Predictive Analysis: Factors Affecting IBM Employee Attrition

Mitzie Irene P. Conchada, 301258577

Mary Claire C. Doña, 301323966

Karen Ann H. Francisco, 301238093

Jason S. Yap, 301293413

Centennial College

BA 723–Capstone

David Parent

July 11, 2024

Introduction

Employee attrition can be an issue for companies since this translates to higher costs when they hire and train new employees. Companies take on different strategies to keep their employees, especially the talented ones and those who can potentially contribute to the growth of the company. Another important aspect of employee attrition is the impact on employee's morale, productivity and overall organization culture (Conchada et al., 2023). This study investigated the various factors affecting employee attrition, specifically at IBM.

Business Problem Statement

This study used the IBM dataset from Kaggle, which has 1,470 observations, 12 predictors, and 1 target variable. The research question of this study delved into: What are the factors affecting employee attrition at IBM? To answer the research question, the following are the business objectives:

1. Determine the factors affecting employee attrition at IBM using predictive modelling that will represent the relationship between inputs and the target variable attrition. This study uses the following predictive models:
    a. Decision tree
    b. Regression analysis
    c. Neural networks
2. Choose the best model that yields favorable results based on model assessment comparison.
3. Recommend steps related to employee attrition and points for future research

Data Exploration

Dataset Overview

The IBM dataset is hypothetical and was derived from an open-source website, Kaggle. The database has 1,470 observations, 12 input variables, and 1 target variable. The data dictionary below shows the list and description of inputs and target variables used in the study. *Attrition* has been identified as the target variable; thus, set in binary format. Moreover, the following variables were set as nominal as they are categorical in nature: *education, education field,* *marital status, department, environment satisfaction, job satisfaction* and *work-life-balance*. The rest of the variables are interval: *age, distance, monthly income*, *years at company*, and *number of companies worked*.

Variables	Description	Level
Attrition (Target)	1: Employee resigned (yes) 0: Employee stayed (no)	Binary
Age	Age	Interval
Distance from home	Distance from home from work in kilometres	Interval
Education	1: Below College 2: College 3: Bachelor 4: Master 5: Doctor	Nominal
Education field	Life Sciences Medical Marketing Technical Degree Human Resources Other	Nominal
Marital status	Married Single Divorced	Nominal
Monthly income	Monthly income	Interval
Years at company	Employees' total working years at IBM	Interval
Number of companies worked	Number of companies worked at prior to IBM	Interval
Department	Research and development Sales Human Resources	Nominal
Environment satisfaction	1: Low 2: Medium 3: High 4: Very High	Nominal
Job satisfaction	1: Low 2: Medium 3: High 4: Very High	Nominal
Work-life-balance	1: Bad 2: Good 3: Better 4: Best	Nominal

Summary Statistics

The distribution of target class is highly imbalanced since the percentage of attrition and non-attrition instances is 83.9% and 16.1%, respectively. This was addressed in the study and a discussion follows in the succeeding section.

The following discussion is an overview of the raw dataset. IBM employees are 18 to 60 years old with an average age of 37. The average distance from home is 9 kilometers. Moreover, most employees attained college or higher-level education, mostly in Life Sciences. The average tenure at IBM is 7 years, while the average number of companies previously worked on before IBM is 3. With respect to monthly income, it averages USD 6,503. For survey questions, the following results were generated: the average rating for *environment satisfaction* is 2.7 or 3 (high); the average rating for *job satisfaction* is 2.7 or 3 (high); and the average rating for *work life balance* is 2.7 or 3 (better). The figure below shows the summary of descriptive statistics.

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum
Age	INPUT	36.92381	9.135373	1470	0	18	36	60
DistanceFromHome	INPUT	9.192517	8.106864	1470	0	1	7	29
Education	INPUT	2.912925	1.024165	1470	0	1	3	5
EnvironmentSatisfaction	INPUT	2.721769	1.093082	1470	0	1	3	4
JobSatisfaction	INPUT	2.728571	1.102846	1470	0	1	3	4
MonthlyIncome	INPUT	6502.931	4707.957	1470	0	1009	4908	19999
NumCompaniesWorked	INPUT	2.693197	2.498009	1470	0	0	2	9
WorkLifeBalance	INPUT	2.761224	0.706476	1470	0	1	3	4
YearsAtCompany	INPUT	7.008163	6.126525	1470	0	0	5	40

Data Preparation

As preparation for modeling, issues such as imbalanced target class, misclassified data type, missing values, skewness, and non-numeric inputs were identified and addressed. These are essential steps since modeling techniques like regression and neural network use prediction formula to train the model and score new cases; thus, they are sensitive to data distribution and completeness.

