Attrition Rate Prediction

Group 4: Data Mining Final Project

ISM6136.901F22

By

Mounica Pothureddy (U96174850)

Prasuna Challa (U99156895)

Sree Krishnakanth Gurram (U74957812)

JayChandra Yadlapalli (U61385249)

Contents

Introduction:

Currently, Workplaces are going through talent drain resulting in high Attrition rates. One of the most effected areas of work is Tech Industry. This high attrition rates leads to lose of expertise, stagnation, scaling problems in hiring fresh talent, increased training costs etc. There may be many causes driving the employees to leave their job in pursuit of others. Identifying some of these causes and build robust predictive models to aid the employers to bring these migration forms the crux of our motivation to pursue this project.

Motivation:

1. In the wake of COVID-19 pandemic, phenomenon “Great Resignation”, “Quiet-Quitting” gathered traction where employees voluntarily resigned from their jobs resulting in high attrition rates in numerous companies across sectors.
2. In this project, we attempted to understand the factors driving this behavior among employees.
3. Identify major factors where employers can focus to retain their best talent.

Source:

Data-Mining-Employee-Attrition-Project/Attrition.csv at main · quocduyenanhnguyen/Data-Mining-Employee-Attrition-Project · GitHub

Methodology and Evaluation Metrics:

Data mining is a valuable technique for identifying hidden relationships and trends in our data. It also aids in the prediction of the prospective outcomes of various data analysis applications. Here we are using different DM methods of classification, regression, neural network to predict the employee attrition.

Based on our problem statement we need to understand the factors and attributes related to problem and gather data around it and some attributes could be derivate which can be ignored, and they would have correlation among them and might impact the model. Once we gather the data, we should understand what our dependent and independent variables are based on which our next further analysis would be started. Here in our project our dependent variable would be Attrition.

This dependent variable is binary either 1(Attrition) or 0(No Attrition) is predicted based on the given attributes. Here we are using 3 different tools Azure ML Studio, SAS Enterprise Miner to analyze our data and models. Azure ML Studio and SAS Enterprise Miner are used to build models and compare them.

Data Description:

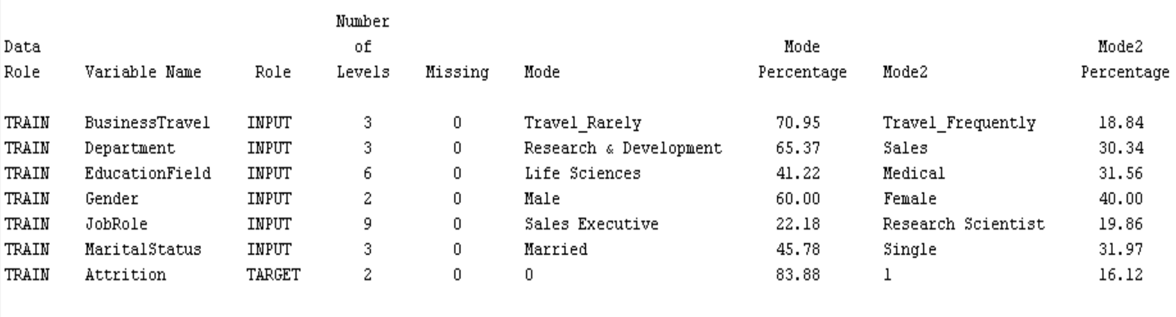
For the project, we downloaded a dataset from the GITHUB (dataset is an extension from the one in Kaggle) The dataset contains the information about the employee details, their experiences, and about attrition of an anonymous company. There are 31 variables in which 29 are independent variables ,1 is a dependent variable and the other is employee Identifier. The data contains around 2941 rows. The attributes are being explained below. The target variable is the **Attrition**

Attribute	Description
Employee Number	Employee Identifier
Attrition	Target variable 1- Yes attrition, 0- No attrition
Age	Age of the employee
Business Travel	Travel commitments for the job (categorical)
Department	Employee Department (categorical)
Distance from Home	Distance from work to home (in km)
Education	1-Below College, 2-College, 3-Bachelor, 4-Master,5-Doctor (categorical)
Education Field	Field of Education (categorical)
Environment Satisfaction	1-Low, 2-Medium, 3-High, 4-Very High (categorical)
Gender	Employee's gender (categorical)
Job Involvement	1-Low, 2-Medium, 3-High, 4-Very High (categorical)
Job Level	Level of job (1 to 5)
Job Role	Job Roles (categorical)
Job Satisfaction	1-Low, 2-Medium, 3-High, 4-Very High (categorical)
Marital Status	Marital Status (categorical)
Monthly Income	Monthly Salary
NumCompaniesWorked	Number of companies worked at
Over18	Over 18 years of age? (categorical)
Overtime	Overtime? (categorical)
Percent Salary Hike	The percentage increase in salary last year
Performance Rating	1-Low, 2-Good, 3-Excellent, 4-Outstanding (categorical)
Relationship Satisfaction	1-Low, 2-Medium, 3-High, 4-Very High (categorical)

