Comparative Study of the Machine Learning Techniques for Predicting the Employee Attrition

Kashyap Bhuva[1*], Kriti Srivastava[2]
[1]Student, D.J. Sanghvi College of Engineering, Mumbai, India
[2]Assistant Professor, Department of Computer Science, D.J. Sanghvi College of Engineering, Mumbai, India

Abstract: Employee Attrition is one of the biggest business problems in HR Analytics**.** Companies invest a lot in the training of the employees keeping in mind the returns they would provide to the company in the future**.** If an employee leaves the company, it is the loss of opportunity cost to the company**.** Attrition is particularly prevailing in the mass recruiting companies**.** In this paper, we have taken a sample of the employee database of IBM USA**.** Using the information value concept, we found out that the characteristics of employees like Job Role, overtime, job level affect the attrition largely**.** We implemented various classification algorithms like logistic regression, LDA, ridge classification, lasso classification, decision trees, random forests to predict the probability of attrition of any new employee and simultaneously tested them. Using various model evaluation metrics we made a comparative analysis of the models & found out that LDA gave the highest accuracy, logistic gave the highest precision and ridge gave the highest recall. The comparative study enables the organization to select a model depending on its business requirements. It will help an organization find out the probability of attrition of any new candidate which they might be hiring**.** Also, it would help companies decide the salary hike & changes in perks for the existing employees and tweak it in such a way that the employee is retained somehow**.**

IndexTerms **- Ridge, Lasso, Logistic Regression, Decision Tree, Random Forest, Naïve Bayes, Linear Discriminant Analysis, Support Vector Machine**

I. INTRODUCTION

Employee Attrition is one of the major problem faced by any organization. In this age of cut-throat competition there are many factors which lead to dissatisfaction in employee. long working hours, peer pressure, job location, job role, travelling time, office space, amenities in the office, perks and many more reasons could be a factor for employee attrition. It is very important for the HR department to understand employee satisfaction level. Sometimes the employee many not have any problem in the company but others many offer a better profile with better pay package. So, the employee may be willing to leave. Retaining one employee needs a lot of insight in many areas. In this research we try to find out important factors that lead to employee attrition. The results of our model can be used by HR department to plan a strategy before the employee sends his resignation. This paper involves a comparative study of various algorithms using the model evaluation metrics like accuracy, precision, recall, FN Rate, F-measure unlike the majority of the existing research work which focuses on a single algorithm for solving a business problem. The reason for using multiple algorithms is that each algorithm has its special advantages and pitfalls. For eg. when the event rate is low, logistic model underperforms while random forests outperform. When there is a large number of significant attributes out of the total attributes, ridge outperforms lasso. When the data is highly non-linear, decision tree outperforms the linear models. Thus, comparative analysis would take into account all the advantages & pitfalls enabling the organizations to select to best model for their data.

The analytics project consists of various steps like data preprocessing, feature selection & scaling, modelling using various algorithms and finally evaluating the models using model evaluation metrics. After evaluation, the best model is deployed on the data to make predictions.

Fig 1. The analytics project workflow

II. RAW DATASET

The dataset we selected has 35 different attributes like the Age, Business Travel, Daily Rate, Department, Distance , Marital Status, Monthly Income, Number of companies worked, Over18, Over Time, Percent Salary Hike, Performance rating, Relationship

Satisfaction, standard working hours, Stock option level, Employee field, Environment Satisfaction, Gender, Hourly Rate, Job Involvement, Job Level, Job Role, Job Satisfaction , Total working years, training times last year, work-life balance, Years at company , Years in current role, Year since last promotion, years with current manager. There are around 1470 observations, each observation corresponding to an employee.

III. DATA PREPROCESSING

The process that involves transforming the raw data into a model-able format is called data pre-processing. Real-world data is often incomplete and inconsistent. Data pre-processing has become a separate topic of study in the literature. Zhu Yan-li,& Zhang Jia [1] has done research on data pre-processing for credit card behaviour mining.

The following are the steps for the data pre-processing:

Data Cleaning: Data is cleaned through processes such as getting rid of the missing values, smoothening the noisy data, or resolving the issues in the data.

Data Integration: Data with varied representations are put together to resolve the conflicts between them.

Data Transformation: To transform the data so as to make it fulfil the modelling assumptions.

Data Reduction: To reduce the dimensions of the data.

Data Discretization: Involves the reduction of a number of values of a continuous attribute by dividing the range of attribute intervals.

IV. FEATURE SELECTION

Feature selection is one of the most important processes of the data analysis. Isabelle Guyon & Andre Elisseeff [2] gives good insights about feature selection in their work. Popular methods for feature selection are correlation analysis & chi-square analysis, exploratory bivariate analysis and information value analysis. Correlation analysis is used for the numeric variables & Chi-square analysis is used for the categorical variables. A high correlation or a chi-square value proves a feature significant.