I. Data Cleaning

i. Data Type

Assigning the right data type for each variable based on what it represents is crucial in models' performance and in interpreting results. For example, variables such as *education*, *work-life-balance, job satisfaction, and environment* were coded as numeric values; however, they represent categorical data. Hence, they were changed from interval to nominal level.

ii. Missing Values

Since the dataset has no missing values, imputation was not deemed necessary.

iii. Skewness

Based on the generated results, the following interval variables have skewness beyond the cut-off of 1: *Years at Company, Monthly Income, Number of Companies Worked,* and *Distance from Home*. Hence, regularization of skewed data is warranted. To mitigate the impact of extreme variables, the following were performed:

- **Cap & Floor** - the first approach performed was Cap and Floor which used a standard deviation of 3 as a cut-off basis for the replacement values. Cap & Floor reduced the skewness of some variables; however, there are still some with skewness greater than 1.
- **Transformation** – the next approach performed to address the remaining skewed variables was log transformation; after which, all variables were managed to have skewness of 1 and below.

iv. Non-numeric Inputs

Simulation was performed to examine the impact of reduced non-numeric levels or categorical inputs. Education, Education Field, and Marital Status were recoded and a total of 2,107 records were affected.

v. Imbalanced Target Class

Imbalanced target class distribution was addressed by employing stratified random sampling technique, extracting 20% of sample from each department group. From 1,233 non-attrition observations, it was reduced to 246 to match with attrition class and achieve an approximate of 50-50 distribution. This was achieved through stratified random sampling, where a random sample was generated within each department to match the number of observations with 'yes' attrition.

II. Input Selection

To ensure optimal model performance, only relevant input variables were selected. The modeling techniques employed, except for Neural Network, have built-in input selection methods. Decision Tree uses split-search algorithm while Regression uses sequential selection methods. Neural Network, however, uses the variables selected by the best Regression model. Below is the summary of input selection per model.

Model	Method	Threshold Basis
Decision Tree	Split-search algorithm	logworth of >=0.7
Regression	Sequential Selection • Forward • Backward • Stepwise	p-value of <= 0.05
Neural Network	Based on the input selected of the best Regression Model	

III. Data Partition

The dataset was partitioned, allocating 50% for both training and validation. The training data was used ‘for fitting the model,’ while the validation data was used ‘for empirical validation’ (Parent, 2023). This is especially important for the model to have the right amount of flexibility (not overfit nor underfit) and give us the best generalization from the results. With smaller raw data sets, model stability can become a critical issue. In this case, increasing the number of cases devoted to the training partition can be a reasonable course of action. But since our model is stable, given its results, we think there is no need to increase the number of cases for the training partition.

Modeling Exploration

Predictive modeling is a statistical technique to predict future occurrences based on historical data. There are different predictive models widely used in industries and applications to solve various issues such as prediction of employee churn. All models follow essential tasks below:

- Predict new cases that leads to a decision, rank, or estimate
- Select useful inputs
- Optimize model complexity

In this study, the model helps predict the factors that contribute to employee attrition, takes all relevant variables and no redundant inputs, and adjusts the complexity to avoid underfitting or overfitting that might lead to bias results.

This section aims to identify the most suitable model to address the research question: What are the factors affecting employee attrition? The following models were analyzed: Decision Tree, Logistic Regression and Neural Networks.

I. Decision tree

The first predictive model is the decision tree, which is a tree-like structure to model decision and the possible outcomes. The goal of a decision tree model is to predict the value of the target variable based on the values of the predictors. This is the most common model as it is easy to visualize, understand, and interpret results; however, this may not perform as powerful as other predictive modeling techniques in complex datasets. Decision trees help in identifying significant relationships between input and target variables in a dataset.

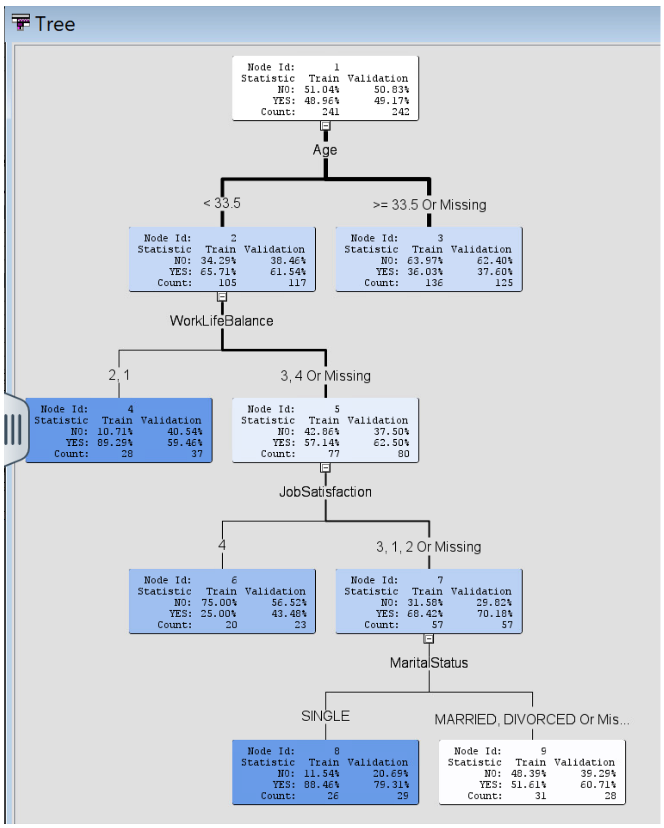
Several types of trees are called in to observe different insights with the following details below:

Decision Trees			
	Maximal Tree	**Misclassification Tree**	**ASE Tree**
Method	Largest	Assessment	Assessment

