Survival Analysis of an Employee and Churn Prediction

Ved Deshpande

Nilkamal School of Mathematics, Applied Statistics and Analytics

Contents

Abstract

Employee attrition is a key issue for firms globally, affecting productivity, morale, and profitability. Understanding the elements that contribute to attrition and forecasting future departure rates are critical for successful labor management. This study examines the strategies used in HR attrition analysis, with an emphasis on both parametric and non-parametric approaches to survival analysis. Parametric methods entail fitting specific distributions to employee tenure data, but non-parametric techniques such as Kaplan-Meier estimation and the Cox regression model provide information about survival probability without assuming a specific distribution. In addition, classification methods such as Logistic Regression is used in churn prediction to identify important drivers of attrition. Organizations may use these analytical tools to design focused retention strategies, reduce attrition rates, and foster a stable and engaged workforce that will lead to long-term success.

1 Introduction

Employee attrition and survival analysis are critical in today's corporate environment, where talent acquisition and retention are major drivers of organisational success. With the job market becoming more competitive and employees having more options than ever, organizations must understand why employees leave and how long they stay. Employee attrition analysis assists firms in identifying trends, patterns, and underlying issues that drive turnover, allowing them to build focused retention strategies to keep top personnel. Employee survival analysis. Employee survival analysis, on the other hand, allows businesses to forecast future attrition rates, identify at-risk individuals, and put proactive measures in place to keep staff longer. In today's dynamic market, where workforce stability and engagement are critical, firms must use attrition and survival analysis to retain a talented and motivated workforce, drive organizational performance, and stay ahead of the competition.

2 Problem Statement

Understanding staff attrition and survival trends is critical in today's business environment, when talent acquisition and retention are significant drivers of organizational success. With the job market becoming more competitive and employees having more options than ever before, companies must examine why employees leave and how long they stay. This data assists companies in developing focused retention strategies and forecasting future attrition rates, allowing them to maintain a talented and motivated staff while driving organizational performance.

3 Methodology

Employee survival analysis uses both parametric and nonparametric methodologies to simulate the survival distribution based on available data. Parametric approaches estimate survival probability by fitting certain distributions to the data, such as the exponential or Weibull distribution. Non-parametric approaches, such as the Kaplan-Meier estimator and the Cox regression model, are utilized when the underlying distribution is unknown or complex. These methods provide useful information about employee tenure and the possibility of attrition over time.

Classification algorithms are critical in churn prediction since they help discover elements that contribute to attrition. Logistic Regression and Linear Discriminant Analysis are prominent methods for categorizing employees as churn or non-churn based on a variety of predictor variables. By examining indicators such as job satisfaction, tenure, performance evaluations, and demographic characteristics, these algorithms can accurately anticipate employee attrition and identify the most relevant reasons causing churn inside the firm. This allows HR managers to proactively address retention issues and apply tailored initiatives to reduce turnover rates.

4 Data

Employee turnover prediction was based on data from a github project. It consisted of 31 variables and 1470 instances. It includes four time random variables: years at the company, years in the current role, years with the current manager, and years since last promotion. Total working years is not considered in our analysis because it was not very useful to analyze that variable. The data description is given as follow,

Table 1: Variable Table

Name	Class	Values
Age	integer	Num: 18 to 60
Attrition	character	
BusinessTravel	character	
DailyRate	integer	Num: 102 to 1499
Department	character	
DistanceFromHome	integer	Num: 1 to 29
Education	integer	Num: 1 to 5
EducationField	character	
EnvironmentSatisfaction	integer	Num: 1 to 4
Gender	character	
HourlyRate	integer	Num: 30 to 100
JobInvolvement	integer	Num: 1 to 4
JobLevel	integer	Num: 1 to 5
JobRole	character	
JobSatisfaction	integer	Num: 1 to 4
MaritalStatus	character	
MonthlyIncome	integer	Num: 1009 to 19999
MonthlyRate	integer	Num: 2094 to 26999
NumCompaniesWorked	integer	Num: 0 to 9
OverTime	character	
PercentSalaryHike	integer	Num: 11 to 25
PerformanceRating	integer	Num: 3 to 4
RelationshipSatisfaction	integer	Num: 1 to 4
StockOptionLevel	integer	Num: 0 to 3
TotalWorkingYears	integer	Num: 0 to 40
TrainingTimesLastYear	integer	Num: 0 to 6
WorkLifeBalance	integer	Num: 1 to 4
YearsAtCompany	integer	Num: 0 to 40
YearsInCurrentRole	integer	Num: 0 to 18
YearsSinceLastPromotion	integer	Num: 0 to 15
YearsWithCurrManager	integer	Num: 0 to 17

5 Survival Analysis

5.1 Parametric

The parametric approach to survival analysis for employee churn data entails modeling the distribution of survival times using specified probability distributions. This method implies that survival times have a known mathematical form, such as the exponential, Weibull, or log-normal distribution.

To use the parametric approach, the data is first reviewed to determine whether it is suitable for fitting a specific distribution. The chosen distribution's parameters that best describe employee survival times are then estimated using statistical approaches such as maximum likelihood estimation (MLE) or Bayesian inference.

The parametric technique has various advantages, including simplicity, efficiency in parameter estimation, and the capacity to extrapolate survival probability beyond the observed data. However, its success is dependent on the assumption that the chosen distribution adequately depicts the underlying survival process.

