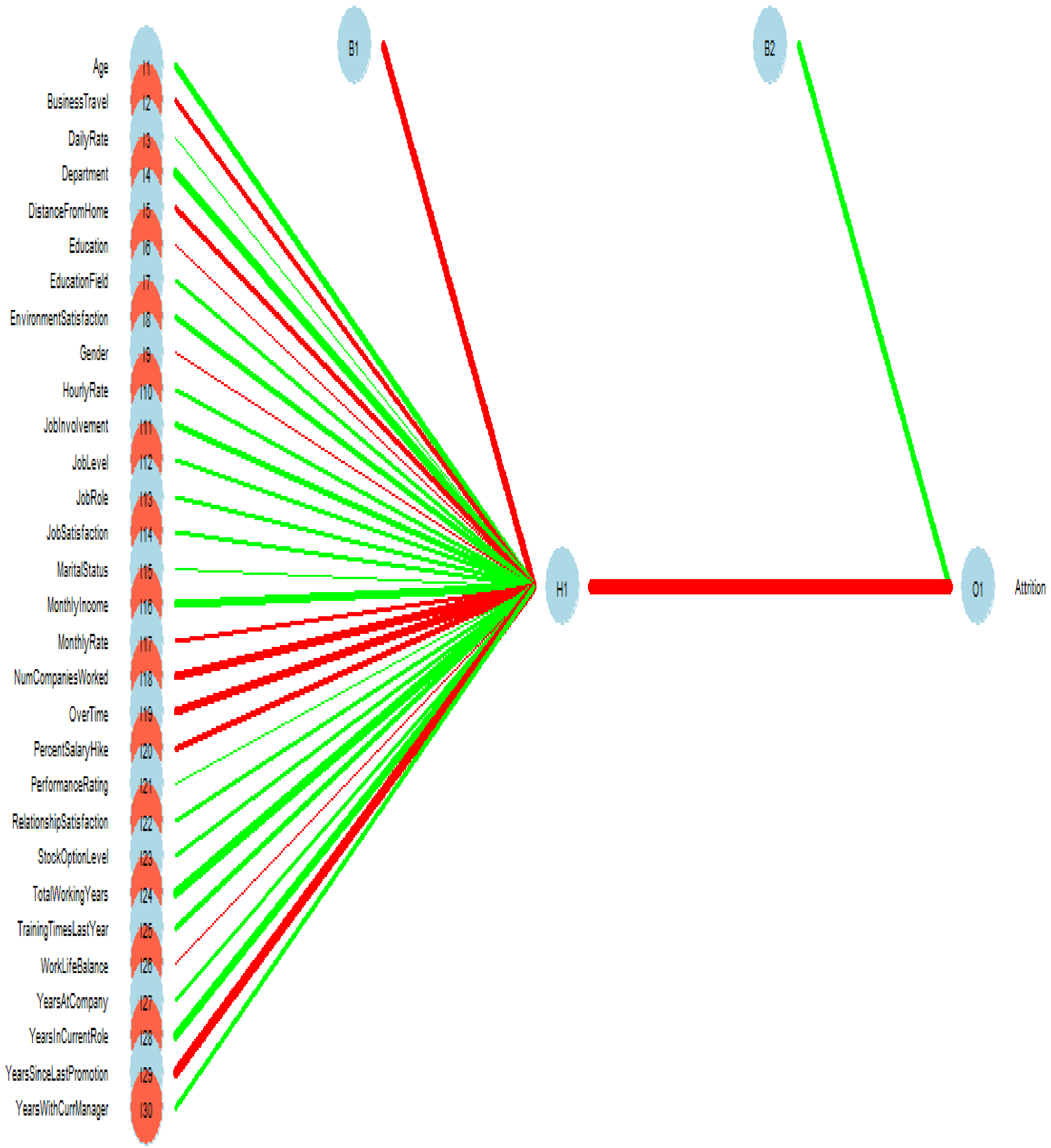


- Each individual neural-weight can be obtained by function *neuralweights(nn)*. Neuralweights are useful to catch no linear patterns. [FIG-5.2-A].
- To summarise, although *'neuralnet'* is a modern algorithm with the flexibility of tuning several hyper-parameters, in this study it gave marginally poor results compared to other algorithms being discussed below.

5.2 Results obtained by 'nnet' algorithm

- As a trial-error approach in case of *'neuralnet'* based modelling suggested, a better method would be to first obtain a smaller subset of *tuned* hyper-parameters and then use them.
- So, as described in section 4.2, *cv.nn ()* function was used to obtain '*tuned'* values of hyper-parameters '*size'* and '*decay'*. Then '*nnet'* algorithm was used with top-5 values. The results obtained are in table 4.1. *The model with one hidden neuron and decay= 13e^{-4} was the best.* Its diagram, obtained by *plotnet()* function, is shown in Figure 5.5 below:

FIGURE- 5.5 NN created by best 'nnet' model



- The variable-significance obtained by *'nnet'* is shown in figure 5.6 below. Using it top 6 positive and top 6 negative predictors were obtained (Table 5.5 and 5.6)

Figure 5.6 -Variable Importance Plot By best 'nnet' based model

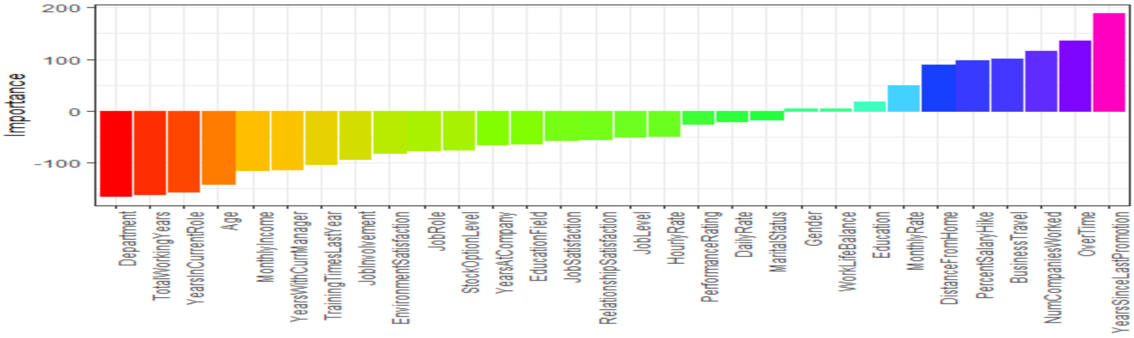


TABLE- 5.5 Top6 positive predictors 'nnet'

	Top Positive Predictors i.e. as the value of predictor increases, attrition-tendency also increases	Whether matches with the pattern identified by visual-analytics
1	YearsSinceLastPromotion	No
2	OverTime	Yes
3	NumCompaniesWorked	Yes
4	business travel	Yes
5	PercentSalaryHike	Yes
6	DistanceFromHome	Yes

TABLE- 5.6 Top6 negative predictors 'nnet'

	Top Negative Predictors i.e. as the value of predictor increases, attrition-tendency decreases	Whether matches with the pattern identified by visual-analytics
1	Department	Yes
2	TotalWorkingYears	Yes
3	YearsInCurrentRole	Yes
4	Age	Yes
5	monthly income	Yes
6	YearsWithCurrentManager	Yes

- So nnet based model also has good metrics (sensitivity 0.9584 and F1_Score 0.9117) than 'neuralnet' based model. *Particularly rise of sensitivity by 7% is important*, as it is metric measuring *"number of attritions which are correctly predicted as attrition by model"*. At the same time, *variable-significance* found matches with patterns identified by visual analytics again for 11out of 12 top predictors.

- Lekprofile for top-6 predictors is shown in Fig 5.7 below. It can be used for detailed sensitivity analysis of individual predictors. Similarly, ROC plot is shown in Fig. 5.8:

FIG- 5.7 Lekprofile for top6 predictors 'nnet'

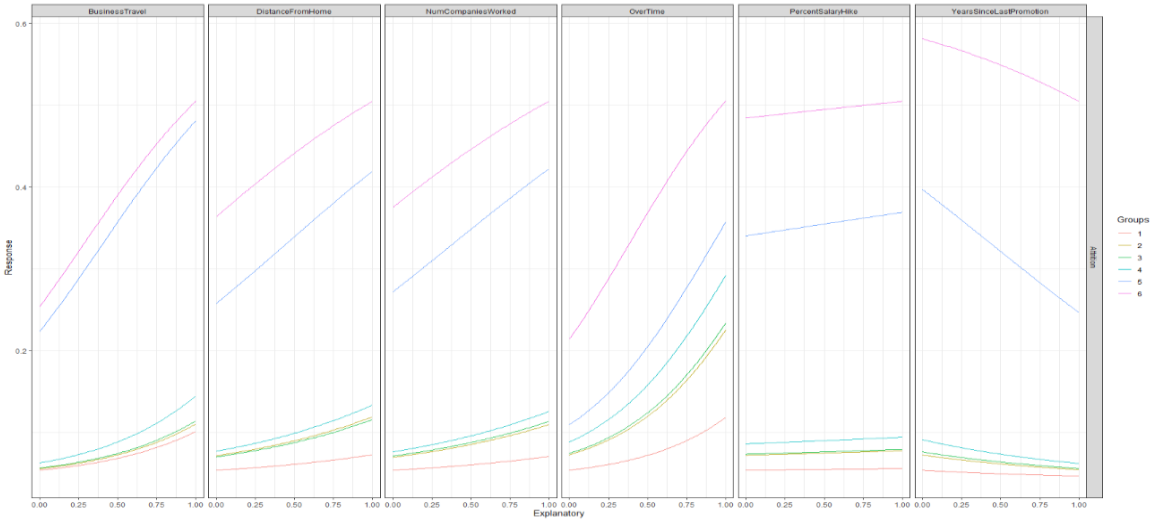
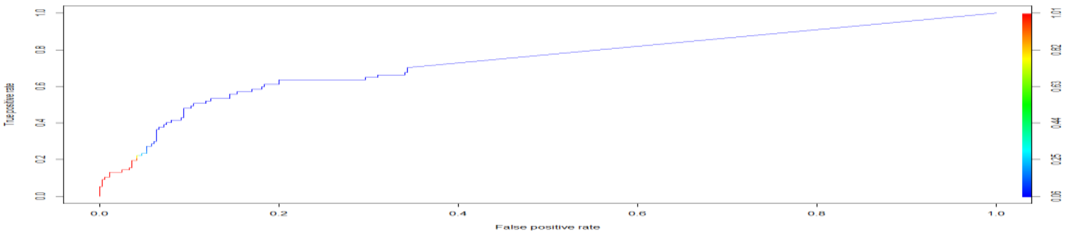


FIG 5.8- ROC plot (tpr Vs fpr) obtained for 'nnet' model



- Comparison of metrics obtained by the best *'neuralnet'* based model and the best *'nnet'* based model is shown below.

TABLE- 5.6- B Comparison of 'neuralnet' based and 'nnet' based model performance

Metric	accuracy	F1	sensitivity	AUC	MAE
'neuralnet'	0.8503	0.9002	0.8811	0.784	0.1496
'nnet'	0.8485	0.9117	0.9584	0.762	0.1519

Thus, for the two-key metrics (F1_Score, Sensitivity) 'nnet' based model gives better results. *Particularly >7% rise of sensitivity is a big improvement for this specific problem.*

5.3. Results obtained by 'caretStack' algorithm and their analysis

- A stacked model using library '*CaretEnsemble*' was made as described in section 4.5. The metrics obtained by using base-classifiers and CaretStack () function are in Table 5.7.

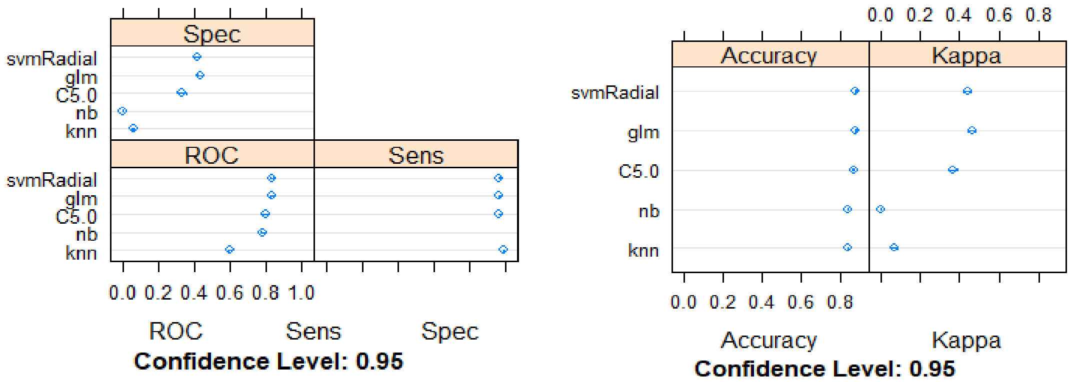
Table- 5.7 Performance of base-classifiers and caret-stack on entire data

Base-Classifier	Median ROC	Median SENSITIVITY	Median SPECIFICITY	Median ACCURACY	Median KAPPA
C 5.0	0.8001	0.9675	0.3483	0.8606	0.3543
NB	0.7870	1.0000	0.0000	0.8390	0.0000
GLM	0.8345	0.9675	0.4331	0.8796	0.4797
KNN	0.6017	0.9839	0.0662	0.8375	0.0730
svmRadial	0.8436	0.9675	0.4261	0.8789	0.4534
Best_RF_Stack	0.8696	0.9692	0.4761	0.9156	0.6476

Table 5.7 shows that very good sensitivity is obtained for all base-classifiers. However NB is perhaps just classifying every employee as a quitter (‘Yes’). Its zero *specificity* and *kappa* point to the same fact. However, when it was tried to remove NB and keep only four base-classifiers, the performance of stack suffered a bit.

- The table also confirms the result of many earlier studies, that *the performance of stacked-model is better than the best of base-classifier*. Dot plots below show above results.

FIG 5.9 Base classifier metrics for caretStack



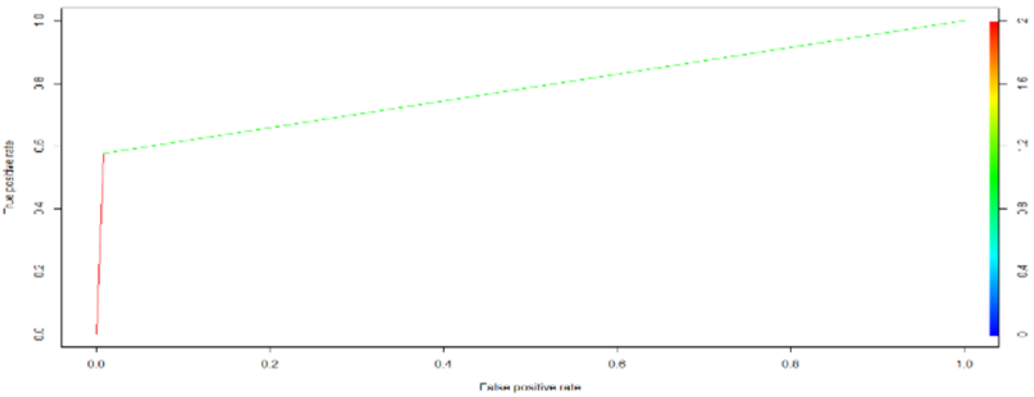
- Table 5.8 below describes the metrics found by best performing stacked model using RF as meta-learner and above variables as base-learners by using it to obtain predictions on test-set while figure 5.10 shows ROC plot for base classifiers.