4.1 Correlation

Mark. A. Hall [3] research work is a good example of the use of correlation for the numeric features selection. The following Pearson correlation plot shows that there is a strong correlation between some attributes such as monthly income and job level, job level and total working years, total working years and monthly income, however our targeted attribute attrition showed poor correlation with other attributes, thus attrition is more likely to depend on a combination of attributes rather than on a single one. The attrition is positively correlated with Age, Monthly Income, Total Working Years, Years at Company and Years with Current Manager, and no significant negative correlation is found . The correlation plot is shown in the figure 2.

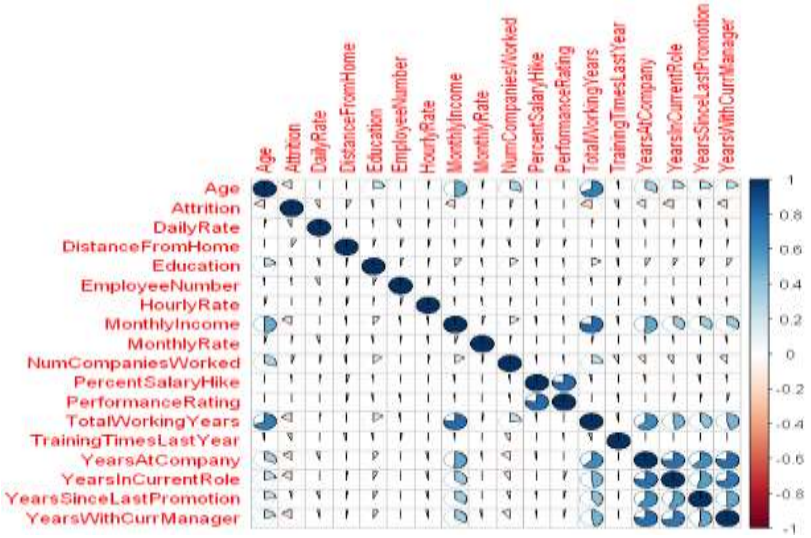

Fig 2. Correlation plot

4.2 Chi-Square Analysis

Chi-square is used to determine whether there is a commendable difference between the expected and observed values of the categories of the categorical variables. Xin Jin, Anbang Xu1, Rongfang Bie1, Ping Guo [4] has used chi-square analysis for feature selection of the cancer classification problem.

The purpose of chi-square is to evaluate how likely the observations which are made would be assuming that the null hypothesis is true. Chi-square test is also called as the goodness of fit because it measures how well the observed distribution of the data fits. The following are the requirements when you perform a chi-square test:

1. Quantitative data.
2. One or more categories.
3. Independent observations.
4. Adequate sample size

The actual formula for running a chi-square is as below:

$$\chi 2 = \frac{\sum(0-E)^2}{E}$$

We refer to the **degrees of freedom**, usually labelled as *df* for short, and is defined for the chi-square as the number of categories minus 1. Due to the nature of the chi-square test, we will use the number of categories minus 1 to find the degrees of freedom. The reason to do this is that there is an assumption that the sample data is biased, which helps shift your scores to allow for error which occurred.

We will then locate it in the chi-square distribution table. Using degrees of freedom, we shall locate the *p*-value typically, the *p*-value is 0.05. In our dataset except for Gender and relationship satisfaction, all other attributes appear to be significant.

4.3 Bivariate Analysis

We have plotted bivariate graphs which help us analyze the comparison of expected attrition/event rate and the actual attrition in each category of categorical variables. Figure 3 and figure 4 shows the bivariates for business travel and job involvement respectively.

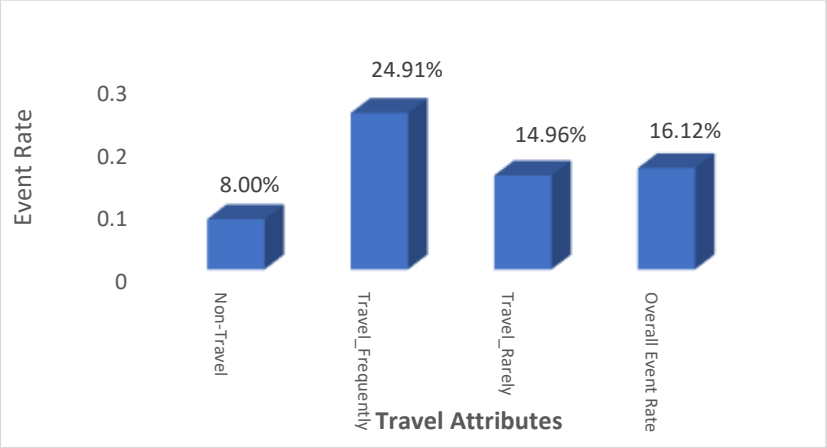

Fig 3. Bivariate for Business Travel

Thus, the higher the difference found between the expected and actual attrition, the more important the feature is.