For the given data the histogram of the time random variables were as follow,

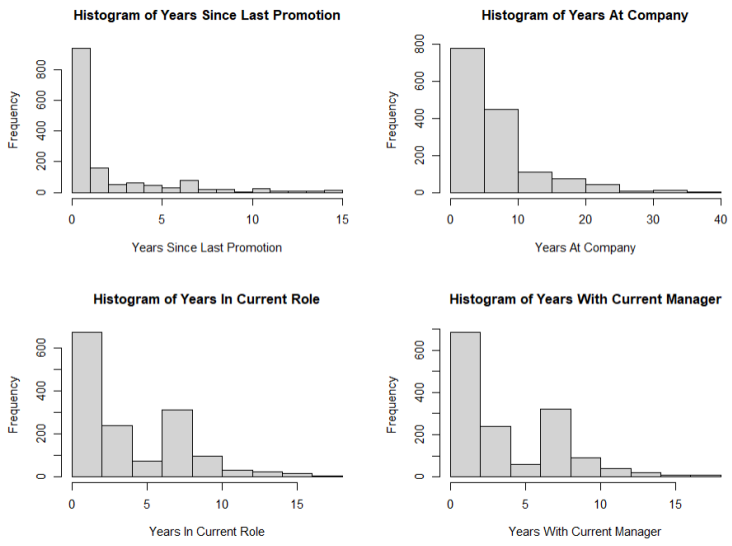

Figure 1: Histogram Plot of Time Random Variables

and the goodness of statistic for the parametric distributions were given as follow,

```
$YearsAtCompany
Goodness-of-fit statistics
                                  1-mle-exp 2-mge-weibull 3-mge-gamma 4-mge-lnorm
Kolmogorov-Smirnov statistic 0.1155826     0.06927675  0.07293563    0.1084143
```

```
Cramer-von Mises statistic    3.8058290     1.08515295   1.12589904    2.0812070
Anderson-Darling statistic          Inf            Inf          Inf          Inf

Goodness-of-fit criteria
                                1-mle-exp 2-mge-weibull 3-mge-gamma 4-mge-lnorm
Akaike's Information Criterion  8666.402            Inf          Inf          Inf
Bayesian Information Criterion  8671.695            Inf          Inf          Inf

$YearsInCurrentRole
Goodness-of-fit statistics
                                1-mle-exp 2-mge-weibull 3-mge-gamma 4-mge-lnorm
Kolmogorov-Smirnov statistic 0.1720438     0.1659864    0.1659864    0.1659864
Cramer-von Mises statistic   7.2218441     5.8627602    5.8699426    6.3506527
Anderson-Darling statistic         Inf           Inf          Inf          Inf

Goodness-of-fit criteria
                                1-mle-exp 2-mge-weibull 3-mge-gamma 4-mge-lnorm
Akaike's Information Criterion  7181.554            Inf          Inf          Inf
Bayesian Information Criterion  7186.847            Inf          Inf          Inf

$YearsSinceLastPromotion
Goodness-of-fit statistics
                                1-mle-exp 2-mge-weibull 3-mge-gamma 4-mge-lnorm
Kolmogorov-Smirnov statistic  0.3952381     0.3952381    0.3952381    0.3952381
Cramer-von Mises statistic   42.3552126    32.3240679   32.3781698   32.3246200
Anderson-Darling statistic          Inf           Inf          Inf          Inf

Goodness-of-fit criteria
                                1-mle-exp 2-mge-weibull 3-mge-gamma 4-mge-lnorm
Akaike's Information Criterion  5243.655           -Inf         -Inf          Inf
Bayesian Information Criterion  5248.948           -Inf         -Inf          Inf

$YearsWithCurrManager
Goodness-of-fit statistics
                                1-mle-exp 2-mge-weibull 3-mge-gamma 4-mge-lnorm
Kolmogorov-Smirnov statistic 0.1789116     0.1789116    0.1789116    0.1789116
Cramer-von Mises statistic   7.3056351     6.1417360    6.1415592    6.7221748
Anderson-Darling statistic         Inf           Inf          Inf          Inf

Goodness-of-fit criteria
                                1-mle-exp 2-mge-weibull 3-mge-gamma 4-mge-lnorm
Akaike's Information Criterion  7106.840            Inf         -Inf          Inf
Bayesian Information Criterion  7112.133            Inf         -Inf          Inf
```

The goodness-of-fit results may indicate that the exponential distribution is the best fit for our data, based on statistical tests and criteria. However, upon visual analysis of the histogram, we see anomalies indicating that the exponential distribution may not adequately represent the underlying survival process of our employee churn dataset. These contradictions emphasize the parametric distribution approach's limitations and the possibility of mistakes when representing complicated real-world processes.

To overcome this , we use non-parametric methods like the Kaplan-Meier estimator. Non-parametric techniques, unlike parametric methods, make few assumptions about the distribution of survival times, allowing for a more flexible and data-driven approach to predicting survival probability. The Kaplan-Meier estimator helps us to better capture the underlying survival patterns and account for the intricacies inherent in employee churn data by taking observed event times into account directly. Thus, using non-parametric techniques improves the reliability and robustness of our survival analysis, resulting in more accurate insights and informed decisions about staff attrition management.

