**CAPSTONE PROJECT REPORT**

**Predicting Employee Attrition**

*Submitted By:*

*Riba Khan*

*M.S. in Business Analytics*

*Kent State University*

2022-07-31

**Abstract**

Employee turnover has been deemed a major problem for businesses due to its detrimental effects on workplace productivity and long-term expansion plans. Organizations use machine learning approaches to anticipate employee turnover to address this issue. Organizations can act for employee retention or succession planning owing to accurate predictions. This paper implemented five machine learning models and the performance of each of the models is analyzed by checking the accuracy. Additionally, the best model which turned out to be Random Forest with the accuracy of 97.56% is chosen to make the prediction and saved.

Keywords: Attrition, Data Visualisation, Model selection, Random Forest

Table of Contents

[Abstract 2](#_Toc110416747)

[CHAPTER 1: INTRODUCTION 4](#_Toc110416748)

[CHAPTER 2: RELATED WORK 7](#_Toc110416755)

**CHAPTER 3: DATA PREPARATION AND UNDERSTANDING………………………………………………………………………………….9**

**CHAPTER 4: MODEL VALIDATION………………………………………………………………………………………19**

[CHAPTER 5: KEY FINDINGS 27](#_Toc110416759)

[CHAPTER 7: REFERENCES 29](#_Toc110416776)

[Works Cited 29](#_Toc110416777)

# CHAPTER 1: INTRODUCTION

## Title & Objective of the study

***Title***: Predicting Employee Attrition

***Objective***: To predict employee attrition rate by choosing the best machine learning algorithms and to evaluate which factors contribute most to the attrition of employees.

## Need of the Study

Employee Attrition is an important way to measure the overall growth and management of the organization. A business's top priority should be lowering staff turnover. To find, hire, and train a replacement employee might cost up to twice the pay of the current employee. Additionally, turnover can lower production and morale among your remaining staff, making it more difficult to find fresh talent. Therefore, it is important to study employee attrition.

## What is Employee Attrition?

Employee attrition is the term used to describe an employee leaving an organization due to unforeseen circumstances like resignation due to personal or professional reasons. In many situations, the employer has no influence over why workers are departing the job at a faster rate than they are being hired. In many situations, the employer has no influence over why workers are departing the job at a faster rate than they are being hired.

Employee Attrition can be classified into two types: <https://www.betterup.com/blog/employee-attrition>

1. Involuntary Attrition

Involuntary Attrition happens when instead of the employee choosing to quit the decision to fire the employee is made by the organization. It’s when the company decides to eliminate a particular position. The choice to eliminate a posy might have been made to cut workforce costs or maintain stability.

1. Voluntary Attrition

When an employee chooses to quit the company, its is termed as voluntary attrition. There could be numerous reasons behind an employee leaving the organization, such as migrating to a new company, changing careers, personal and family issues, etc. Additionally, a terrible work environment and a lack of work-life balance can occasionally result in employees quitting a company quickly.

## 1.4 Data Sources:

The dataset is taken from Kaggle. The dataset contains 35 rows and 1470 columns. This is a fictional data set created by IBM data scientists.

## 1.5 Tools & Techniques

*Tools*:

*Libraries used:* Some basic libraries used for analysis & model building are mentioned below

* *library(ggplot2)* – To create complex plots from a data.
* *library(pROC)* – To build ROC curves.
* *library(rpart)* – To build classification trees.
* *library(rpart.plot)* – To plot and adjust the displayed tree for best fit.
* *library(caret)* – To solve any classification machine learning problem.
* *library(dplyr)* – To apply filters , sorting or making adjustments to the data.
* *library(nnet)* – To predict the value and see the accuracy.
* *library(randomForest)* – To build random forest machine learning algorithm.
* *library(class)* – To return the values of the class attributes.
* *library(lattice)* – To provide high-level data visualization system.
* *library(e1071)* – To provide functions for statistical algorithms like naïve bayes.
* *library(ROCR)* - To evaluate and visualize the performance of scoring classifiers.

The analysis is performed on R programming.