TABLE 5.8- Metrics found by caretStack RF ensemble

METRIC	VALUE
Confusion Matrix	No Yes # No 361 29 # Yes 8 42
Accuracy	0.9159
95% CI	(0.886, 0.9401)
No Information Rate	0.8386
P-Value [Acc > NIR]	1.422e^{-06}
Kappa	0.6694
Mcnemar's Test P-Value:	0.001009
Sensitivity/recall	0.9783
Specificity	0.5915
Pos Pred Value/precision	0.9256
Neg Pred Value	0.8400
Prevalence	0.8386
Detection Rate	0.8205
Detection Prevalence	0.8864
Balanced Accuracy	0.7849
F1_Score	0.9514

- Table 5.8 shows that excellent results can be obtained by stacking. Compared to 'nnet' based model, accuracy is ≈7% more, Sensitivity is ≈2% more and F1_score is 4% more. *Although computing time and hardware resources needed for training base models and meta-classifier is more for a stacked model*, it does not need much data preparation. However, it needs more programming effort than neural-networks. *Also, the training process for stack was very slow even after* using *parallelisation*.

Fig 5.10 ROC plot [TPR Vs FPR]**for Caret based ensemble stack with RF as meta classifier**



- Although excellent in performance, Caret-Stacked model was found to be more resource-hungry than '*nnet'* based model. It was also more difficult to find variable-significance from this model. It was obtained by filterVarImp () function of caret library. A limitation of this function is that it gives only the absolute value of significance but not its direction. Hence, to compare this variable-significance with two neural-network (NN) based models, the method used was to consider top 6 positive and top 6 negative predictors from NN models and see how many match with stack-predicted top 12 predictors. Still, among the top 12 predictors, 10 out of the top 12 variables (Greyed out in table 5.9) are common between neuralnet-based models and 'caretStack' based model. Further, most of the identified top variables match with patterns identified by visual-analytics.

TABLE-5.9 Variable-Significance by CaretStack

Order	Variable
1	StockOptionLevel
2	marital status
3	NumCompaniesWorked
4	work-life balance
5	JobInvolvement
6	YearsAtCompany
7	YearsSinceLastPromotion
8	Age
9	RelationshipSatisfaction
10	job satisfaction
11	hourly rate
12	job role

- As discussed in point 10 of section 4.5, CaretStack could be used to predict whether an individual *case_employee* will predict or not but still model was found wanting in explainability. Hence H2o based stack was created as described in section 4.5.

5.4 Results obtained by 'h2o_stack' and their analysis

- As discussed in point 12 of section 4.6, once explainer objects for each base-classifier and the stacked classifier are obtained, they were used to obtain comparison plots and box plots for each classifier. These plots are shown in figure 5.11 and 5.12 below. Fig-5.11 shows that all classifiers reach 100% residual (1), but on different paths, indicating that each classifier learns about data in a different manner and maybe each learns different aspects of data. It also shows that stack (green) works in the most efficient manner (indicated by its fast descent and lowest graph in the most part).

FIGURE-5.11 Comparison plot of residuals for h20 stack

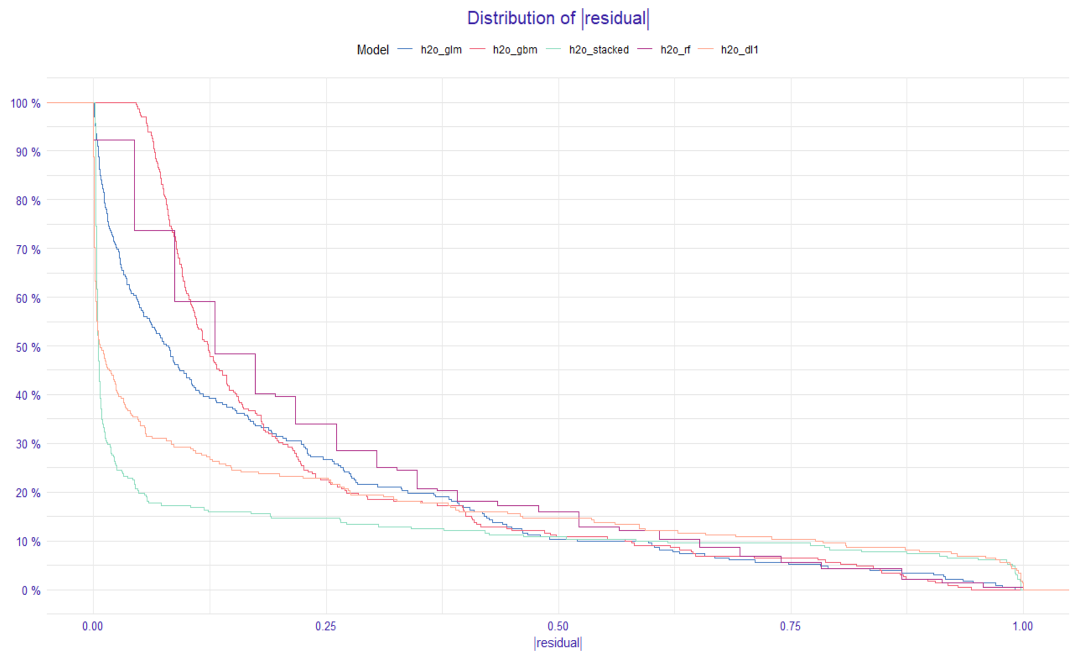
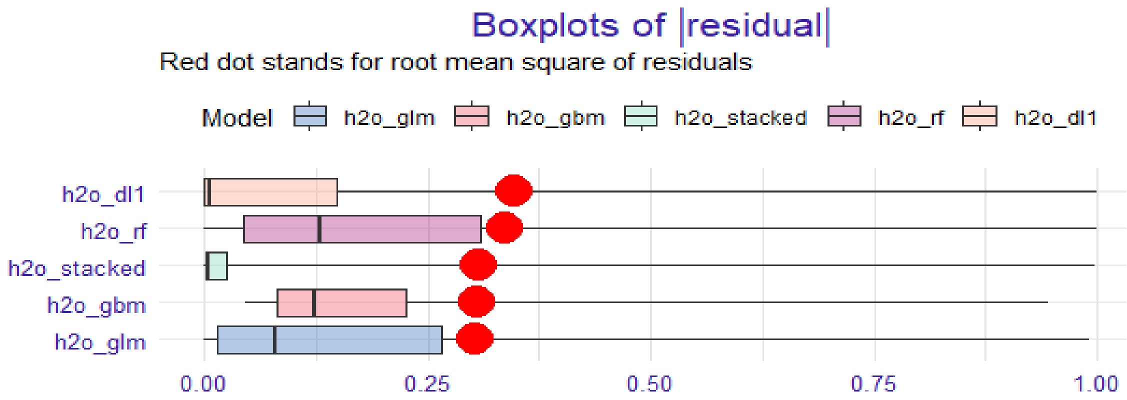


FIGURE-5.12 Box plot of residuals



- Variable Importance plot for all variables is shown in figure 5.13 and top 6 positive and negative predictors are shown in Table- 5.10. Only 7 predictors in Top 12 match with the pattern identified by visual analytics. However, the hugely positive contribution of 'overtime' is noteworthy and matches with the study reviewed earlier.

FIG 5.13: Variable importance plot all variables by h2o stack

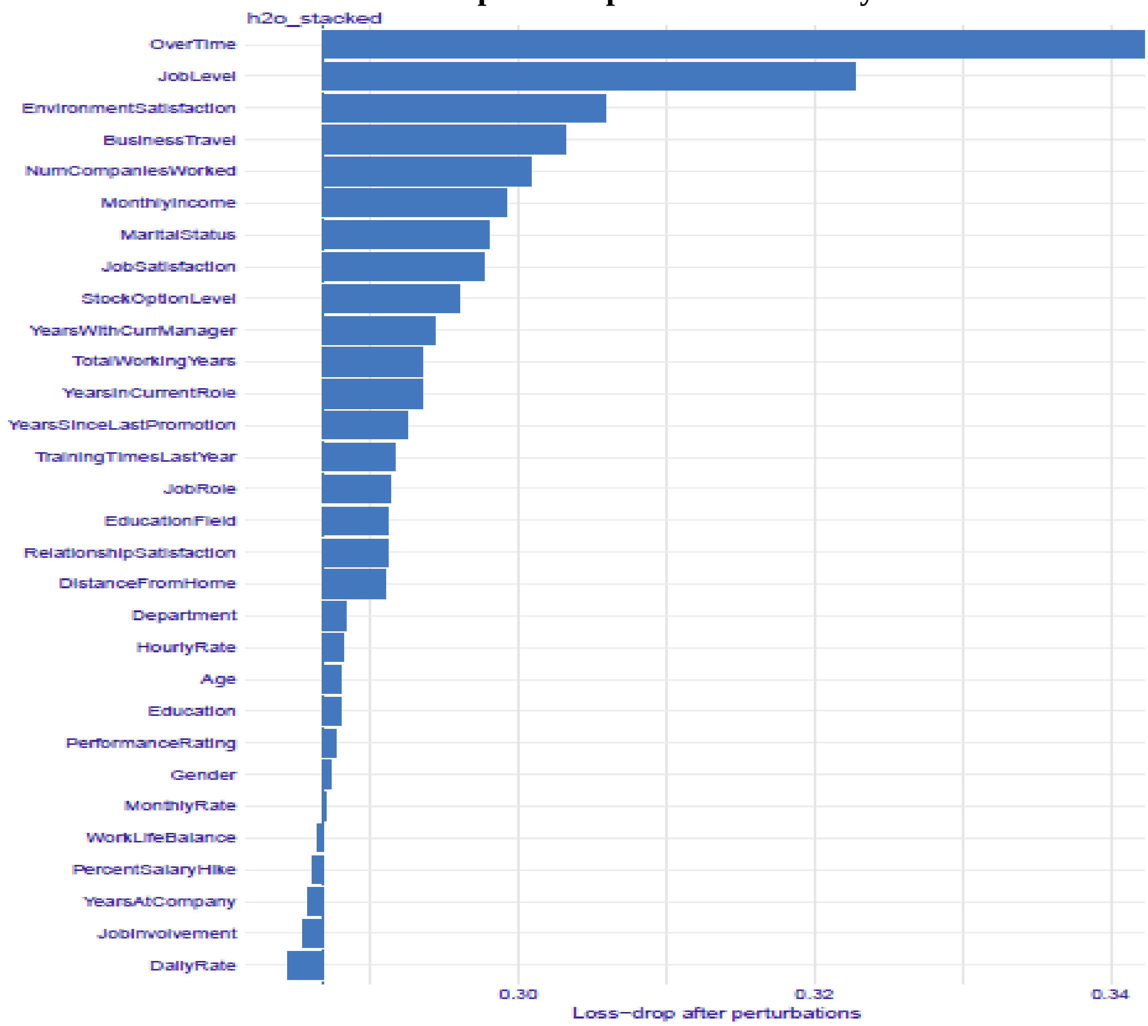


TABLE-5.10 Top 6 positive and negative predictors by h2o-stack

TOP 6 POSITIVE PREDICTORS	IF THEY MATCH WITH PATTERN FOUND BY VISUAL-ANALYTICS		TOP 6 NEGATIVE PREDICTORS	IF THEY MATCH WITH PATTERN FOUND BY VISUAL-ANALYTICS
OverTime	Yes		Gender	Yes
job role	Yes		YearsInCurrentRole	Yes
StockOptionLevel	No		daily rate	No
JobInvolvement	No		YearsWithCurrentManager	Yes
NumCompaniesWorked	Yes		monthly rate	No
monthly income	NO		YearsSinceLastPromotion	Yes

- The real power of this model is realised in predicting if an individual case_employee will quit as described in point 15 of Section 4.6. The prediction (probability 0.171) and top 10 predictors responsible for the *'decision'* (which of course still depends on cut-off chosen by the organisation) along with their relative contribution to cumulative-probability are visually represented nicely as to figure 5.14 shows. Table 5.11 shows the contribution of all 30 predictors for the same case if one wants to dig deeper and study effect of each predictor.

FIG 5.14: The Quit or not decision of a case_employee

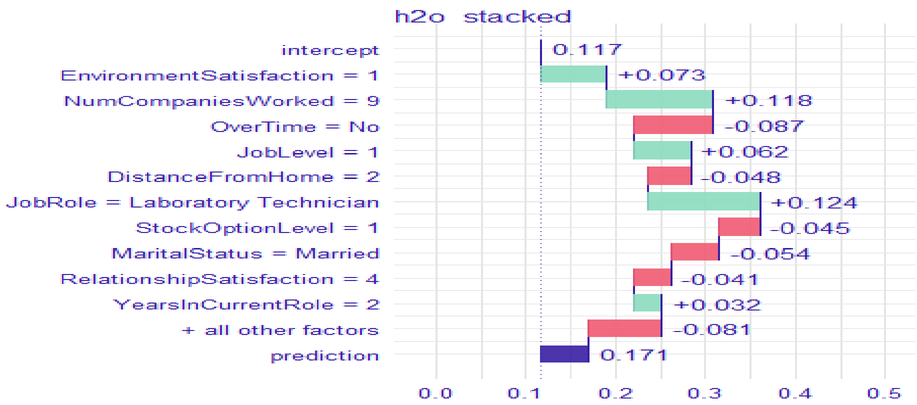


Table 5.11: Digging deeper in individual quit decision

Predictor	Contribution
Intercept	0.117
EnvironmentSatisfaction = 1	0.073
NumCompaniesWorked = 9	0.118
OverTime = No	-0.087
JobLevel = 1	0.062
DistanceFromHome = 2	-0.048
JobRole = Laboratory Technician	0.124
StockOptionLevel = 1	-0.045
MaritalStatus = Married	-0.054
BusinessTravel = Travel_Rarely	-0.003
RelationshipSatisfaction = 4	-0.041
JobSatisfaction = 2	0.008

MonthlyIncome = 3500	-0.030
EducationField = Medical	-0.023
JobInvolvement = 3	-0.022
DailyRate = 590	0.012
Age=27	0.015
Education=1	0.015
YearsAtCompany=2	-0.025
YearsSinceLastPromotion = 2	0.008
Gender = Male	0.008
TrainingTimesLastYear = 3	-0.004
WorkLifeBalance=3	-0.009
PercentSalaryHike = 12	0.006
Department = Research & Development	-0.010
YearsWithCurrManager = 2	-0.002
TotalWorkingYears = 6	0.002
PerformanceRating = 3	-0.003
HourlyRate = 40	-0.009
MonthlyRate = 17000	-0.014
YearsInCurrentRole = 2	0.032
Cumulative Prediction	0.171

- Another interesting analysis is presented in Figure 5.15. It shows how each base model and the stack would have predicted the same employee in a comparative manner. It shows two important facts: *One: For different classifiers the top 5 variables, their order and their contribution changes. This validates conclusion of some earlier studies that variable significance is classifier-dependent and role of old-fashioned EDA and visual analytics should not be underestimated*. *Two:* Although in AUC of best base classifier (GLM 0.8284) and ensemble stack (0.8358) there is an only small difference, the cumulative prediction made by them differ a lot (0.385 for GLM and 0.170 for stack). So, if the cut-off is (say) 0.365 then GLM will predict that employee will quit but the stack will predict employee will not. So, going with higher AUC can totally change prediction. *This underlines the need for using more than one method (like neural network and stack) for mission-critical employees*.