Assessment Measure	Decision	Misclassification	Average Score Error

Maximal Tree

The maximal tree is the full potential of the tree that is based on the statistical measure of split logworth on the training data. The following properties were used: Subtree Method – Largest, Assessment Measure - Decision. The Largest option provided an independent way to generate the Maximal Tree. Below is the generated maximal tree:



- Number of leaves: 5
- Variable with the highest logworth: *age;* split into three branches *work life balance*, *job satisfaction*, and *marital status.* The Logworth value determines the split: the highest logworth is the best split.
- Fit Statistics showed an Average Squared Error (ASE) value of **0.247598**. The

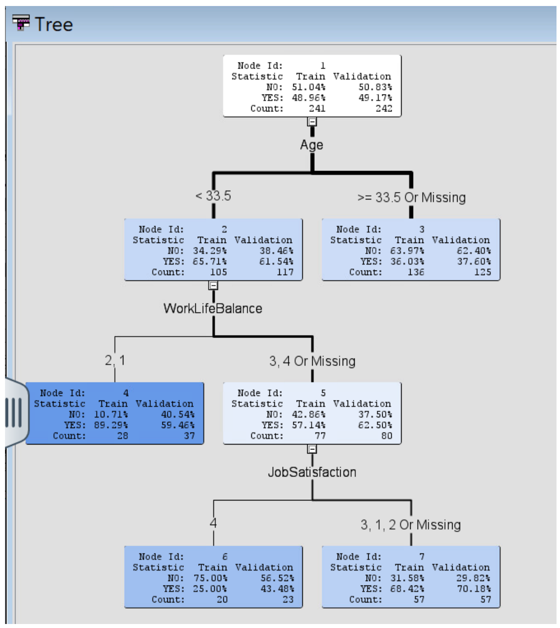
ASE is the average squared difference between the predicted/estimated and actual value. The smaller the value, the better since it signifies that the actual values are very close to the predicted value.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
Attrition	Attrition	NOBS	Sum of Frequencies	241	242
Attrition	Attrition	MISC	Misclassification Rate	0.311203	0.367769
Attrition	Attrition	MAX	Maximum Absolute Error	0.892857	0.892857
Attrition	Attrition	SSE	Sum of Squared Errors	96.33988	119.8376
Attrition	Attrition	ASE	Average Squared Error	0.199875	0.247598
Attrition	Attrition	RASE	Root Average Squared Error	0.447074	0.497592
Attrition	Attrition	DIV	Divisor for ASE	482	484
Attrition	Attrition	DFT	Total Degrees of Freedom	241	.

Misclassification Tree

To prune the maximal tree, Assessment Subtree method was used, and Assessment Measure was set to Misclassification that specifies the optimality measure used to select the best tree in the sequence. The default number of maximum branches is two and the following result is generated:



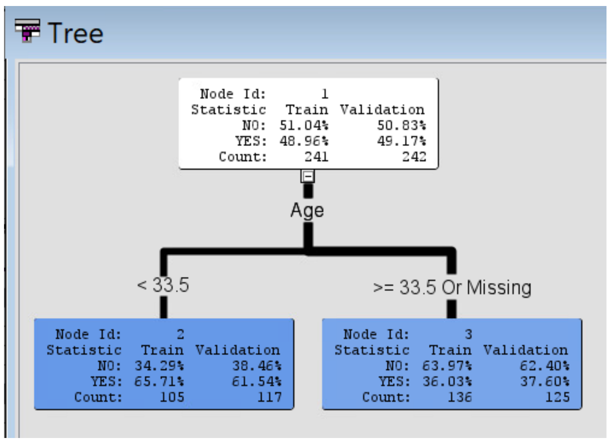
- Number of leaves: 4
- Variable with highest logworth is: *age;* split into two branches*: work life balance*, and *job satisfaction.*
- Fit Statistics showed an Average Squared Error (ASE) value of **0.247744**.

Fit Statistics

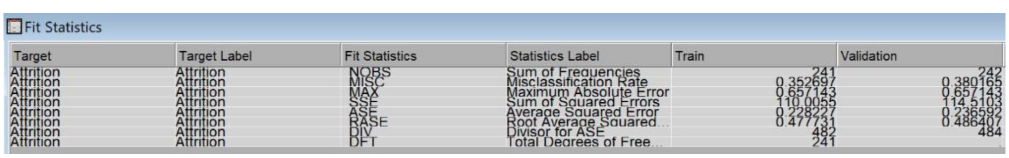
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
Attrition	Attrition	NOBS	Sum of Frequencies	241	242
Attrition	Attrition	MISC	Misclassification Rate	0.311203	0.367769
Attrition	Attrition	MAX	Maximum Absolute Error	0.892857	0.892857
Attrition	Attrition	SSE	Sum of Squared Errors	100.1799	119.9083
Attrition	Attrition	ASE	Average Squared Error	0.207842	0.247744
Attrition	Attrition	RASE	Root Average Squared...	0.455897	0.497739
Attrition	Attrition	DIV	Divisor for ASE	482	484
Attrition	Attrition	DFT	Total Degrees of Free...	241	.