*Techniques*:

Techniques:

This report follows the technique of predictive analysis. To better understand the variables exploratory data analysis is performed. The variables that contribute to the most to high attrition rate are also identified.

The predictive analysis is performed in the following stages:

* **Data Understanding**– understanding the data and variables that can have an impact on the prediction.
* **Data Exploration (EDA)** – looking at categorical and numerical variables by using different plots and diagrams to make inferences about the data
* **Data Cleaning** – imputing missing values in the data and eliminating the variables that don’t have any significance.
* **Data Splitting** – Splitting the data set into the train (70%) and test (30%) data sets.
* **Feature Engineering** –changing existing variables to factors as needed for analysis.
* **Model Building** – using five machine learning algorithms
* **Model Selection** – after building multiple predictive models, the model with high accuracy is selected and implemented

.

## 1.6 Challenges faced during study

* The dataset that was collected was biased. The target variable (Attrition) is unbalanced.
* The dataset has a lot of variables that are factors but are imported as numbers.

# CHAPTER 2: RELATED WORK

In their study, (Hong, Wei, & Chen, 2007)investigated the viability of using the Logit and Probit models, which have been successfully used to handle nonlinear classification and regression issues, to predict employee voluntary turnover. With a valid sample size of 132, a numerical example using voluntary turnover data of 150 professional employees from a central Taiwanese motor marketing company was used. The modeling data set and the testing data set were separated from the overall data set. The logit and probit models were evaluated using the modeling data set. The testing data set was utilized to estimate model performance when applied to future data rather than for model creation or model selection. According to the empirical findings of their study, the proposed models have strong predictive power, and the two models (logit and probit) also offer a potential option for forecasting employee turnover in human resource administration. The authors proposed that turnover research should take new tacks based on newer technologies, which would give rise to fresh dilemmas and difficulties (such as the use of neural networks and support vector machines to conduct classification problems for detecting who will stay and leave.

(Usha.P.M, 2019) conducted study on the decision tree algorithm J48 and a Naive Bayes classifier were evaluated in for forecasting the likelihood of an employee leaving the organization. Particularly, the effectiveness of percentage split 70 and tenfold cross-validation were assessed for each algorithm. Tenfold cross-validation yielded an accuracy of 82.4 percent and an inaccurate classification rate of 17.6 percent with J48, whereas percentage split 70 yielded an accuracy of 82.7 percent and an incorrect classification rate of 17.3 percent. Using tenfold cross-validation, the Naive Bayes classifier obtained an accuracy of 78.8% and an inaccurate classification of 21.2%; meanwhile, percentage split 70 produced an accuracy of 81.0% and an incorrect classification of 19%.

Features include things like employee performance, the typical number of hours worked each month, the number of years an employee has been with the organization, and more were studied by (Shankar, Rajanikanth, Sivaramaraju, & Murthy, 2018). An experimental comparison of Naive Bayes, Logistic Regression, and Multi-layer Perceptron (MLP) Classifier showed that k-Nearest Neighbors performs significantly better than its rival methods. Several well-known classifiers are used to the human resource data to reduce employee attrition, including Decision Tree, Logistic Regression, SVM, KNN, Random Forest, and Naive Bayes approaches.

The Performance Assessment of Data Balancing Techniques was suggested by (Anil Jadhav, 2022) for the classification task . In many classification issues, including intrusion detection, fraud detection, anomaly detection, and many others, the imbalanced dataset is a significant difficulty. To get the high accuracy outcomes from the prediction models, data balancing was used. The research project used an unbalanced dataset to empirically address the performance difficulties. We looked at a number of balancing strategies, including hybrid sampling (HS), synthetic minority over sampling (SMOTE), under sampling (US), random over sampling examples (ROSE), over sampling (OS), and clustering-based under sampling (CBUS). The performance factor was the imbalance ratio (IR). The findings of the study experiments demonstrate that data balance helps to enhance the performance of the used classifiers. The outcomes also show that there is no discernible performance difference in balancing technique.