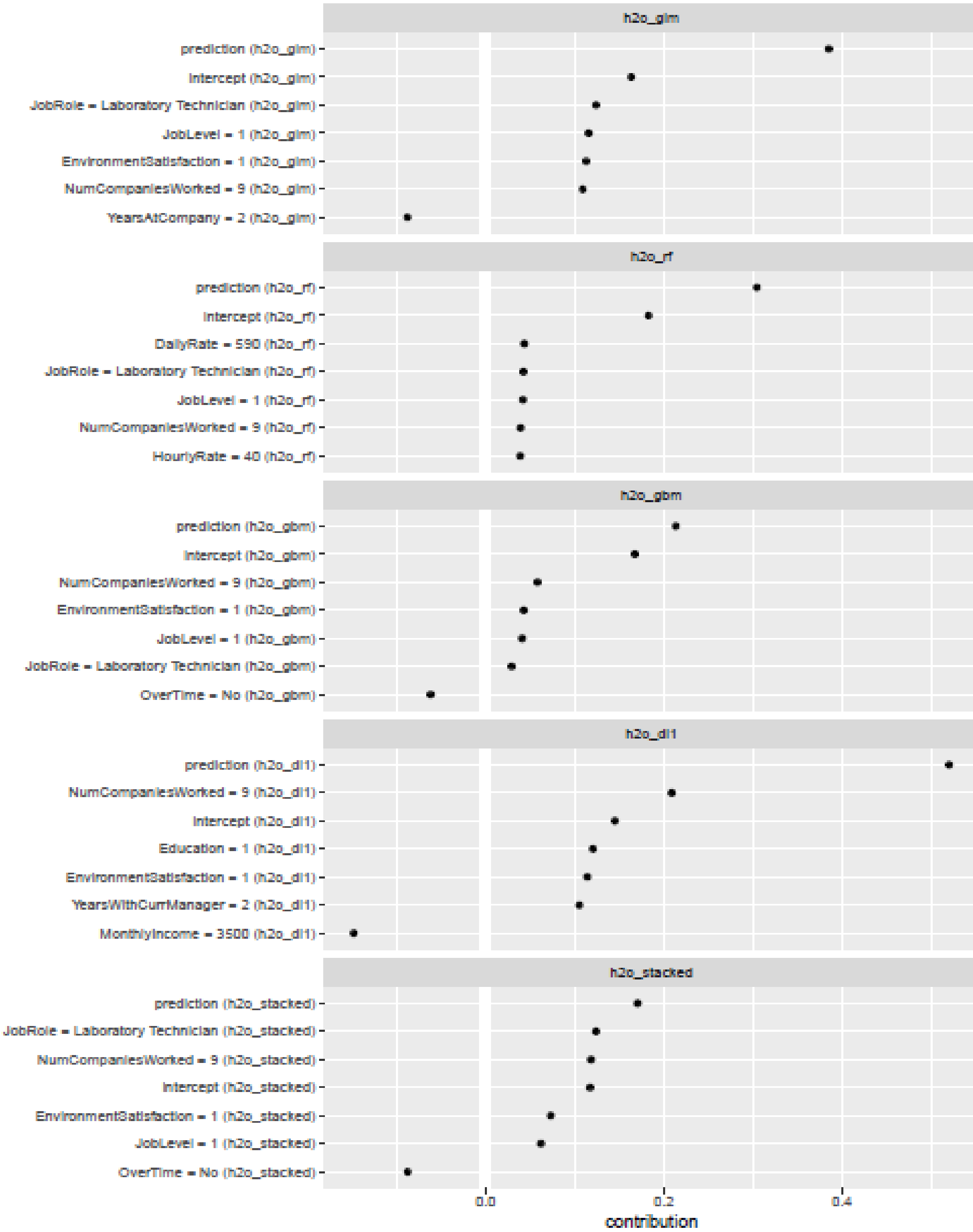
- Comparison of metrics obtained by two stacked-models:

Table 5.12: Comparison of caretStack and h2ostactack

Model	Accuracy	F1	Sensitivity	AUC	KAPPA
CaretStack	0.9199	0.9534	0.9757	0.7363	0.6696
H2oStack	0.8868	0.9340	0.9340	0.8358	0.5377

Thus Caret-Stack is the most accurate model but H2o_Stack perhaps more than adequately compensates by its much superior explainability and interpretability (Although [DALEX]() is model agnostic and can be combined with models created by 'caret' researcher could not find a method to combine it with models created by 'caretStack')

Figure 5.15 Quit/Not decision by each model



Next chapter will be: Conclusions, limitations and future directions.

CHAPTER 6 CONCLUSIONS, LIMITATIONS, FUTURE DIRECTION

6.1 Conclusions:

1. The 'nnet' based model showed that, first obtaining a 'fit' for hyperparameters and then using them in training NN can *simplify architecture, reduce training time, guarantee convergence and improve performance*.
2. Comparison of CaretStack with H2oStack showed that *H2oStack has marginally poor performance but much better explainability and interpretability*. Overall, the study concluded that NN and stack are no longer black-box methods.
3. All four models handled problems like imbalance, outliers, collinearity etc. well. *So, it is worth trying this approach of feeding given data with minimal pre-processing into such models and check performance first*. If satisfactory performance is not obtained, then one can try more pre-processing methods.
4. No feature-engineering or domain-expertise was required showing the power of ANN and stack in specific and the data-driven approach in general.
5. Use of ANN *could find non-linear patterns* which would remain uncaught by many conventional methods. It was perhaps due to this reason the study achieved much better metrics than past studies.
6. The conclusion of earlier studies that "*relative predictor-importance estimates only have relevance within each model, whereas only the rankings (e.g., least, most important) can be compared between models*" [1], was validated but the study could find *more commonality even in this aspect between models than previous studies*. E.g. both NN models' top-12 predictors were common with patterns identified by visual analytics, which underlined importance of visual analytics and EDA as supporting tools. In caret-stack 10 out of 12 matched with top-12 found by NN. Only the H2o stack had poor performance in finding variable-significance.

7. Although ANN and Stack belong to two totally different classes of models and it is not fair to compare ANN with Stack, the metrics obtained in the study are in Table-6.1; showing that stacking is a powerful prediction method.

TABLE-6.1 ALL FOUR MODELS COMPARED

Metric	accuracy	F1	Sensitivity	AUC
'neuralnet'	0.8503	0.9002	0.8811	0.7840
'nnet'	0.8485	0.9117	0.9584	0.7620
CaretStack	0.9199	0.9534	0.9757	0.7363
H2oStack	0.8868	0.9340	0.9340	0.8358

6.2 Limitations:

1. A very 'simple' dataset was used in this study. It was structured, tabular, static, mid-sized and free from NAs, NULL, duplicate rows, etc. The models trained on such a data may not perform equally well on bigdata, streaming data, unstructured data or 'dirty' data. However, in literature-review two studies 2.2.2.4 and 2.2.2.9 were seen where model performance for larger real-world datasets was also good. Further study of Yue_Zhao et al. proved that *data-size and source do not affect* ROC. *So, maybe more complex datasets would need more preprocessing but basic techniques discussed here would be still usable*.
2. The explainability of two NN based model and the caret-based model was not as good as H2o-based ensemble. The researcher could not explore more packages like 'IML' and 'LIME' due to time constraints. Similarly, for stacking 'SUPERLEARNER' and H2o's 'AUTOML' packages could not be explored.
3. The 'neuralnet' based model could have been tuned for better performance subject to availability of time.

6.3 Contribution to knowledge:

1. Study successfully showed the power of stacking which needs little data preprocessing but gives excellent metrics (Table-6.1). Further H2o Stack also provided much better explainability and interpretability.

2. The study also got much better metrics (mainly sensitivity and F1_Score) using ANN than many earlier ANN based studies. The variable significance, hyperparameter tuning and sensitivity analysis (again all of which touch the explainability and interpretability aspect) are also explained much clearly here.
3. For added rigour and extra validation results of modern methods were compared with those of visual analytics.
4. The study drills down *up to the level of individual employee* and also *to the level of attrition vs each individual predictor* and *also explains how nonlinearity can be detected by generalized neural weights*.

6.4 FUTURE-DIRECTIONS:

1. A Shiney-app and dashboard can be developed to hide lower-level code from HR-Managers and to make models more user-friendly.
2. Models can be validated on other data frames to see if it can handle the 5 V's (Volume, velocity, variety, veracity and value) of big data.
3. More detailed sensitivity-analysis by 'lekprofile' and 'PDP' can be done.
4. H2o Stack had marginally poor metrics than caret-stack and variable-significance given by it also differed more from the other three models. So, a question *"Does increase in interpretability and explainability come at the cost of accuracy?"* can be of interest for expert data scientists. There is already considerable literature available on this vast subject including journals and even entire books. [75] [76]

REFERENCES OR BIBLIOGRAPHY

I. REFERENCES-I: LITERATURE-REVIEW FOR ML BASED POST-2008 STUDIES

1) Beck, M.W., 2018. NeuralNetTools: Visualization and Analysis Tools for Neural Networks. Journal of Statistical Software; Vol 1, Issue 11, 2018,

https://doi.org/10.18637/jss.v085.i11

2) Chen, G., Ployhart, R.E., Thomas, H., Anderson, N. and Bliese, P.D. (2011) The Power of Momentum: A New Model of Dynamic Relationships between Job Satisfaction Change and Turnover Intentions. Academy of Management Journal, 54, 159-181. http://dx.doi.org/10.5465/AMJ.2011.59215089

3) Deep Sanghavi, Jay Parekh, Shaunak Sompura, Pratik Kanani, "Data Visualisation and Improving Accuracy of Attrition Using Stacked Classifier", International Journal of Engineering Development and Research (IJEDR), ISSN:2321-9939, Volume.6, Issue 4, pp.284-293, November 2018, Available at: http://www.ijedr.org/papers/IJEDR1804054.pdf

4) Džeroski, S., Ženko, B., 2004. Is combining classifiers with stacking better than selecting the best one? Machine Learning 54, 255–273. https://doi.org/10.1023/ B: MACH.0000015881.36452.6e

5) Fan, C.-Y., Fan, P.-S., Chan, T.-Y., Chang, S.-H., 2012. Using hybrid data mining and machine learning clustering analysis to predict the turnover rate for technology professionals. Expert Systems with Applications 39, 8844–8851. https://doi.org/10.1016/j.eswa.2012.02.005

6) Felps, W., Mitchell, T.R., Hekman, D.R., Lee, T.W., Holtom, B.C., Harman, W.S., 2009. Turnover Contagion: How Co-workers' Job Embeddedness and Job Search Behaviours Influence Quitting. AMJ 52, 545–561. https://doi.org/10.5465/amj.2009.41331075

7) Frauke Günther and Stefan Fritsch,2010, volume 2:1, pages 30-38. "neuralnet: Training of Neuralnetworks", The R journal Vol. 2, June 2010, ISSN:2073-4859, https:/ /doi.org/10.32614/RJ-2010-00

8) Funda Güneş, Russ Wolfinger, and Pei-Yi Tan, 2017, Stacked Ensemble Models for Improved Prediction Accuracy, Paper SAS-2017, SAS Institute Inc.

9) Günther, F., Fritsch, S., 2010. neuralnet: Training of Neural Networks. The R Journal 2,

30–38. https://doi.org/10.32614/RJ-2010-006

10) HMN Yousaf, 2016. "Analysing which factors are of influence in predicting the employee turnover". Research Paper Business Analytics. Faculty of Sciences. 1081 HV Amsterdam.
11) Hom, P.W., Tsui, A.S., Wu, J.B., Lee, T.W., Zhang, A.Y., Fu, P.P., Li, L., 2009. Explaining employment relationships with social exchange and job embeddedness. Journal of Applied Psychology 94, 277–297. https://doi.org/10.1037/a0013453
12) Jain, D., Buckley, M.B.,2017. Evaluation of Employee Attrition by Effective Feature Selection using Hybrid Model of Ensemble Methods MSc Research Project Data Analytics, School of Computing, National College of Ireland.
13) Kaiming Ting, Ian H Witten, 1999, Journal of Artificial Intelligence Research, Volume 10, Issue 1, 271-289, ISSN 1076-9757, https://doi.org/10.1613/jair.594

14) Tud.ttu.ee. (2020). [online] Available at: http://www.tud.ttu.ee/im/Vladimir.Viies/materials/L%C3%95PETAMINE/17MAGhannesele/Thesis_Mari_Maisuradze.pdf [Accessed 12 Feb. 2020].

15) Nyberg, A., 2010. Retaining your high performers: Moderators of the performance –job satisfaction– voluntary turnover relationship. Journal of Applied Psychology 95-3, 440–453. https://doi.or g/10.1037/a0018869.
16) Quinn, A., Rycraft, J.R., Schoech, D., 2002. Building a Model to Predict Caseworker and Supervisor Turnover Using a Neural Network and Logistic Regression. Journal of Technology in Human Services 19, 65–85. https://doi.org/10.1300/J017v19v04_05

17) Sexton, R.S., McMurtrey, S., Michalopoulos, J.O., Smith, A.M., 2005. Employee-turnover : a neural network solution. Computers & Operations Research 32, 2635–2651. https://doi.org/10.1016/j.cor.2004.06.022
18) Somers, M., 1999. Application of two neural network paradigms to the study of voluntary employee turnover. The Journal of applied psychology 84, 177–85. https://doi.org/10.1037//0021-9010.84.2.177
19) Spackman, K.A., 1992. Combining logistic regression and neural networks to create predictive models. Proc Annual Symposium Computer Applied Medical Care 456–459.
20) Trevor, C.O., Nyberg, A.J., 2008. Keeping Your Headcount When All About You Are Losing Theirs: Downsizing, Voluntary Turnover Rates, and The Moderating Role of HR Practices. AMJ 51, 259–276. https://doi.org/10.5465/amj.2008.31767250

21) U. Soni, N. Singh, Y. Swami, P. Deshwal, 2018. A Comparison Study between ANN and ANFIS for the Prediction of Employee Turnover in an Organization. Presented at the 2018 International Conference on Computing, Power and Communication Technologies (GUCON), pp. 203–206. https://doi.org/10.1109/GUCON.2018.8674886

22) Wolpert, D.H., 1992. Stacked generalization. Neural Networks 5, 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

23) ZEHRA ÖZGE KISAOGLU, Sept,2014. "Employee-turnover prediction using machine learning based methods" Master Thesis for MS in Computer Engineering. GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES OF MIDDLE EAST TECHNICAL UNIVERSITY. http://etd.lib.metu.edu.tr/upload/12617686/index.pdf

24) Zhao, Y., Hryniewicki, M., Cheng, F., Fu, B., Zhu, X., 2018. Employee Turnover Prediction with Machine Learning: A Reliable Approach. https://doi.org/10.1007/978-3-030-01057-7

II. REFERENCES-II: FOR NON-ML BASED METHODS AND PRE-2008 METHODS

25) Abelson, M.A., Sheridan, J.E., 1983. Cusp Catastrophe Model of Employee Turnover. The Academy of Management Journal 26, 418–436. https://doi.org/10.2307/256254

26) Allen, D.G., Griffeth, R.W., 2001. Test of a mediated performance–turnover relationship highlighting the moderating roles of visibility and reward contingency. Journal of Applied Psychology 86, 1014–1021. https://doi.org/10.1037/0021-9010.86.5.1014

27) Allen, D.G., Weeks, K.P., Moffitt, K.R., 2005. Turnover Intentions and Voluntary Turnover: The Moderating Roles of Self-Monitoring, Locus of Control, Proactive Personality, and Risk Aversion. Journal of Applied Psychology 90, 980–990. https://doi.org/10.1037/0021-9010.90.5.980

28) Arthur Jr., W., Bell, S.T., Villado, A.J., Doverspike, D., 2006. The use of person-organization fit in employment decision making: An assessment of its criterion-related validity. Journal of Applied Psychology 91, 786–801. https://doi.org/10.1037/0021-9010.91.4.786

29) Barrick, M., Zimmerman, R., 2005. Reducing Voluntary, Avoidable Turnover Through Selection. The Journal of applied psychology 90, 159–66. https://doi.org/10.1037/0021-9010.90.1.159

30) Barrick, M.R., Mount, M.K., 1996. Effects of impression management and self-deception on the predictive validity of personality constructs. Journal of Applied Psychology 81, 261–272. https://doi.org/10.1037/0021-9010.81.3.261

31) Bauer, T.N., Erdogan, B., Liden, R.C., Wayne, S., 2006. A Longitudinal Study of the Moderating Role of Extraversion: Leader-Member Exchange, Performance, and Turnover During New Executive Development. Article in Journal of Applied Psychology. https://doi.org/10.1037/0021-9010.91.2.298

32) Beck, M.W., 2018. NeuralNetTools: Visualization and Analysis Tools for Neural Networks. Journal of Statistical Software; Vol 1, Issue 11 (2018).