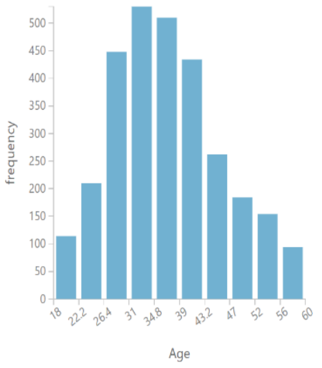
Standard Hours	Standard Hours
StockOptionLevel	Stock Option Level
TotalWorkingYears	Total Work Experience
TrainingTimesLastYear	Number of trainings attended last year
WorkLife Balance	1-Low, 2-Good, 3-Excellent, 4-Outstanding
YearsAtCompany	Years at Company
YearsInCurrentRole	Years in the current role
YearsSinceLastPromotion	Years since the last promotion
YearsWithCurrManager	Years with the current manager

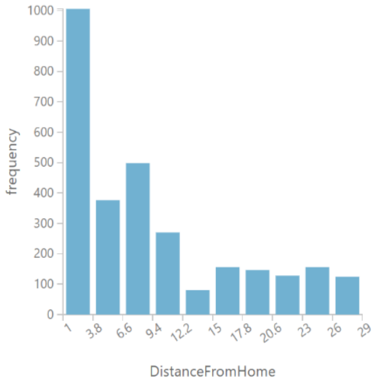
Exploratory Analysis

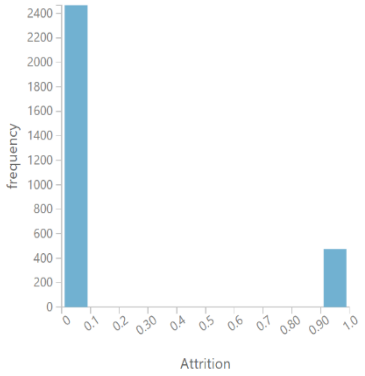
We conducted a preliminary analysis of the whole data set to get a sense of the data using Azure ML Studio and SAS Enterprise Miner and got the below results.

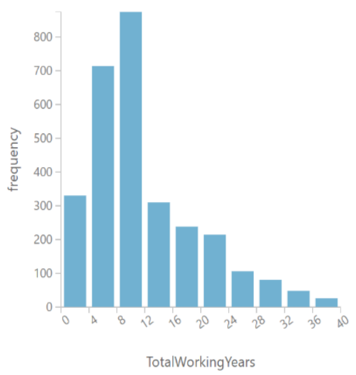


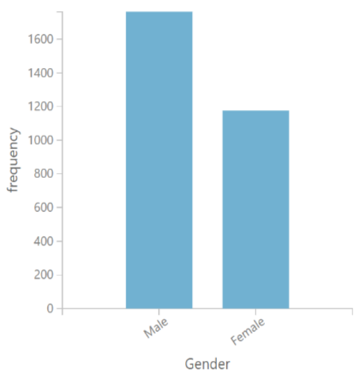
```
                                     Number
Data                                   of                                         Mode                          Mode2
Role     Variable Name      Role     Levels   Missing   Mode                   Percentage   Mode2              Percentage

TRAIN    BusinessTravel     INPUT       3         0     Travel_Rarely             70.95     Travel_Frequently     18.84
TRAIN    Department         INPUT       3         0     Research & Development    65.37     Sales                 30.34
TRAIN    EducationField     INPUT       6         0     Life Sciences             41.22     Medical               31.56
TRAIN    Gender             INPUT       2         0     Male                      60.00     Female                40.00
TRAIN    JobRole            INPUT       9         0     Sales Executive           22.18     Research Scientist    19.86
TRAIN    MaritalStatus      INPUT       3         0     Married                   45.78     Single                31.97
TRAIN    Attrition          TARGET      2         0     0                         83.88     1                     16.12
```