To forecast employee turnover, a three-stage method based on pretreatment, processing, and post-processing techniques was proposed from (Saeed Najafi-Zangeneh, 2021). The system was tested and trained using the IBM HR employee dataset. For the dimension reduction stage, the max-out feature selection method was used. The method of logistic regression was used to forecast staff attrition. The model's output had an accuracy rating of 81%. The parameters of the framework were verified.

# CHAPTER 3: DATA PREPARATION AND UNDERSTANDING

* 1. **Data Exploration**

#---------------Structure and summary of data----------------  
str(Employee\_Attrition)

## 'data.frame': 1470 obs. of 35 variables:  
## $ Age : int 41 49 37 33 27 32 59 30 38 36 ...  
## $ Attrition : chr "Yes" "No" "Yes" "No" ...  
## $ BusinessTravel : chr "Travel\_Rarely" "Travel\_Frequently" "Travel\_Rarely" "Travel\_Frequently" ...  
## $ DailyRate : int 1102 279 1373 1392 591 1005 1324 1358 216 1299 ...  
## $ Department : chr "Sales" "Research & Development" "Research & Development" "Research & Development" ...  
## $ DistanceFromHome : int 1 8 2 3 2 2 3 24 23 27 ...  
## $ Education : int 2 1 2 4 1 2 3 1 3 3 ...  
## $ EducationField : chr "Life Sciences" "Life Sciences" "Other" "Life Sciences" ...  
## $ EmployeeCount : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ EmployeeNumber : int 1 2 4 5 7 8 10 11 12 13 ...  
## $ EnvironmentSatisfaction : int 2 3 4 4 1 4 3 4 4 3 ...  
## $ Gender : chr "Female" "Male" "Male" "Female" ...  
## $ HourlyRate : int 94 61 92 56 40 79 81 67 44 94 ...  
## $ JobInvolvement : int 3 2 2 3 3 3 4 3 2 3 ...  
## $ JobLevel : int 2 2 1 1 1 1 1 1 3 2 ...  
## $ JobRole : chr "Sales Executive" "Research Scientist" "Laboratory Technician" "Research Scientist" ...  
## $ JobSatisfaction : int 4 2 3 3 2 4 1 3 3 3 ...  
## $ MaritalStatus : chr "Single" "Married" "Single" "Married" ...  
## $ MonthlyIncome : int 5993 5130 2090 2909 3468 3068 2670 2693 9526 5237 ...  
## $ MonthlyRate : int 19479 24907 2396 23159 16632 11864 9964 13335 8787 16577 ...  
## $ NumCompaniesWorked : int 8 1 6 1 9 0 4 1 0 6 ...  
## $ Over18 : chr "Y" "Y" "Y" "Y" ...  
## $ OverTime : chr "Yes" "No" "Yes" "Yes" ...  
## $ PercentSalaryHike : int 11 23 15 11 12 13 20 22 21 13 ...  
## $ PerformanceRating : int 3 4 3 3 3 3 4 4 4 3 ...  
## $ RelationshipSatisfaction: int 1 4 2 3 4 3 1 2 2 2 ...  
## $ StandardHours : int 80 80 80 80 80 80 80 80 80 80 ...  
## $ StockOptionLevel : int 0 1 0 0 1 0 3 1 0 2 ...  
## $ TotalWorkingYears : int 8 10 7 8 6 8 12 1 10 17 ...  
## $ TrainingTimesLastYear : int 0 3 3 3 3 2 3 2 2 3 ...  
## $ WorkLifeBalance : int 1 3 3 3 3 2 2 3 3 2 ...  
## $ YearsAtCompany : int 6 10 0 8 2 7 1 1 9 7 ...  
## $ YearsInCurrentRole : int 4 7 0 7 2 7 0 0 7 7 ...  
## $ YearsSinceLastPromotion : int 0 1 0 3 2 3 0 0 1 7 ...  
## $ YearsWithCurrManager : int 5 7 0 0 2 6 0 0 8 7 ...

summary(Employee\_Attrition)