ASE Tree

ASE tree used the following pruning properties: Subtree Method – Assessment; Assessment Measure – Average Squared Error. In generating optimal probability estimates, the Average Square Error is the appropriate assessment measure. The change in assessment measure generated changes in the optimal selection, in this case, the ASE tree pruning sequence is based on the lowest average square error on the validation sample. The generated tree map is below:



- Number of leaves: 2
- Variable with the highest logworth is *age.*
- Fit Statistics revealed an Average Squared Error (ASE) value of **0.236592**



Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
Attrition	Attrition	NOBS	Sum of Frequencies	241	242
Attrition	Attrition	MISC	Misclassification Rate	0.352697	0.380165
Attrition	Attrition	MAX	Maximum Absolute Error	0.657143	0.657143
Attrition	Attrition	SSE	Sum of Squared Errors	110.0055	114.5103
Attrition	Attrition	ASE	Average Squared Error	0.228227	0.236592
Attrition	Attrition	RASE	Root Average Squared...	0.477731	0.486407
Attrition	Attrition	DIV	Divisor for ASE	482	484
Attrition	Attrition	DFT	Total Degrees of Free...	241	

The table below summarizes the ASE values for each of the decision tree models generated.

Decision Trees			
	Maximal Tree	**Misclassification Tree**	**ASE Tree**
Method	Largest	Assessment	Assessment
Assessment Measure	Decision	Misclassification	Average Score Error

ASE Value	0.247598	0.247744	**0.236592**

Based on the ASE values, the **best model** is the **ASE Tree** since it has the lowest ASE value of **0.236592.** From this, it revealed that the variable ***age*** is the **most significant variable** in the model. From the other tree models, other variables *work life balance, job satisfaction,* and *marital status* are also identified as significant splits based on log order.

Significant Splits (Split Node)		
Maximal Tree	**Misclassification Tree**	**ASE Tree**
Age Work Life Balance Job Satisfaction Marital Status	Age Work Life Balance Job Satisfaction	Age

Considering all the significant splits generated in the maximal tree, the following are the observations:

- The input variable ***age*** turned out to be the most significant variable. Employees who are less than 33 years of age are 61.54% more likely to resign compared to those who are older than 33 years old, only 37.60%.
- From the employees who are less than 33 years old and maintain high work life balance (3 and 4), 62.50% more likely to resign compared to those who have less work life balance of 1 and 2, which is at 59.46%.
- Another significant variable is job satisfaction. Even if employees maintain a better work life balance of 3 and 4, but are less satisfied with the job, 70% of them are most likely to leave the company.
- Marital status is another factor for attrition. The employees who are single, aged less than 33 years old are also seen to be more likely to leave the company (79.31%), compared to those who are married or divorced.
- In summary, employees who are single, less than 33 years old, and are not satisfied with the job tend to leave the company and look for other opportunities. Young graduates tend to move from one job to another as they try to figure out the best career that will work out

for them. On the other hand, employees who are married are more likely to stay since having a secure job is more important for them because of their family responsibilities.

- The characteristics of employees who are more likely to resign are: younger employees.
- The characteristics of employees who are more likely to stay: older employees and married.

II. Logistic Regression Model

To evaluate whether and how attrition is influenced by various employee characteristics, a predictive model was established using a logistic regression approach. The target outcome is binary, employee attrition Yes or No, and results were interpreted based on the odds ratio. The logistic regression model included demographic variables such as *Marital Status, Age, Education, Education Field, Monthly Income, Years at Company, Distance from Home*. It also included survey indicators pertaining to: *Environment Satisfaction, Job Satisfaction, and Work Life Balance.*

After data preparation for regression modelling, the results for the Full Regression Model are summarized in the table below. The regression model was first run on a set of non-recoded variables before it was run on a set of recoded variables.

Input Selection and Model Optimization

The study explored sequential methods such as *Forward, Backward, and Stepwise* to identify the best set of input variables and optimize the model complexity of the *Full Regression* model.

In the following sections, we reviewed the results of the sequential regression models without and with recoding.

WITHOUT RECODE

Forward Selection

Forward selection without recode reached until **Step 4**. **Four input variables** (Marital Status, log value of Years at Company, Education Field and Job Satisfaction) were qualified since its p-value fell below the cut-off of <.05. The final input combination of Forward regression is comprised of **11 degrees of freedom** or parameter estimates resulting to overall p-value of <.0001.

Summary of Forward Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq	Validation Error Rate
1	MaritalStatus	2	1	12.2047	0.0022	317.5
2	LOG_REP_YearsAtCompany	1	2	7.6304	0.0057	298.4
3	EducationField	5	3	14.4063	0.0132	301.7
4	JobSatisfaction	3	4	8.3646	0.0390	298.1

he selected model, based on the error rate for the validation data, is the model trained in Step 4. It consists of the following effects:

ntercept EducationField JobSatisfaction LOG_REP_YearsAtCompany MaritalStatus

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
333.993	290.581	43.4120	11	<.0001

Backward Selection

Summary of Backward Elimination

Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq	Validation Error Rate
1	LOG_REP_NumCompaniesWorked	1	11	0.0589	0.8082	306.4
2	LOG_REP_MonthlyIncome	1	10	0.1633	0.6862	308.5
3	EducationField	5	9	6.5050	0.2601	316.2
4	REP_Age	1	8	1.4496	0.2286	320.3
5	LOG_REP_DistanceFromHome	1	7	2.3104	0.1285	320.1
6	EnvironmentSatisfaction	3	6	6.1600	0.1041	322.3
7	Education	4	5	7.9452	0.0936	299.2
8	WorkLifeBalance	3	4	5.3069	0.1507	300.7

e selected model, based on the error rate for the validation data, is the model trained in Step 7. It consists of the following effects:

tercept Department JobSatisfaction LOG_REP_YearsAtCompany MaritalStatus WorkLifeBalance

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
333.993	291.045	42.9482	11	<.0001

The Summary of Backward Elimination indicates that the model reached step 7 with the following variable inputs: Department, Job Satisfaction, log value of Years at Company, Marital Status and Work Life Balance. with p-values more than the stay cut-off of <.05. This model underwent a total of 9 steps.

