**INDEX**

|  |  |  |
| --- | --- | --- |
| Sr. No. | Topic | Page No. |
| 1 | Introduction | 3 |
| 2 | Motivation | 3 |
| 3 | Objectives | 4 |
| 5 | Data Source | 4 |
| 6 | Data Description | 4 |
| 7 | Data Pre-processing | 7 |
| 8 | Methodology :   * Kaplan-Meier estimates of S(t) * Cox-PH model * Random Survival Forest * Logistic Regression | 12 |
| 9 | Analysis | 16 |
| 10 | Conclusion/Findings | 28 |
| 11 | Recommendations and Future Scope | 28 |
| 12 | References | 29 |

**Introduction**

Employee turnover, or employee turnover rate, is the measurement of the number of employees who leave an organization during a specified time period, typically one year.

While an organization usually measures the total number of employees who leave, turnover can also apply to subcategories within an organization like individual departments or demographic groups.

Voluntary turnover is any instance in which an employee actively chooses to leave.

Involuntary turnover is when an employer chooses to terminate an employee or remove them permanently from the group in question, possibly because of poor performance, toxic behavior, or other reasons.

**Motivation**

In human resource management, employee turnover prediction is very important for company operation since the leave of key employees can bring great loss to companies.

The human resources department spends a lot of money and efforts on dealing with employee turnover, since the leave of excellent employees will cause huge losses to the company. Therefore, it is important to study and predict the turnover behavior of employees.

**Objectives**

* To predict employee turnover
* To determine the variables which affect employee turnover the most

using Cox proportional hazards model, Random Survival Forest and other

machine learning techniques like Logistic regression

**Data Sources**

* For this project, an HR dataset named ‘IBM HR Analytics Employee Attrition & Performance’, has been picked, which is available on IBM website.

**Data Description**

* The data contains records of 1,470 employees and 35 variables.
* It has information about employee’s current employment status, the total number of companies worked for in the past, Total number of years at the current company and the current roles, their education level, distance from home, monthly income, etc.
* Users information mainly consist of four parts, Namely:

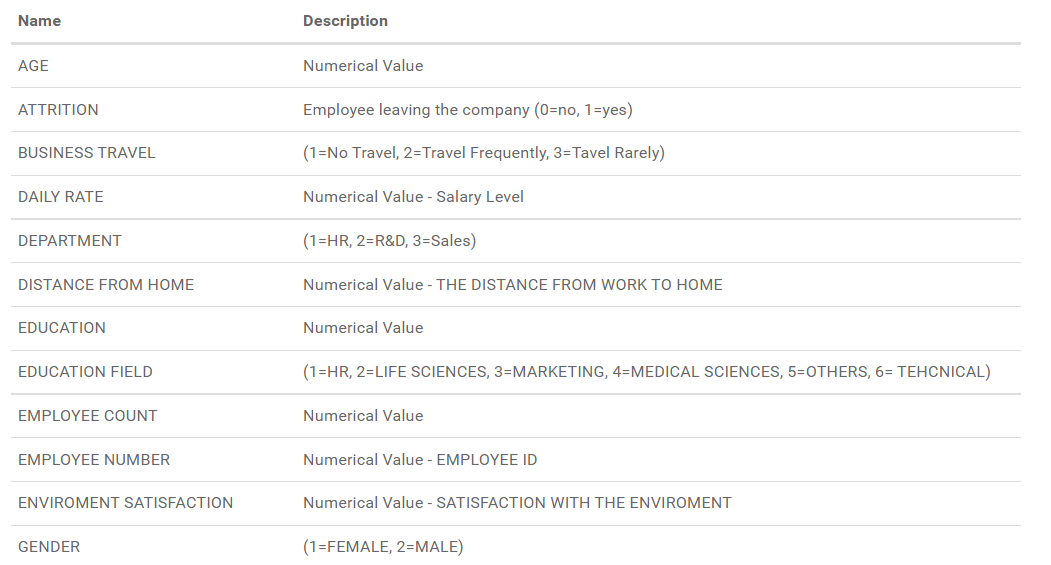
i) **Demographic information**: Age, Gender, Distance from Home, Marital Status, etc;

ii) **Current job information**: Department, Business Travel, Job Role, Monthly Income, etc

iii) **Work experience information**: Years At Company, Years In Current Role, Number of Companies Worked, Total Working Years, etc