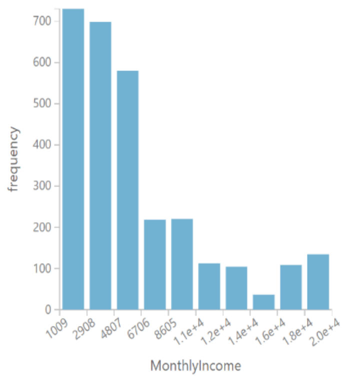








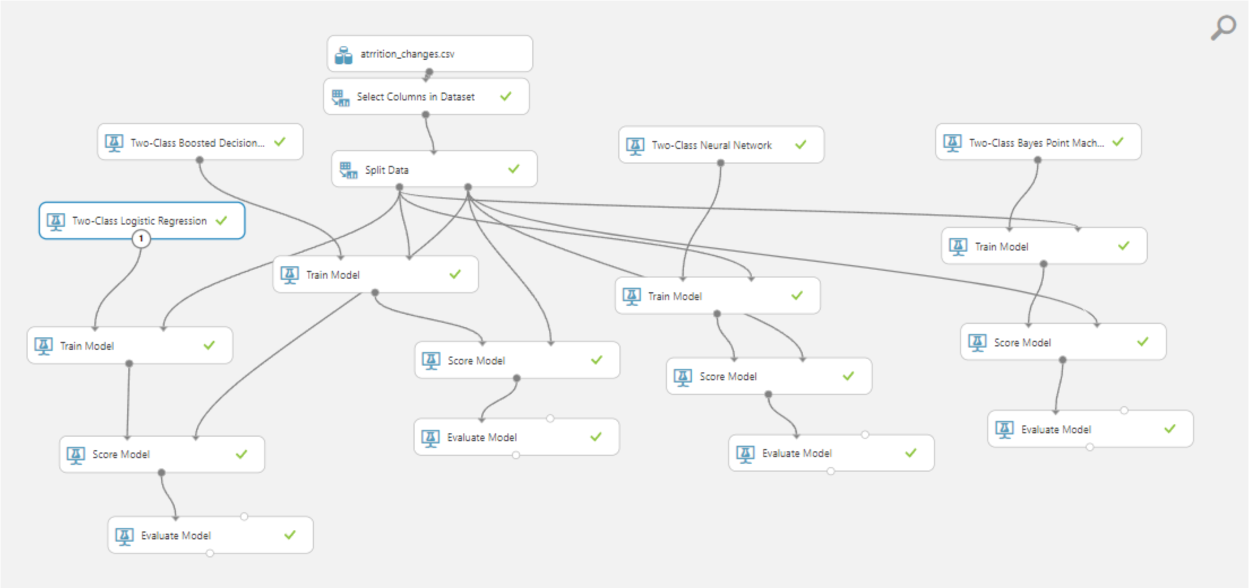




1. Missing Value column showing ‘0’, indicating no missing values in the dataset.
2. Most of the employees are younger people and the average age of the data set is 37.
3. Most employees commute around 1-3miles every day and the average commute distance is 9miles.
4. The current attrition rate is around 16.1% which is high, and the company needs to address some issues to bring this down.
5. The dataset fairly represents both male and female employees.

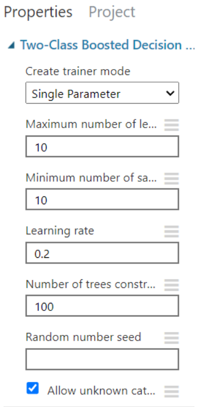
Algorithms & Results

1. In the Select column of the dataset node, All the additional columns created at the data preprocessing are included and the original columns corresponding to the columns created are omitted to maintain the integrity of the source dataset.
2. Data is split across train and test in the ratio of 7:3



Two-Class Boosted Decision Tree

Two-Class Boosted Decision Tree module creates a machine learning model that is based on the boosted decision trees algorithm. A boosted decision tree is an ensemble learning method in which the second tree corrects for the errors of the first tree, the third tree corrects for the errors of the first and second trees, and so forth. Predictions are based on the entire ensemble of trees together that makes the prediction.



Properties Project

Two-Class Boosted Decision ...