33) Bloom, M., Michel, J.G., 2002. The Relationships Among Organizational Context, Pay Dispersion, and Among Managerial Turnover. AMJ 45, 33–42. https://doi.org/10.5465/3069283.

34) Burton, M.D., Beckman, C.M., 2007. Leaving a Legacy: Position Imprints and Successor Turnover in Young Firms. American Sociological Review 72, 239–266. https://doi.org/10.1177/000312240707200206

35) Chen, X.-P., Hui, C., Sego, D., 1998. The Role of Organizational Citizenship Behavior in Turnover: Conceptualization and Preliminary Tests of Key Hypotheses. [Article]. Journal of Applied Psychology December 1998;83(6):922-931 83. https://doi.org/10.1037/0021-9010.83.6.922

36) Crossley, C.D., Bennett, R.J., Jex, S.M., Burnfield, J.L., 2011. "Development of a global measure of job embeddedness and integration into a traditional model of voluntary turnover": Clarification to Crossley et al. (2007). Journal of Applied Psychology 96, 1316–1316. https://doi.org/10.1037/a0025569

37) Donnelly, D., Quirin, J., 2006. An extension of Lee and Mitchell's unfolding model of voluntary turnover. Journal of Organizational Behavior 27, 59–77. https://doi.org/10.1002/job.367

38) Eisenberger, R., Stinglhamber, F., Vandenberghe, C., Sucharski, I.L., Rhoades, L., 2002. Perceived supervisor support: Contributions to perceived organizational support and employee retention. Journal of Applied Psychology 87, 565–573. https://doi.org/10.1037/0021-9010.87.3.565

39) Elvira, M.M., Cohen, L.E., 2001. Location Matters: A Cross-Level Analysis of the Effects of Organizational Sex Composition on Turnover. AMJ 44, 591–605. https://doi.org/10.5465/3069373

40) Friedman, R.A., Holtom, B., 2002. The effects of network groups on minority employee turnover intentions. Human Resource Management 41, 405–421. https://doi.org/10.1002/hrm.10051

41) Graen, G., Liden, R., Hoel, W., 1982. Role of leadership in employee withdrawal process. Journal of Applied Psychology 67, 868–872. https://doi.org/10.1037/0021-9010.67.6.868

42) Griffeth, R.W., Allen, D.G., Steel, R.P., Bryan, N., 2005. The development of a multidimensional measure of job market cognitions: The Employment Opportunity Index (EOI). Journal of Applied Psychology 90, 335–349. https://doi.org/10.1037/0021-9010.90.2.335

43) Harter, J.K., Schmidt, F.L., Hayes, T.L., 2002. Business-unit-level relationship between employee satisfaction, employee engagement, and business outcomes: A meta-analysis. Journal of Applied Psychology 87, 268–279. https://doi.org/10.1037/0021-9010.87.2.268

44) Holtom, B.C., Lee, T.W., Tidd, S.T., 2002. The relationship between work status congruence and work-related attitudes and behaviours. Journal of Applied Psychology 87, 903–915. https://doi.org/10.1037/0021-9010.87.5.903

45) Holtom, B.C., Mitchell, T.R., Lee, T.W., Inderrieden, E.J., 2005. Shocks as causes of turnover: What they are and how organizations can manage them. Human Resource Management 44, 337–352. https://doi.org/10.1002/hrm.20074

46) Hom, P.W., Roberson, L., Ellis, A.D., 2008. Challenging conventional wisdom about who quits: revelations from corporate America. Journal of Applied Psychology 93, 1–34. https://doi.org/10.1037/0021-9010.93.1.1

47) Kammeyer-Mueller, J.D., Wanberg, C.R., Glomb, T.M., Ahlburg, D., 2005. The Role of Temporal Shifts in Turnover Processes: It's About Time. Journal of Applied Psychology 90, 644–658. https://doi.org/10.1037/0021-9010.90.4.644

48) Koys, D.J., 2001. The effects of employee satisfaction, organizational citizenship behavior, and turnover on organizational effectiveness: A unit-level, longitudinal study. Personnel Psychology 54, 101–114. https://doi.org/10.1111/j.1744-6570.2001.tb00087.x

49) Lee, T.W., Mitchell, T.R., Holtom, B.C., McDaneil, L.S., Hill, J.W., 1999. The Unfolding Model of Voluntary Turnover: A Replication and Extension. AMJ 42, 450–462. https://doi.org/10.5465/257015

50) Lee, T.W., Mitchell, T.R., Wise, L., Fireman, S., 1996. An Unfolding Model of Voluntary Employee Turnover. The Academy of Management Journal 39, 5–36. https://doi.org/10.2307/256629

51) Maertz Jr, C.P., Griffeth, R.W., Campbell, N.S., Allen, D.G., 2007. The effects of perceived organizational support and perceived supervisor support on employee turnover. Journal of Organizational Behavior 28, 1059–1075. https://doi.org/10.1002/job.472

52) Maertz, C.P., Campion, M.A., 2004. PROFILES IN QUITTING: INTEGRATING PROCESS AND CONTENT TURNOVER THEORY. Academy of Management Journal 18.

53) March, J.G., Simon, H.A., 1958. Organizations., Organizations. Wiley, Oxford, England

54) Meyer, J., Stanley, D., Herscovitch, L., Topolnytsky, L., 2002. Affective, Continuance, and Normative Commitment to the Organization: A Meta-analysis of Antecedents, Correlates, and Consequences. Journal of Vocational Behavior 61, 20–52. https://doi.org/10.1006/jvbe.2001.1842

55) Mirvis, P.H., Lawler III, E.E., 1984. Accounting for the quality of work life. Journal of Organizational Behavior 5, 197–212. https://doi.org/10.1002/job.4030050304

56) Mitchell, T.R., Holtom, B.C., Lee, T.W., Sablynski, C.J., Erez, M., 2001. Why People Stay: Using Job Embeddedness to Predict Voluntary Turnover. AMJ 44, 1102–1121. https://doi.org/10.5465/3069391

57) Mobley, W.H., 1977. Intermediate linkages in the relationship between job satisfaction and employee turnover. Journal of Applied Psychology 62, 237–240. https://doi.org/10.1037/0021-9010.62.2.237.

58) Mobley, W.H., 1982. Some Unanswered Questions in Turnover and Withdrawal Research. AMR 7, 111–116. https://doi.org/10.5465/amr.1982.4285493

59) Dalton, D.R., Todor, W.D., 1979. Turnover Turned over: An Expanded and Positive Perspective. The Academy of Management Review 4, 225–235, https://doi.org/10.2307/257776.

60) Pfeffer, J., 1985. Organizational Demography: Implications for Management. California Management Review 28, 67–81. https://doi.org/10.2307/41165170

61) Porter, L. W., & Steers, R. M. (1973). Organizational, Work, and Personal Factors in Employee Turnover and Absenteeism. Psychological Bulletin, 80, 151-176. https://doi.org/10.1037/h0034829

62) PRICE, J.L., 1989. The Impact of Turnover on the Organization. Work and Occupations 16, 461–473. https://doi.org/10.1177/0730888489016004005

63) Price, J.L., Mueller, C.W., 1981. A Causal Model of Turnover for Nurses. The Academy of Management Journal 24, 543–565. https://doi.org/10.2307/255574. Also, P-36, ch-2, https://shodhganga.inflibnet.ac.in/bitstream/10603/18571/13/13_chapter%202.pdf

63-b) Peter W. Hom, David G. Allen, Rodger W. Griffeth. “Employee Retention and Turnover: Why Employees Stay or Leave. 1st Edition - ISBN 9781138503816, CRC Press. (P-183).

64) Rafferty, A., Griffin, M., 2006. Perceptions of Organizational Change: A Stress and Coping Perspective. The Journal of applied psychology 91, 1154–62. https://doi.org/10.1037/0021-9010.91.5.1154

65) Salamin, A., Hom, P., 2005. In Search of the Elusive U-Shaped Performance-Turnover Relationship: Are High Performing Swiss Bankers More Liable to Quit? The Journal of applied psychology 90, 1204–16. https://doi.org/10.1037/0021-9010.90.6.1204

66) Settoon, R.P., Mossholder, K.W., 2002. Relationship quality and relationship context as antecedents of person- and task-focused interpersonal citizenship behavior. Journal of Applied Psychology 87, 255–267. https://doi.org/10.1037/0021-9010.87.2.255

67) Simons, T.L., Roberson, Q., 2003. Why Managers Should Care About Fairness: The Effects of Why Managers Should Care About Fairness: The Effects of Aggregate Justice Perceptions on Organizational Outcomes Aggregate Justice Perceptions on Organizational Outcomes. https://doi.org/10.1037/0021-9010.88.3.432

68) Sims, C., Drasgow, F., Fitzgerald, L., 2005. The Effects of Sexual Harassment on Turnover in the Military: Time-Dependent Modeling. The Journal of applied psychology 90, 1141–52. https://doi.org/10.1037/0021-9010.90.6.1141

69) Staw, B.M., 1980. The Consequences of Turnover. Journal of Occupational Behaviour 1, 253–273.

70) Steel, R.P., Lounsbury, J.W., 2009. Turnover process models: Review and synthesis of a conceptual literature. Human Resource Management Review. https://doi.org/10.1016/j.hrmr.2009.04.002

71) Sturman, M.C., Trevor, C.O., 2001. The implications of linking the dynamic performance and turnover literatures. Journal of Applied Psychology 86, 684–696. https://doi.org/10.1037/0021-9010.86.4.684

72) Trevor, C., 2001. Interactions among Actual Ease-of Movement Determinants and Job Satisfaction in the Prediction of Voluntary Turnover. Academy of Management Journal 44, 621–638. https://doi.org/10.2307/3069407

73) Vandenberghe, C., Bentein, K., Stinglhamber, F., 2004. Affective commitment to the organization, supervisor, and work group: Antecedents and outcomes. Journal of Vocational Behavior 64, 47–71. https://doi.org/10.1016/S0001-8791(03)00029-0

74) Wanberg, C.R., Banas, J.T., 2000. Predictors and outcomes of openness to changes in a reorganizing workplace. Journal of Applied Psychology 85, 132–142. https://doi.org/10.1037/0021-9010.85.1.132

REFERENCES-III: MISCELLANEOUS REFERENCES

75) Patrick Hall and Navdeep Gil, (2018), An Introduction to Machine Learning Interpretability. Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

76) Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F. & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*. 51 (5).

77) Hennig, P., Kiefel, M., (2013), Quasi-Newton Methods: A New Direction. Journal of Machine Learning Research 14, 843-865.

78) Some reliable web-links:

a. Ocean Tomo. (2020). *Intangible Asset Market Value Study | Ocean Tomo*. [online] Available at: https://www.oceantomo.com/intangible-asset-market-value-study/ [Accessed 12 Feb. 2020].

b. Bls.gov. (2020). [online] Available at: https://www.bls.gov/news.release/pdf/tenure.pdf [Accessed 12 Feb. 2020].

c. Futureofwork.kornferry.com. (2020). [online] Available at: https://futureofwork.kornferry.com/wp-content/uploads/2018/04/Furture-of-Work-Talent-Crunch-online.pdf [Accessed 12 Feb. 2020].

d. Talentkeepers.com. (2020). [online] Available at: https://www.talentkeepers.com/wp-content/uploads/2015/03/Talentkeepers-The-Dirty-Truth-Employee-Turnover-Cost-Whitepaper.pdf [Accessed 12 Feb. 2020].

e. PositivePsychology.com. (2020). *Big Five Personality Traits: The OCEAN Model Explained [2019 Upd.]*. [online] Available at: https://positivepsychology.com/big-five-personality-theory/ [Accessed 12 Feb. 2020].

79) Links for Data and its License used in thesis

A. IBM LICENSE PAGE LINK:

a) IBM Developer. (2020). *Data science process pipeline to solve employee attrition*. [online] Available at: https://developer.ibm.com/patterns/data-science-life-cycle-in-action-to-solve-employee-attrition-problem/ [Accessed 12 Feb. 2020].

B. DATA-HOSTED AT:

b) Kaggle.com. (2020). *IBM HR Analytics Employee Attrition & Performance*. [online] Available at: https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset [Accessed 12 Feb. 2020].

80) LINK TO ALL CODE FILES ON GOOGLE-DRIVE:

APPENDICES

APPENDIX-I The pre 2008 era of attrition research

For studies of pre-1995 era, here only prominent studies are listed (Fig-I-I)

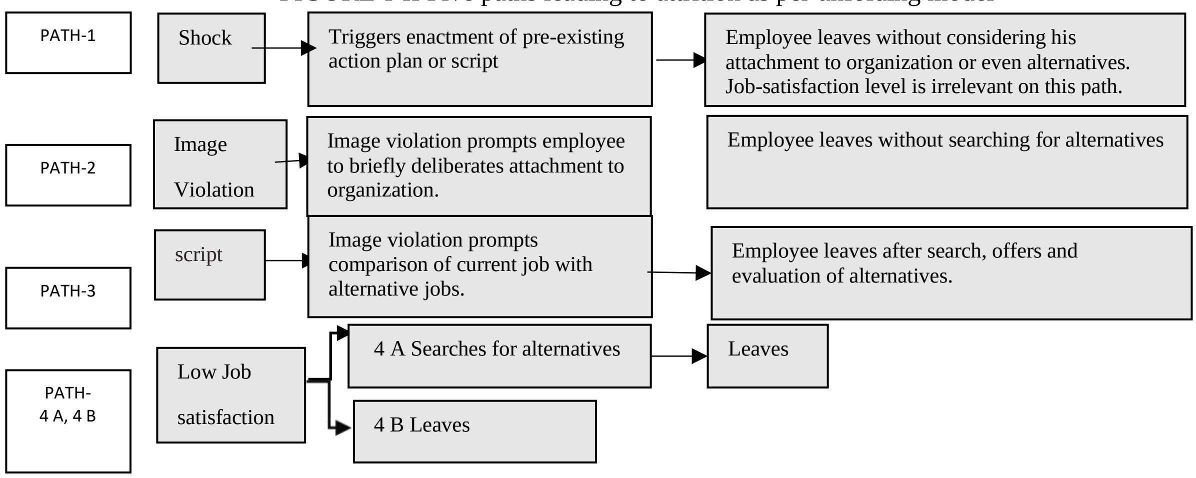
FIGURE-I-I Important studies from pre-1995 era

- 'Theory of organizational equilibrium' by March and Simon (1986) [53]
- 'Theory of met expectations' by Porter and Steers (1973) [61]
- 'Intermediate linkages model' by Mobley (1977) [57]
- 'Structural model' by Price and Muller (1981 to 1986) [63]
- 'Cusp catastrophe model' by Sheridan and Abelson (1983) [25]
- A model by Graen, Liden, and Hoel (1982) introducing new variable 'quality of leader–member exchange relationship' [41]
- A model by Pfeffer (1983) introducing new variable 'demographic fit' [60]
- Models focussing on consequences of attrition (e.g., Dalton & Todor [59], 1979; Mirvis & Lawler [55], 1977; Mobley, 1982[58]; Staw, 1980[69])