iv) **Educational experience information**: Education, Education Field, etc

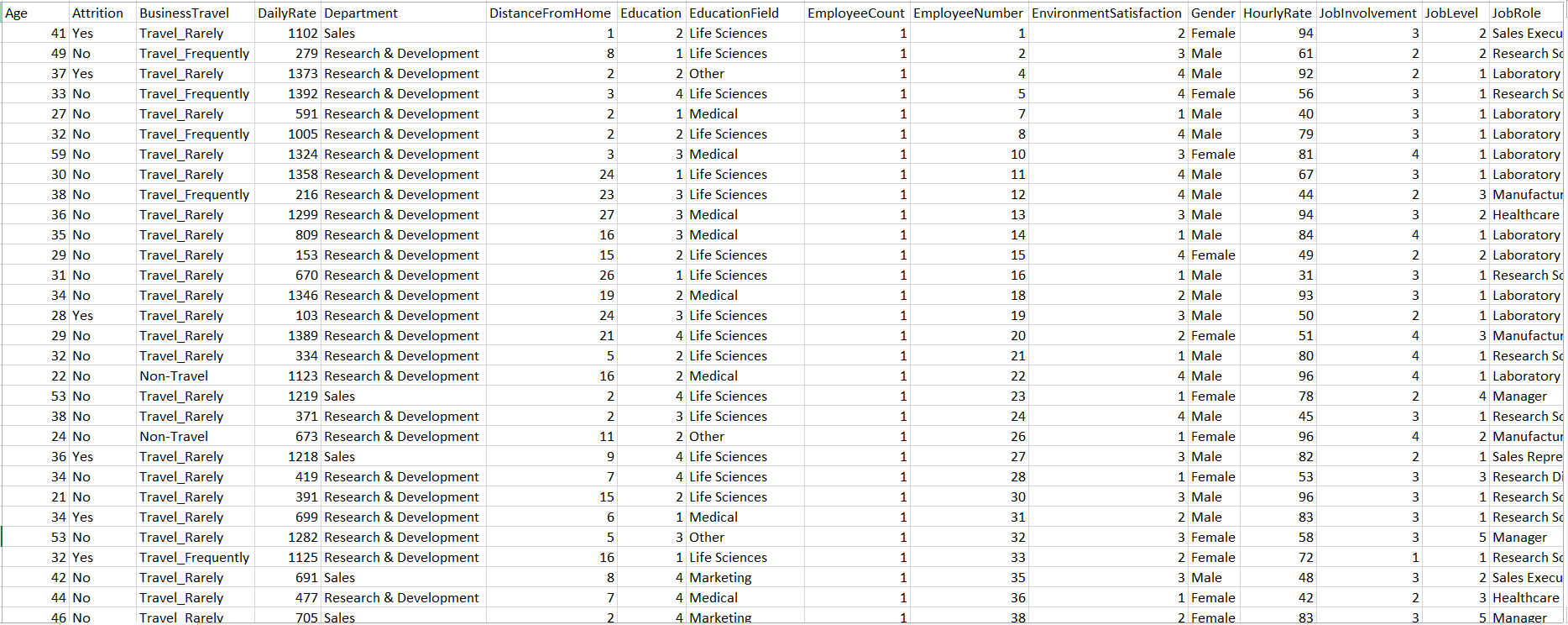
**Briefing of the variables:**

****

****

****

**This is a sample of the dataset:**



**Data Preprocessing**

* **Feature selection:** Process wherein those features are selected, which contribute most to the prediction variable or output. (Only used in some methods)
* **Benefits of feature selection**:
* Improves the performance
* Improves Accuracy
* Provides better understanding of data
* Dropping non-relevant variables
* **Label Encoding**:

Label Encoding refers to converting the categorical variables into numeric form, so as to convert it into the machine-readable form. It is an important pre-processing step for the structured dataset in supervised learning.

