# **CAPSTONE PROJECT IN BUSINESS ANALYTICS**

# **IBM EMPLOYEE ATTRITION**

# Table of Content

Problem Statement ……………………………………………………………………………………….1

Introduction ………………………………………………………………………………………………….1

Overview of the Dataset …………………………………………………………………………………2

Exploratory Data Analysis ………………………………………………………………………………3

Modelling Approach ……………………………………………………………………………………….12

Performance Evaluation ………………………………………………………………………………..14

Issues Encountered and Proposed Solutions ……………………………………………………17

Conclusions …………………………………………………………………………………………………….17.

Reference ……………………………………………………………………………………………………..18

# Problem Statement

Develop a predictive model for analyzing employee attrition using HR analytics dataset. This project aims to analyze the provided dataset, including employee-related features such as age, pay, work satisfaction, performance ratings, and more, to forecast whether an employee is likely to leave the company in the future. By building an accurate prediction model and adopting retention strategies, the organization wants to boost employee satisfaction among workers and reduce staff turnover. I’m going to predict the attrition rate of the employees in IBM company by using advanced analysis techniques like Random Forest, Decision Tree and XGBoost.

# Introduction

This is a hypothetical dataset created by IBM data scientists. This contains background details of each employee and whether they are still in the company or whether they have resigned. The dataset is imbalanced since 84% of the employees are currently employed and only 16% are resigned. In descriptive analysis we got that age, marital status, distance from work, field of education monthly income, daily rate, and stock option level overtime, environment satisfaction, job satisfaction, department and job role variables had the biggest contribution towards the attrition.

Employee attrition is a significant business concerns in today’s knowledge-driven marketplace, where employees are the most important human capital assets. The organization having capacity to withstand its long-lasting relationship with employees would survive in the marketplace and other would fade away in the long run. The reasons for the employees to leave the organization was endless but the reason why employees leave the organization is vary according to some matters of the business.

# Overview to the dataset:

Table 1: About the dataset

|  |  |
| --- | --- |
| Dataset | IBM employees |
| Variables that remove from the dataset | * EmployeeCount * EmployeeNumber(Alternative for the Application ID) * Application.ID * Over18 * StandardHours * Employee.Source |
| Total number of records | 23436 |
| No of Missing Values | 133 |
| No of incompatible data | 79 |
| Total Removed Records | 212 |
| Remaining Records | 23224 |
| Training Set | 18580 |
| Test Set | 4644 |

A screenshot of a computer

Description automatically generatedA list of jobs with words

Description automatically generated

# Exploratory Data Analysis:

In this analysis, the main question to answer was why employees leave? The brief answer to that is employees tend to stay in the company until some force causes them to leave.

It can be divided into three main groups. One relevant factor is outside the company and two relevant factors are within the company. Outside the company, we must first consider employee’s **personal details** and how it influences attrition. Age, gender, marital status, education, and distance from home belong to personal details. Within the company, there are two factors, one is the **working environment** and the other is the income details of the employees. Employee’s satisfaction level (both job and work environment), job details, department, business travels, and their relationship with managers belong to the working environment. Employee’s monthly salary, percent salary hike, and stock option level fall under **income details**. Depending on these three main factors employee’s perceptions regarding outside job opportunities might change.

A graph with a blue and green rectangle

Description automatically generated

Figure 1:Response Variable (Customer Attrition)

Attrition is the response variable. It can be observed that 84% of the employees are currently employed, and only 16% of them have resigned. Many companies with low turnover rates assume that their employees are pleased with their jobs. It might not be the case. There is no perfect correlation between job dissatisfaction and the turnover rate. A low rate may be an effect of a tight job market or perhaps because of the financial benefits given by the company. So, it is important to separately analyze their environment satisfaction, job satisfaction and attrition rate. In itself, the fact that an employee stays on a payroll is meaningless; the company must also know why they stay. Carelessly conceived methods of maintaining a low turnover rate can be detrimental to the financial health of a company and the mental health of its employees.