5.2 Non Parametric

Non-parametric survival analysis is a statistical method that examines time-to-event data without making assumptions about the probability distribution. In contrast to parametric procedures, which require providing a functional form for the survival distribution, non-parametric methods estimate survival probabilities directly from observed data.

The Kaplan-Meier estimator is a popular nonparametric tool in survival analysis. The Kaplan-Meier estimator computes the probability of survival at various time points using the observed survival times of people in the dataset. It considers censored data in which the event of interest (e.g., staff attrition) did not occur by the conclusion of the study period. Non-parametric survival analysis approaches are useful because of their flexibility and robustness in examining data with complex survival patterns or when the underlying distribution is unknown or difficult to characterize. The KM estimator formula is given as,

$$\hat{S} = \prod_{t_i \leq t} (1 - \frac{d_i}{n_i})$$

$\hat{S}$ is the estimated survival probability at time t, t_i is the observed time event, d_i is the number of event(attrition) at time t_i, n_i is the number of individual at risk just before time t_i.

Censoring: Censoring in survival analysis refers to missing or reduced observations in time-to-event data. It occurs when some individuals do not experience the event of interest (such as death, failure, or attrition) by the conclusion of the study period or are lost to follow-up. Censoring is represented by adding the observed time till censorship in the dataset but not observing the event. Handling censoring is critical in survival analysis since it influences the estimation of survival probabilities and can bias results if not properly accounted for. Censoring is accurately accounted for and analyzed using a variety of statistical methods, including Kaplan-Meier estimation and Cox proportional hazards models.

To make our study easier and less complicated, we will simply look at the Years in Current firm for the survival analysis approach, and the identical study can be done to the other time random variables to acquire greater insights.

For the given time random variable years at company the survival probabilities and hazard function were obtained as follow

Time	Survival Probabilities	Cumulative Hazard
0	0.9891156	0.0109440
1	0.9481915	0.0531988
2	0.9277922	0.0749475
3	0.9113420	0.0928371
4	0.8940265	0.1120199
5	0.8729314	0.1358983
6	0.8616110	0.1489514
7	0.8462749	0.1669110
8	0.8318498	0.1841034
9	0.8169953	0.2021219
10	0.7768152	0.2525528
11	0.7704996	0.2607161
12	0.7704996	0.2607161
13	0.7627946	0.2707664
14	0.7541265	0.2821951
15	0.7493536	0.2885444
16	0.7439235	0.2958171
17	0.7380193	0.3037853
18	0.7317114	0.3123690
19	0.7246758	0.3220310
20	0.7168835	0.3328419
21	0.7060217	0.3481093
22	0.6924443	0.3675274
23	0.6737296	0.3949264
24	0.6544802	0.4239139
25	0.6544802	0.4239139
26	0.6544802	0.4239139
27	0.6544802	0.4239139
29	0.6544802	0.4239139
30	0.6544802	0.4239139
31	0.6135752	0.4884525
32	0.5663771	0.5684952
33	0.5097394	0.6738557
34	0.5097394	0.6738557
36	0.5097394	0.6738557
37	0.5097394	0.6738557

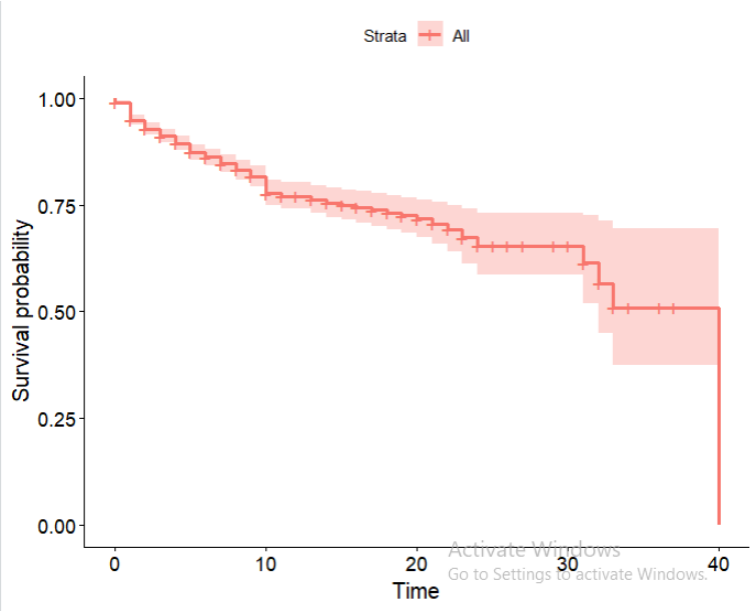

Figure 2: Survival Plot

The cross on the graphs represent the censored data, the above graphs gives us the survival probabilities of an employee with respect to time.

The Kaplan-Meier (KM) survival analysis graph swiftly declines to zero since each event detected at a certain time reduces the number of individuals at risk at succeeding time points. As the study advances, the number of people at risk lowers as events unfold over time. As a result, the likelihood of surviving after each time point is more influenced by the occurrence of events, resulting in a rapid fall in the projected survival probability. This quick reduction represents the cumulative effect of occurrences on survival probability, demonstrating how the likelihood of encountering the event increases over time as more people are affected. Thus, the KM estimator gives a dynamic representation of survival probability over time, reflecting the shifting risk landscape as events unfold.