* **Scaling the features:**

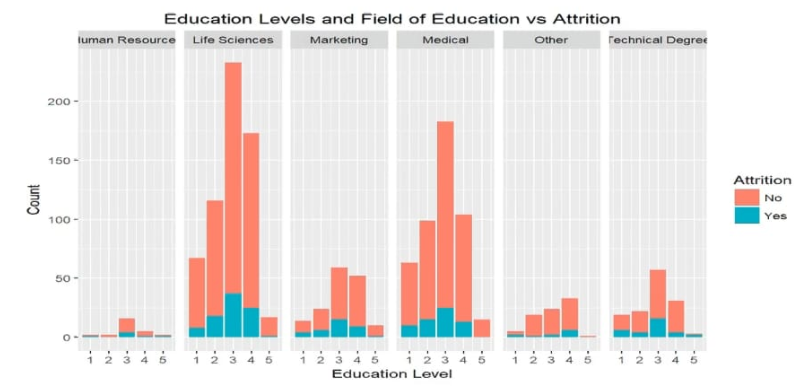
When the numeric data is of different units, scaling the data is important to make the values lie between the 3 sigma Confidence Interval. This makes it easier for the algorithm to compare the values having different units

* **Splitting data into train and test:**

The process of modeling means training a machine learning algorithm to predict the labels from the features, tuning it for the business need, and validating it on holdout data. This is highly useful in testing the accuracy of the built models.

**Exploratory Data Analysis**

**Education Analysis**

****

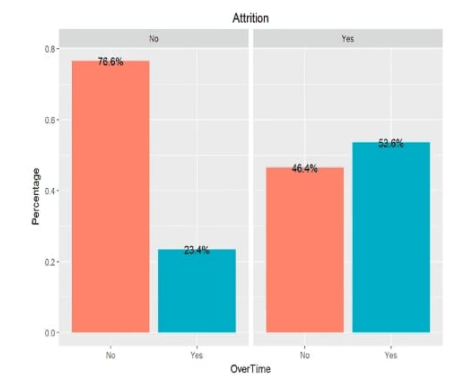
* The highest number of employees, who have attritions, are the ones with bachelor (Education level=3) and master (Education level=4) degrees in the Life Sciences and Medical Fields.

**Performance Analysis**

****

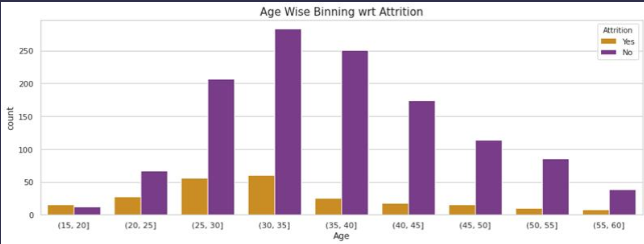
* No difference between the performance rating and attrition levels. Both the groups have similar percentages

**OverTime Analysis**

****

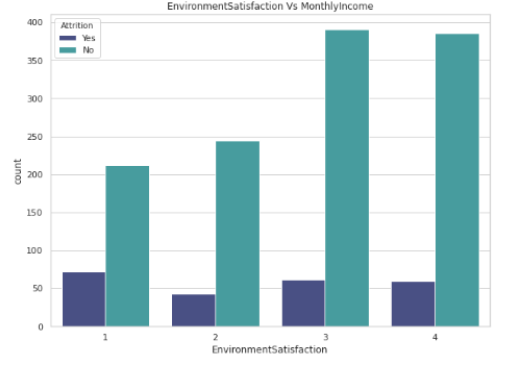
* In the group of employees who don’t have attrition, the employees who work overtime promote faster than the ones who don’t work overtime.

**Age Analysis**

****

* There is maximum attrition in the Age group 30-35, and minimum in the group 55-60 which is logically sound.

**Environment Satisfaction vs Attrition**

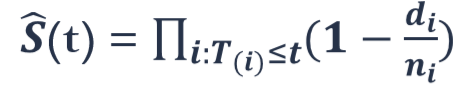


* Lesser Environment Satisfaction leads to higher Attrition

**Methodology**

* **Kaplan Meier Estimator :**

Kaplan-Meier estimate is one of the best options to be used to measure the fraction of subjects living for a certain amount of time after treatment. The Kaplan–Meier estimator, also known as the product limit estimator, is a non-parametric statistic used to estimate the survival function from lifetime data. It is often used to measure the fraction of patients living for a certain amount of time after treatment.



* **Cox Proportional Hazard (CoxPH)**

The purpose of the model is to evaluate simultaneously the effect of several factors on survival. In other words, it allows us to examine how specified factors influence the rate of a particular event happening (e.g., infection, death) at a particular point in time. This rate is commonly referred as the hazard rate. Predictor variables (or factors) are usually termed covariates in the survival-analysis literature.

