Regression Analysis Project

Subject Code:MTH416A

Modelling the Probability of Attrition in a Company Using Logistic Regression

Submitted by
Sampriti Dutta-211366
Bidhisha Ghosh-211423
Shreya Karmakar-211382

under supervision of
Dr.Sharmishtha Mitra

Department:STATISTICS
IIT KANPUR

Acknowledgement

We would like to express a deep sense of thanks and gratitude to Dr. Sharmishtha Mitra for providing us the opportunity to prepare this project and constantly motivating us with constructive advices. It has been a great learning experience building practical insights of the theoretical knowledge gathered during course lectures.

Last, but not the least, our parents provided us with continuous encouragement and extensive support throughout the session. So , with due regards we express our gratitude to them for completion of the project within the stipulated time-period.

ABSTRACT

Attrition (employees leaving, either on their own or because they got fired) is harmful for any company as it causes delay of ongoing projects, requires new employee recruitment and often involves training the fresh talent. This project targets to understand what factors a company should focus on, in order to curb attrition, using logistic regression model. At first Here we deal with the missing values and adopted Mean Imputation Technique to impute the missing values of the dataset.To address the data imbalance issue, we apply Synthetic Minority Oversampling Technique (SMOTE) to oversample the minority class labels and after that we perform Exploratory Data Analysis.Then we check for Variance Inflation Factor (VIF) in order to deal with the multicollinearity problem,and then in order to select the important variables we adopted Lasso Regression.Then we bulid an appropriate Logistic Regression Model on our cleaned dataset as well as on the original dataset and finally analyze and evaluate the performance of the two models using several metrices such as accuracy,Precision,Recall etc.The results thus obtained will be used to understand what changes a company should make to their workplace, in order to get most of their employees to stay.

Contents

1 Introduction

Attrition is the departure of employees from the organization for any reason (voluntary or involuntary), including resignation, termination, death or retirement.There are mainly two types of attrition i.e Voluntary attrition ,when a employee chooses to leave the company another one is involuntary attrition ,when a company decides to part ways from the employee.Generally attrition can happen for several reasons, including pay, lack of growth and progression,declining work environment,lack of feedback and recognition,and being overworked etc.Here in our project work we are concerned about voluntary attrition of employees.The direct impacts of attrition are delay of ongoing projects,investment of money for training of fresh talents etc.But during the recruitment process an increase of workload can happen in the case of other employees which results in overtime cost and also hampers their own productivity also.So, as a part of HR Analytics team of a firm we would like to find out the principle factors behind employee attrition in that particular company.Here our main object is to find the probability of employee attrition in a company considering the different factors that will have impact on that employees professional work life.The conclusive part of the project work will be focused on different ways of curbing attrition .The results thus obtained will be used to understand that what changes a company should make in order to make most of their employees to stay.

1.1 Objectives Of The Study:

In this project we want to predict the probability of attrition in a company.The key steps involved are as follows–

- 1.Dealing with the problem of Missing Values using suitable Data Imputation Techniques.
- 2.Performing Exploratory Data Analysis to extract the main features of the data.
- 3.Modeling the probability of attrition using logistic regression.
- 4.Comparing the accuracy of the model.

2 Methodology:

In the previous section we have briefly explained about the problem statement of employee attrition in a company.In this section,we will provide step by step procedure of curbing attrition in a company .At first we introduce the In_time data set and Out_time data set containing the different factors involved in employee attrition.Then we focused on preprocessing the data which involves the most important step i.e dealing with missing values of the data and adopting a method for imputing them in the data if they appears in large number ,otherwise for the sake of simplicity we can also drop the missing values if the number is small enough.Next we build a logistic model on our data set and explain how we train our data using this model.Later ,we analyzed the performance of the logistic model using certain metrics like accuracy,precision and recall.

The data set is about predicting the probability of attrition of an employee in a particular company.This data is collected from www.kaggle.com.There are total of 5 data sets I)Employee Survey Data,II)General Data on Employees III)In_time data of employees IV)Out_time data of employees V)Manager survey data on employees.This data is collected from an firm over the year 2015.As this data set is based on a recent years data

so it will be very helpful in predicting the probability of attrition during the recent time period of that firm.The 5 different kind of data sets are explained below–

2.1 Dataset Description:

1. **Employee Survey Data**–Provides information on an approximate overview of the company's environment,jobsatisfaction of the employee and work life balance of that particular employee.