**Personal life:**

Figure :Response Variable (Customer Attrition)

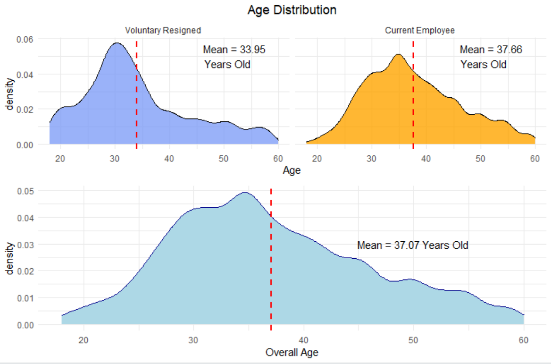


Figure 2:Age Distribution

According to the age distribution of the IBM employees, the majority of them are between the age of 30 and 40. Here it can be observed that the majority of the resigned employees are younger than the rest of the employees in the company. Employees who are between the age of 28-34 have a higher attrition rate than employees who are between the age of 34-40. Several studies have suggested that employees between 35-55 are relatively stable, focused, and loyal to the organization. Similarly, they have noted that younger employees (35 years and under) are more likely to move to new jobs than employees aged 35 to 55. The reason for this difference is younger employees want new challenges, promotions, and higher pay. On the other hand, older employees can access fewer openings and more likely to stay remain with their organization.

In Figure 3, Male employee percentage in the company is higher than the female percentage in the company. The attrition rate is approximately equal for both genders. It is a common thing in companies to have a disparity in working environment satisfaction between genders. It is more reasonable to compare environment satisfaction between departments for two genders.

Figure 3:Gender Distribution

A graph with blue and red dots

Description automatically generated

Figure 4:Department, Working Environment Satisfaction vs Gender.

It can be observed that females have scored lower than the males in the research and development department and human resources department. Females have scored higher in the sales department than males. The HR department has the lowest environment satisfaction scores for both males and females. The reason for the females to have a higher environment satisfaction in the sales department might be because in the sales department employees tend to work outside of the company more than they work inside.

A graph of different colored rectangles

Description automatically generated

Figure 5:Marital Status vs Attrition

According to the figure 5, single employees have a higher attrition rate than the married and divorced groups. When the marital status is compared with employees' age, it can be observed that as expected, single employees who are younger than the others are more likely to be in the resigned group. As it was discussed before the reason might be that they are more willing to take risks and look for new opportunities.

A graph of a bar chart

Description automatically generated

Figure 6:Education Field vs Attrition

According to figure 6, employees from the technical field have the highest turnover rate. This is a common issue in the tech field. Tech has the highest employee turnover of any business sector, with an attrition rate of 13.2%. The main reasons for the higher attrition rate in the tech space are lack of growth, intense competition, and company culture.

A graph of a person with blue and red squares

Description automatically generated

Figure :Education type vs Attrition

According to figure 7, the highest attrition rate can be observed among the employees with a lower educational level. This might be due to lack of growth in their career.

**Working Environment:**

A graph of a graph with a red and blue bar

Description automatically generated

Figure : Environmental Satisfaction vs Attrition

There is a negative relationship between attrition and environment satisfaction. Employees whose environmental satisfaction is low have a higher intent of voluntary resignation while the employees with very high environmental satisfaction have the lowest percentage of voluntary resignation. That means, in order to an employee to work comfortably working environment have to be good.

A graph with blue and red squares

Description automatically generated

Figure 9:Department vs Attrition

In Figure 9, Research and Development department and Human Resources department show a higher attrition rate. And here we can say that the high attrition in HR department is due to the high environment dissatisfaction.