## Age Attrition BusinessTravel DailyRate   
## Min. :18.00 Length:1470 Length:1470 Min. : 102.0   
## 1st Qu.:30.00 Class :character Class :character 1st Qu.: 465.0   
## Median :36.00 Mode :character Mode :character Median : 802.0   
## Mean :36.92 Mean : 802.5   
## 3rd Qu.:43.00 3rd Qu.:1157.0   
## Max. :60.00 Max. :1499.0   
## Department DistanceFromHome Education EducationField   
## Length:1470 Min. : 1.000 Min. :1.000 Length:1470   
## Class :character 1st Qu.: 2.000 1st Qu.:2.000 Class :character   
## Mode :character Median : 7.000 Median :3.000 Mode :character   
## Mean : 9.193 Mean :2.913   
## 3rd Qu.:14.000 3rd Qu.:4.000   
## Max. :29.000 Max. :5.000   
## EmployeeCount EmployeeNumber EnvironmentSatisfaction Gender   
## Min. :1 Min. : 1.0 Min. :1.000 Length:1470   
## 1st Qu.:1 1st Qu.: 491.2 1st Qu.:2.000 Class :character   
## Median :1 Median :1020.5 Median :3.000 Mode :character   
## Mean :1 Mean :1024.9 Mean :2.722   
## 3rd Qu.:1 3rd Qu.:1555.8 3rd Qu.:4.000   
## Max. :1 Max. :2068.0 Max. :4.000   
## HourlyRate JobInvolvement JobLevel JobRole   
## Min. : 30.00 Min. :1.00 Min. :1.000 Length:1470   
## 1st Qu.: 48.00 1st Qu.:2.00 1st Qu.:1.000 Class :character   
## Median : 66.00 Median :3.00 Median :2.000 Mode :character   
## Mean : 65.89 Mean :2.73 Mean :2.064   
## 3rd Qu.: 83.75 3rd Qu.:3.00 3rd Qu.:3.000   
## Max. :100.00 Max. :4.00 Max. :5.000   
## JobSatisfaction MaritalStatus MonthlyIncome MonthlyRate   
## Min. :1.000 Length:1470 Min. : 1009 Min. : 2094   
## 1st Qu.:2.000 Class :character 1st Qu.: 2911 1st Qu.: 8047   
## Median :3.000 Mode :character Median : 4919 Median :14236   
## Mean :2.729 Mean : 6503 Mean :14313   
## 3rd Qu.:4.000 3rd Qu.: 8379 3rd Qu.:20462   
## Max. :4.000 Max. :19999 Max. :26999   
## NumCompaniesWorked Over18 OverTime PercentSalaryHike  
## Min. :0.000 Length:1470 Length:1470 Min. :11.00   
## 1st Qu.:1.000 Class :character Class :character 1st Qu.:12.00   
## Median :2.000 Mode :character Mode :character Median :14.00   
## Mean :2.693 Mean :15.21   
## 3rd Qu.:4.000 3rd Qu.:18.00   
## Max. :9.000 Max. :25.00   
## PerformanceRating RelationshipSatisfaction StandardHours StockOptionLevel  
## Min. :3.000 Min. :1.000 Min. :80 Min. :0.0000   
## 1st Qu.:3.000 1st Qu.:2.000 1st Qu.:80 1st Qu.:0.0000   
## Median :3.000 Median :3.000 Median :80 Median :1.0000   
## Mean :3.154 Mean :2.712 Mean :80 Mean :0.7939   
## 3rd Qu.:3.000 3rd Qu.:4.000 3rd Qu.:80 3rd Qu.:1.0000   
## Max. :4.000 Max. :4.000 Max. :80 Max. :3.0000   
## TotalWorkingYears TrainingTimesLastYear WorkLifeBalance YearsAtCompany   
## Min. : 0.00 Min. :0.000 Min. :1.000 Min. : 0.000   
## 1st Qu.: 6.00 1st Qu.:2.000 1st Qu.:2.000 1st Qu.: 3.000   
## Median :10.00 Median :3.000 Median :3.000 Median : 5.000   
## Mean :11.28 Mean :2.799 Mean :2.761 Mean : 7.008   
## 3rd Qu.:15.00 3rd Qu.:3.000 3rd Qu.:3.000 3rd Qu.: 9.000   
## Max. :40.00 Max. :6.000 Max. :4.000 Max. :40.000   
## YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager  
## Min. : 0.000 Min. : 0.000 Min. : 0.000   
## 1st Qu.: 2.000 1st Qu.: 0.000 1st Qu.: 2.000   
## Median : 3.000 Median : 1.000 Median : 3.000   
## Mean : 4.229 Mean : 2.188 Mean : 4.123   
## 3rd Qu.: 7.000 3rd Qu.: 3.000 3rd Qu.: 7.000   
## Max. :18.000 Max. :15.000 Max. :17.000