2. **General Data On employees**–Provides information about the employees educational,professional,marital status etc.

3. **In_Time Data Of Employees**–Gives detailed information about when an employee enter in that firm every working day.

4. **Out_Time Data Of Employees**–Gives specific timing when a employee leaves the firm after working hour in a specific working day.

5. **Manager Survey Data On Employees**–Provides information about an employees performance and involvement in a specific job.

   The data set in summarized below in the table 1:

Variable name	Description
X_1	Age of the employee
X_2	How frequently the employee travelled for business purpose in the last year.
X_3	Department of the company.
X_4	Distance from home in kilometers.
X_5	Education level.
X_6	Field of education.
X_7	Employee count.
X_8	Employee ID.
X_9	Work environment satisfaction level.
X_{10}	Gender of the Employee.
X_{11}	Job level at company on a scale of 1 to 5.
X_{12}	Name of a job role in company.
X_{13}	Job satisfaction level.
X_{14}	Marital status of the employee.
X_{15}	Monthly income in rupees per month.
X_{16}	Total number of companies the employee has worked for.
X_{17}	Whether the employees is above 18 years of age or not.
X_{18}	Percentage of salary hike for last year.
X_{19}	Performance satisfaction for last year.
X_{20}	Relationship satisfaction level.
X_{21}	Standerd hours of work for the employee.
X_{22}	Stock option level of the employee.
X_{23}	Total no. of years the employee has worked so far.
X_{24}	Number of times training was conducted for this employee.
X_{25}	Worklife balance level.
X_{26}	Total number of years spent at the company by the employee.
X_{27}	Number of years since last promotion.
X_{28}	Number of years under current manager.

Table 1:Description of Explanatory Variables

As shown in the Table 1 there are total 29 features numbered X_1 to X_{29} .15 of them are continuous variable and 14 of them are categorical variable.Here we are using only the data set of the year 2015.

2.2 Dataset Quality Assessment

Before moving on to assessing the quality of the dataset, we first standardized our dataset as we can see that our dataset have large differences between their ranges which can possibly cause a lot of trouble to build a model. So, to prevent this problem, transforming features to comparable scales using standardization is the solution.

2.2.1 Missing Data

First, we look at some statistics of missing values. We count the number of values of each regressor variable. We notice that the feature X26 , i.e. Work life balance level has the highest number of missing values.

We have visually seen the sparsity in the data. Now, let us see how much of data is actually missing. In Table 1 shown below, the second column shows the total number of instances in each dataset, and third column shows the number of instances or rows with missing values for at least one of the features. A naive approach of dealing with missing values would be to drop all such rows as in Listwise deletion. But dropping all such rows leads to a tremendous data loss. Column 4 shows the number of instances that would remain in each dataset if all rows with missing values were dropped. Column 5 shows the percent of data loss if all the rows with missing data values were indeed dropped. As the data loss in most of the datasets is over 50%, it is now clear that we cannot simply drop the rows with missing values, as it leads to severe loss in the representativeness of data.

Explanatory Variables	**Total no. of Instances**	**Instances with missing Values**	**Instances that would remain if all rows with missing values were dropped**	**Data loss if rows with missing values were dropped**
Environment Satisfaction	4410	25	4385	0.56
Job Satisfaction	4410	20	4390	0.45
Work Life balance	4410	38	4372	0.86
NumCompanies Worked	4410	19	4391	0.43
Total Working Years	4410	9	4401	0.20

Table 2:Sparcity matrix of the data set

2.2.2 Dealing With Missing Values:

The major problems created by missing data is mentioned below–

1.The absence of data reduces the statistical power,as a result the efficiency get reduced.

2.The lost data introduces substantial amount of bias in the model.

3.It reduces representativeness of the sample observations.

Dropping all the missing observations is not a feasible choice at all, as it may introduce bias in the model and can eventually lead to invalid results.So,here we need to impute missing data using suitable imputing techniques.