5.3 Cox Regression

Having ascertained the likelihood of an employee's survival, we can now examine the variables that impact that probability. Here, Cox Regression or the Cox-proportional hazard model is used to determine the factors affecting the attrition.

```
Call:
  n= 1470, number of events= 237

                                      coef exp(coef) se(coef)      z Pr(>|z|)
Age                                -0.06218   0.93971  0.01120 -5.551 2.84e-08 ***
Education                          -0.06299   0.93896  0.06502 -0.969 0.332664
EnvironmentSatisfaction            -0.20724   0.81283  0.05915 -3.504 0.000459 ***
Gender                              0.35622   1.42792  0.13834  2.575 0.010026 *
JobInvolvement                     -0.37561   0.68687  0.08631 -4.352 1.35e-05 ***
JobLevel                           -0.58227   0.55863  0.20728 -2.809 0.004969 **
JobSatisfaction                    -0.23106   0.79369  0.05839 -3.957 7.59e-05 ***
OverTime                            1.20824   3.34759  0.13427  8.999  < 2e-16 ***
PercentSalaryHike                   0.01722   1.01737  0.02921  0.590 0.555420
PerformanceRating                  -0.24941   0.77926  0.30110 -0.828 0.407489
RelationshipSatisfaction           -0.13664   0.87228  0.06144 -2.224 0.026146 *
StockOptionLevel                   -0.42217   0.65562  0.09158 -4.610 4.03e-06 ***
TrainingTimesLastYear              -0.13100   0.87722  0.05915 -2.215 0.026778 *
WorkLifeBalance                    -0.18700   0.82944  0.09038 -2.069 0.038548 *
BusinessTravelTravel_Frequently     1.15125   3.16213  0.31919  3.607 0.000310 ***
BusinessTravelTravel_Rarely         0.65469   1.92455  0.30321  2.159 0.030835 *
DailyRate                          -0.07465   0.92807  0.06726 -1.110 0.267036
HourlyRate                         -0.05037   0.95087  0.06978 -0.722 0.470341
DistanceFromHome                    0.26478   1.30315  0.06400  4.137 3.51e-05 ***
MonthlyIncome                      -0.35942   0.69808  0.24950 -1.441 0.149714
NumCompaniesWorked                  0.49196   1.63552  0.06455  7.622 2.51e-14 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1  1

                                   exp(coef) exp(-coef) lower .95 upper .95
Age                                   0.9397     1.0642     0.9193     0.9606
Education                             0.9390     1.0650     0.8266     1.0666
EnvironmentSatisfaction               0.8128     1.2303     0.7239     0.9127
Gender                                1.4279     0.7003     1.0888     1.8727
JobInvolvement                        0.6869     1.4559     0.5800     0.8135
JobLevel                              0.5586     1.7901     0.3721     0.8386
JobSatisfaction                       0.7937     1.2599     0.7079     0.8899
OverTime                              3.3476     0.2987     2.5730     4.3553
PercentSalaryHike                     1.0174     0.9829     0.9608     1.0773
PerformanceRating                     0.7793     1.2833     0.4319     1.4060
RelationshipSatisfaction              0.8723     1.1464     0.7733     0.9839
StockOptionLevel                      0.6556     1.5253     0.5479     0.7845
TrainingTimesLastYear                 0.8772     1.1400     0.7812     0.9850
WorkLifeBalance                       0.8294     1.2056     0.6948     0.9902
BusinessTravelTravel_Frequently       3.1621     0.3162     1.6916     5.9111
BusinessTravelTravel_Rarely           1.9246     0.5196     1.0623     3.4868
DailyRate                             0.9281     1.0775     0.8134     1.0588
HourlyRate                            0.9509     1.0517     0.8293     1.0902
DistanceFromHome                      1.3031     0.7674     1.1495     1.4773
MonthlyIncome                         0.6981     1.4325     0.4281     1.1384
NumCompaniesWorked                    1.6355     0.6114     1.4412     1.8561

Concordance= 0.863  (se = 0.012 )
Likelihood ratio test= 430.2  on 21 df,   p=<2e-16
Wald test            = 371.4  on 21 df,   p=<2e-16
Score (logrank) test = 408.8  on 21 df,   p=<2e-16
```

The significant factors are given by * with respect to given level of significance. Significant codes are: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 "

The concordance is 0.863 which tells us that applying cox regression is very useful to analyse the factors affecting the survival of an employee in a company. The concordance gives us the goodness of fit for our cox regression model.

5.3.1 Stepwise Regression

Since we now have many factors in our regression model we will use stepwise regression approach to find out the best fit model.

One statistical technique for choosing the most pertinent independent variables to include in a regression model is stepwise regression. It entails a methodical procedure for including or eliminating variables in accordance with their statistical importance. Usually, the process starts with a first model that includes every possible predictor. Iterative stages are then used to find the variables that are most important in explaining the variation in the dependent variable.

```
The main model
Start:  AIC=2760.02
Surv_obj ~ Age + Education + EnvironmentSatisfaction + Gender +
    JobInvolvement + JobLevel + JobSatisfaction + OverTime +
    PercentSalaryHike + PerformanceRating + RelationshipSatisfaction +
    StockOptionLevel + TrainingTimesLastYear + WorkLifeBalance +
    BusinessTravelTravel_Frequently + BusinessTravelTravel_Rarely +
    DailyRate + HourlyRate + DistanceFromHome + MonthlyIncome +
    NumCompaniesWorked

The reduced model
Step:  AIC=2753.2
Surv_obj ~ Age + EnvironmentSatisfaction + Gender + JobInvolvement +
    JobLevel + JobSatisfaction + OverTime + RelationshipSatisfaction +
    StockOptionLevel + TrainingTimesLastYear + WorkLifeBalance +
    BusinessTravelTravel_Frequently + BusinessTravelTravel_Rarely +
    DistanceFromHome + MonthlyIncome + NumCompaniesWorked
```