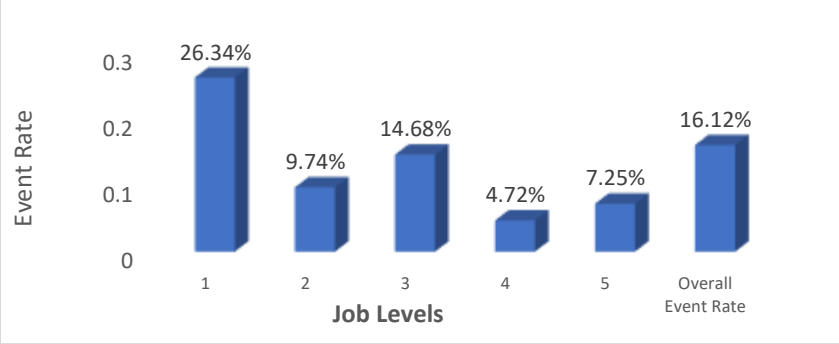

Fig 4. Bivariate for Job Level

From the all the graphs, we can conclude that the following employees characteristics have a high probability of attrition:

1) No Business Travel *(*maybe because they are bored at the same place*)*
2) Medical Education field *(*maybe because they are opening up their own clinic*)*
3) High Job Involvement *(*Maybe because they are saturated now*)*
4) Higher job level (maybe because ample of job opportunities are available in the market at an executive level with lucrative salary)
5) Doing less overtime（Strange, but that could be due to sampling bias.

But sometimes it does happen that only one of the category is notably different from the expected and others are almost inline. It is a confusion in this case whether to include this variable. Also, sometimes, there are less than five observations in a category. Such a small sample could easily be biased and could throw misleading results on the graphs.

Hence, a more advanced method of feature selection called the information value analysis is used to identify the significant variables.

4.4 Information Value

Information value considers both the difference between expected and the actual value in a category of a categorical variable and the number of observations in that category. Thus, it is the most credible source of finding how significant a variable

is.
The formula for the information value is:

$$\text{IV} = \sum_{i=1}^{n}(Distribution\ Good - Distribution\ Bad)\ \text{x}\ \ln(\frac{Distribution\ Good}{Distribution\ Bad})$$

where,
Distribution Good/Bad = Proportion of good (non- attrition)/bad (attrition) employees out of the total good employees in a particular category of a variable.
n= number of categories of a variable.
One can refer here for the calculations:
https://github.com/kashyapbhuva/IBM-Attrition-HR-Analytics/blob/master/IV.xlsx
Final results are as shown in Table 1.

Table 1: Information values in the Descending order:

Attributes	Information Value
Job Role	0.4909
Overtime	0.4001
Job Level	0.3841
Monthly income	0.3408
Age	0.3337
Stock Option Level	0.3190
Marital Status	0.2188
Years in current role	0.1680
No of companies worked	0.1309
Job Involvement	0.1259
Business Travel	0.1208
Environment Satisfaction level	0.0998
Years with current Manager	0.0921
Job Satisfaction level	0.0876
Hourly rate	0.0860
Training times last year	0.0741
Education Field	0.0727
Work-life Balance	0.0669
Department	0.0521
Monthly rate	0.0280
Relationship Satisfaction	0.0245
Education level	0.0165
Years since last promotion	0.0132
Per cent Salary Hike	0.0124
Gender	0.0064
Performance Rating	0.0001

The above table shows the information values of each variable in the dataset in a descending order. Thus, Job Role is the most significant variable affecting the attrition & performance rating is the least significant. The numeric variables are split into the bins and then, their information value was calculated.

V. MODELLING

Once we have the significant features identified, and data is in a model ready format, we could start the predictive analytics part of the project. Classification & Regression Trees are the most commonly used algorithms. There are also some more unique and advanced algorithms like ridge & lasso, naïve Bayes, linear discriminant analysis, support vector machines which are used in this project to compare the results. Since few of the algorithms are affected due to the different scaling of the numeric variables, a feature scaling is always recommended before the analysis. Algorithms like Lasso & Ridge are highly affected due to Feature scaling, while other CART algorithms are not.

The modelling starts by splitting the available data into a training set and a testing set. Algorithms are deployed on the training set and tested on the testing set. Randomly 80 % of the data is assigned to the training set and 20% of the data is assigned to the testing set.

We would be using classification algorithms since our response variable is binary. The “Yes” and “No” characters in the Attrition variable are converted into 1 and 0 respectively for convenience.