Imputation is a technique used for replacing the missing data with some substitute value to retain most of the data information of the data set.There are several method of data imputation like Regression Imputation,Paiwise Deletion,K- Nearest Neighbours,Mean Imputation etc.Here we have adopted the Mean Imputation technique to deal with the missing data problem–

Mean Imputation Technique:

Mean Imputation is a process that replaces missing values of a certain variable by the mean of the non missing cases of the variable.Mean imputation attenuates any correlations involving the variables that are imputed. This is because, in cases with imputation, there is guaranteed to be no relationship between the imputed variable and any other measured variables. Thus, mean imputation has some attractive properties for univariate analysis but becomes problematic for multivariate analysis.In our data set we have missing values for 4 features and we replaced the missing value of that feature by the mean of the other non-missing value of that particular feature.

2.3 Data Imbalance:

Imbalanced data refers to those types of datasets where the target class has an uneven distribution of observations, i.e one class label has a very high number of observations and the other has a very low number of observations.Here one of the shortcomings of our dataset is it is highly imbalanced so, we need to deal with the problem of data imbalance.

Data imbalance can be treated using different techniques such as Random undersampling,Random oversampling,cluster based over sampling,Synthetic Minority Over-sampling Technique for imbalanced data (SMOTE),Modified synthetic minority oversampling technique (MSMOTE) etc.Oversampling and Undersampling are opposite and roughly equivalent techniques of dealing with Data Imbalance, where they adjust the class distribution of a dataset.The process of Oversampling increases the class distribution of the minority class label whereas Undersampling decreases the class distribution of the majority class label. In our project, we have used Synthetic Minority Oversampling Technique or SMOTE to deal with the data imbalance problem.

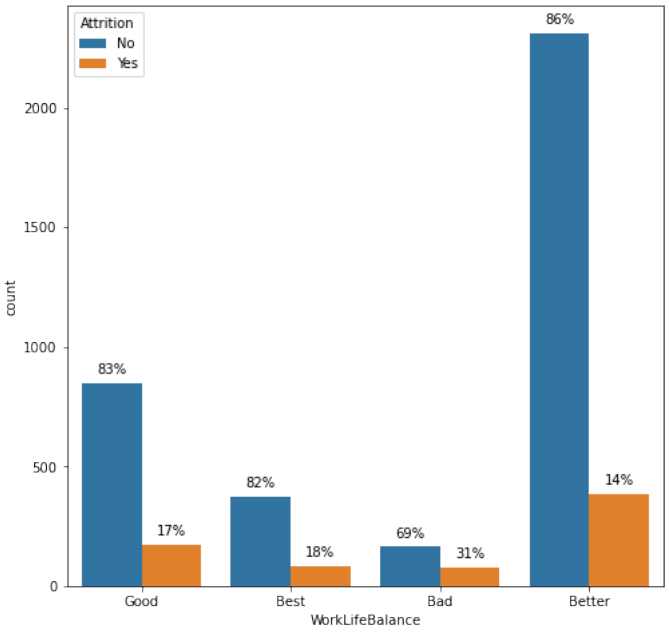
Synthetic Minority Over-sampling Technique for imbalanced data (SMOTE):

One of the widely used oversampling technique is 'SMOTE'.To illustrate how this technique works consider some training data which has s samples, and f features in the feature space of the data. For simplicity, assume the features are continuous. As an example, let us consider a dataset of birds for clarity. The feature space for the minority class for which we want to oversample could be beak length, wingspan, and weight. To oversample,take a sample from the dataset, and consider its k nearest neighbors in the feature space. To create a synthetic data point, take the vector between one of those k neighbors, and the current data point. Multiply this vector by a random number x which lies between 0, and 1. Adding this to the current data point will create the new synthetic data point. SMOTE was implemented from the imbalanced-learn library.

3 Exploratory Data Analysis:

Before proceeding to the theoretical analysis,we perform **Exploratory Data Analysis** to analyse the main characteristics of the data visually.