Progress in main research-directions of pre-1994 era during 1995-2008:

A landmark event in attrition research was 'unfolding model'[50][49], proposed by Lee and Mitchell's (1994). It mentioned 5 paths leading to attrition. Major components of this model were: '*shocks*', '*scripts*', '*image-violations*', '*job-dissatisfaction*' and '*job-search*'. Fig I-II shows these 5 paths by 1,2,3,4-A and 4-B and how they lead to attrition in different manners**.**

FIGURE-I-II Five paths leading to attrition as per unfolding model



In addition to 'unfolding-model' attrition studies before 1995 identified some major research-directions. In decade of 1995-2005 each trend was further expanded as listed in table I-I:

Table I.I: Major research-directions and brief introduction to works in each direction

Name Of researchers	Conclusion/summary
Research Direction -1 prediction models increasingly started including *individual differences among employees* in addition to *organizational factors*	
Barrick and Zimmerman (2005) [29]	Proposed that *during recruitment process itself, certain individual qualities can be measured which negatively correlate with attrition.* Still today it is confirmed by major HR experts that most important reason for attrition-problem is *bad hiring strategy in first place.*
Barrick and Mount (1996) [30]	Among the big 5 personality constructs[g], 'conscientiousness' *negatively relates to attrition*.
Maertz and Campion (2004) [52]	Combined various *'content-models' and 'process-models' of attrition. Proposed 'eight attrition-motives'* and suggested that these are related to '*four types of quitters'* and also proposed that each group of quitters is driven by dissimilar forces. Claimed to have identified '*eight nearby causes of attrition-cognitions' which are the best predictors of attrition* and suggested that *these eight causes mediate the effects of all other main constructs in the literature*.
Bauer, Erdogan, Liden, and Wayne (2006) [31]	Proposed that *there is negative correlation between Leader-Member- Exchange (LMX) and attrition*.
Research Direction -2 The studies on *effect of work-stress and employee's adaptability to change* came to forefront of research	

Name Of researchers	Conclusion/summary
Wanberg & Banas, 2000 [74]	Proved that *employees with high adaptation to changes, had more job-satisfaction and hence less attrition*.
Rafferty & Griffin, 2006 [64]	Showed that *frequent, badly planned, or transformational changes induce uncertainty which may lead to attrition*. This emphasizes the fact that effective retention management is required during the time of big organizational transformation.
Sims, Drasgow, and Fitzgerald (2005) [68]	Showed that *experiences with sexual harassment is strong predictor of attrition even in jobs with high job satisfaction*
Research direction -3: Studies on the unfolding model.	
Lee, Mitchell, Wise & Fireman, 1996; [50]	Tested 'unfolding model' empirically for the first time and *demonstrated that up to 91% people in their model do follow one of above 5 paths* (Figure-2.3) when quitting
Holtom, Mitchell, Lee & Inderrieden, (2005) [45]	Reported that, *more often shocks are immediate cause of attrition than job-dissatisfaction.*
Donnelly and Quirin (2006) [37]	This study did independent tests on unfolding model. Their important conclusions were: *One*: Economic consequences are more important to path 2 and 4B leavers than leavers in other paths. *Two*: 33% of leavers and 83% stayers indicated economic considerations are important to their decision. *Three*: Women experience more shocks than men and follow paths 1, 2 and 3 more.
Research direction -4: Increased limelight on "contextual variables" and "interpersonal relationships"	
Harter, Schmidt, & Hayes (2002) [43]	Reported that *low employee-satisfaction aggregated at unit level leads to higher attrition and vice-versa*.
Koys (2001) [48]	Reported that *unit-level attrition from one year, negatively predicted customer satisfaction and unit profit in subsequent year*.
Elvira and Cohen (2001) [39]	Found that *while determining the effect of gender-diversity on attrition; percentage of employees of one's own gender at various strata of organization is crucial*.
Bloom and Michel (2002) [33]	Found that a *firm's salary-distribution affects attrition*. Low salary-differentiation prompts outstanding employees to leave a company.
Eisenberger et al. (2002) [38] and Maertz et al (2007) [51]	A new variable called "Perceived Organizational Support" (POS) was introduced in studies that indicated that POS *had significant impact on attrition.*
Hom et al. (2007) [46]	In a comprehensive study of 20 US corporations with more than 4,50,000 professionals and managers, found that *incumbents of jobs that are typically held by a greater number of African Americans and Hispanics were at a higher risk of attrition*.
Friedman and Holtom (2002) [40]	Examined effects of minority network groups on minority attrition and *confirmed importance of social-embeddedness in predicting attrition*.

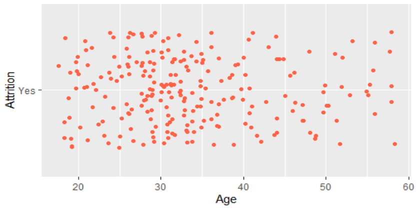
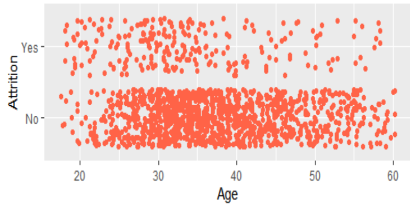
Name Of researchers	Conclusion/summary
Simons and Roberson (2003) [67]	Proved that *significant and sequential linkages exist from 'procedural and interactional justice' to 'employee commitment' to 'intention to remain' and 'attrition'*
Bauer et al., 2006 [31] and Arthur et al. 2006 [28]	*Did more studies on variables 'LMX' and 'P-O fit'*. Found that '*P–O*' fit predicts attrition but its effect becomes half when mediated by job-attitudes and cognitions
Holtom, Lee, & Tidd (2002) [44]	Showed that *flexible working hours leads to reduced attrition*.
Mossholder et al., 2005 [66], Chen, Hui & Sego, (1998) [35]	Found that *employees who exhibited lower levels of supervisor-rated "organizational citizenship behaviours" were more likely to quit*.
Burton and Beckman (2007) [34]	Proposed novel idea of *"position imprinting*" and proved that *employees who were different from their position's creator were more likely to quit than employees who were similar to creators*.
2005, Griffeth et al. [42]	Introduced concept *of "Employment Opportunity Index",* a 5 -dimensional scale which explained attrition very satisfactorily.
Research direction -5: Studies proposing that there should be increased focus on *factors responsible for staying* and not just on factors for quitting	
Mitchell et al. (2001) [56]	Introduced concept of 'job-embeddedness' which includes a collection of factors affecting a person's staying/quitting a job. Study reported some important facts about job-embeddedness. For example, *one:* It was measured as an aggregated score across items and it negatively correlated with intention to leave and predicted attrition. *Second*: It significantly predicted attrition after controlling for certain factors.
Crossley et al. (2007) [36]	Re-conceptualized job embeddedness into two types: 'composite' and 'general' and tested how these two could be integrated into Mobley-type attrition variables. *general job-embeddedness* significantly related to the intention to search, intention to quit and attrition. In contrast, *composite job -embeddedness* only significantly related to intention to search and intention to quit, but not to attrition
Research-direction -6: A dynamic modelling of attrition considering that attrition predictors like job-satisfaction are not static but change with time	
Sturman and Trevor (2001) [71]	Found that quitters' performance over time did not significantly change while stayers' performance slope was positive. Also, performance over last two months and all prior months were negatively related to attrition.
Steel (2002) [63-b]	Proposed an evolutionary search model of attrition. It proposed *three distinct job-search phases* ("passive-scanning", "focused-search", and "contacting prospective employers"), and two job-search gateways ("financial considerations" and "spontaneous job offers").

Name Of researchers	Conclusion/summary
Kammeyer-Mueller et al (2005) [47]	Directly compared a static attrition model to a dynamic model. Dynamic model fit data better than static model. Found that leavers became less committed and less satisfied over time and had increased levels of work withdrawal and alternative-search.
Research-direction -7: Expansion of understanding of previously identified relationships	
Meyer, Stanley, Herskovits and Topolnytsky (2002) [54]	The study applied meta-analysis and *reported weighted correlations between 'attrition', and 'affective-commitment' (0.17), 'normative-commitment' (0.16), and 'continuance-commitment' (0.10).*
Vandenberghe, Bentein, and Stinglhamber (2004) [73]	Reported that "*affective-commitment to supervisor and group predicted affective-commitment to organization*"; which in turn, predicted "*intention to quit*", which predicted "*actual attrition*".
Salamin & Hom (2005) [65]	Showed that relationship *between work-performance and attrition was curvilinear such that low and high performers were more likely to quit.*
Trevor (2001) [72]	Extending March and Simon's (1958) work found that *general job-availability*, *movement-capital* and *job-satisfaction* interacted *with each other simultaneously to affect attrition*.
Allen et al., 2005 [27]	Proved that employees with low self-observation, low risk-aversion, and an internal centre of control had stronger tendency to convert attrition-intention into attrition.
Allen and Griffeth (2001) [26]	Reported that promptness and magnitude of reward for good work moderates job performance–job satisfaction–attrition linkages such that performance–satisfaction link was positive for high rewards and negative for low rewards.

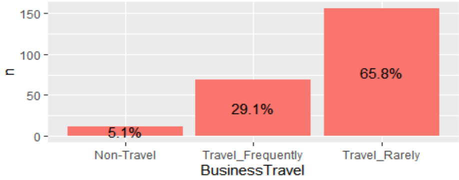
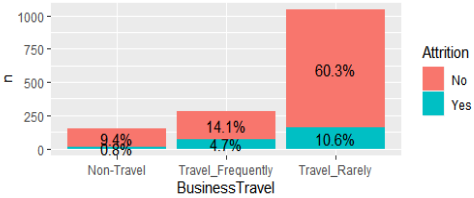
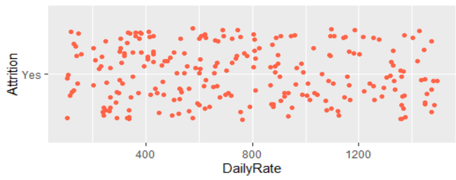
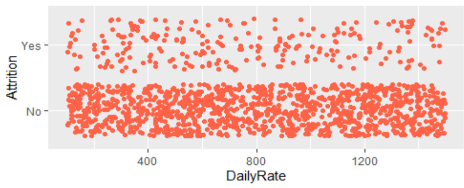
APPENDIX-II Exploratory data analysis

TABLE II-I Shows variation of attrition with each variable

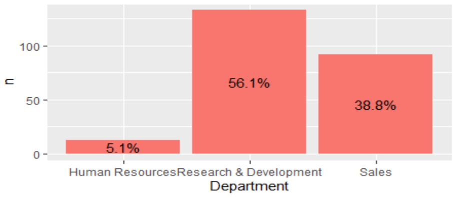
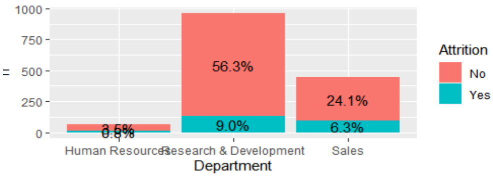
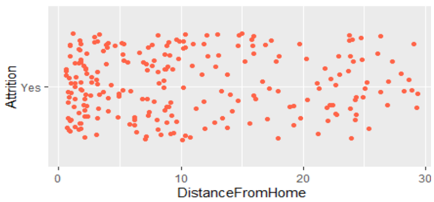
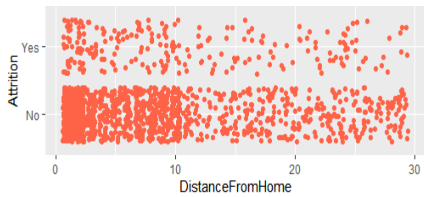
[P.T.O]

	FROM DATAFRAME ‘attrited’ SHOWING ABSOLUTE TREND	FROM ENTIRE DATAFRAME SHOWING PROPORTIONATE TREND
1		

Age bin	18-24	24-30	30-36	36-42	42-48	48-54	54-60
% in given data	6.05	19.66	28.03	19.93	12.38	8.71	4.69
% in ‘attrited’	14.35	26.16	27.85	11.39	8.02	5.91	4.64

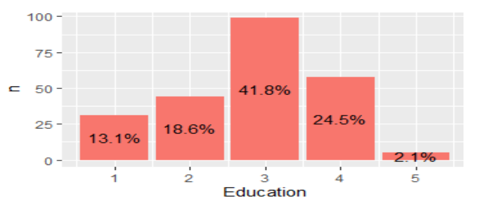
2		
3		

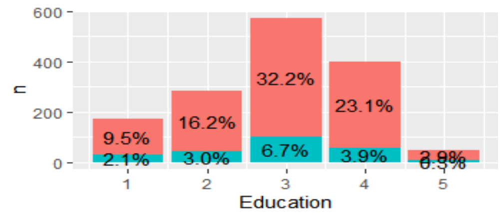
DailyRateBin	100-300	300-500	500-700	700-900	900-1100	1100-1300	1300-1500
% in given data	13.40	14.15	16.46	13.13	13.40	14.69	14.76
% in ‘attrited’	14.35	18.57	17.30	13.50	12.66	9.70	13.92

4		
5		

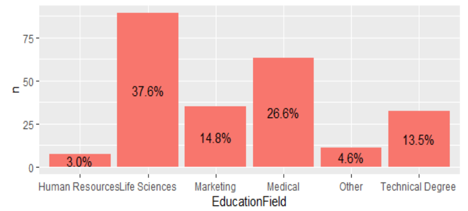
Distance Bin	0-6	6-12	12-18	18-24	24-30
% in given data	47.01	26.12	9.80	9.46	7.62
% in 'attrited'	39.66	25.32	13.08	13.92	8.02

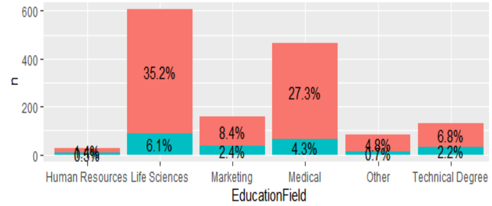
6



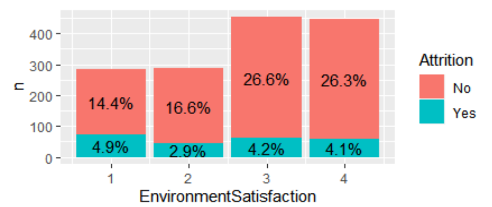


7

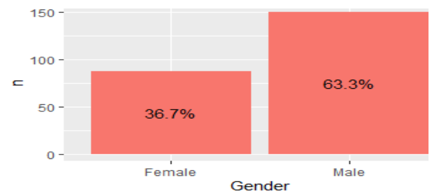


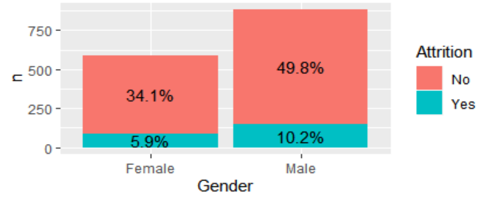


8

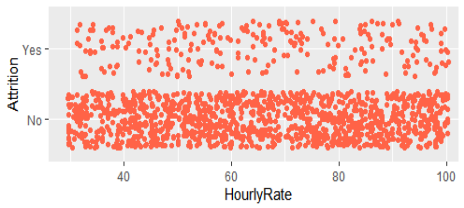


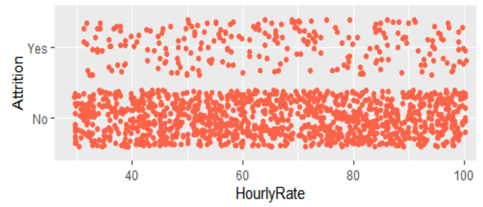
9





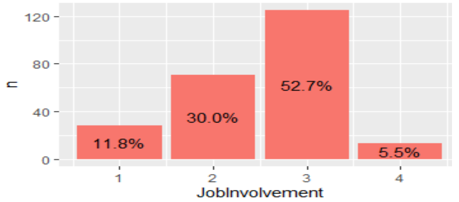
10

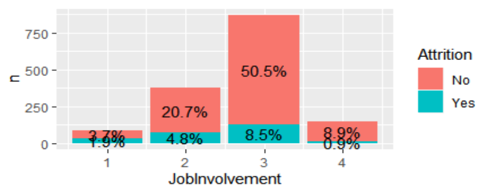




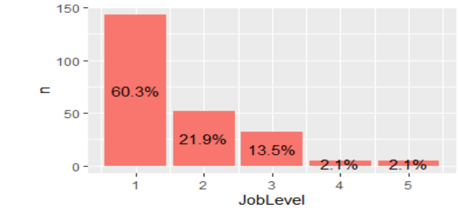
Bin	30-40	40-50	50-60	60-70	70-80	80-90	90-100
% in given data	11.70	14.83	14.83	12.93	14.83	14.69	14.90
% in 'attrited'	13.92	13.50	16.03	15.61	12.24	13.92	14.77