Stepwise Selection

The summary above is the same as the results in Forward Selection.

Summary of Stepwise Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Validation Error Rate
1	MaritalStatus	2	1	12.2047		0.0022	317.5
2	LOG_REP_YearsAtCompany	1	2	7.6304		0.0057	298.4
3	EducationField	5	3	14.4063		0.0132	301.7
4	JobSatisfaction	3	4	8.3646		0.0390	298.1

he selected model, based on the error rate for the validation data, is the model trained in Step 4. It consists of the following effects:

ntercept EducationField JobSatisfaction LOG_REP_YearsAtCompany MaritalStatus

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
333.993	290.581	43.4120	11	<.0001

WITH RECODE

Forward Selection

Summary of Forward Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq	Validation Error Rate
1	LOG_REP_YearsAtCompany	1	1	8.4077	0.0037	312.6
2	JobSatisfaction	3	2	8.6019	0.0351	308.1
3	REP_Age	1	3	3.9636	0.0465	303.4

he selected model, based on the error rate for the validation data, is the model trained in Step 3. It consists of the following effects:

ntercept JobSatisfaction LOG_REP_YearsAtCompany REP_Age

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
333.993	312.783	21.2101	5	0.0007

Forward selection with recode reached until **Step 3**. **Three input variables** (Job Satisfaction, log value of Years at Company, and Rep Age) were qualified since its p-value fell below the cut-off of <.05. The final input combination of Forward regression is comprised of **5 degrees of freedom** or parameter estimates resulting to overall p-value of <.0001.

Backward Selection

Summary of Backward Elimination

Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq	Validation Error Rate
1	LOG_REP_NumCompaniesWorked	1	11	0.3579	0.5497	305.2
2	REP_EducationField	3	10	2.8077	0.4222	310.2
3	REP_MaritalStatus	1	9	0.6645	0.4150	312.8
4	LOG_REP_MonthlyIncome	1	8	0.7816	0.3766	317.4
5	LOG_REP_DistanceFromHome	1	7	1.7809	0.1820	318.0
6	LOG_REP_YearsAtCompany	1	6	2.7570	0.0968	331.3
7	EnvironmentSatisfaction	3	5	6.6805	0.0828	335.8
8	JobSatisfaction	3	4	7.2295	0.0649	344.2
9	REP_Education	2	3	5.8056	0.0549	327.1

ıe selected model, based on the error rate for the validation data, is the model trained in Step 1. It consists of the following effects:

ıtercept Department EnvironmentSatisfaction JobSatisfaction LOG_REP_DistanceFromHome LOG_REP_MonthlyIncome LOG_REP_YearsAtCompany REP_Age REP_Education REP_EducationField REP_MaritalStatus WorkLifeBalance

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
333.993	281.369	52.6240	21	0.0002

The Summary of Backward Elimination indicates that the model sequentially removed the input variables with p-values more than the stay cut-off of <.05 and retained the following:

Department, Environment Satisfaction, Job Satisfaction, Log rep Distance from Home, Log rep Monthly Income, Rep Age, Rep Education, Rep Educational Field, Rep Marital Status and Work Life Balance. This model underwent only one step.

Stepwise Selection

```
                                    Summary of Stepwise Selection

                     Effect                                 Number      Score        Wald                  Validation
Step    Entered                 Removed                DF      In    Chi-Square   Chi-Square   Pr > ChiSq   Error Rate

   1    LOG_REP_YearsAtCompany                          1       1       8.4077                    0.0037        312.6
   2    JobSatisfaction                                 3       2       8.6019                    0.0351        308.1
   3    REP_Age                                         1       3       3.9636                    0.0465        303.4
   4                            LOG_REP_YearsAtCompany  1       2                     3.7149      0.0539        316.3

he selected model, based on the error rate for the validation data, is the model trained in Step 3. It consists of the following effects:

ntercept  JobSatisfaction  LOG_REP_YearsAtCompany  REP_Age


   Likelihood Ratio Test for Global Null Hypothesis: BETA=0

   -2 Log Likelihood          Likelihood
 Intercept     Intercept &         Ratio
     Only       Covariates    Chi-Square       DF     Pr > ChiSq

   333.993         312.783       21.2101        5         0.0007
```

The summary above is exactly the same as the results in Forward Selection.

Selection of the Best Model

With the aim of optimizing model complexity, sequential-based models were explored to configure the best configuration of input variables. Below is the comparative summary of all regression models:

FIT STATISTICS	Full Regression	Forward Regression	Backward Regression	Stepwise Regression
Validation ASE (without recode)	0.220747	0.21345	0.215446	0.21345
Validation ASE (with recode)	0.222621	0.218034	0.219602	0.218034

The results above indicate Forward Regression (without recode) is the best model since it had the lowest average square error of 0.21345. The next section presents the results and analysis of the odds ratio from the Forward Regression (without recode).