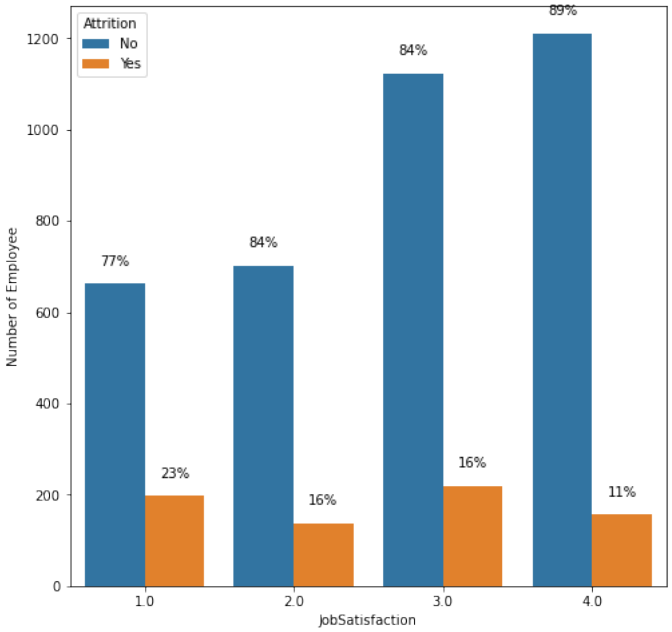
Exploratory Data Analysis is an approach of analyzing data sets that employs a variety of technique to :
I)maximize insight into a data set, II)uncover underlying structure III)extract important variables, IV)detect outliers and anomalies, V)develop parsimonious models.
while exploring the explanatory variables of the model through EDA and taking clue from the pictorial analysis,we can make a guess of lineup of our analysis other than the formal modelling.
Since, our data set is consisting of both the continous(Numeric) and Categorical variable,first ,we explore continuous variables graphically.



m

Fig.1:Bar Plot of WorkLifeBalance versus Count

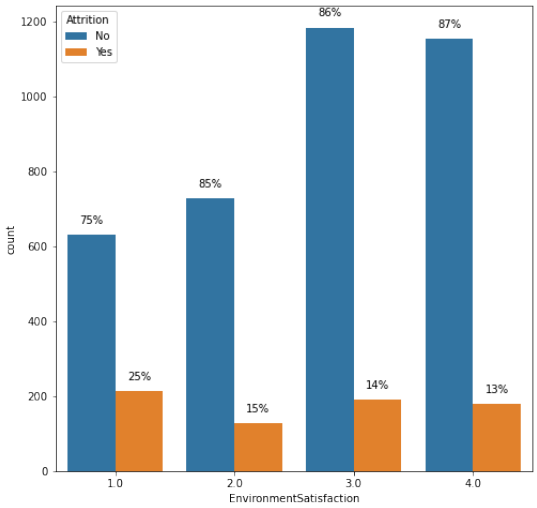
Insights:
Total employees left in the previous year is 701 in number.Among them who believe that WorkLifeBalance is Better in the company were 52% and who believes that WorkLifeBalance is Good in company were 25% of the total population.



m

Fig.2:Bar Plot of JobSatisfaction Versus Number Of Employee

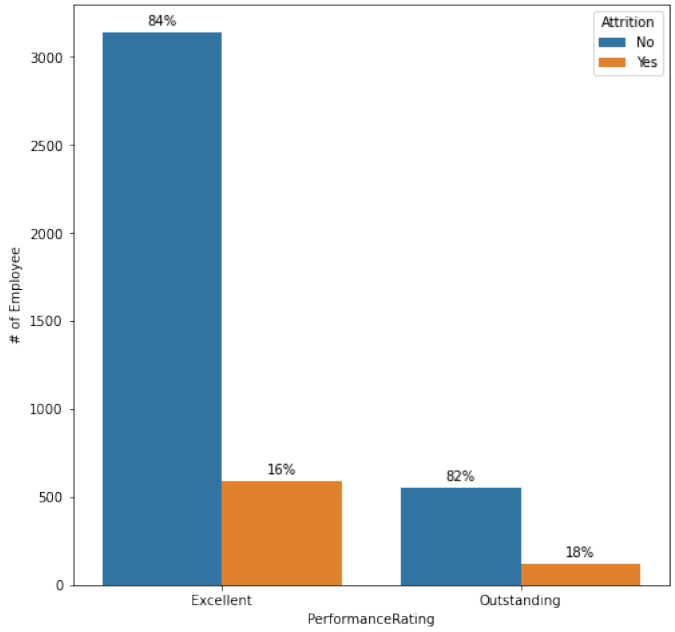
Insights: Among the total employees who left in the previous year 29% of them was having a low job satisfaction level and 30% of them was having a high job satisfaction level.