5.1 Logistic Regression

The statistical method for the purpose of analysis of this dataset is "Logistic Regression" with the help of one or more number of independent variables by predicting or determining an outcome. Chao-Ying Joanne Peng, Kuk Lida Lee & Gary M. Ingersoll [5] has given a lucid explanation of logistic regression in their research work. The main reason for using Logistic Regression method over here is the outcome variable (i.e. Attrition of an employee) is dichotomous or binary (i.e. '1' represents TRUE value (employee is in attrition) and '0' represents FALSE value (employee is not in attrition. Firstly, the entire dataset is divided or split into train and test datasets randomly in the ratio of 80% train and 20% test. Then, the logistic function is applied by using the "glm" command (for both with and without cross-validation) on the dependent or outcome variable (Attrition) with all the other variables on the train dataset in order to find the best model or the best variables in the dataset that are significant or impacts the dependent variable. Then , with the use of this best fit or best model predictions are made on the test dataset to predict whether the given new employee will be in attrition or not on the basis of the probability values that will range between '0' and '1' and further by taking a particular threshold of 0.5(in this case) the predictions which are made are snapped into '0' and '1' classifications by considering probability value of more than 0.5 as '1' classification or TRUE value (i.e. given employee will be in attrition) and probability value of less than 0.5 as '0' classification or FALSE value (i.e. given employee will not be in attrition). For the purpose of checking the accuracy, confusion matrix is created which basically compares the actual attrition (i.e. dependent variable) values of our test dataset with the prediction which we have made of the attrition values on the test dataset with the use of best model that we have found on train dataset by random split of overall dataset into train and test in ratio of 80% train and 20% test. With the help of this confusion matrix as shown in Table 2, we are able to find out the number of values which are misclassified in our predictions on test dataset and then further various model evaluation metrics are found out.

Table 2: Confusion matrix & evaluation for logistic model

	Actual	
Predicted	0	1
0	230	31
1	10	23
Model Evaluation Metrics		
Accuracy	0.8605	
Precision	0.8812	
Recall	0.9583	
FN Rate	0.1091	
F-measure	0.9182	

Precision denotes how precise the model is in predicting the attrition. It is the ratio of the number of truly predicted positive values to the total positively predicted values. The recall is out of the actual positive values, how many were predicted to be positive. Precision and recall are inversely related to each other.

Lastly, the ROC curve has been plotted. The ROC curve is a fundamental tool for the purpose of diagnostic test evaluation. In a ROC curve, the true positive rate (Sensitivity) is plotted as a function of the false positive rate (100-Specificity) for the different cut-off points of a parameter. Each point on the ROC curve which represents a sensitivity/specificity pair corresponding to a particular value of decision threshold. The area under the ROC curve (AUC) is a measure of how well a parameter can distinguish between two diagnostic groups (i.e. attrition or not in this case of a given employee) . ROC curve for the logistic regression model without cross-validation is as shown on the next page in Figure 5.

5.2 Linear Discriminant Analysis

Linear discriminant analysis (LDA) (also known as discriminant function analysis) is a method which is used for the generalization of the Fisher's linear discriminant, a method which is used in statistics, for the purpose of pattern recognition and machine learning for finding a linear combination of features which characterizes or separates two or more classes of objects or events. EI Altman [6] has successfully used the discriminant analysis for predicting the corporate bankruptcy. for Basically, in the linear discriminant analysis, the outcome or the dependent variable can be continuous and can contain an infinite number of possible values and finds the linear combination of features which best characterize or separate the two or more classes. So, LDA is applied to this dataset for finding the linear combination of features for best characterizing or separating 2 groups that whether the given employee will be in attrition or not. Firstly, the entire dataset is divided or split into train and test datasets randomly in the ratio of 80% train and 20% test.