As we can see the reduced model has lower AIC value than our main model, we can say that the stepwise regression yields the best fit model considering the factors which play a vital role in the hazard rate of an employee. The summary statistics of the reduced model is given as,

```
Call:
  n= 1470, number of events= 237

                                  coef exp(coef) se(coef)      z Pr(>|z|)
Age                           -0.06554   0.93657  0.01106 -5.928 3.08e-09 ***
EnvironmentSatisfaction       -0.20769   0.81246  0.05866 -3.541 0.000399 ***
Gender                         0.35359   1.42416  0.13823  2.558 0.010526 *
JobInvolvement                -0.36869   0.69164  0.08525 -4.325 1.53e-05 ***
JobLevel                      -0.57122   0.56483  0.20722 -2.757 0.005840 **
JobSatisfaction               -0.22432   0.79906  0.05796 -3.870 0.000109 ***
OverTime                       1.19829   3.31444  0.13401  8.942  < 2e-16 ***
RelationshipSatisfaction      -0.13366   0.87489  0.06147 -2.174 0.029672 *
StockOptionLevel              -0.42535   0.65354  0.09073 -4.688 2.76e-06 ***
TrainingTimesLastYear         -0.12631   0.88134  0.05873 -2.151 0.031498 *
WorkLifeBalance               -0.18689   0.82954  0.08951 -2.088 0.036802 *
BusinessTravelTravel_Frequently  1.13077   3.09804  0.31790  3.557 0.000375 ***
BusinessTravelTravel_Rarely    0.64754   1.91083  0.30249  2.141 0.032297 *
DistanceFromHome               0.26440   1.30265  0.06312  4.189 2.80e-05 ***
MonthlyIncome                 -0.36731   0.69259  0.25024 -1.468 0.142144
NumCompaniesWorked             0.49208   1.63571  0.06422  7.662 1.83e-14 ***
---
Signif. codes:  0     ***     0.001     **     0.01     *     0.05     .     0.1         1

                              exp(coef) exp(-coef) lower .95 upper .95
Age                              0.9366     1.0677     0.9165     0.9571
EnvironmentSatisfaction          0.8125     1.2308     0.7242     0.9114
Gender                           1.4242     0.7022     1.0862     1.8673
JobInvolvement                   0.6916     1.4458     0.5852     0.8174
JobLevel                         0.5648     1.7704     0.3763     0.8478
JobSatisfaction                  0.7991     1.2515     0.7132     0.8952
OverTime                         3.3144     0.3017     2.5488     4.3100
RelationshipSatisfaction         0.8749     1.1430     0.7756     0.9869
StockOptionLevel                 0.6535     1.5301     0.5471     0.7807
TrainingTimesLastYear            0.8813     1.1346     0.7855     0.9889
WorkLifeBalance                  0.8295     1.2055     0.6961     0.9886
```

```
BusinessTravelTravel_Frequently     3.0980      0.3228     1.6615     5.7767
BusinessTravelTravel_Rarely         1.9108      0.5233     1.0562     3.4570
DistanceFromHome                    1.3027      0.7677     1.1511     1.4742
MonthlyIncome                       0.6926      1.4439     0.4241     1.1310
NumCompaniesWorked                  1.6357      0.6114     1.4422     1.8551

Concordance= 0.863  (se = 0.012 )
Likelihood ratio test= 427.1  on 16 df,   p=<2e-16
Wald test            = 368  on 16 df,   p=<2e-16
Score (logrank) test = 404.2  on 16 df,   p=<2e-16
```

The above regression model shows that factors such as overtime, environment satisfaction, job involvement and others significantly affects the hazard rate. The concordance ratio remains the same which tells us that our reduced model is also a good fit for the given data and the stepwise approach did not generate any loss of information in our data. The other temporal random variable might be treated similarly statistically to provide additional insights.

5.4 conclusion

Employee survival probabilities were determined using non-parametric techniques, such as the Kaplan-Meier (KM) estimator, because parametric approaches were unable to adequately represent the distribution of our data. After survival probabilities were established, we used the Cox regression technique to pinpoint the variables affecting the hazard rate. Using stepwise regression, we were able to simplify our first regression model—which included many variables—by keeping only the most significant variables that had an impact on the hazard rate. With the help of this method, we were able to efficiently determine and rank the major factors that influence employee turnover, which improved the usability and interpretation of our regression model.

6 Classification

Sorting data points into predetermined classes or categories according to their attributes is a basic machine learning activity called classification. It is frequently employed for jobs like picture identification, medical diagnosis, and spam detection. A labeled dataset is used to train a model for classification, in which each data point has a class label assigned to it. In order to enable the model to correctly categorize unseen data, the objective is to learn a mapping from the input features to the relevant class labels.