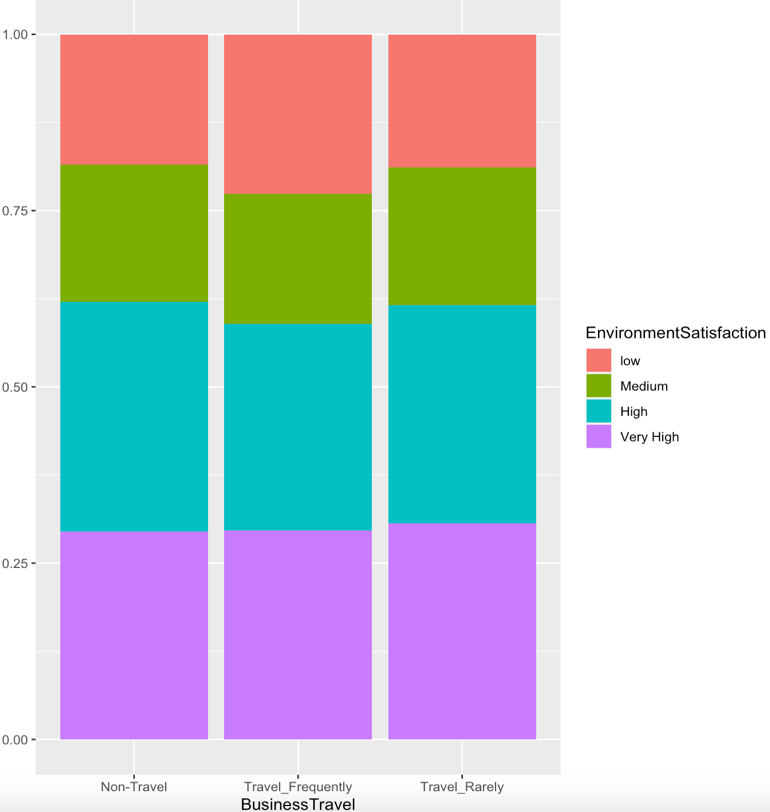


Figure :Business Travel vs Environmental Satisfaction

Most of the ones who go on business travels have a lower environment satisfaction than others. At the same time ones who do not travel frequently have a relatively higher satisfaction. This intends that ones who get a chance to participate on a business trip have a higher temp to give up the job since environmental satisfaction directly effects to the attrition.

**Income details:**

**A graph of a graph

Description automatically generated**

Figure 12:Percent Salary Hike

Figure 11:Monthly Income

Average Monthly Income value is around 6428 and majority of employees get less than 10000 salary per month. Any employee considers about their monthly income mainly. In CEO perspective he/she mainly consider about their profit and try to hold on their employees with giving them a good salary. Otherwise, employee tend to leave the company.

Most of the employees get less than 14% of salary hike. In a company, more low-level workers compared to no of high-level executives. Company cannot give higher salary increase for the lower-level workers as it needs more money. So, they give small salary increment for lower-level workers and higher salary increments for the manager board.

Table :Summary statistic values

|  |  |  |
| --- | --- | --- |
| **Variable** | **Attrition(Based on mean value)** | |
| **Volontary Resign** | **Current Employees** |
| Monthly Income | 5530.750 | 6594.729 |
| Hourly Rate | 66.41105 | 65.73101 |
| Daily Rate | 749.1994 | 811.3137 |
| Monthly Rate | 14192.52 | 14298.02 |

There is not much more influence from the variables, Monthly rate, and Hourly rate to the attrition rate of the employees. Here Monthly Rate, Daily Rate and hourly rate are depending on no of working years and Monthly Income of the employee. But however, average monthly income, average hourly rate, average daily rate, and average monthly rate values of current employees are greater than the voluntary resign employees.

A graph of different colored bars

Description automatically generated

Figure :Education, Monthly Income vs Attrition

Figure 13 shows the relationship between Attrition, Monthly income, and the education level. Current employees get more monthly income than the resign employees. But when compare with doctorate current employees with doctorate resign employees, current employees get less salary than resign employees. Salary variation in resign employees is lower than the current employee’s variation as well. Almost all the resign doctorate employees get less than 12000 monthly income value. But some of the other lower education level employees get higher monthly income than them. So, they left the company because of this reason.

A screenshot of a graph

Description automatically generated

Figure :Department, Monthly Income vs Attrition

According to the figure 14, Sales department employees get the highest mean monthly income value, and this result is same for both current employees and the Attrited employees. But salary variation is high in Human Resource department than the other departments. In sales and R & D department resign employees get higher salaries than the HR department resign employees. This happens because in sales and R&D departments employees work in the field. That can be harder than what they work in an office. So, the monthly income for those two departments may not be sufficient for them.