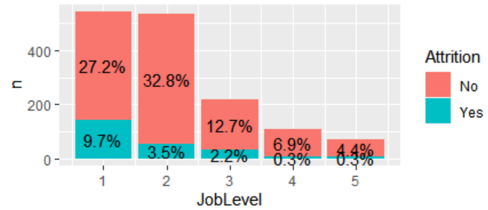
11



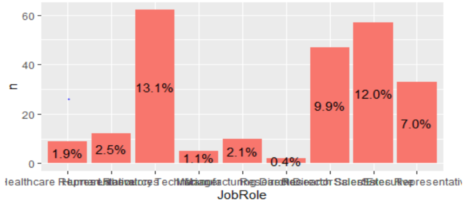


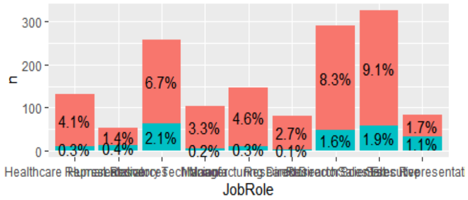
12





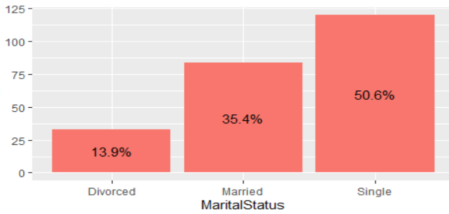
13

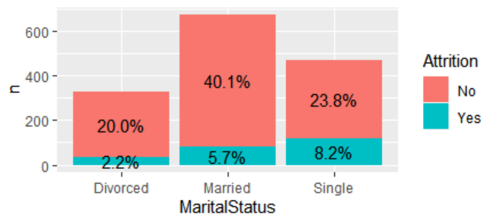




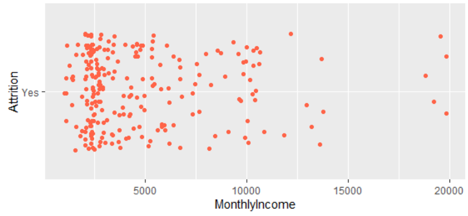
14

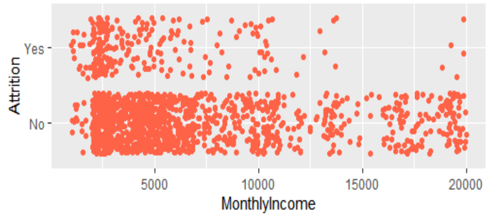
15





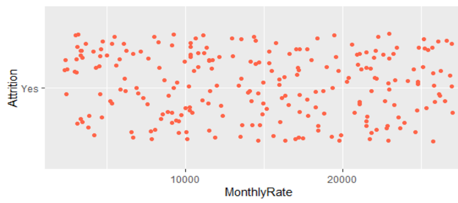
16

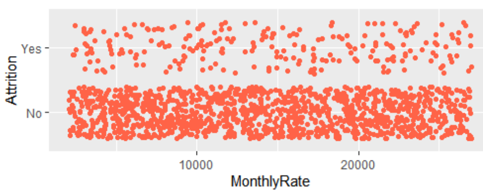




Bin MonthlyIncome	1000-5000	5000-9000	9000-13000	13000-17000
%in given data	50.95	26.39	10.07	5.99
% in 'attrited'	68.78	16.88	10.55	1.69

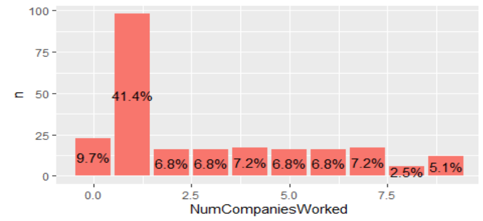
17

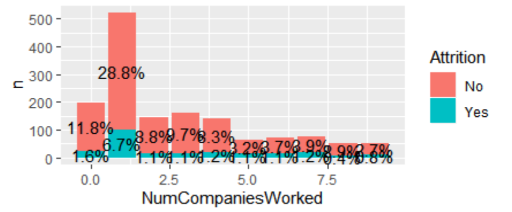




Bin	2000-7000	7000-12000	12000-17000	17000-22000	22000-27000
% in given data	20.61	20.54	19.59	20.27	18.98
% in 'attrited'	19.41	22.36	17.72	20.25	20.28

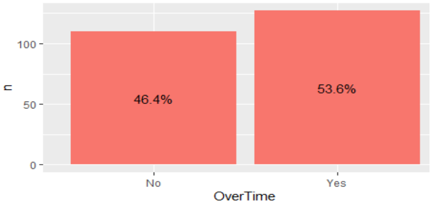
18

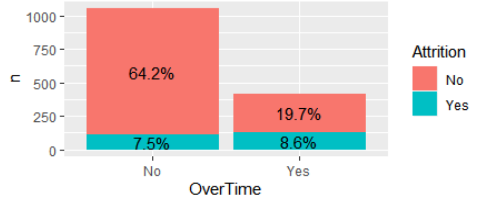




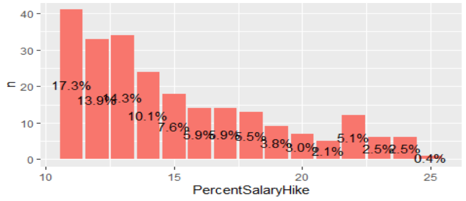
Bin	2000-7000	7000-12000	12000-17000	17000-22000	22000-27000
% in given data	20.61	20.54	19.59	20.27	18.98
% in 'attrited'	19.41	22.36	17.72	20.25	20.28

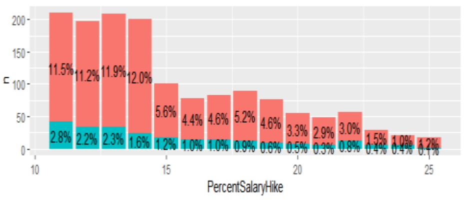
19





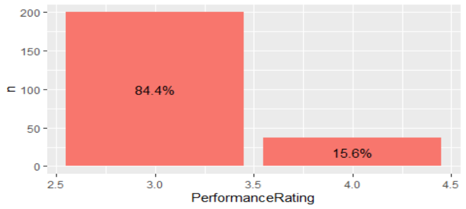
20

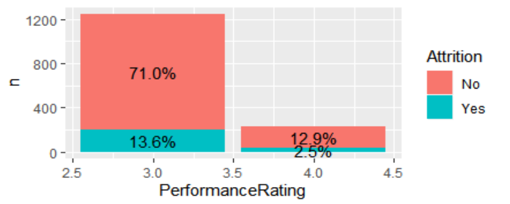




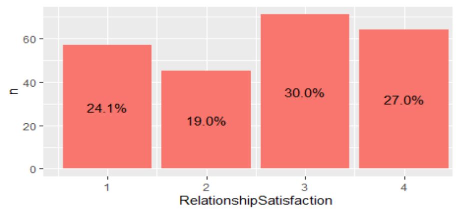
%Hike Bins	11-13	13-15	15-17	17-19	19-21	21-23	23-25
% in given data	27.69	20.54	10.88	11.22	7.01	5.71	2.65
% in 'attrited'	28.27	17.72	11.81	9.28	5.06	7.59	2.95

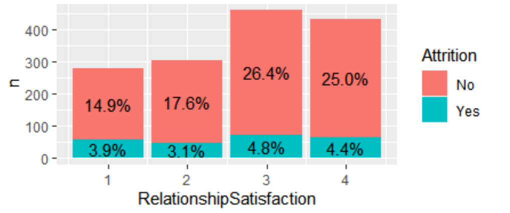
21



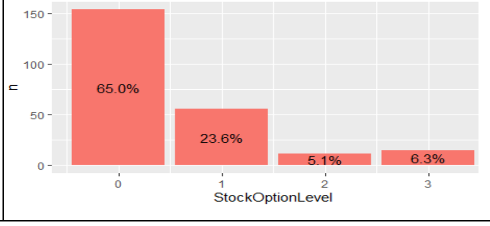


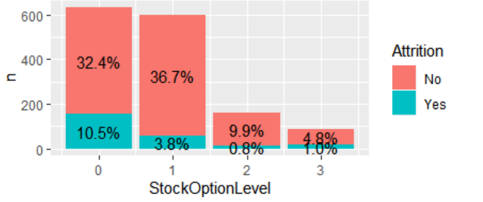
22

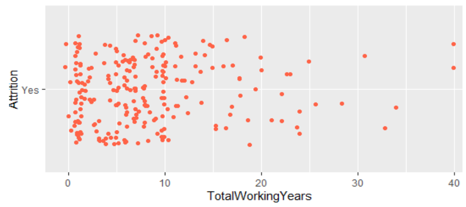
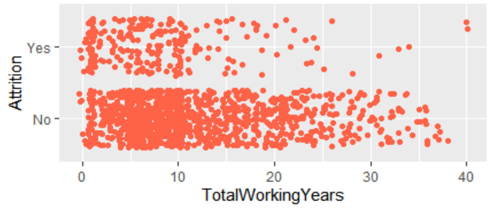
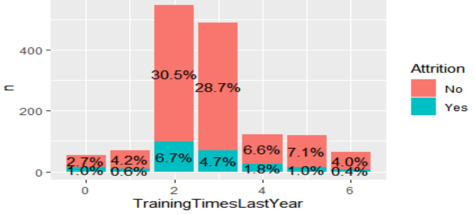




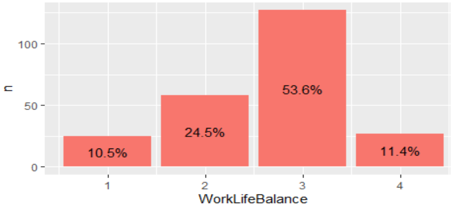
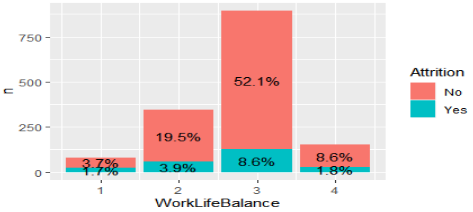
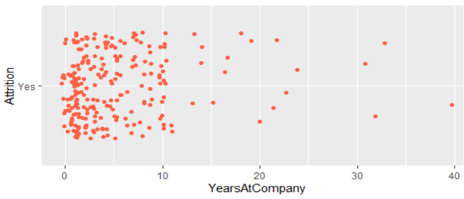
23



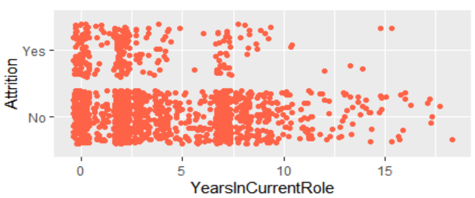


24		
25		

Years	0-4	4-8	8-12	12-16	16-20	20-24	24-28	28-32	32-36	36-40
%in given data	14.76	27.01	25.99	9.80	7.62	6.46	3.33	2.38	1.43	0.48
% in 'attrited'	29.54	30.38	19.83	6.33	5.06	3.38	1.27	0.42	0.84	0.84

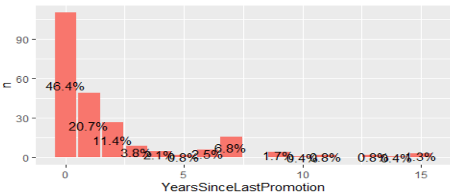
26		
27		

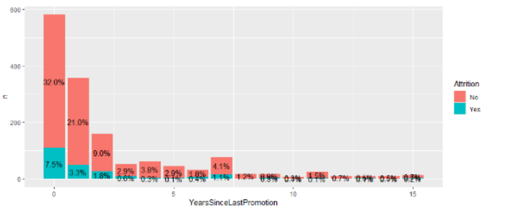
Years	0-4	4-8	8-12	12-16	16-20	20-24	24-28	28-32	32-36	36-40
%in given data	36.46	30.07	16.87	5.03	4.08	2.52	0.68	0.61	0.54	0.14
% in 'attrited'	52.74	21.10	11.81	2.53	1.69	1.69	0.00	0.84	0.42	0.42

28		

YrCurRole		0-2	2-4	4-6	6-8	8-10	10-12	12-14	14-16	16-18
% in given data		29.18	16.26	4.97	21.16	6.53	2.18	1.70	1.02	0.41
% in ‘attrited’		33.33	13.08	1.27	16.03	3.38	0.42	0.42	0.84	0.00

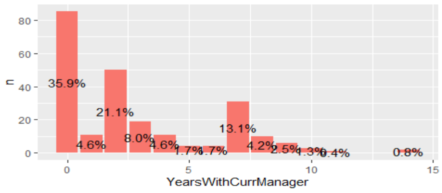
29

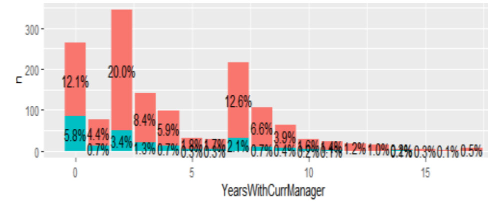




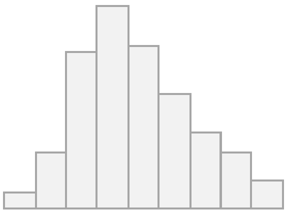
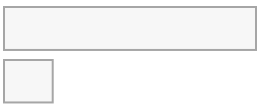
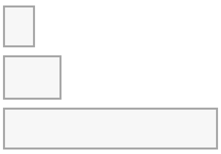
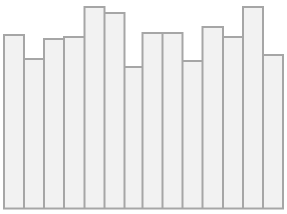
YrLastPromo	0-2	2-5	5-8	8-11	11-15
% in given data	38.64	9.39	7.55	2.72	2.18
% in ‘attrited’	74.63	10.75	8.57	3.20	1.97