The Cox model is expressed by the hazard function denoted by h(t). Briefly, the hazard function can be interpreted as the risk of dying at time t.

The model is given as:



It is used to estimate the effect of the Covariates on Time and to predict the Survival probability at a certain time point.

One important Assumption of Cox-PH model is the Assumption of Proportional hazard. This says that the hazard rate is in proportion to the covariates, which means that the influence of covariates on hazards does not change with time.

* **Survival Random Forest:**

Random survival forests is an ensemble tree method for analysis of right-censored survival data.

Recently it has been shown by Breiman (2001) that ensemble learning can be improved further by injecting randomization into the base learning process, an approach called random forests. Random survival forests (RSF) methodology extends Breiman’s random forests (RF) method.

Extending random forests to right-censored survival data is of great value. Survival data are commonly analyzed using methods that rely on restrictive assumptions such as proportional hazards. Further, because these methods are often parametric, nonlinear effects of variables must be modeled by transformations or expanding the design matrix to include specialized basis functions. Often ad hoc approaches, such as stepwise regression, are used to determine if nonlinear effects exist. Identifying interactions, especially those involving multiple variables, is also problematic. This must be done by brute force (examining all two-way and threeway interactions, e.g.), or must rely on subjective knowledge to narrow the search. In contrast, these difficulties are handled automatically using forests.

So basically it combines the well known classification method Random Forest with Survival Analysis​.

It helps to identify which of the features are important and also predicts the survival probability at a given time point.

* **Logistic Regression:**

Logistic regression is used to obtain odds ratio in the presence of more than one explanatory variable. The procedure is quite similar to multiple linear regression, with the exception that the response variable is binomial. The result is the impact of each variable on the odds ratio of the observed event of interest.

**Assumptions:**

There are four assumptions associated with a Logistic regression model.

* Independence of errors
* linearity in the logit for continuous variables
* absence of multicollinearity
* lack of strongly influential outliers

**What is misclassification?**

Misclassification (or classification error) occurs when a subject is placed into the wrong population subgroup or category because of observational or measurement error. When this occurs, the actual link between exposure and outcome is distorted.

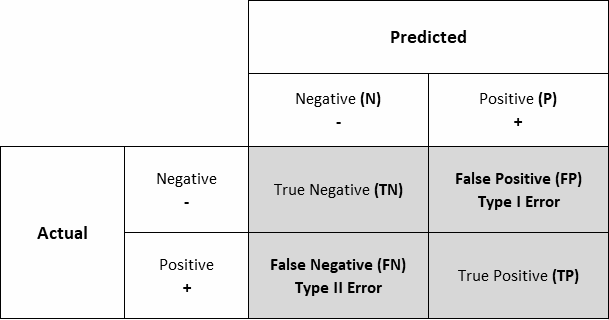
Subjects might be placed into the wrong groups because of:

* Incomplete records.
* Recording errors.
* Misinterpretation of records.
* Errors in records.

Although we take care to minimize the impact of these errors, they are largely unavoidable because human error is innate to any study involving people.

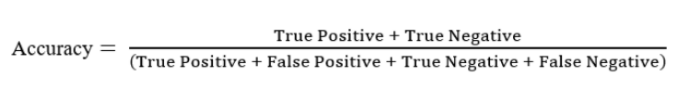
**Confusion Matrix**

Confusion matrix is a performance measurement for machine learning classification problems where output can be two or more classes. It is a table with 4 different combinations of predicted and actual values. It is extremely useful for measuring Recall, Precision, Specificity, Accuracy, and most importantly AUC (area under the curve) and ROC (receiver operating characteristic) curves.



Confusion matrices represent counts from predicted and actual values. The output “TN” stands for True Negative which shows the number of negative examples classified accurately. Similarly, “TP” stands for True Positive which indicates the number of positive examples classified accurately. The term “FP” shows False Positive value, i.e., the number of actual negative examples classified as positive; and “FN” means a False Negative value which is the number of actual positive examples classified as negative.