For classification, a variety of techniques are employed, such as neural networks, logistic regression, decision trees, support vector machines (SVM), and k-nearest neighbors (KNN). Every algorithm has advantages and disadvantages, and the selection of an algorithm is influenced by various elements, including the type of data, the size of the dataset, and the required degree of interpretability.

After the model has been trained, a classification model may be used to automatically identify patterns and make decisions by predicting the class labels of fresh data points. Evaluation measures, such recall, accuracy, precision, and F1-score, are used to make sure that classification models perform well and are useful in practical applications.

In this research, we will model the provided attrition data using logistic regression as our classification model.

6.1 Data processing

One hot encoding Categorical variables can be transformed into a numerical representation that is appropriate for machine learning algorithms using a technique called one-hot encoding. With this approach, a binary vector is used to represent each category or level of a categorical variable, with each dimension corresponding to a distinct category. All other dimensions are set to 0, with the exception of the dimension corresponding to the observation's category, which is set to 1. This makes it possible to express categorical variables as sparse binary matrices, which machine learning algorithms can handle with ease. In order to handle categorical data and allow machine learning models to learn from non-numeric variables, one-hot encoding is necessary.

Converting into factors One preprocessing stage in data analysis, especially in statistical modeling and machine learning, is converting variables into factors. Variables known as factors are used to describe categorical data, in which each level or category is assigned a unique label. When variables are transformed into factors, these labels are applied to the distinct values found in the data, making handling and understanding simpler. Factors facilitate the ability of statistical software to identify the categorical nature of the data and carry out the necessary analysis, including the construction of prediction models, hypothesis testing, and contingency tables.

Oversampling In machine learning, oversampling is a technique used to rectify class imbalance, in which one class has substantially less data than the other. When oversampling occurs, the minority class—the one with fewer sam-

ples—is purposefully made larger by creating new samples or by copying samples that already exist, continuing until the distribution of classes is balanced. This enhances the model's capacity to correctly categorize cases from both classes and helps reduce bias towards the majority class. SMOTE technique was used in our project to remove the imbalance in the data.

6.2 Logistic Regression

When attempting to predict the likelihood that an observation will belong to a specific class, one statistical technique called logistic regression is widely used. In contrast to linear regression, which forecasts continuous values, logistic regression models the likelihood that the outcome variable will fall into one of two classes. Because of its ease of use and interpretability, it is extensively employed in a variety of industries, including marketing, finance, and medicine.

The logistic function, sometimes referred to as the sigmoid function, is the foundation of the logistic regression model. It converts the result of a linear feature combination into a probability value between 0 and 1. The logistic function can be found using,

$$P(y = 1|x) = \frac{1}{1 + e^{-z}}$$

where $P(y = 1|x)$ is the probability that the outcome variable y is 1 given the input variables x, z is the linear combination of the input features and their corresponding coefficients(it is the log-odds of the probability of the positive class)

$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$

$x_1, x_2, x_3, ..., x_n$ are the input features.

6.2.1 Assumptions

No Multicollinearity: The independent variables should not be highly correlated with each other. Multicollinearity can lead to unstable estimates of the regression coefficients. This can be checked with the help of the condition number which is given by the eigen values of the $(X'X)^{-1}Xy$ matrix. The formula is,

$$\lambda = \frac{\lambda_{max}}{\lambda_{min}}$$

If the condition number is between 0-100 then no multicollinearity exist, if it is between 100-1000 then maybe some multicollinearity exist and if it is greater than 1000 then severe multicollinearity exist. For our data the condition number came out to be 96.28167 which suggests that there is no multicollinearity between the independent variables. Since the assumptions were getting satisfied, we proceed we model building after splitting the data into train and test with 80-20 ratio.