*As observed from the dataset, the target variable is the Attrition column which is mentioned as ‘Yes’ and ‘No’. The dataset is a mix of numerical and categorical data.*

**3.2 Most significant variables in the dataset**

df$variable.importance

## OverTime JobRole TotalWorkingYears   
## 149.1185480 90.4091879 87.1986316   
## MonthlyIncome YearsAtCompany Age   
## 60.9431696 32.2064100 30.5826958   
## NumCompaniesWorked YearsInCurrentRole JobLevel   
## 27.6066587 22.2679743 17.2957267   
## JobInvolvement EmployeeNumber MaritalStatus   
## 17.1026359 16.3554752 14.0360770   
## JobSatisfaction EducationField YearsWithCurrManager   
## 12.1006830 7.4826803 6.3687320   
## DailyRate HourlyRate PercentSalaryHike   
## 5.0120100 4.7816058 3.9839734   
## MonthlyRate TrainingTimesLastYear WorkLifeBalance   
## 3.8381760 3.4058774 2.7924653   
## YearsSinceLastPromotion Department Education   
## 2.1186531 1.8616435 0.7640068

As we can see that there are six variables that have the highest significance in attrition of employees which are ***Over Time, Job Role, Total Working Years, Monthly Income, Years at a company and Age.***

* 1. **Data Cleaning**

Data cleaning is the process of removing duplicate, corrupted, incorrectly formatted, inaccurate, or corrupted data from a dataset. When combining data from many sources, there are many possibilities for data to be duplicated or incorrectly categorized. Therefore, it becomes important to clean the data before building the model.

* Checking for any null values

## Cleaning the data  
#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# checking null values in the dataset   
sum(is.na.data.frame(Employee\_Attrition1))

## [1] 0

*There are no Null Values which is good.*

* Removing columns which don’t have any significance   
    
  # Removing coloumns that are not needed for the analysis  
  Employee\_Attrition1$Over18 <- NULL  
  Employee\_Attrition1$EmployeeCount <- NULL  
  Employee\_Attrition1$StandardHours <- NULL  
  Employee\_Attrition1$EmployeeNumber<-NULL  
  Employee\_Attrition1$DistanceFromHome<-NULL

**3.4 Data Preparation**

#Prparing the data for ML Algorithms  
#\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
   
# convert certain integer variable to factor variable as they make more sense as factor data type  
convert\_factor<- c("Education", "EnvironmentSatisfaction", "JobInvolvement",   
 "JobLevel","Department","EducationField","JobSatisfaction",   
 "PerformanceRating", "RelationshipSatisfaction",  
 "StockOptionLevel","TrainingTimesLastYear","WorkLifeBalance",  
 "BusinessTravel","Gender","JobRole","MaritalStatus","OverTime" )  
  
Employee\_Attrition1[, convert\_factor] <- lapply((Employee\_Attrition1  
 [, convert\_factor]), as.factor)  
Employee\_Attrition1[, convert\_factor] <- lapply((Employee\_Attrition1  
 [, convert\_factor]), unclass

**3.5 Data Visualisations**

* Attrition

Chart, bar chart

Description automatically generated

Figure:1

The above graph shows that the attrition data is highly imbalanced. It is visible that there are 1233 employees who have not leave the company while only 237 attired. The imbalance in the dataset needs to be fixed and can be seen in further steps in the report.

Since we balance our dataset , the graph looks different than Figure 1 now.