**Accuracy**



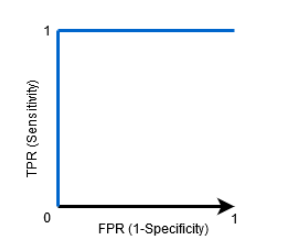
Accuracy is the ratio of the total number of correct predictions and the total number of predictions. Using accuracy as a defining metric for our model does make sense intuitively, but more often than not, it is always advisable to use Precision and Recall too. Although we do aim for high precision and high recall value, achieving both at the same time is not possible.

**Receiver Operator Characteristic (ROC Curve)**

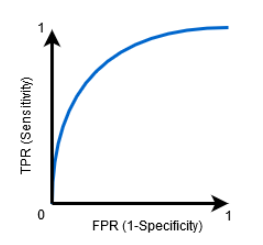
The Receiver Operator Characteristic (ROC) curve is an evaluation metric for binary classification problems. It is a probability curve that plots the TPR against FPR at various threshold values and essentially separates the ‘signal’ from the ‘noise’.

**Area Under Curve (AUC curve)**

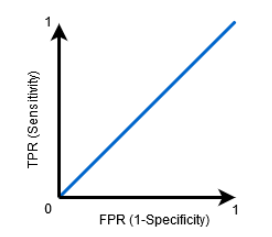
Area Under Curve or AUC is one of the most widely used metrics for model evaluation. It is generally used for binary classification problems. AUC measures the entire two-dimensional area present underneath the entire ROC curve. AUC of a classifier is equal to the probability that the classifier will rank a randomly chosen positive example higher than that of a randomly chosen negative example.



When AUC = 1, then the classifier is able to perfectly distinguish between all the Positive and the Negative class points correctly.



When 0.5<AUC<1, there is a high chance that the classifier will be able to distinguish the positive class values from the negative class values.



When AUC=0.5, then the classifier is not able to distinguish between Positive and Negative class points. Meaning either the classifier is predicting random class or constant class for all the data points.

So, the higher the AUC value for a classifier, the better its ability to distinguish between positive and negative classes.

**Analysis**

* **Logistic Regression:**

The following is the best subset model we obtained to fit the Logistic regression for which the AIC value is minimum at 651.7.

Call: glm(formula = Attrition ~ Age + BusinessTravel + DistanceFromHome +

EducationField + EnvironmentSatisfaction + Gender + JobInvolvement +

JobRole + JobSatisfaction + MaritalStatus + NumCompaniesWorked +

OverTime + PercentSalaryHike + RelationshipSatisfaction +

TotalWorkingYears + TrainingTimesLastYear + WorkLifeBalance +

YearsAtCompany + YearsInCurrentRole + YearsSinceLastPromotion +

YearsWithCurrManager, family = "binomial", data = TrainingData)

Coefficients:

(Intercept) Age

2.39506 -0.02582

BusinessTravelTravel\_Frequently BusinessTravelTravel\_Rarely

2.27473 1.22619

DistanceFromHome EducationFieldLife Sciences

0.04415 -0.64408

EducationFieldMarketing EducationFieldMedical

0.31719 -0.78963

EducationFieldOther EducationFieldTechnical Degree

-0.57202 0.24200

EnvironmentSatisfaction GenderMale

-0.43843 0.43233

JobInvolvement JobRoleHuman Resources

-0.52721 1.52115

JobRoleLaboratory Technician JobRoleManager

1.33100 0.41330

JobRoleManufacturing Director JobRoleResearch Director

0.47421 -15.78918

JobRoleResearch Scientist JobRoleSales Executive

0.58586 0.75873

JobRoleSales Representative JobSatisfaction

1.42601 -0.34296

MaritalStatusMarried MaritalStatusSingle

0.47517 1.45219

NumCompaniesWorked OverTimeYes

0.19962 1.92209

PercentSalaryHike RelationshipSatisfaction

-0.04655 -0.35975

TotalWorkingYears TrainingTimesLastYear

-0.11388 -0.22353

WorkLifeBalance YearsAtCompany

-0.34294 0.13557

YearsInCurrentRole YearsSinceLastPromotion

-0.19590 0.14191

YearsWithCurrManager

-0.10590

Degrees of Freedom: 1024 Total (i.e. Null); 990 Residual

Null Deviance: 891.3

Residual Deviance: 581.7 AIC: 651.7

* The model was then fitted on the train data, we obtain the summary of the model as follows:

Deviance Residuals:

Min 1Q Median 3Q Max

-1.7800 -0.4785 -0.2389 -0.0688 3.3088

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 2.39506 1.54588 1.549 0.121306