A graph of different colored squares

Description automatically generated

Figure :Monthly Income, Stock option level vs Attrition

Current employees get much more variation in monthly income than the resign employees when consider their stock option levels. In figure 4.3.6 result that most of resign employees are with level 0. Stock option level 0 means they have no stocks in the company. So, they cannot get extra money as well. That implies their monthly income is also lower than the others.

# Modelling Approach

A screenshot of a graph

Description automatically generated

Figure 16:Correlogram

According to the correlation plot it can be observed that strong correlation exists between years since last promotion, years in current role, years at company, total working years and years with current manager. strong correlation also exists between monthly income and total working years. Due to multicollinearity in the predictor variables, it is not accurate to fit non regularized regression models to the dataset. Ridge regression, lasso regression and elastic-net regression models were fitted as a remedy for multicollinearity.

**Statistical Models:**

Also, this dataset is an **imbalanced dataset**. 84% of the employees are currently employed and only 16% have voluntary resigned. Due to this there is higher chance for the models to be biased towards current employee group. First ridge regression, lasso regression and elastic-net regression were fitted without performing any remedial procedure. **Voluntary resigned group** was considered as the **positive class** (indicated by **0**). The regularization models were fitted using **caret** package.

Table 3:Initial Model Performance

|  |  |  |  |
| --- | --- | --- | --- |
| **Measures** | **Ridge** | **Lasso** | **Elastic-net** |
| Accuracy | 0.8449 | 0.8447 | 0.8411 |
| Sensitivity | 0.08163 | 0.07891 | 0.09388 |
| Specificity | 0.98849 | 0.98874 | 0.98158 |

As expected, **sensitivity** of the three models is **very low**. It implies that these models are less likely to identify employees that are going to resign. This places a huge disadvantage for the company. Main objective of the company is to identify the employees that are going to resign. So, it is important to improve the sensitivity of our models. As a remedy for this problem oversampling, under sampling, synthetic sampling techniques were used.

# Performance Evaluation

Table 4:Model Performance with oversampling

|  |  |  |  |
| --- | --- | --- | --- |
| **Measures (Oversampling)** | **Ridge** | **Lasso** | **Elastic-net** |
| Accuracy | 0.6759 | 0.6806 | 0.6804 |
| Sensitivity | 0.6735 | 0.6639 | 0.6626 |
| Specificity | 0.6763 | 0.6837 | 0.6837 |

Table 5:Model Performance with Under sampling

|  |  |  |  |
| --- | --- | --- | --- |
| **Measures**  **(Under sampling)** | **Ridge** | **Lasso** | **Elastic-net** |
| Accuracy | 0.6244 | **0.6397** | 0.6196 |
| Sensitivity | 0.7279 | **0.7184** | 0.7401 |
| Specificity | 0.6049 | **0.6249** | 0.5970 |

Table :Model Performance with Synthetic sampling

|  |  |  |  |
| --- | --- | --- | --- |
| **Measures (Synthetic)** | **Ridge** | **Lasso** | **Elastic-net** |
| Accuracy | 0.6621 | 0.6720 | 0.6703 |
| Sensitivity | 0.6803 | 0.6585 | 0.6667 |
| Specificity | 0.6586 | 0.6745 | 0.6709 |

According to the above tables, accuracy, sensitivity, and specificity levels are approximately equal for all remedial techniques. It is more appropriate to select the model with the lowest number of predictor variables to increase interpretability. The **lasso model** obtained from **under-sampling** has the lowest number of predictor variables among all the models. Therefore, the lasso model was the best among them.

**Machine learning models:**

When developing a data product, prioritizing the accuracy of the results takes precedence over model interpretability. As a result, non-interpretable models, including decision trees, random forests, and extreme gradient boosting (XGBoost) were employed to fit the dataset. The same notations were retained as in the previous approach.