Effect		**Point Estimate**
Education Field	Human Resources vs Technical Degree	4.248
Education Field	Life Science vs Technical Degree	0.381
Education Field	Marketing vs Technical Degree	0.857
Education Field	Medical vs Technical Degree	0.681
Education Field	Other vs Technical Degree	0.219
Job Satisfaction	1 vs 4	3.256
Job Satisfaction	2 vs 4	1.730
Job Satisfaction	3 vs 4	2.001
Log_rep_yearsatcompany		0.603
Marital Status	Divorced vs single	0.201
Marital Status	Married vs single	0.452

The interpretations of odds ratio estimates are as follows:

- Employees having degree in Human Resource are most likely to resign among all educational fields: 4 times more likely than Technical. On the other hand, those in Technical field are more likely to resign by 62%, 14%,31%, and 78% compared with Life Science, Marketing, Medical, and others, respectively.
- With respect to satisfaction surveys, low degree of satisfaction in terms of job satisfaction increases the probability of attrition. Below is the extent of high attrition likelihood versus **4** being the highest rating.

Rating	**Job Satisfaction**
1	3 times
2	73%
3	2 times

- The longer the years of employees' tenure multiplied to a factor of 2.74, the lesser likelihood of resignation by **40%**
- Single employees are more likely to resign by **80%** and **55%** vs. Divorced and Married employees, respectively.

III. Neural Network

Neural network models require a complete record for estimation and scoring, and transformations and replacements are not necessarily needed, but it takes advantage of these two. In this study, Cap and Floor and Transformation of variables were used as discussed in the data preparation and Neural Network with N hidden units ranging from two (2) to eight (8) were called in show varying performance as the model number increases.

Though neural networks in general have no selection of inputs, the models used all those input variables selected by the Forward Regression model, which is the optimal regression model, in preparation for the neural network model that uses hidden units.

The following screenshots show a summary of the Fit Statistics for both the Without Recode model and With Recode model.

WITHOUT RECODE

Fit Statistics
Model Selection based on Valid: Roc Index (_VAUR_)

Selected Model	Model Node	Model Description	Valid: Roc Index	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error	Valid: Misclassification Rate
Y	Neural7	NN Transform	0.757	0.21005	0.28216	0.20965	0.30165
	Neural4	NN Cap&Floor	0.742	0.18892	0.28631	0.20790	0.32231
	Neural9	NN BReg 6H	0.734	0.20715	0.32365	0.21106	0.30992
	Neural	NN BReg 2H	0.730	0.21045	0.32365	0.21467	0.31818
	Neural3	NN BReg 8H	0.725	0.21606	0.34440	0.21257	0.32231
	Reg2	Forward Regression	0.723	0.20779	0.36100	0.21345	0.32645
	Reg4	Stepwise Regression	0.723	0.20779	0.36100	0.21345	0.32645
	Neural5	NN BReg 4H	0.721	0.18548	0.29046	0.21756	0.33884
	Neural6	NN BReg 3H	0.718	0.21273	0.33610	0.21842	0.32231
	Reg	Full Regression	0.718	0.18394	0.28216	0.22075	0.34298
	Reg3	Backward Regression	0.714	0.20527	0.26971	0.21545	0.35950
	Neural2	NN BReg 7H	0.714	0.21548	0.34440	0.21545	0.33058
	Neural8	NN BReg 5H	0.706	0.20569	0.33610	0.22001	0.34711
	Tree	Maximal Tree	0.622	0.19988	0.31120	0.24760	0.36777
	Tree3	ASE Tree	0.620	0.22823	0.35270	0.23659	0.38017
	Tree2	Misclassification Tree	0.617	0.20784	0.31120	0.24774	0.36777

WITH RECODE

Fit Statistics
Model Selection based on Valid: Roc Index (_VAUR_)

Selected Model	Model Node	Model Description	Valid: Roc Index	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error	Valid: Misclassification Rate
Y	Neural9	NN Transform	0.757	0.21005	0.28216	0.20965	0.30165
	Neural5	NN Cap&Floor	0.742	0.18892	0.28631	0.20790	0.32231
	Neural4	NN BReg 8H	0.742	0.17414	0.26971	0.20928	0.32645
	Neural2	NN BReg 6H	0.734	0.19351	0.26141	0.21526	0.33471
	Neural8	NN BReg 3H	0.726	0.20761	0.32780	0.21283	0.33471
	Neural3	NN BReg 7H	0.726	0.20519	0.31120	0.21858	0.31405
	Neural7	NN BReg 4H	0.721	0.18530	0.25311	0.21252	0.34711
	Neural10	NN BReg 5H	0.714	0.23061	0.39834	0.21818	0.33884
	Reg2	Forward Regression	0.711	0.22689	0.33610	0.21803	0.33884
	Reg4	Stepwise Regression	0.711	0.22689	0.33610	0.21803	0.33884
	Reg3	Backward Regression	0.708	0.19714	0.31120	0.21960	0.36777
	Neural6	NN Recode	0.707	0.20025	0.29876	0.22595	0.30992
	Reg	Full Regression	0.701	0.19683	0.29461	0.22262	0.35537
	Neural	NN BReg 2H	0.696	0.20736	0.29046	0.22321	0.36364
	Tree	Maximal Tree	0.622	0.19988	0.31120	0.24760	0.36777
	Tree3	ASE Tree	0.620	0.22823	0.35270	0.23659	0.38017
	Tree2	Misclassification Tree	0.617	0.20784	0.31120	0.24774	0.36777

Based on the results, the best Neural Network in terms of ROC is Neural Network Transform (with recode) with 0.757. The model with the lowest ASE value is Neural Network Cap&Floor (with recode) with 0.207902**.** While neural network models are a natural extension of a regression model, it faces interpretability challenges. Hence, no further interpretation was done for this model.