m

Fig.3:Bar Plot of EnvironmentSatisfaction versus Count

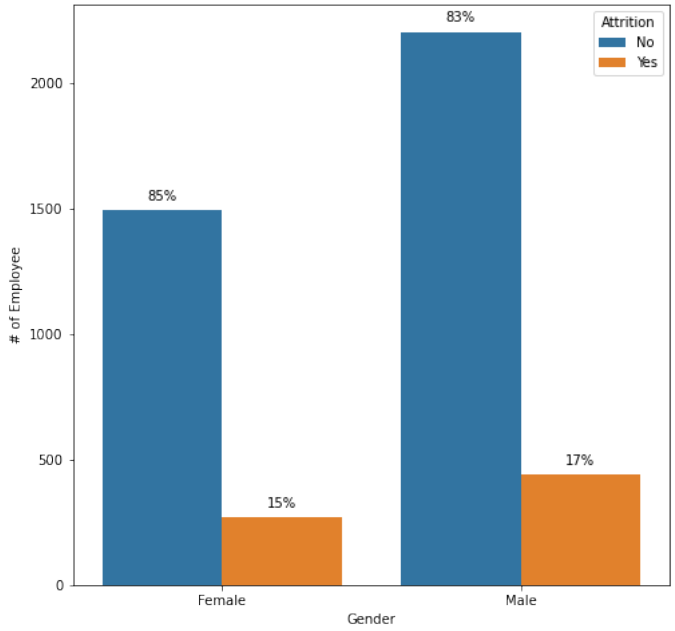
Insights Among the total employees who left in the previous year 30% of them was not satisfied about the job environment and 35% of them was highly satisfied regarding the job environment.



m

Fig.4:Bar Plot of PerformanceRating versus # of Employee

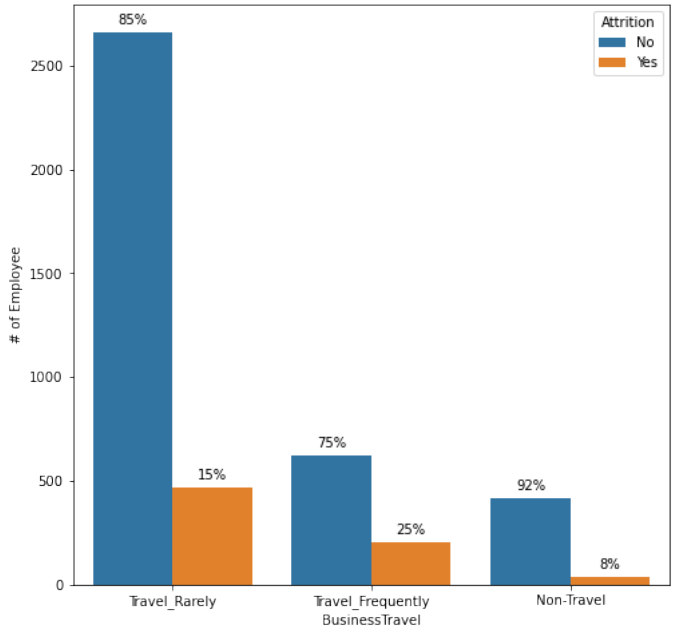
Insights:
People who left in the previous year & Performance rating was Excellent in the company were 83% of population who left in the previous year.



m

Fig.5:Bar Plot of Gender versus #of Employee

Insights: Among the total employees who left the company 15% of them were female emoployee and 17% of them were male employee.



m

Fig.6:Bar Plot of BusinessTravel Versus #of Employee

Insights: Employees who left in the previous year 5% of them had not participated in the business travel and 29% of them participated in the business travel frequently.