Table : ML model comparison (Without any sampling technique)

|  |  |  |  |
| --- | --- | --- | --- |
| **Model (No sampling techniques)** | **Metrics (As a percentage)** | | |
| Sensitivity | Specificity | Overall accuracy |
| Decision Tree | 29.54 | 98.31 | 8732 |
| Random Forest | 72.79 | 100 | 95.69 |
| XGBoost | 31.70 | 98.51 | 87.94 |

Table :ML model comparison (With over sampling technique)

|  |  |  |  |
| --- | --- | --- | --- |
| **Model (Over sampling techniques)** | **Metrics (As a percentage)** | | |
| Sensitivity | Specificity | Overall accuracy |
| Decision Tree | 60.95 | 75.02 | 72.8 |
| Random Forest | 64.90 | 100 | 94.44 |
| XGBoost | 80.27 | 71.16 | 72.60 |

Table :ML model comparison (With under sampling technique)

|  |  |  |  |
| --- | --- | --- | --- |
| **Model (Under sampling techniques)** | **Metrics (As a percentage)** | | |
| Sensitivity | Specificity | Overall accuracy |
| Decision Tree | 75.10 | 64.53 | 66.21 |
| Random Forest | 83.95 | 99.03 | 96.64 |
| XGBoost | 83.27 | 72.62 | 74.31 |

Upon evaluating all the employed statistical and machine learning models, it is evident that the Random Forest model with under-sampling technique outperforms the others. It demonstrates commendable overall accuracy, sensitivity, and specificity values. Let’s proceed to analyze performance metrics for this model.

**Variable Importance plot for the best model:**

**A graph of a graph of different types of numbers

Description automatically generated with medium confidence**

Figure 17: Variable Importance Plot - Random Forest Model with under sampling technique

According to this figure, the most important variables that determines employee attrition are age, daily rate, monthly rate, distance from home, job role, hourly rate, and education. Number of companies worked; gender have the lowest impact.

**ROC curve for the best model:**

A graph with a blue line

Description automatically generated

Figure 18: ROC curve for the Random forest model with under sampling technique

The ROC curve serves as a visual representation of a binary classifier's performance, effectively illustrating the trade-off between sensitivity and specificity. In this case, the classifier is focused on distinguishing between Attrited and existing employees, as the response variable is employee attrition. An AUC (Area Under the Curve) value of 0.91 has been obtained, which indicates a substantial level of discrimination power. This high AUC value implies that the classifier has a robust ability to differentiate between the two classes, making it proficient in predicting customer attrition. Essentially, the model's performance is excellent, and the higher the AUC value, the better the classifier's overall accuracy. With such a remarkably high AUC value, we can be highly confident in the classifier's capacity to precisely classify employee attrition.

# Issues Encountered and Proposed solutions.

This was the first issue that needed a solution was data imbalance. When the dataset is imbalanced, models tend to be biased towards the majority class. Algorithms try to minimize the overall error to which the minority class contributes only a little. That will make the accuracy of the minority class to be reduced significantly. As a remedy oversampling, under-sampling and synthetic sampling techniques in ROSE package were used. In under-sampling it removes the observations from the majority class randomly, in oversampling it replicates the observations from the minority class to make the dataset balanced and in synthetic sampling it overcome imbalances by generating artificial data. Major disadvantage of the under-sampling technique is that it can remove important details from the majority class, and it reduces the accuracy of that class. Due to this positive predicted value of the lasso model was very low. This means that among the predicted employees in resigned group only 30% were observed as resigned. So, to gain more accuracy in both classes random forest techniques were used.

# Conclusion

Model obtained from regularized regression techniques was the most interpretable model. Model obtained using random forest techniques was the most accurate model. For a company it is more important to identify the employees accurately. Due to that when building the data product random forest model was used as the final model.

# References

# 

<https://en.wikipedia.org/wiki/IBM>

<https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/>

<https://machinelearningmastery.com/data-sampling-methods-for-imbalanced-classification/>

<https://machinelearningmastery.com/types-of-classification-in-machine-learning/>