Age -0.02582 0.01623 -1.591 0.111630

BusinessTravelTravel\_Frequently 2.27473 0.54975 4.138 3.51e-05 \*\*\*

BusinessTravelTravel\_Rarely 1.22619 0.51430 2.384 0.017117 \*

DistanceFromHome 0.04415 0.01299 3.400 0.000675 \*\*\*

EducationFieldLife Sciences -0.64408 0.97046 -0.664 0.506894

EducationFieldMarketing 0.31719 1.03345 0.307 0.758906

EducationFieldMedical -0.78963 0.97159 -0.813 0.416378

EducationFieldOther -0.57202 1.07902 -0.530 0.596023

EducationFieldTechnical Degree 0.24200 0.99122 0.244 0.807121

EnvironmentSatisfaction -0.43843 0.10244 -4.280 1.87e-05 \*\*\*

GenderMale 0.43233 0.22638 1.910 0.056162 .

JobInvolvement -0.52721 0.14711 -3.584 0.000339 \*\*\*

JobRoleHuman Resources 1.52115 0.81078 1.876 0.060634 .

JobRoleLaboratory Technician 1.33100 0.54552 2.440 0.014692 \*

JobRoleManager 0.41330 0.79632 0.519 0.603753

JobRoleManufacturing Director 0.47421 0.62340 0.761 0.446848

JobRoleResearch Director -15.78918 703.79094 -0.022 0.982101

JobRoleResearch Scientist 0.58586 0.54412 1.077 0.281608

JobRoleSales Executive 0.75873 0.55817 1.359 0.174041

JobRoleSales Representative 1.42601 0.63391 2.250 0.024479 \*

JobSatisfaction -0.34296 0.09875 -3.473 0.000515 \*\*\*

MaritalStatusMarried 0.47517 0.31249 1.521 0.128365

MaritalStatusSingle 1.45219 0.32090 4.525 6.03e-06 \*\*\*

NumCompaniesWorked 0.19962 0.04821 4.140 3.47e-05 \*\*\*

OverTimeYes 1.92209 0.23446 8.198 2.44e-16 \*\*\*

PercentSalaryHike -0.04655 0.03199 -1.455 0.145653

RelationshipSatisfaction -0.35975 0.10090 -3.565 0.000363 \*\*\*

TotalWorkingYears -0.11388 0.03519 -3.236 0.001213 \*\*

TrainingTimesLastYear -0.22353 0.09093 -2.458 0.013964 \*

WorkLifeBalance -0.34294 0.14989 -2.288 0.022138 \*

YearsAtCompany 0.13557 0.04963 2.731 0.006309 \*\*

YearsInCurrentRole -0.19590 0.05846 -3.351 0.000805 \*\*\*

YearsSinceLastPromotion 0.14191 0.05045 2.813 0.004909 \*\*

YearsWithCurrManager -0.10590 0.05871 -1.804 0.071286 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

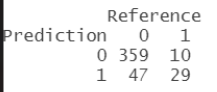
Null deviance: 891.31 on 1024 degrees of freedom

Residual deviance: 581.72 on 990 degrees of freedom

AIC: 651.72

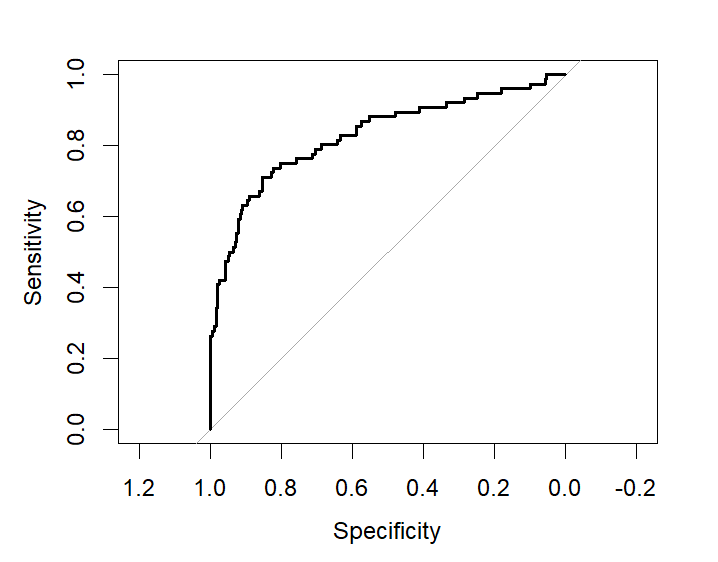
Number of Fisher Scoring iterations: 17

* We can see which variables are significant in predicting the attrition with the help of the displayed p-value. Variables for which p-value<0.05 are significant.
* The confusion Matrix was obtained as follows:



**Accuracy** **= 87.19%**

* The model is thus a good fit
* The ROC curve was obtained as follows:



**AUC = 0.8286**

* **Random Survival Forest:**

The rfsrc function was used to fit the Random Survival Model, we can see the output as follows:

Sample size: 1029

Number of deaths: 167

Number of trees: 1000

Forest terminal node size: 15

Average no. of terminal nodes: 38.671

No. of variables tried at each split: 6

Total no. of variables: 31

Resampling used to grow trees: swor

Resample size used to grow trees: 650

Analysis: RSF

Family: surv

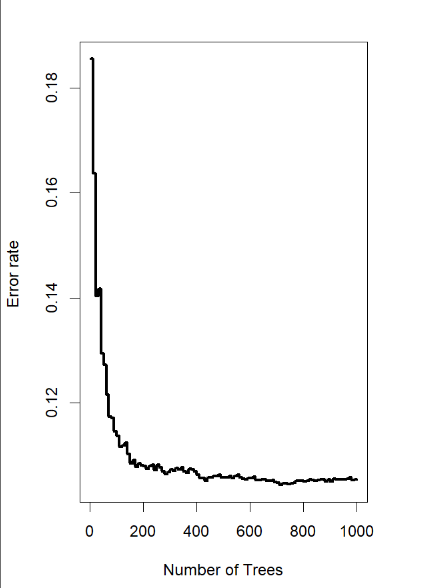
Splitting rule: logrank \*random\*

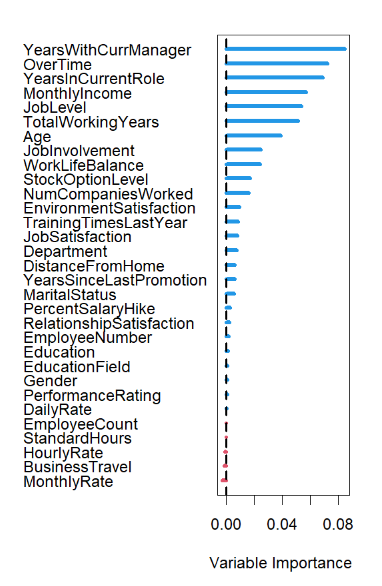
Number of random split points: 3

(OOB) CRPS: 0.12424134

(OOB) Requested performance error: 0.10536098

* Plotting this fit we get the following graphs:



****

* The first graph shows that as the number of trees used was increased the error decreases. The second graph shows the feature importance. Blue lines indicate that the feature is of positive importance(the more the length, the more important is the feature), the interpretation for the red lines is completely opposite.
* Using vimp() we can get the feature importance scores, Higher value indicates more importance.

RelationshipSatisfaction StandardHours StockOptionLevel

-1.518085e-04 0.000000e+00 4.835392e-03

TotalWorkingYears TrainingTimesLastYear WorkLifeBalance

1.979782e-02 9.291575e-04 2.000250e-03

YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager

3.515992e-02 2.490619e-03 3.672183e-02

PerformanceRating PercentSalaryHike Age

-9.493306e-06 -5.482933e-04 8.794892e-03

BusinessTravel DailyRate Department

1.558302e-03 -5.124202e-04 2.656650e-03

DistanceFromHome Education EducationField

2.044043e-03 -5.077722e-05 5.098888e-04

EmployeeCount EmployeeNumber EnvironmentSatisfaction

0.000000e+00 -2.373661e-04 5.088565e-03

Gender HourlyRate JobInvolvement

4.103877e-04 -7.339144e-04 3.563872e-03

JobLevel JobSatisfaction MaritalStatus

2.194941e-02 1.483397e-03 3.285577e-03

MonthlyIncome MonthlyRate NumCompaniesWorked

2.185625e-02 -9.056352e-04 5.485707e-03

OverTime

2.950199e-02

* Now we predict for the Test data using pred(), the output is as follows:

Sample size of test (predict) data: 441

Number of grow trees: 1000

Average no. of grow terminal nodes: 38.671

Total no. of grow variables: 31

Resampling used to grow trees: swor

Resample size used to grow trees: 279

Analysis: RSF

Family: surv

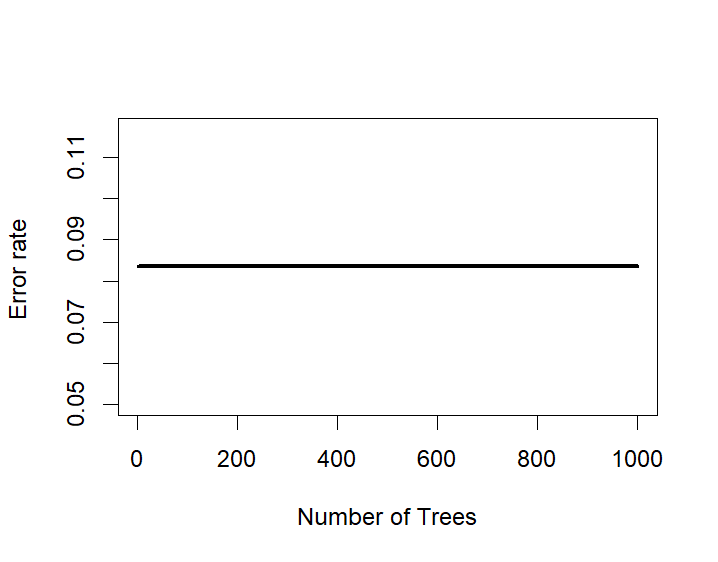
CRPS: 0.20975399

Requested performance error: 0.08350962

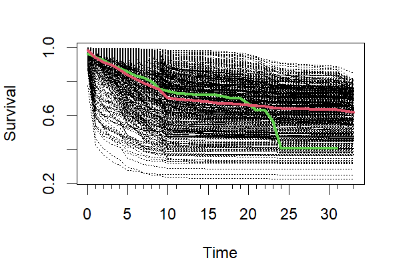
* We get the head of the predicted Survival time as follows:

10.148533 13.656142 7.286356 4.832067 12.665874 15.954437

* If we plot the predicted values we get the following graph:



* As the number of trees increases the Error rate remains constant.
* The graph for Survival function was obtained as follows:

****

* From the graph we can say that on average the survival time of an individual in the company is around 15 years.

**Key Findings**

|  |  |
| --- | --- |
| **Method** | **Accuracy** |
| **Logistic Regression** | **87.19%** |

* **We had seen from the Feature importance that YearsWithCurrManager, OverTime, YearsInCurrentRole, MonthlyIncome, JobLevel, TotalWorkingYears are the most important variables contributing to Attrition.**

**Recommendations**

* Plan and allocate projects in such a way to avoid the use of overtime.
* Gather information on industry benchmarks to determine if the company is providing competitive wages.
* The people at the managerial role should be polite and give reasonable deadlines and perks to the workers.
* Lastly, if three of the above things are not possible, the company can make their own guidelines on whom to hire looking at these important variables so that there is as little Attrition as possible.

**Future Scope**

* Help organizations to find better ways to prevent attrition or to plan in advance the hiring of new candidate
* This project can be extended to companies in all industries
* More detailed Analysis using Survival Analysis and other Machine Learning classifier Algorithms

**References:**

* CoxRF: Employee Turnover Prediction based on Survival Analysis

Link: <https://www.semanticscholar.org/paper/CoxRF%3A-Employee-Turnover-Prediction-Based-on-Zhu-Shang/bb798dc641d058472544b90497a81a4756f611e7>

In this paper we came across the following references:

[1] H. Li, Y. Ge, H. Zhu, H. Xiong, and H. Zhao, “Prospecting the career

development of talents: A survival analysis perspective.” ACM, 2017,

pp. 917–925.

[2] Z. R. Mulla, K. Kelkar, M. Agarwal, S. Singh, and N. E. Sen,

“Engineers’ voluntary turnover: application of survival analysis,” The

Indian Journal of Industrial Relations, pp. 328–341, 2013.

[3] J. Wang, Y. Zhang, C. Posse, and A. Bhasin, “Is it time for a career

switch?” ACM, 2013, pp. 1377–1388.