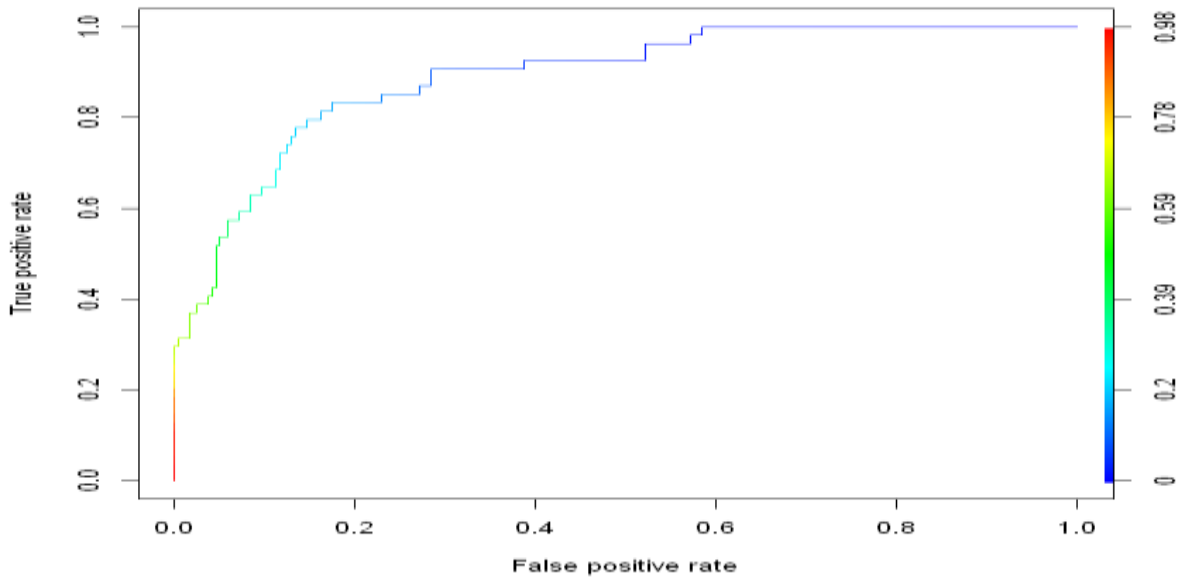

Fig 5. ROC Curve for Logistic Regression

Then, the "lda" command is applied on the training dataset (for both with and without cross-validation) on the dependent or outcome variable of "Attrition" with all other independent variables in the dataset. Then , by using this LDA function predictions are made on the test dataset to predict whether the given new employee will be in attrition or not on the basis of the probability values that will range between '0' and '1' and further by taking a particular threshold of 0.5(in this case) the predictions which are made are snapped into '0' and '1' classifications by considering probability value of more than 0.5 as '1' classification or TRUE value (i.e. given employee will be in attrition) and probability value of less than 0.5 as '0' classification or FALSE value (i.e. given employee will not be in attrition). For the purpose of checking the accuracy, confusion matrix is created which basically compares the actual attrition (i.e. dependent variable) values of our test dataset with the prediction which we have made of the attrition values on the test dataset with the use of "lda" model that we have found on train dataset by random split of overall dataset into train and test in ratio of 80% train and 20% test.

Table 3: Confusion matrix & evaluation for LDA

	Actual	
Predicted	0	1
0	232	32
1	8	22
Model Evaluation Metrics		
Accuracy	0.8639	
Precision	0.8788	
Recall	0.9667	
FN Rate	0.1118	
F-measure	0.9206	

With the help of this confusion matrix, we are able to find the number of values which are misclassified in our predictions on test dataset and then further finding the accuracy of our predictions on the test dataset. Rest all procedures are same as logistic regression. With the help of this confusion matrix as shown in Table 3, we are able to find out the number of values which are misclassified in our predictions on test dataset and then further various model evaluation metrics are found out.

5.3 Lasso & Ridge Classification

Lasso (least absolute shrinkage and selection operator) is a regression analysis method that performs both variable & regularization to enhance the prediction accuracy & interpretability of the Model it produces. Lasso was formulated for least square models it is used for the estimator, including its relationship to ridge regression. It is the best subset selection method.

N. Rao, R. Nowak, C. Cox and T. Rogers [7] has explained the algorithm of lasso in more detail. Lasso & Ridge shrinks the coefficients of the variables. Thus, the problem of over-fitting on account of multiple attributes is overcome, and only the key performing variables survive.

Shuangge Ma, Jian Huang [8] has successfully used penalized feature selection in the domain of bioinformatics. Ridge Regression is used for analyzing multiple regression data that suffer from multicollinearity.in multicollinearity, least squares estimates

are unbiased, and their variances are very large may be far from the true value. By the addition of a degree of bias to the regression estimates, ridge regression curbs down the standard errors. The net effect will be to give estimates that are more reliable.

Multicollinearity can lead to inaccurate estimates of the regression coefficients, thereby increase the number of standard errors of the regression coefficients, deflate the partial t-tests for the regression coefficients, give false, nonsignificant, p - values, and degrade the predictability of the model.

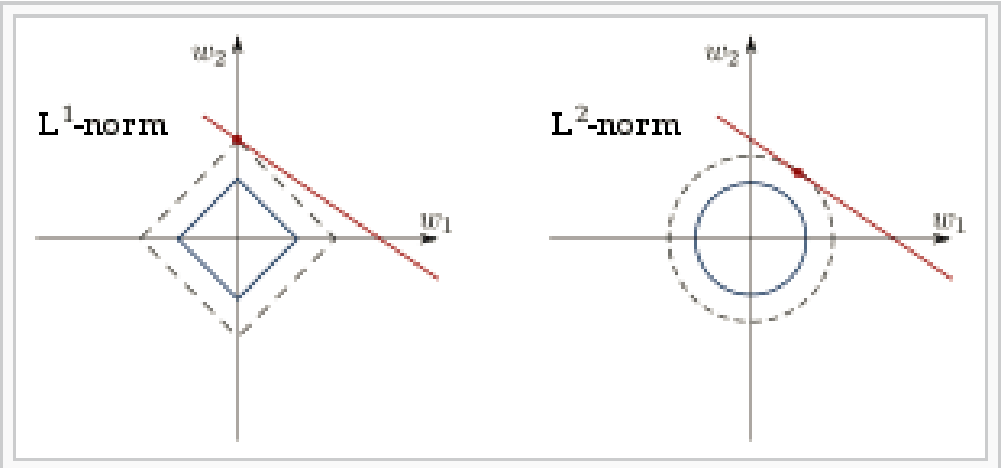

Fig 6. Lasso vs Ridge (Source: Wikiwand)

In lasso, coefficients can set to zero, whereas in ridge regression, which appears superficially similar, cannot. Because of difference in the shape of the constraint boundaries in the two cases. Both lasso and ridge regression can be interpreted as minimizing the same objective function. With the help of this confusion matrix as shown in Table 4 and Table 5, we are able to find out the number of values which are misclassified in our predictions on test dataset and then further various model evaluation metrics are found out for Ridge & Lasso Classification respectively.