Create trainer mode: Single Parameter

Maximum number of le...: 10

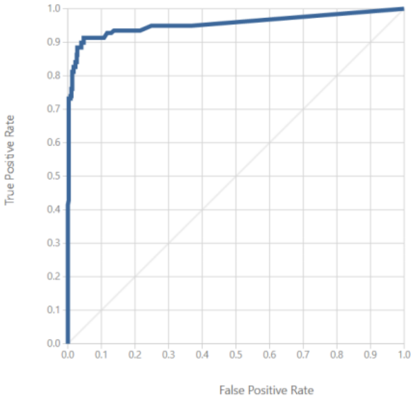
Minimum number of sa...: 10

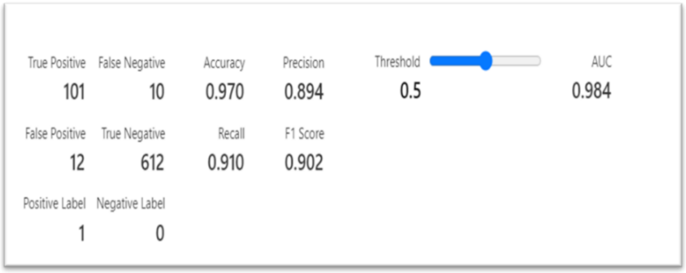
Learning rate: 0.2

Number of trees constr...: 100

Random number seed:

☑ Allow unknown cat...

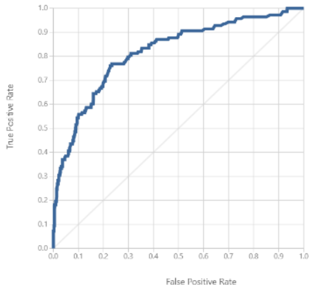


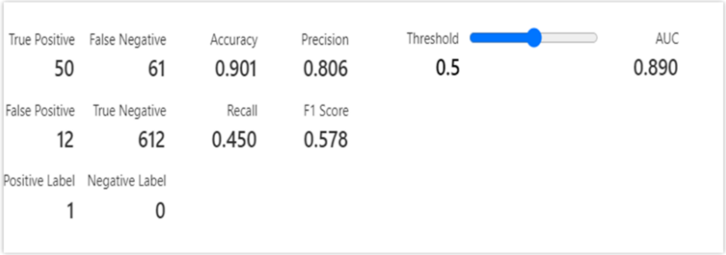


True Positive	False Negative	Accuracy	Precision	Threshold	AUC
101	10	0.970	0.894	0.5	0.984
False Positive	True Negative	Recall	F1 Score		
12	612	0.910	0.902		
Positive Label	Negative Label				
1	0				

Two-Class Logistic Regression

Logistic regression is designed for two-class problems, modeling the target using a binomial probability distribution function. The class labels are mapped to 1 for the positive class or outcome and 0 for the negative class or outcome.

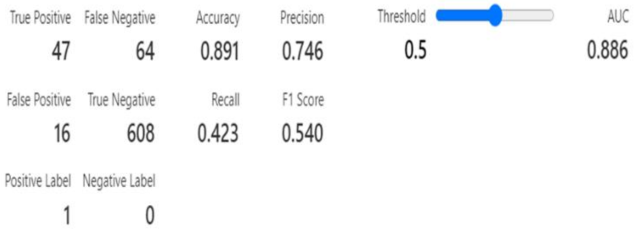




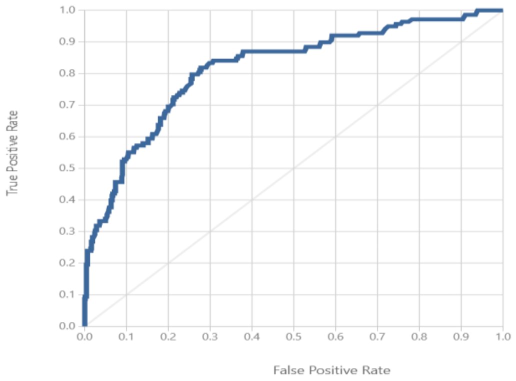
True Positive	False Negative	Accuracy	Precision	Threshold	AUC
50	61	0.901	0.806	0.5	0.890
False Positive	True Negative	Recall	F1 Score		
12	612	0.450	0.578		
Positive Label	Negative Label				
1	0				

Two-Class Bayes Point Machine

Naive Bayes classifiers are simple probabilistic classifiers based on applying Bayes theorem with strong independence assumption between the features.