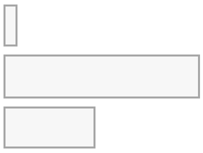
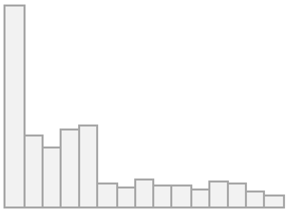
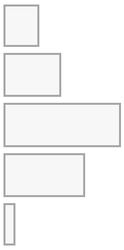
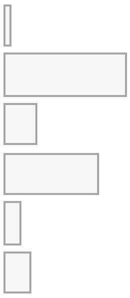
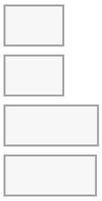
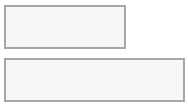
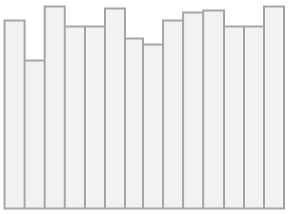
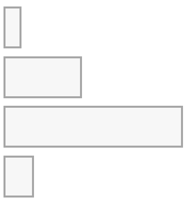
30

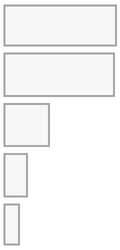
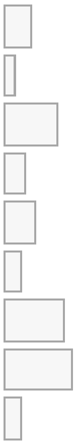
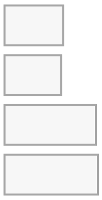
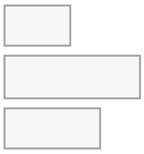
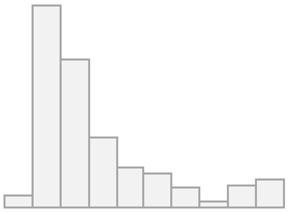
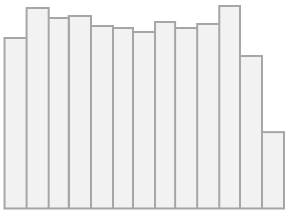


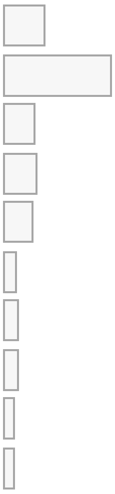
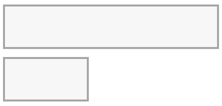
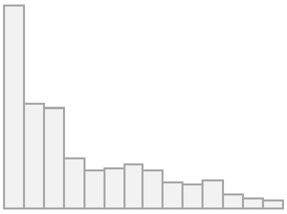
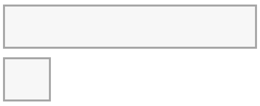
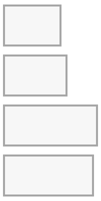
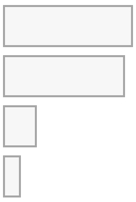
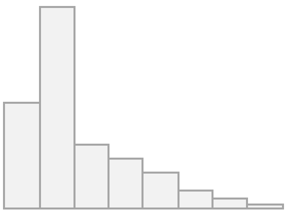


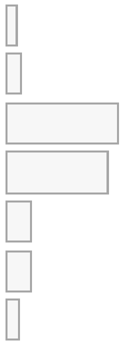
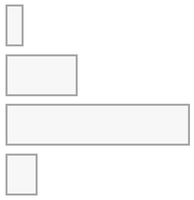
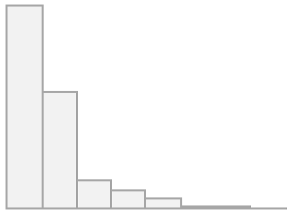
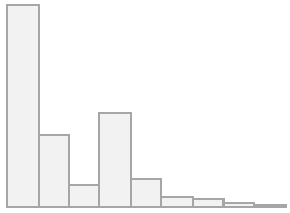
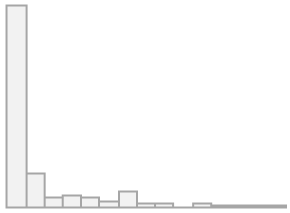
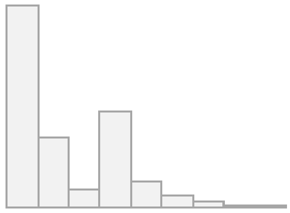
APPENDIX-III Univariate analysis by dfSummary ()

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	Age [integer]	Mean (sd): 36.9 (9.1) min < med < max: 18 < 36 < 60 IQR (CV): 13 (0.2)	43 distinct values		1470 (100%)	0 (0%)
2	Attrition [factor]	1. No 2. Yes	1233 (83.9%) 237 (16.1%)		1470 (100%)	0 (0%)
3	BusinessTravel [factor]	1. Non-Travel 2. Travel_Frequently 3. Travel_Rarely	150 (10.2%) 277 (18.8%) 1043 (71.0%)		1470 (100%)	0 (0%)
4	DailyRate [integer]	Mean (sd): 802.5 (403.5) min < med < max: 102 < 802 < 1499 IQR (CV): 692 (0.5)	886 distinct values		1470 (100%)	0 (0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
5	Department [factor]	1.Human Resources 2.Research & Development 3. Sales	63 (4.3%) 961 (65.4%) 446 (30.3%)		1470 (100%)	0 (0%)
6	DistanceFromHome [integer]	Mean (sd): 9.2 (8.1) min < med < max: 1 < 7 < 29 IQR (CV): 12 (0.9)	29 distinct values		1470 (100%)	0 (0%)
7	Education [integer]	Mean (sd): 2.9 (1) min < med < max: 1 < 3 < 5 IQR (CV): 2 (0.4)	1 : 170 (11.6%) 2 : 282 (19.2%) 3 : 572 (38.9%) 4 : 398 (27.1%) 5 : 48 (3.3%)		1470 (100%)	0 (0%)
8	EducationField [factor]	1.HR 2. Life Sciences 3. Marketing 4. Medical 5. Other 6.Technical Degree	27 (1.8%) 606 (41.2%) 159 (10.8%) 464 (31.6%) 82 (5.6%) 132 (9.0%)		1470 (100%)	0 (0%)
9	EnvironmentSatisfaction [integer]	Mean (sd): 2.7 (1.1) min < med < max: 1 < 3 < 4 IQR (CV): 2 (0.4)	1 : 284 (19.3%) 2 : 287 (19.5%) 3 : 453 (30.8%) 4 : 446 (30.3%)		1470 (100%)	0 (0%)
10	Gender [factor]	1. Female 2. Male	588 (40.0%) 882 (60.0%)		1470 (100%)	0 (0%)
11	HourlyRate [integer]	Mean (sd): 65.9 (20.3) min < med < max: 30 < 66 < 100 IQR (CV): 35.8 (0.3)	71 distinct values		1470 (100%)	0 (0%)
12	JobInvolvement [integer]	Mean (sd): 2.7 (0.7) min < med < max: 1 < 3 < 4 IQR (CV): 1 (0.3)	1 : 83 (5.6%) 2 : 375 (25.5%) 3 : 868 (59.1%) 4 : 144 (9.8%)		1470 (100%)	0 (0%)

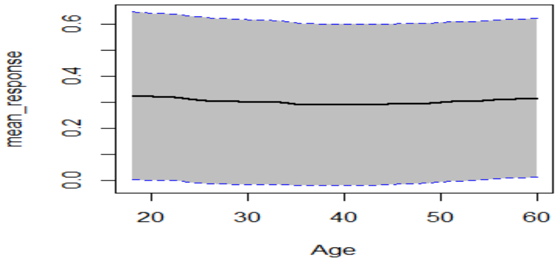
No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
13	JobLevel [integer]	Mean (sd): 2.1 (1.1) min < med < max: 1 < 2 < 5 IQR (CV): 2 (0.5)	1 : 543 (36.9%) 2 : 534 (36.3%) 3 : 218 (14.8%) 4 : 106 (7.2%) 5 : 69 (4.7%)		1470 (100%)	0 (0%)
14	JobRole [factor]	1.Healthcare Representative 2.Human Resources 3.Laboratory Technician 4. Manager 5. Manufacturing Director 6.Research Director 7.Research Scientist 8. Sales Executive 9.Sales Representative	131 (8.9%) 52 (3.5%) 259 (17.6%) 102 (6.9%) 145 (9.9%) 80 (5.4%) 292 (19.9%) 326 (22.2%) 83 (5.6%)		1470 (100%)	0 (0%)
15	JobSatisfaction [integer]	Mean (sd): 2.7 (1.1) min < med < max: 1 < 3 < 4 IQR (CV): 2 (0.4)	1 : 289 (19.7%) 2 : 280 (19.1%) 3 : 442 (30.1%) 4 : 459 (31.2%)		1470 (100%)	0 (0%)
16	MaritalStatus [factor]	1. Divorced 2. Married 3. Single	327 (22.2%) 673 (45.8%) 470 (32.0%)		1470 (100%)	0 (0%)
17	MonthlyIncome [integer]	Mean (sd): 6502.9 (4708) min < med < max: 1009 < 4919 < 19999 IQR (CV): 5468 (0.7)	1349 distinct values		1470 (100%)	0 (0%)
18	MonthlyRate [integer]	Mean (sd): 14313.1 (7117.8) min < med < max: 2094 < 14235.5 < 26999 IQR (CV): 12414.5 (0.5)	1427 distinct values		1470 (100%)	0 (0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
19	NumCompaniesWorked [integer]	Mean (sd): 2.7 (2.5) min < med < max: 0 < 2 < 9 IQR (CV): 3 (0.9)	0 : 197 (13.4%) 1 : 521 (35.4%) 2 : 146 (9.9%) 3 : 159 (10.8%) 4 : 139 (9.5%) 5 : 63 (4.3%) 6 : 70 (4.8%) 7 : 74 (5.0%) 8 : 49 (3.3%) 9 : 52 (3.5%)		1470 (100%)	0 (0%)
20	OverTime [factor]	1. No 2. Yes	1054 (71.7%) 416 (28.3%)		1470 (100%)	0 (0%)
21	PercentSalaryHike [integer]	Mean (sd): 15.2 (3.7) min < med < max: 11 < 14 < 25 IQR (CV): 6 (0.2)	15 distinct values		1470 (100%)	0 (0%)
22	PerformanceRating [integer]	Min: 3 Mean: 3.2 Max: 4	3 : 1244 (84.6%) 4 : 226 (15.4%)		1470 (100%)	0 (0%)
23	RelationshipSatisfaction [integer]	Mean (sd): 2.7 (1.1) min < med < max: 1 < 3 < 4 IQR (CV): 2 (0.4)	1 : 276 (18.8%) 2 : 303 (20.6%) 3 : 459 (31.2%) 4 : 432 (29.4%)		1470 (100%)	0 (0%)
24	StockOptionLevel [integer]	Mean (sd): 0.8 (0.9) min < med < max: 0 < 1 < 3 IQR (CV): 1 (1.1)	0 : 631 (42.9%) 1 : 596 (40.5%) 2 : 158 (10.8%) 3 : 85 (5.8%)		1470 (100%)	0 (0%)
25	TotalWorkingYears [integer]	Mean (sd): 11.3 (7.8) min < med < max: 0 < 10 < 40 IQR (CV): 9 (0.7)	40 distinct values		1470 (100%)	0 (0%)

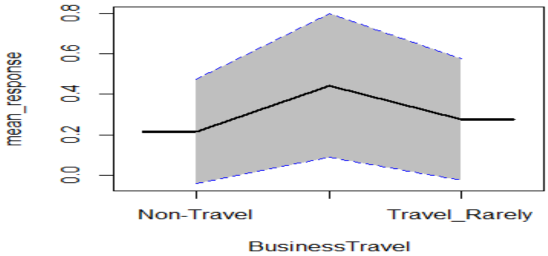
No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
26	TrainingTimesLastYear [integer]	Mean (sd): 2.8 (1.3) min < med < max: 0 < 3 < 6 IQR (CV): 1 (0.5)	0 : 54 (3.7%) 1 : 71 (4.8%) 2 : 547 (37.2%) 3 : 491 (33.4%) 4 : 123 (8.4%) 5 : 119 (8.1%) 6 : 65 (4.4%)		1470 (100%)	0 (0%)
27	WorkLifeBalance [integer]	Mean (sd): 2.8 (0.7) min < med < max: 1 < 3 < 4 IQR (CV): 1 (0.3)	1 : 80 (5.4%) 2 : 344 (23.4%) 3 : 893 (60.8%) 4 : 153 (10.4%)		1470 (100%)	0 (0%)
28	YearsAtCompany [integer]	Mean (sd): 7 (6.1) min < med < max: 0 < 5 < 40 IQR (CV): 6 (0.9)	37 distinct values		1470 (100%)	0 (0%)
29	YearsInCurrentRole [integer]	Mean (sd): 4.2 (3.6) min < med < max: 0 < 3 < 18 IQR (CV): 5 (0.9)	19 distinct values		1470 (100%)	0 (0%)
30	YearsSinceLastPromotion [integer]	Mean (sd): 2.2 (3.2) min < med < max: 0 < 1 < 15 IQR (CV): 3 (1.5)	16 distinct values		1470 (100%)	0 (0%)
31	YearsWithCurrManager [integer]	Mean (sd): 4.1 (3.6) min < med < max: 0 < 3 < 17 IQR (CV): 5 (0.9)	18 distinct values		1470 (100%)	0 (0%)

APPENDIX- IV PDP for each predictor

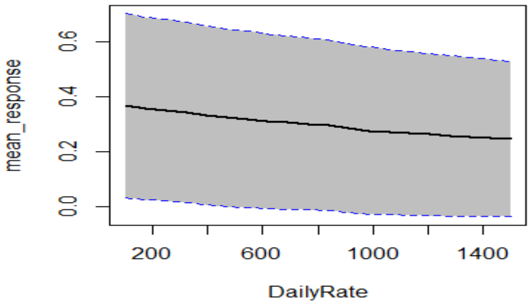
<table>
<tr><th>No:</th><th>Partial Dependence Plot-PDP
RESPONSE(‘Attrition’) Vs Predictor</th><th>INTERPRETATION
Type of realtion with attrition (i.e. linear/non-linear, positive or negative (inverse) and how strong it is.</th></tr>
<tr><td>1</td><td>



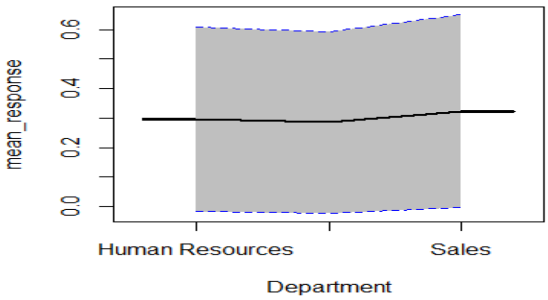
</td><td>Nearly linear very weak trend. Age is weak attrition-predictor.</td></tr>
<tr><td>2</td><td>