Since lasso reduces the no of the variable which is not significant we can see of total variables 55 variables, 24 variables are insignificant hence removed also from confusion matrix one can see the accuracy of the model is 84.6% which is pretty good.

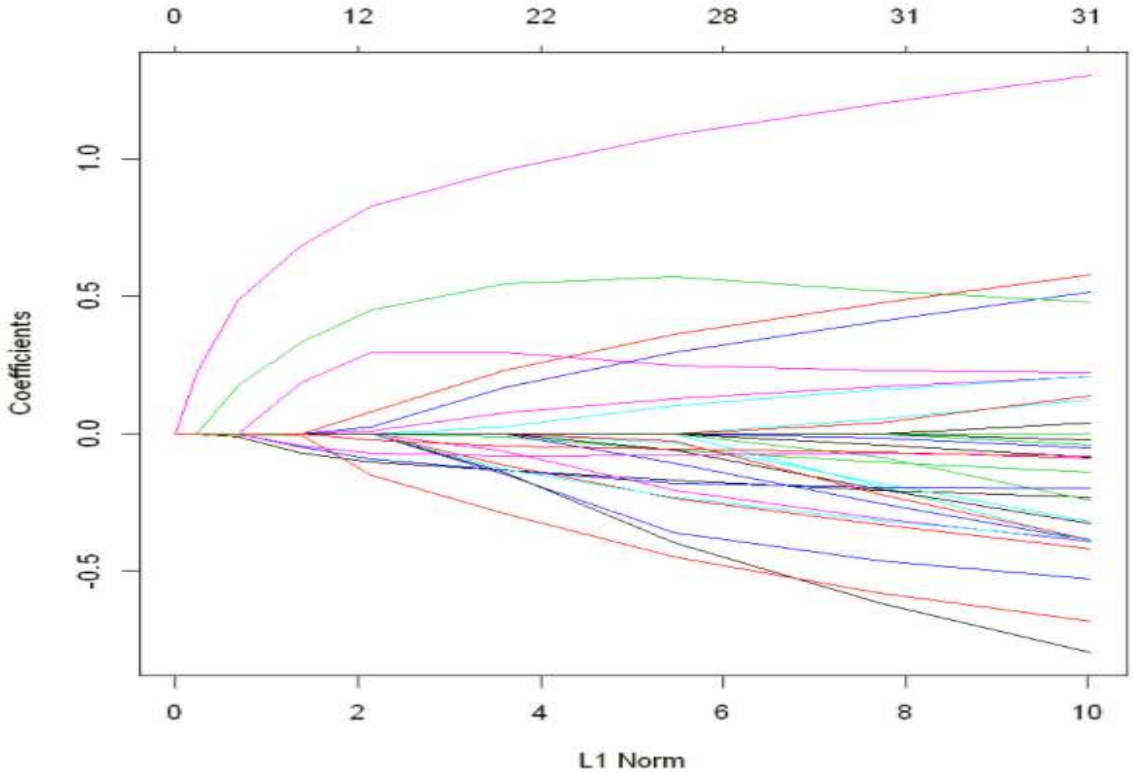

Fig 7. Shrinking of coefficients at various values of λ For Lasso

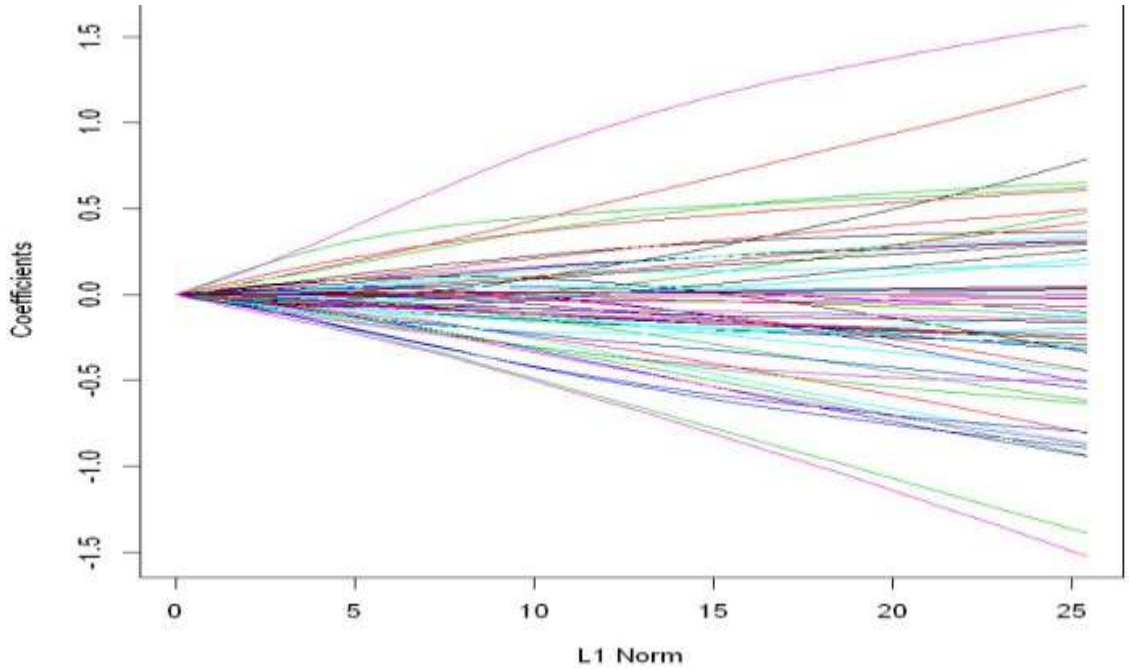

Fig 8. Shrinking of coefficients at various values of λ For Ridge

Table 4: Confusion Matrix & Evaluation for the Ridge model

	Actual	
Predicted	0	1
0	240	44
1	0	10
Model Evaluation Metrics		
Accuracy	0.8503	
Precision	0.8451	
Recall	1.0	
FN Rate	0.1496	
F-measure	0.9160	

Table 5: Confusion Matrix & Evaluation for the Lasso model

	Actual	
Predicted	0	1
0	239	44
1	1	10
Model Evaluation Metrics		
Accuracy	0.8469	
Precision	0.8445	
Recall	0.9958	
FN Rate	0.1501	
F-measure	0.9140	

5.4 Decision Tree

A decision tree is basically used for building classification or regression models in the form of a tree-like structure. Imas Sukaesih Sitanggang & Mohd Hasmadi Ismail [9] used decision trees to classify hotspot occurrences. It breaks down a dataset into smaller and smaller number of subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with the required decision nodes and leaf nodes. Leaf node (i.e., target or dependent variable "Attrition" for this dataset) represents a classification or decision. The other independent variables represent the decision nodes. Firstly, the entire dataset is divided or split into train and test datasets randomly in the ratio of 80% train and 20% test. Then, the decision tree model is applied on the training dataset with the help of "rpart" command on the dependent or outcome variable of "Attrition" with all the other independent variables in the dataset. The entire decision tree model on this training dataset is plotted with the help of command