**Visualizations features with higher significance**

* **Overtime**

**Graphical user interface

Description automatically generated**

**Figure:2**

*From the above image it looks like employees who are not doing over time are staying with the company for longer period.*

* **Job Role**

**Chart, bar chart

Description automatically generated**

**Figure:3**

*Job Role of laboratory Technician and Sales Execute have high attrition rate.*

* **Total Working Years**

Chart, bar chart

Description automatically generated

**Figure:4**

*Employees who are working in the company for a shorter span of time have attired rather than the ones who are working for long.*

* **Monthly Income**

Chart, histogram

Description automatically generated

**Figure:5**

*Employees with monthly salary less than 5000 have left the organization.*

* **Age**

Chart, bar chart

Description automatically generated

**Figure:6**

*Employees in the age group of 30-35 have leave the company more than any other age group.*

* **Years at The Company**

**Graphical user interface

Description automatically generated**

**Figure: 7**

*Employee with less no of years spent working in the organization tend to quit.*

**3.6 Checking for Class Imbalance**

Chart, bar chart

Description automatically generated

**Figure:8**

Imbalanced classification is a supervised learning issue in which one class significantly outdoes the other class. Compared to multi-level classification problems, binary classification problems are more likely to encounter this issue. Therefore, it is important to look at the sampling methods that can be used in the study to balance the data. Researchers argue three main methods could be of great benefit to balance the dataset. In brief, these three methods are as follows:

* Under Sampling

Working with the majority class is this way. Balancing the data set lowers the number of observations from the dominant class. When the data set is large and the number of training samples is decreased, run time and storage issues are improved, making this method one of the choices.

* Up Sampling

To present a better representation of the minority class in the sample, oversampling increases the number of instances in the minority class by randomly repeating them.

* SMOTE (Synthetic Minority Over-sampling Technique).

The SMOTE algorithm generates false data based on feature space (rather than data space) commonalities from minority samples. To skew the classifier's learning bias in favor of the minority class, one may additionally state that it generates a random set of minority class observations.

# Observation : Attrition NO> Attirition YES  
# Data in not balanced so We will use UpSampling Techique  
  
#--------------UP Sampling---------------------#  
yes<- which(Employee\_Attrition$Attrition=="Yes")  
No<- which(Employee\_Attrition$Attrition=="No")  
length(yes)

## [1] 237

length(No)

## [1] 1233

Yes.upsam <- sample(yes, length(No), replace= T)  
Employee\_Attrition1<- Employee\_Attrition[c(No, Yes.upsam),]  
table(Employee\_Attrition1$Attrition)

##   
## No Yes   
## 1233 1233

Chart, bar chart

Description automatically generated

Figure:9

*Since we balance our dataset , the graph looks different than Figure 8 now.*

*After looking at the three sampling methods, I decided to choose the Up Sampling technique to balance the dataset. In comparison, to the other methods, one advantage that Up Sampling provides is that it prevents the model from losing any information.*

**CHAPTER 4: MODEL VALIDATION**

The process of ensuring that a model truly serves its intended function is known as model validation. This usually entails verifying that the model is accurate in the circumstances of its intended application.

* Splitting the dataset into train and test. The study divides 70% of the data for the training and 30% for the test.
* Five machine learning algorithms are used to make further predictions.
  1. **Fitting Models to The Data**
* ***Decision Tree***

It is a machine learning algorithm with uses in several different fields. Both classification and regression issues can be solved using decision trees. The term itself implies that it displays the predictions that come from a sequence of feature-based splits using a flowchart that resembles a tree structure. The decision is made by the leaves at the end, which follow the root node.

Diagram

Description automatically generated

**Figure:10**

## For Desicion Tree   
## Accuracy = 77.7027   
## Prediction Error= 22.2973   
## Percision= 0.7869318   
## Recall= 0.7547684

Chart, line chart

Description automatically generated

**Figure:11**

*Area Under the Curve : 0.802*

* ***Logistic Regression***

One of the most often used Machine Learning algorithms, within the category of Supervised Learning, is logistic regression. Using a predetermined set of independent factors, it is used to predict the categorical dependent variable.