6.2.2 Model

When we apply the logistic regression model on our data we get the following R-output which incorporates all the significant variables, the model deviance, null deviance, AIC value.

```
Call:
glm(formula = as.factor(Attrition) ~ ., family = binomial, data = Train_scaled)

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                      2.59752    1.82080   1.427 0.153701
Age                             -0.04095    0.01563  -2.620 0.008794 **
Education                        0.02981    0.09784   0.305 0.760638
EnvironmentSatisfaction         -0.49253    0.09386  -5.247 1.54e-07 ***
Gender                           0.28277    0.20364   1.389 0.164952
JobInvolvement                  -0.43514    0.14307  -3.041 0.002355 **
JobLevel                        -0.24404    0.35150  -0.694 0.487515
JobSatisfaction                 -0.39459    0.09044  -4.363 1.28e-05 ***
OverTime                         1.92133    0.21544   8.918  < 2e-16 ***
PercentSalaryHike               -0.04855    0.04381  -1.108 0.267857
PerformanceRating                0.22380    0.44155   0.507 0.612260
RelationshipSatisfaction        -0.22665    0.09252  -2.450 0.014292 *
StockOptionLevel                -0.16851    0.16785  -1.004 0.315414
TotalWorkingYears               -0.03659    0.03413  -1.072 0.283649
TrainingTimesLastYear           -0.19516    0.08299  -2.352 0.018688 *
WorkLifeBalance                 -0.32977    0.13585  -2.427 0.015204 *
YearsAtCompany                   0.09271    0.04321   2.145 0.031933 *
YearsInCurrentRole              -0.14478    0.04962  -2.918 0.003528 **
YearsSinceLastPromotion          0.19460    0.04750   4.097 4.18e-05 ***
YearsWithCurrManager            -0.14521    0.05493  -2.644 0.008205 **
BusinessTravelTravel_Frequently  2.06500    0.49244   4.193 2.75e-05 ***
BusinessTravelTravel_Rarely      1.30334    0.45940   2.837 0.004553 **
MaritalStatusMarried             0.13676    0.29956   0.457 0.648004
MaritalStatusSingle              1.04121    0.38413   2.711 0.006716 **
EducationFieldLife.Sciences     -0.32071    0.82268  -0.390 0.696655
EducationFieldMarketing          0.11769    0.87638   0.134 0.893173
EducationFieldMedical           -0.33454    0.81996  -0.408 0.683277
EducationFieldOther             -0.54811    0.89339  -0.614 0.539534
EducationFieldTechnical.Degree   0.77136    0.83839   0.920 0.357548
JobRoleHuman.Resources           1.51914    0.73970   2.054 0.040002 *
JobRoleLaboratory.Technician     1.57439    0.55000   2.863 0.004203 **
JobRoleManager                  -0.21190    0.92518  -0.229 0.818838
JobRoleManufacturing.Director    0.50997    0.58173   0.877 0.380680
JobRoleResearch.Director        -1.10821    1.07963  -1.026 0.304669
JobRoleResearch.Scientist        0.65171    0.56399   1.156 0.247872
JobRoleSales.Executive           1.10444    0.51399   2.149 0.031655 *
JobRoleSales.Representative      2.19315    0.62153   3.529 0.000418 ***
DailyRate                       -0.14747    0.09970  -1.479 0.139104
HourlyRate                       0.03704    0.10221   0.362 0.717048
DistanceFromHome                 0.42260    0.10024   4.216 2.49e-05 ***
MonthlyIncome                    0.14491    0.43063   0.337 0.736482
NumCompaniesWorked               0.52115    0.10851   4.803 1.56e-06 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1        1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1050.01  on 1175  degrees of freedom
Residual deviance:  689.43  on 1134  degrees of freedom
AIC: 773.43

Number of Fisher Scoring iterations: 7
```

As we can see few variables such as Age, Environment Satisfaction, Overtime, JobSatisfaction and many others are turning out to be significant but there are many factors which are not significant and therefore we apply the stepwise regression to get a reduced model.

6.2.3 Stepwise Logistic Model

```
Call:
glm(formula = as.factor(Attrition) ~ Age + EnvironmentSatisfaction +  JobInvolvement
      + JobSatisfaction + OverTime + RelationshipSatisfaction +
    TotalWorkingYears + TrainingTimesLastYear + WorkLifeBalance +
    YearsAtCompany + YearsInCurrentRole + YearsSinceLastPromotion +
    YearsWithCurrManager + BusinessTravelTravel_Frequently +
    BusinessTravelTravel_Rarely + MaritalStatusSingle + EducationFieldTechnical.
        Degree +
    JobRoleHuman.Resources + JobRoleLaboratory.Technician + JobRoleResearch.Director
        +
    JobRoleResearch.Scientist + JobRoleSales.Executive + JobRoleSales.Representative
        +
    DailyRate + DistanceFromHome + NumCompaniesWorked, family = binomial,
    data = Train_scaled)

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                      1.99266    0.96725   2.060 0.039386 *
Age                             -0.03943    0.01514  -2.604 0.009209 **
EnvironmentSatisfaction         -0.49340    0.09249  -5.334 9.59e-08 ***
JobInvolvement                  -0.42180    0.14087  -2.994 0.002752 **
JobSatisfaction                 -0.39960    0.08902  -4.489 7.16e-06 ***
OverTime                         1.89949    0.21303   8.916  < 2e-16 ***
RelationshipSatisfaction        -0.20248    0.09056  -2.236 0.025356 *
TotalWorkingYears               -0.04410    0.02975  -1.482 0.138226
TrainingTimesLastYear           -0.21006    0.08178  -2.569 0.010207 *
WorkLifeBalance                 -0.33717    0.13450  -2.507 0.012178 *
YearsAtCompany                   0.08385    0.04105   2.042 0.041114 *
YearsInCurrentRole              -0.13026    0.04847  -2.688 0.007197 **
YearsSinceLastPromotion          0.18461    0.04581   4.030 5.58e-05 ***
YearsWithCurrManager            -0.14275    0.05308  -2.689 0.007165 **
BusinessTravelTravel_Frequently  2.17457    0.48253   4.507 6.59e-06 ***
BusinessTravelTravel_Rarely      1.39970    0.45143   3.101 0.001931 **
MaritalStatusSingle              1.11067    0.20562   5.402 6.61e-08 ***
EducationFieldTechnical.Degree   1.05736    0.30255   3.495 0.000474 ***
JobRoleHuman.Resources           1.60053    0.50974   3.140 0.001690 **
JobRoleLaboratory.Technician     1.49398    0.35393   4.221 2.43e-05 ***
JobRoleResearch.Director        -1.25120    0.89622  -1.396 0.162689
JobRoleResearch.Scientist        0.58698    0.36246   1.619 0.105349
JobRoleSales.Executive           1.10100    0.33861   3.252 0.001148 **
JobRoleSales.Representative      2.17385    0.43099   5.044 4.56e-07 ***
DailyRate                       -0.16243    0.09757  -1.665 0.095942 .
DistanceFromHome                 0.41392    0.09830   4.211 2.55e-05 ***
NumCompaniesWorked               0.50985    0.10708   4.761 1.92e-06 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1050.01  on 1175  degrees of freedom
Residual deviance:  698.37  on 1149  degrees of freedom
AIC: 752.37

Number of Fisher Scoring iterations: 6
```

As we can see that the total number of variables are reduced and the AIC value is also less for the reduced model as compared to the whole model, so we can conclude that our reduced model is the best fit model which considers all the factors which play a significant role in attrition of an employee.