"fancyRpartPlot".With the use of this decision tree model of train dataset , the required predictions are made on the test dataset to predict whether the given new employee will be in attrition or not on the basis of the probability values that will range between '0' and '1' and further by taking a particular threshold of 0.5(in this case) the predictions which are made are snapped into '0' and '1' classifications by considering probability value of more than 0.5 as '1' classification or TRUE value (i.e. given employee will be in attrition) and probability value of less than 0.5 as '0' classification or FALSE value (i.e. given employee will not be in attrition).

Table 6: Confusion Matrix & Evaluation for the Decision Tree

	Actual	
Predicted	0	1
0	226	37
1	14	17
Model Evaluation Metrics		
Accuracy	0.8265	
Precision	0.8593	
Recall	0.9417	
FN Rate	0.1321	
F-measure	0.8986	

Table 7: Confusion Matrix & Evaluation for the Random Forests

	Actual	
Predicted	0	1
0	238	46
1	2	8
Model Evaluation Metrics		
Accuracy	0.8367	
Precision	0.8380	
Recall	0.9917	
FN Rate	0.1575	
F-measure	0.9084	

5.5 Random Forests

The Random Forest method, first introduced by Breiman in 2001 can be grouped under the category of ensemble models. Why ensemble? The building block of a Random Forest is the ubiquitous Decision Tree. The decision tree as a standalone model is often considered a "weak learner" as its predictive performance is relatively poor. Provost F, Hibert C & Malet J P [10] have used random forests to classify endogenous seismic sources in a landslide. However, a Random Forest gathers a group (or ensemble) of decision trees and uses their combined predictive capabilities to obtain relatively strong predictive performance - "strong learner".This principle of using a collection of "weak learners" to come together to create a "strong learner" underpins the basis of ensemble methods which one regularly comes across in Machine learning. RF samples randomly training data with replacement on constructing each decision tree that is called bagging. Each decision tree returns a class and then bagging combines them to reach a unique decision (Breiman, 2001). Random forest is an ensemble of the decision trees is expected to perform better and hence give a higher accuracy.