In a categorical dependent variable, the output is predicted via logistic regression. As a result, the finding must be a discrete or categorical value. Rather of providing the exact values of 0 and 1, it provides the probabilistic values that fall between 0 and 1.

## For Logistic Regression   
## Accuracy = 73.78378   
## Prediction Error= 26.21622   
## Percision= 0.7520436   
## Recall= 0.7282322

Chart

Description automatically generated

**Figure:11**

*Area under the curve:0.832*

* ***Random Forest***

Leo Breiman and Adele Cutler are the creators of the widely used machine learning technique known as random forest, which mixes the output of various decision trees to produce a single outcome. Its widespread use is motivated by its adaptability and usability because it can solve classification and regression issues.

## For Random forest   
## Accuracy = 97.02703   
## Prediction Error= 2.972973   
## Percision= 0.9618529   
## Recall= 0.9778393

Chart

Description automatically generated

**Figure:12**

Chart, line chart

Description automatically generated\

**Figure:13**

*Area under the curve: 0.9935*

* ***KNN***

The k-nearest neighbors algorithm, sometimes referred to as KNN or k-NN, is a supervised learning classifier that employs proximity to produce classifications or predictions about the grouping of a single point of data. Although it can be applied to classification or regression issues, it is commonly employed as a classification algorithm because it relies on the idea that comparable points can be discovered close to one another.

The value of k here is taken as 3.

## For KNN Classifier   
## Accuracy = 77.83784   
## Prediction Error= 22.16216   
## Percision= 0.599455   
## Recall= 0.92827

Chart, line chart

Description automatically generated

**Figure:14**

* ***Naïve Bayes***

It is a classification algorithm built on the Bayes Theorem and based on the idea of predictor independence. A Naive Bayes classifier, to put it simply, believes that the presence of one feature in a class has nothing to do with the presence of any other feature.

Simple to construct and especially helpful for very big data sets is the naive Bayes model. Along with being intuitive, Naive Bayes is known to perform better than even the most complex classification techniques.

## For Naive bayes   
## Accuracy = 64.86486   
## Prediction Error= 35.13514   
## Percision= 0.5558583   
## Recall= 0.6777409

Chart, line chart

Description automatically generated

**Figure:15**

*Area under the curve : 0.749*

**4. 2 Model Selection**

It is the process of finally selecting the best model to make predictions.

**Combine ROC Curves of the models**

Graphical user interface, chart

Description automatically generated

**Figure:16**

**Results of Each Model**

**Text

Description automatically generated**

**Figure:17**

*After, careful consideration of all the machine learning algorithms it is observed that accuracy of Random Forest is the highest. Additionally, prediction error is lowest and other measures are also good.*

* 1. **Deployment of Model**

Soo, we use Random Forest model on the real data.

The results of final predictions turned out to be as follows:

## Final Predictions   
## Accuracy = 98.36735   
## Prediction Error= 1.632653   
## Percision= 0.9886456   
## Recall= 0.9918633

**The confusion matrix turns out to be this on the real data. The results appears as follows:**

**Diagram

Description automatically generated with medium confidence**

Figure:18

Finally, the existing model is being saved and the whole purpose of building the model on the existing data is to better understand current business issues and develop more targeted solutions, the company will be able to utilize the model on future data once it becomes available to predict or prevent employee attrition from happening before it does.

CHAPTER 5: KEY FINDINGS

The following questions can now be answered by completing the final report.

* What factors contribute the most to employee attrition rate?

There were five factors that contributed most to the attrition which were:

Job Role

Overtime

Total Working Years

Monthly Income

Years at company

Age

* Can we predict future terminations?

In a word we can say yes.

* How well can we predict?

We can respond, okay. Simply said, "pretty good." We define accuracy as the number of times the model correctly predicted the actual when evaluating models for predicting categories. As a result, we have a variety of interests. Error matrices are the earliest of them. Cross-tabulating the actual findings with the expected results is what error matrices are used for. Every prediction would be identical to reality if it were a "perfect" 100 percent of the time. This hardly ever occurs. The better the rate of correct predictions and the lower the rate of errors.