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
47	64	0.891	0.746	0.5	0.886
False Positive	True Negative	Recall	F1 Score		
16	608	0.423	0.540		
Positive Label	Negative Label				
1	0				

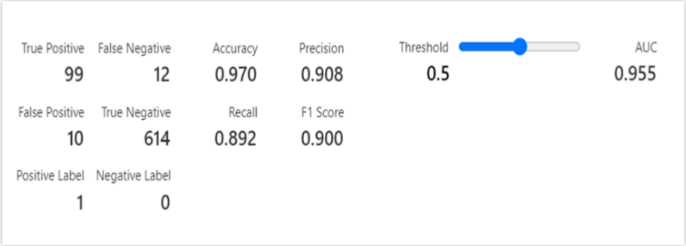


Two-Class Neural Network

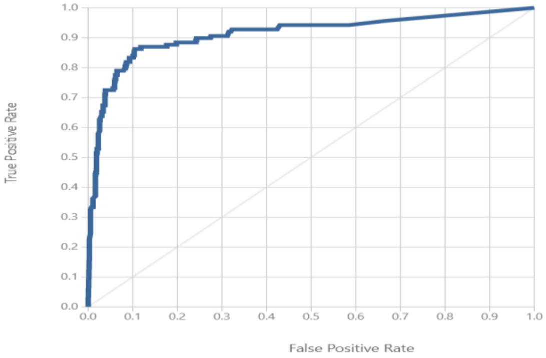
A neural network is a set of interconnected layers. The inputs are the first layer and are connected to an output layer by an acyclic graph comprised of weighted edges and nodes. Between input and output layer you can insert any number of hidden layers. But most of the predictive tasks can be achieved by one or few hidden layers. All nodes in a layer are connected by weighted edges to the next layer. Neural networks are trained iteratively by optimization techniques called gradient descent.

Back-Propagation

This iterative process is the essence of neural network training. It’s a practice of fine tuning of weights based on the error rate obtained in the previous iteration. Proper tuning of weights ensures reliability of model by decreasing the error rate.



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
99	12	0.970	0.908	0.5	0.955
False Positive	True Negative	Recall	F1 Score		
10	614	0.892	0.900		
Positive Label	Negative Label				
1	0				



Evaluation Metrics

Precision

The precision value is affected from false positive from confusion matrix.

P = TP/(TP+FP)

Recall

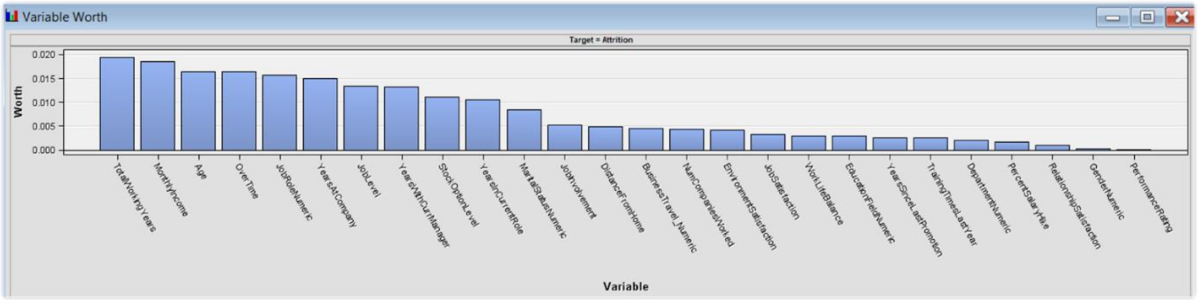
The recall value is affected from false negatives

Recall = TP/(TP+FN)

F1 Score = **2*((Precision*recall)/(precision+recall))**

Model Comparison

Algorithm	Accuracy	Precision	Recall	F-1 Score
Decision Tree	0.970	0.894	0.910	0.902
Neural Network	0.970	0.908	0.892	0.900
Logistic Regression	0.901	0.806	0.450	0.578
Bayes Point Machine	0.891	0.746	0.423	0.540



Conclusion

1. From the above results, both Decision tree and Neural Network are giving almost similar results for this dataset.
2. From the variable worth chart, Factors impacting attrition are given in decreasing order of impact. ***Since, this is predictive modelling causality cannot be explained but these may be the factors where employers can concentrate to retain their best talent.***

Further Study

1. Spend much more time in collecting the data specifically in tech industry. Preferred medium of collecting data would be LinkedIn, Facebook, Quora and WhatsApp.
2. Inclusion of additional attributes like preferred mode of work, and any other inflationary effects.
3. Build robust models on data specific to location like Asia and North America and understand the employee behavior in different parts of world.