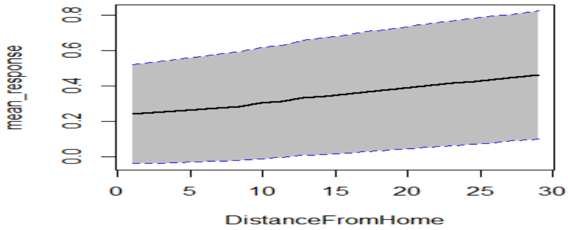
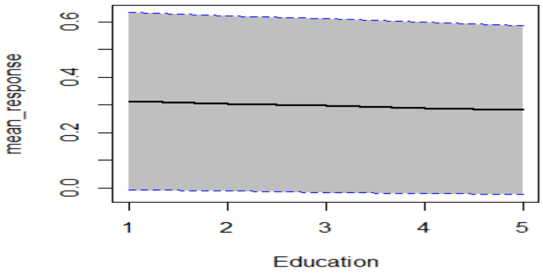
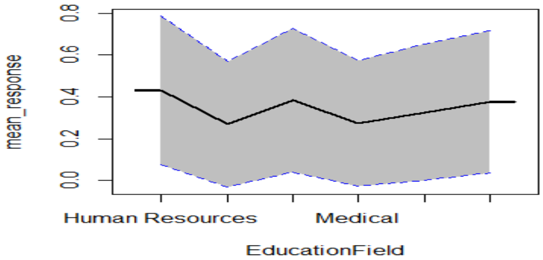
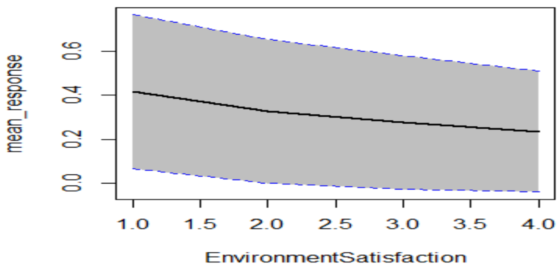
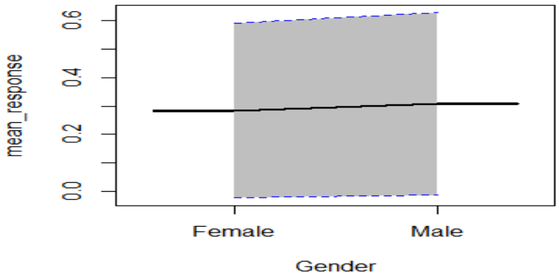
</td><td>Attrition shows clear high response to Travel Frequently (middle). So BusinessTravelFrequently is strong predictor.</td></tr>
<tr><td>3</td><td>

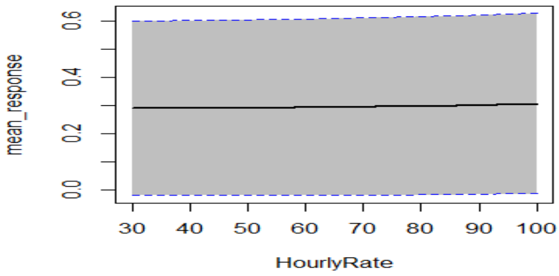
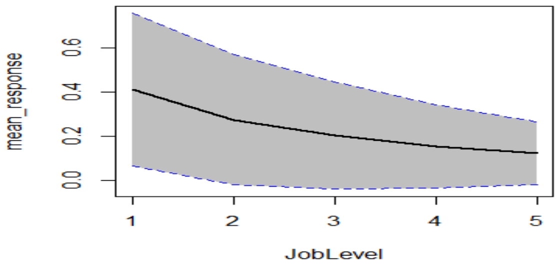
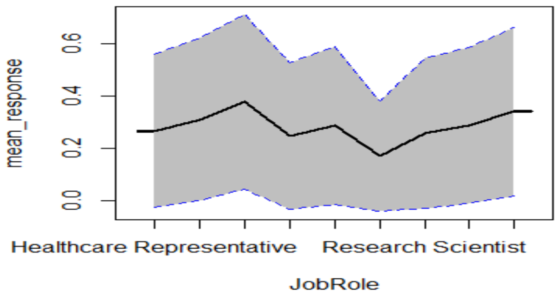


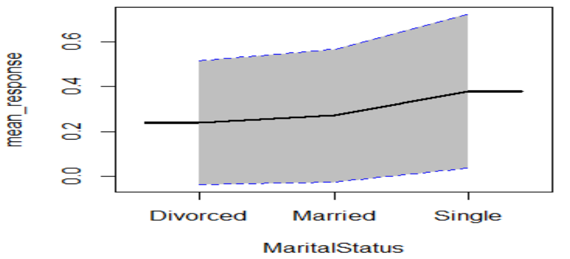
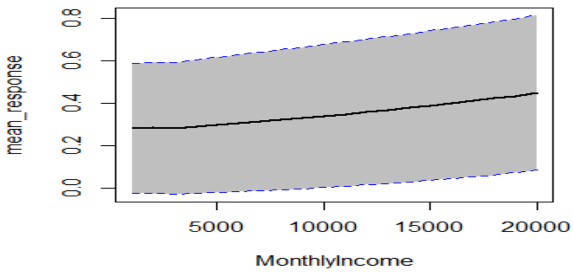
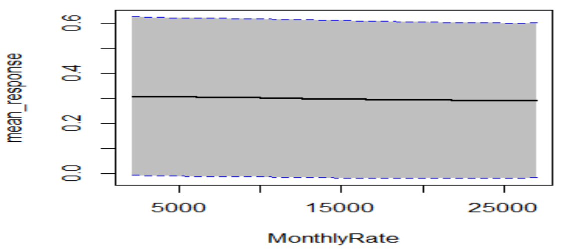
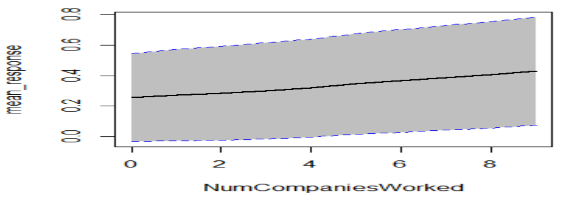
</td><td>mild inverse linear response to increase in daily rate. So, Daily rate is mild negative predictor.</td></tr>
<tr><td>4</td><td>

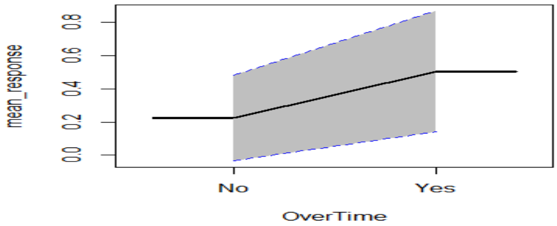
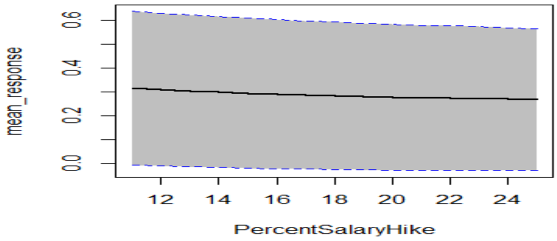
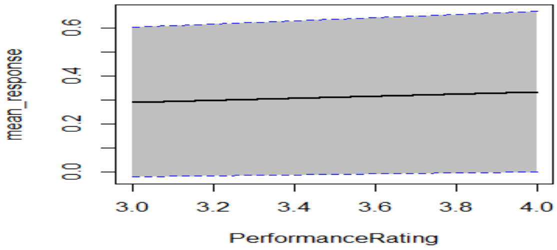
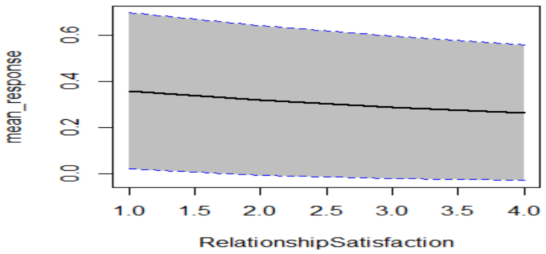
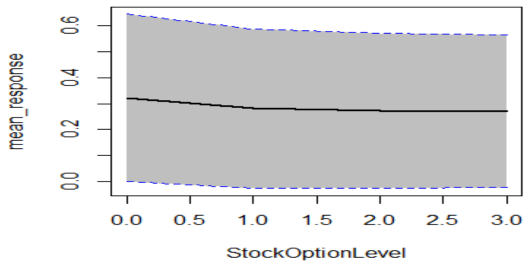


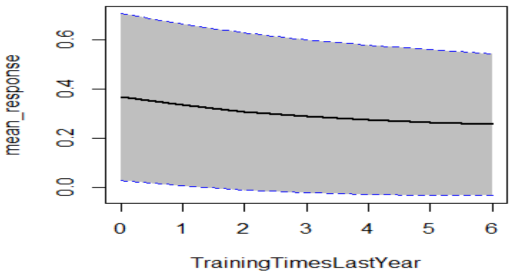
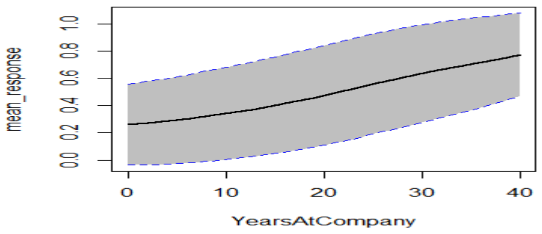
</td><td>mild high response to departments HR and Sales.</td></tr>
</table>

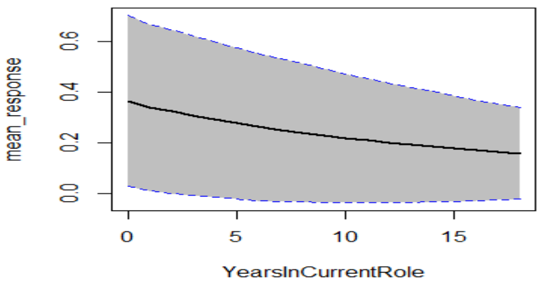
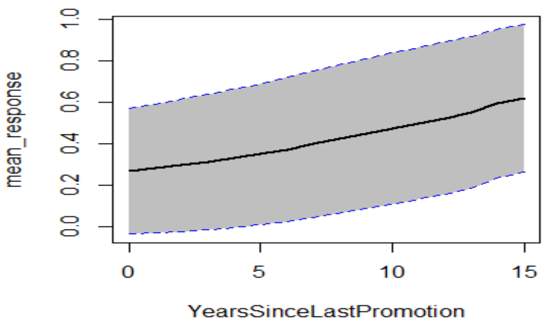
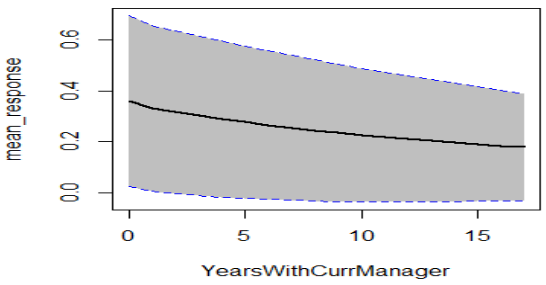
No:	Partial Dependence Plot-PDP RESPONSE('Attrition') Vs Predictor	INTERPRETATION Type of realtion with attrition (i.e. linear/non-linear, positive or negative (inverse) and how strong it is.
5		Llinear response to distance from Home. So, Distance From Home is a strong positive predictor.
6		Very mild inverse linear relationship to education.
7		Attrition shows high response to Education Fields 1,3,6 i.e. HR, Marketing, Technical Degree are more liqely to quit.
8		Inverse linear trend with Environmental Satistaction. So low satisfaction means more attrition and vice versa. Medium strength predictor.
9		Flat linear graph shows, gender is very weak predictor. Males have slightly higher tendency to quit.

No:	Partial Dependence Plot-PDP RESPONSE('Attrition') Vs Predictor	INTERPRETATION Type of realtion with attrition (i.e. linear/non-linear, positive or negative (inverse) and how strong it is.
10		Attrition does not much depend on hourly rate.
11		Inverse linear trend with Job Involvement. So low involvement means more attrition and vice versa.
12		Inverse, non linear trend with Job Level. So low Job Level means more attrition and vice versa. Strong predictor.
13		JobRole '3','7','8','9' show more response to attrition. So, Laboratory Technician, Research Scientist, Sales Executive and Sales Representative.

No:	Partial Dependence Plot-PDP RESPONSE('Attrition') Vs Predictor	INTERPRETATION Type of realtion with attrition (i.e. linear/non-linear, positive or negative (inverse) and how strong it is.
14		Mild inverse linear trend. So, those with high job satisfaction quit less. Weak predictor.
15		Non-linear increasing trend. 'Single; Marital status has highest tendency to quit. Stron predictor.
16		Almost positive linear trend. Those with high monthly income have more quit tendency. Strong Predictor
17		Flat linear trend. No dependence on Monthly Rate
18		Mild increasing linear trend. So, those who worked in 6 or more companies more likely to quit. Strong Predictor

No:	Partial Dependence Plot-PDP RESPONSE('Attrition') Vs Predictor	INTERPRETATION Type of realtion with attrition (i.e. linear/non-linear, positive or negative (inverse) and how strong it is.
19		Nonlinear trend with OverTime 'Yes' clearly causing very high mean response value. Strong predictor.
20		Mild inverse linear trend. Weak predictor.
21		Almost flat trend. Attrition does not depend on performance rating.
22		Mild inverse linear trend. Moderate predictor.
23		Very mild inverse nearly linear trend. Mild predictor.

No:	Partial Dependence Plot-PDP RESPONSE('Attrition') Vs Predictor	INTERPRETATION Type of realtion with attrition (i.e. linear/non-linear, positive or negative (inverse) and how strong it is.
24		Clear inverse non-linear trend. Moderate predictor.
25		Mild inverse non-linear trend. Those with '0' training had maximum quit tendency.
26		Mild inverse non-linear trend. Mild predictor.
27		Strong positive non linear trend. Strong predictor.

No:	Partial Dependence Plot-PDP RESPONSE('Attrition') Vs Predictor	INTERPRETATION Type of realtion with attrition (i.e. linear/non-linear, positive or negative (inverse) and how strong it is.
28		Inverse non-linear trend. Moderate predictor.
29		Strong positive nearly linear trend. Strong predictor.
30		Nearly linear inverse trend. Maximum quitters spend less than 5 years with current manager.

APPENDIX-V Defining performance metrics

- All performance metrics can be obtained by using the confusion-matrix (or as is done in this thesis by library MLmetrics). A confusion-matrix for attrition problem would be as below:

	Predicted Negative	Predicted Positive
Actual Negative	TN= TRUE NEGATIVES No attrition occurred and model also predicted no attrition for that employee	FP= FALSE POSITIVE No attrition occurred but model predicted attrition for that employee
Actual Positive	FN = FALSE NEGATIVE Attrition occurred but model predicted no Attrition for that employee	TP= TRUE POSITIVE Attrition occurred and model also predicted it as attrition for that employee

a) $\text{OverallAccuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} =$

fraction of the total sample that is correctly identified

b) $\text{OverallError} = \frac{\text{FP+FN}}{\text{TP+TN+FP+FN}} = 1 - \text{overall accuracy}$

c) $\text{ExpectedErrorRate Kappa} = \frac{O-E}{1-E} = 1 - \text{overall accuracy}$

d) $Sensitivity/Recall/$TPR $= \frac{\text{TP}}{\text{TP+FN}}$ = Proportion of attritions that are correctly identified as attritions

e) Specificity/TNR = $\frac{\text{TP}}{\text{TN+FP}}$ = Proportion of non-attritions that are correctly identified as non-attritions

f) PPV/Precision = $\frac{\text{TP}}{\text{TP+FP}}$ = $\ The\ fraction\ of\ the\ positive\ predictions\ that\ are\ actually\ positive$

g) $\text{F1}\_Score = \frac{2\times(\text{Precision}\times\text{Recall})}{(\text{Precision+Recall})} = harmonic\ mean\ of\ precision\ and\ recall$

h) ROC CURVE*:* A curve of sensitivity Vs (1 -specificity)

i) AUCROC or AUC= Area under ROC curve. A measure of performance of model.

j) $False\ Positive\ Rate\ FPR = \frac{\text{FP}}{\text{TN+FP}} = TYPE - I\ ERROR\ RATE$

k) $False$ Negative $Rate\ FNR = \frac{\text{FN}}{\text{FN+TP}} = TYPE - II\ ERROR\ RATE$