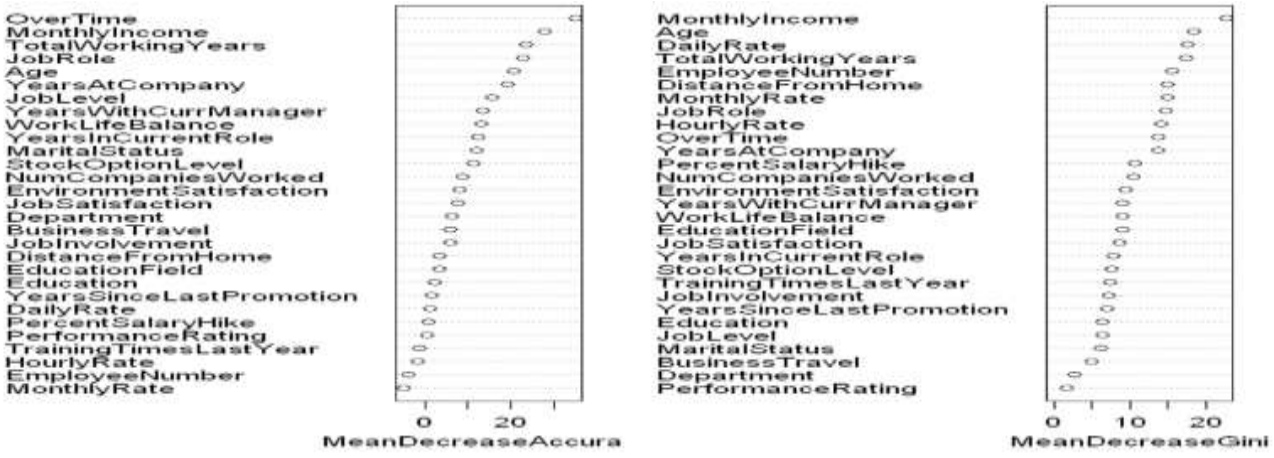

Fig 9. Variable Importance Plot

VI. CONCLUSION

In this paper, employee attrition was predicted on IBM, USA data having 35 data mining techniques and machine learning algorithms. The focus is on using different algorithms and combinations of several target attributes for intelligent and effective employee attrition prediction using data mining as Employee Attrition is one of the biggest Business Problem. Logistic regression, Linear Discriminant Analysis (LDA), Decision Tree Classification, Lasso Regression, Ridge Regression, Random Forest, Naïve Bayes Classification and Support Vector Machine algorithms are performed on the IBM employee attrition data. We analysed that accuracy obtained through Linear Discriminant Analysis Model having 86.39% efficiency outperforms than the mining techniques. Hence, the outcome of the predictive data mining techniques on the same dataset reveals that Linear Discriminant Analysis outperforms than other data mining technique followed by Logistic Regression Model for this particular dataset if accuracy is the metric preferred.

VII. ACKNOWLEDGEMENT

Sincere thanks to Prof. Kriti Srivastava for her valuable guidance.

REFERENCES

[1] Zhu Yan-li, Zhang Jia, Research on Data Preprocessing In Credit Card Consuming Behavior Mining Energy Procedia 17 (2012) 638 – 643, 2012 International Conference on Future Electrical Power and Energy Systems

[2] Isabelle Guyon, Andre Elisseeff," An Introduction to Variable and Feature Selection", Journal of Machine Learning Research 3 (2003) 1157-1182.

[3] Mark. A. Hall," Correlation-based feature selection for discrete and numeric class machine learning". Working Paper Series ISSN 1170-487X.

[4] Xin Jin, Anbang Xu1, Rongfang Bie1, Ping Guo1," Machine Learning Techniques and Chi-Square Feature Selection for Cancer Classification Using SAGE Gene Expression Profiles", Data Mining for Bio-medical applications. LNBI 3916

[5] Chao-Ying Joanne Peng, Kuk Lida Lee & Gary M. Ingersoll," An Introduction to Logistic Regression Analysis and Reporting". The Journal of Educational Research, Volume 96, 2002 - Issue 1

[6] EI Altman," Financial ratios, discriminant analysis and the prediction of corporate bankruptcy", The Journal of finance, 1968 - Wiley Online Library

[7] N. Rao, R. Nowak, C. Cox and T. Rogers, "Classification With the Sparse Group Lasso," in IEEE Transactions on Signal Processing, vol. 64, no. 2, pp. 448-463, Jan.15, 2016

[8] Shuangge Ma, Jian Huang; Penalized feature selection and classification in bioinformatics, Briefings in Bioinformatics, Volume 9, Issue 5, 1 September 2008, Pages 392–403

[9] Imas Sukaesih Sitanggang & Mohd Hasmadi Ismail (2011) "Classification model for hotspot occurrences using a decision tree method", Geomatics, Natural Hazards and Risk, 2:2, 111-121

[10] Provost F, Hibert C, Malet J P, et al. "Automatic classification of endogenous seismic sources within a landslide body using random forest algorithm" [C]//EGU General Assembly Conference Abstracts. 2016, 18: 15705