Model Assessment

Predictive models are compared based on two performance metrics: Receiver Operator Characteristic (ROC) index and Average Squared Error (ASE). As mentioned in the data exploration, this study performed modeling on two sets of datasets: one without recoded categorical variables, and the other was based on recoded variables The table below shows the result for two models: Without Recode and With Recode.

Without Recode

Model Description ▲	Selection Criterion: Valid: Roc Index	Valid: Average Squared Error
ASE Tree	0.62	0.236592.
Backward Regression	0.714	0.215446.
Forward Regression	0.723	0.21345.
Full Regression	0.718	0.220747.
Maximal Tree	0.622	0.247598.
Misclassification Tree	0.617	0.247744.
NN BReg 2H	0.73	0.21467.
NN BReg 3H	0.718	0.218423.
NN BReg 4H	0.721	0.217563.
NN BReg 5H	0.706	0.220008.
NN BReg 6H	0.734	0.211065.
NN BReg 7H	0.714	0.215454.
NN BReg 8H	0.725	0.212575.
NN Cap&Floor	0.742	0.207902.
NN Transform	0.757	0.209651.
Stepwise Regression	0.723	0.21345.

With Recode

Model Description ▲	Selection Criterion: Valid: Roc Index	Valid: Average Squared Error
ASE Tree	0.62	0.236592.
Backward Regressi...	0.708	0.219602.
Forward Regression	0.711	0.218034.
Full Regression	0.701	0.222621.
Maximal Tree	0.622	0.247598.
Misclassification Tree	0.617	0.247744.
NN BReg 2H	0.696	0.223213.
NN BReg 3H	0.726	0.212828.
NN BReg 4H	0.721	0.212515.
NN BReg 5H	0.714	0.218177.
NN BReg 6H	0.734	0.215259.
NN BReg 7H	0.726	0.218577.
NN BReg 8H	0.742	0.209284.
NN Cap&Floor	0.742	0.207902
NN Recode	0.707	0.22595.
NN Transform	0.757	0.209651.
Stepwise Regression	0.711	0.218034.

ROC Index		**Average Squared Error**	
Model Description	***Valid: Roc Index***	***Model Description***	***Valid: ASE***
NN Transform With Recode	0.757	NN Cap&Floor With Recode	0.207902
NN Transform Without Recode	0.757	NN Cap&Floor Without Recode	0.207902
NN BReg 5H With Recode	0.746	NN Transform With Recode	0.209651
NN Cap&Floor With Recode	0.742	NN Transform Without Recode	0.209651
NN Cap&Floor Without Recode	0.742	NN BReg 6H Without Recode	0.211065
NN BReg 6H With Recode	0.735	NN BReg 8H Without Recode	0.212575
NN BReg 6H Without Recode	0.734	NN BReg 8H With Recode	0.213375
NN BReg 2H Without Recode	0.73	Forward Regression With Recode	0.21345
NN BReg 4H Without Recode	0.726	Stepwise Regression Without Recode	0.21345
NN BReg 3H With Recode	0.725	NN BReg 3H With Recode	0.213488
NN BReg 8H Without Recode	0.725	NN Recode	0.213488
NN Recode	0.725	NN BReg 2H Without Recode	0.21467
Forward Regression With Recode	0.723	NN BReg 4H Without Recode	0.215346

Stepwise Regression Without Recode	0.723	NN BReg 6H With Recode	0.215371
NN BReg 8H With Recode	0.72	Backward Regression With Recode	0.215446
Full Regression Without Recode	0.718	NN BReg 7H Without Recode	0.215454
NN BReg 3H Without Recode	0.718	NN BReg 2H With Recode	0.216345
Backward Regression Without Recode	0.714	NN BReg 5H With Recode	0.21818
NN BReg 7H Without Recode	0.714	NN BReg 3H Without Recode	0.218423
NN BReg 2H With Recode	0.712	NN BReg 5H Without Recode	0.22001
NN BReg 7H With Recode	0.71	Full Regression Without Recode	0.220747
Backward Regression With Recode	0.707	Backward Regression Without Recode	0.221015
Full Regression With Recode	0.707	Full Regression With Recode	0.221015
NN BReg 5H Without Recode	0.706	NN BReg 7H With Recode	0.22167
Forward Regression Without Recode	0.69	Forward Regression without Recode	0.226776
Stepwise Regression With Recode	0.69	Stepwise Regression With Recode	0.226776
NN BReg 4H With Recode	0.627	ASE Tree	0.236592
Maximal Tree	0.622	NN BReg 4H With Recode	0.244441
ASE Tree	0.62	Maximal Tree	0.247598
Misclassification Tree	0.617	Misclassification Tree	0.247744

The best model based on ROC is NN Transform with Recode which yielded a ROC index of 0.757. Meanwhile, the **best model based on the ASE is NN Cap & Floor with Recode** with the lowest value of 0.20792. That said, the study shows the overall best model is Neural Network.