We can forecast well thanks to random forest, which proved to be an excellent machine learning method when tested on real data.

CHAPTER 6: RECOMMENDATIONS AND CONCLUSION

* Using basic exploratory data analysis, feature engineering, and learning models like a Random Forest and a Logistic Regression classifier, we have developed a very simple model that predicts employee attrition with an accuracy of 98.36 percent on the real data.
* Practically speaking, the emphasis would be on the organization to make "decisions" about what the data is showing them if this were real data from a real firm.
* The imbalance between work and reward is probably one of the causes of attrition. It should be determined whether our organization has an appropriate overtime policy in this situation because it mostly pertains to those who are working overtime and who frequently earn relatively modest salaries.
* Compared to other jobs, positions for lab technicians and sales representatives have experienced the most attrition. Their excessive working hours (including overtime) and low monthly salary make it justifiable. The HR department should establish a plan to maintain work-life balance and update these employees' salary in accordance with market standards to prevent attrition in these positions.
* I have so attempted to address the research aims, which were to identify the causes generating attrition and predict the attrition phenomena, by conducting this analysis. The study's conclusions suggest that HR professionals can employ statistical methods and machine learning models to develop more informed, fact-based policies and decisions. The results of changes in HR policy can also be directly observed by analysing numerous scenarios using the prediction algorithm. The HR team can be informed as soon as a critical or high-performing employee is more likely to attrition by using such strategies as an early warning system. The HR team can then appropriately reward the employee or address any complaints made.

CHAPTER 7: REFERENCES

I am using a dataset put up by IBM for my analysis. It can be downloaded from:

<https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset>

<https://www.javatpoint.com/logistic-regression-in-machine-learning>

<https://www.ibm.com/topics/knn>

# Works Cited

Wooll, M. (2022, January 24). *Fighting employee attrition: What is within your control?* Retrieved from BetterUP: https://www.betterup.com/blog/employee-attrition

*SVM in R for Data Classification using e1071 Package*. (n.d.). Retrieved from Tech Vidan: https://techvidvan.com/tutorials/svm-in-r/#:~:text=e1071%20Package%20in%20R,support%20vector%20machine%2C%20etc

Tobias Sing, O. S. (n.d.). *ROCR: visualizing classifier performance in R.*

avcontentteam. (2020, July 5). *Practical Guide to deal with Imbalanced Classification Problems in R* . Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2016/03/practical-guide-deal-imbalanced-classification-problems/

Stedman, C. (n.d.). *data cleansing (data cleaning, data scrubbing)* . Retrieved from Tech Target: https://www.techtarget.com/searchdatamanagement/definition/data-scrubbing#:~:text=Data%20cleansing%2C%20also%20referred%20to,removing%20data%20to%20correct%20them

Saini, A. (2021, August 29). *Decision Tree Algorithm – A Complete Guide*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/

n, I. C. (2020, December 2020). *Random Forest*. Retrieved from IBM: https://www.ibm.com/cloud/learn/random-forest

Ray, S. (2017, September 11). *6 Easy Steps to Learn Naive Bayes Algorithm with codes in Python and R*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

Hong, W.-C., Wei, S.-Y., & Chen, Y.-F. (2007). A Comparative Test of Two Employee Turnover Prediction Models. *International Journal of Management;*, 216-229.

Shankar, R. S., Rajanikanth, J., Sivaramaraju, V., & Murthy, K. (2018). PREDICTION OF EMPLOYEE ATTRITION USING DATAMINING. *IEEE*.

Usha.P.M, D. N. (2019). Analysing employee attrition using machine learning. *Karpagam Journal of Computer Science*, 277-282.

Anil Jadhav, S. M. (2022). An Empirical Assessment of Performance of Data Balancing Techniques in Classification Task.". *Applied Sciences*, 3928.

Saeed Najafi-Zangeneh, N. S.-G.-N. (2021). An Improved Machine Learning-Based Employees Attrition Prediction Framework with Emphasis on Feature Selection. *Mathematics*, 1266.