6.2.4 Model Accuracy

Confusion Matrix: A table that is frequently used to assess how well a classification model is performing is called a confusion matrix. By showing the counts of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions, it provides an overview of a classification algorithm's performance. The confusion matrix helps us understand the performance of a classification model in terms of accuracy, precision, recall, F1-score, and other metrics. It provides insights into the model's ability to correctly classify instances and identify areas where the model may need improvement.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 241  25
         1   9  19

               Accuracy : 0.8844
                 95% CI : (0.8422, 0.9186)
    No Information Rate : 0.8503
    P-Value [Acc > NIR] : 0.0566

                  Kappa : 0.4656

 Mcnemar's Test P-Value : 0.0101

            Sensitivity : 0.9640
            Specificity : 0.4318
         Pos Pred Value : 0.9060
         Neg Pred Value : 0.6786
             Prevalence : 0.8503
         Detection Rate : 0.8197
   Detection Prevalence : 0.9048
      Balanced Accuracy : 0.6979

       'Positive' Class : 0
```

For the given data, logistic regression shows a accuracy of 88.44% which is very good and we can conclude that the logistic model is a good fit for our data. Low specificity in a classification model can be attributed to several factors imbalanced dataset, class overlapping,feature selection, model hyperparameters, data preprocessing and many more, we can try to increase the specificity by improving the above factors.

ROC and AUC

Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) are important tools for evaluating the goodness of fit of logistic regression models.

ROC curve is a graphical representation of the trade-off between the true positive rate (Sensitivity) and the false positive rate (1 - Specificity) across different threshold values

AUC quantifies the overall performance of the classifier across all possible threshold values. AUC ranges from 0 to 1, where a higher value indicates better discrimination ability. An AUC of 1 represents a perfect classifier, while an AUC of 0.5 suggests a classifier that performs no better than random guessing.

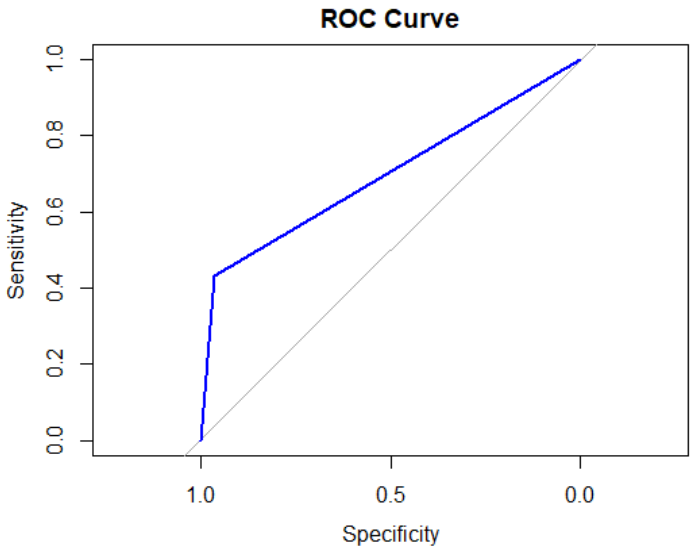

Figure 3: ROC Plot

The above is the ROC plot for the given attrition data and the Area under the curve is 0.6979 which tells us that our model is good fit.

6.2.5 Cross Validation

In machine learning, cross-validation is an essential method for evaluating the effectiveness and capacity for generalization of prediction models. The dataset is divided into several folds, or subsets, and the model is trained on a portion of the data while being validated on the remaining data that has not yet been viewed. Cross-validation helps to evaluate how well a predictive model generalizes to new, unseen data by simulating the model's performance on multiple subsets of the dataset. Cross-validation helps to assess the model's robustness against overfitting by providing a more accurate estimate of its performance on unseen data compared to traditional train-test split methods. After doing the cross validation on the given data the error was 0.1267225. The training Accuracy was 87.74% and the testing accuracy was 85.94 %. Since there is not a huge difference between the training and testing accuracy, we can suggest that there was no overfitting in our model.

7 Future Scope

7.1 Survival Analysis

- We can use methods such as cox regression and risk regression to get the survival predicts for the new hires
- We can further dive into non-parametric approach and use Kernal density or muhaz function to estimate the survival and hazard functions

7.2 Classification

- We saw that the models which we have used needs certain assumptions with respect to data, so we can maybe try some machine learning approach such Random Forest, Decision Tree or SVM to get the attrition.
- We can also find which variable plays the most important role in attrition with the help of variable importance plot.

8 Reference

- Montgomery DC, Peck EA, Vining GG. 2012. Introduction to linear regression analysis.Fifth edition. New York: John Wiley Sons
- Hosmer D.W. and Lemeshow, S. (2000) Applied Logistic Regression. 2nd Edition, Wiley, New York
- Elisa T. Lee and John Wenyu Wang Statistical Methods for Survival Data Analysis