Best Model / Model Recommendation

After comparing the two models, without recode and with recode, the study found out that both had the same ROC and ASE values. Given this, the study chose the simplest model with fewer dummy variables – with recode. The figure below summarizes the results of the model.

Prediction Type	Decision	Ranking	Estimates
Model Description ▲	Valid: Misclassification Rate	Selection Criterion: Valid: Roc Index	Valid: Average Squared Error
ASE Tree	0.380165	0.62	0.236592
Backward Regression	0.359504	0.714	0.215446
Forward Regression	0.326446	0.723	0.21345
Full Regression	0.342975	0.718	0.220747
Maximal Tree	0.367769	0.622	0.247598
Misclassification Tree	0.367769	0.617	0.247744
NN BReg 2H	0.318182	0.73	0.21467
NN BReg 3H	0.322314	0.718	0.218423
NN BReg 4H	0.338843	0.721	0.217563
NN BReg 5H	0.347107	0.706	0.220008
NN BReg 6H	0.309917	0.734	0.211065
NN BReg 7H	0.330579	0.714	0.215454
NN BReg 8H	0.322314	0.725	0.212575
NN Cap&Floor	0.322314	0.742	0.207902
NN Transform	0.301653	0.757	0.209651
Stepwise Regression	0.326446	0.723	0.21345

Based on the assessment measure per prediction type, the following models were identified as best:

- **Decision**- Neural Network Transform for having the lowest Misclassification rate. It is the most accurate in matching decision with outcome.
- **Ranking**- Neural Network Transform for having the highest ROC index. It is the most accurate in ordering primary and secondary outcomes.
- **Estimates** - Neural Network Cap & Floor for having the lowest ASE. It has the lowest variance between the target and estimate.

Due to interpretability issues of Neural Networks, the study checked the next best non-Neural Network model in terms of having the lowest ASE. It turned out that the Forward Regression had the lowest ASE. Based on the odds-ratio estimates, we conclude that the main factors affecting employee attrition at IBM are demographic and environment characteristics*:*

- **Single** - Single employees are more likely to resign by **80%** and **55%** vs. Divorced and Married employees, respectively.
- **Tenure** - The longer the years of employees' tenure multiplied to a factor of 2.74, the lesser likelihood of resignation by **40%**
- **Education Field -** Employees with a Technical Degree are more likely to resign by **78%** than other degrees. However, they are **4 times** more likely to stay than those with Human Resource degree.
- **Job Satisfaction -** With respect to satisfaction surveys, low degree of satisfaction in terms of job satisfaction increases the probability of attrition. Those who gave a rating of 1 are **three times** more likely to resign.

Conclusion

To answer the study's research question on the factors affecting employee attrition at IBM, a predictive modeling approach was employed and used three predictive models namely: decision tree, regression, and neural networks. Two sets of modeling were done: first was based non-recoded dataset (using all categorical variables), and second was based on recoded variables wherein some inputs levels for *Education, Education Field,* and *Marital Status* were combined to reduce the dimensionality of the dataset. Results of the modeling show that the best model based on highest ROC and lowest ASE were **Neural Network Transform** and **Neural Network Cap & Floor** respectively, both based on recoded variables.

Since the prediction results of neural network models are innately challenging to interpret, and for the purpose of interpreting the results in relation to the factors affecting employee attrition, the study focused on the next best non-NN which is the Forward Regression. The study chose the Forward Regression without recode since it had the lowest ASE and highest ROC among regression models. Based on the results, we conclude that the main factors affecting employee attrition at IBM are demographics and working environment conditions. The profile with the highest likelihood of attrition are single employees in the human resource field, with short tenures and low job satisfaction level.

Recommendations

This study gave us significant insights into why employees at IBM leave their jobs. Organizations like IBM may use these data to implement targeted retention strategies, enhance employee satisfaction, and lower attrition. Considering this, the following actions are recommended:

- **Elevate Professional Development Career Programs**
  Since most of those who resign are single employees and have a background in human resources, it could be worthwhile for IBM to invest in professional development programs that will enhance the skills of their employees.

- **Strengthen the long-term loyalty of employees**
  Recognize employees who have been with the company for a long time. It can be in the form of service awards and incentives. In this way, employees will feel that their efforts are being valued; subsequently, it will strengthen their commitment to the company.

- **Enhance Employee Satisfaction**
  It is also equally important for the IBM management to ensure that the company has conducive working conditions, effective job structure, and promotes a culture that values work life balance.

References

Conchada, M., Doña, M. and Francisco, K. (2023). Analysis on factors affecting IBM employee attrition. Analytic Lifecycle Management Final Paper.

Kaggle. (2023). IBM Attrition Dataset. https://www.kaggle.com/datasets/yasserh/ibm-attrition-dataset/code

Parent, D. (2023). SAS Regression Lecture Notes.