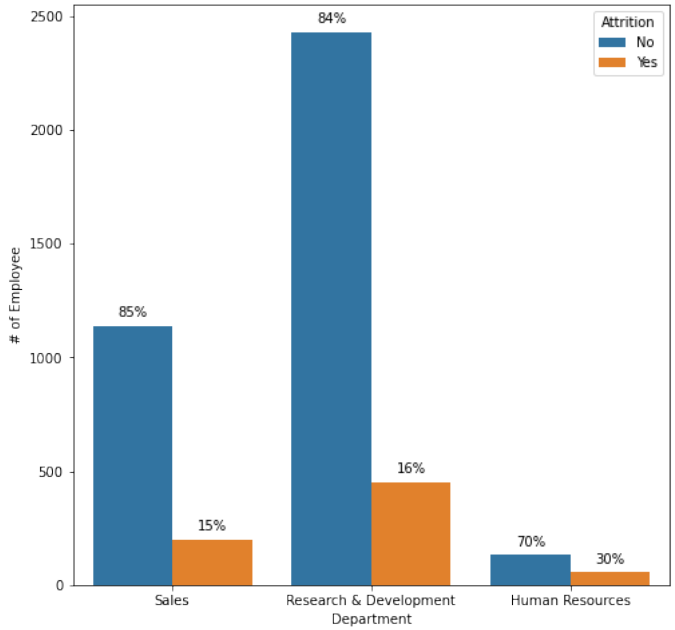


Fig.7:Bar Plot of Department versus #of Employee

Insights: Attrition rate of employees working in the Human & Resources is lower among all the other departments.

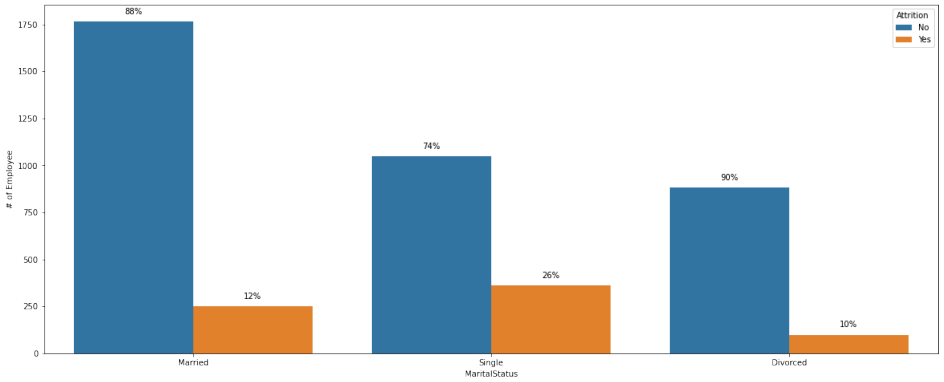


Fig.8:Bar plot of MaritalStatus Versus #of Employee

Insights: Among the employees who left the company in previous year 52% of them were single and 34% of them were married and 14% of them were divorced.

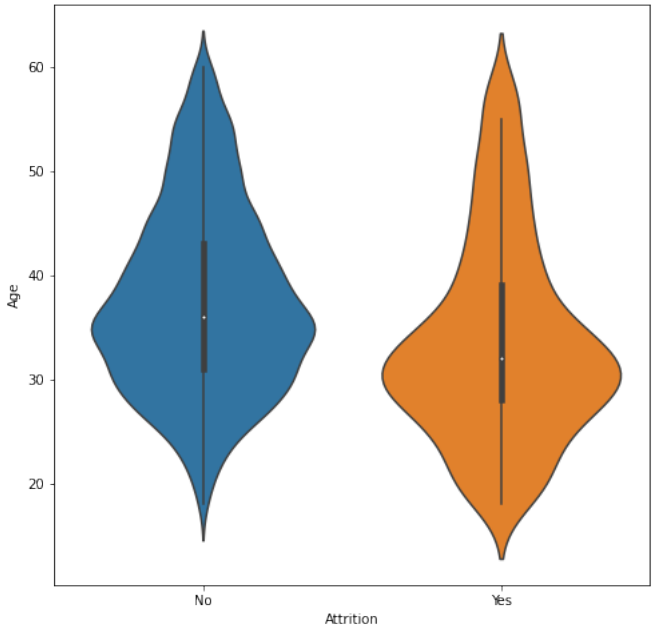


Fig.9:Violin Plot of Attrition Versus Age

Insights: The probability of non attrition is high between the age group 30-40,whereas the probability of attrition is high among some lesser ages.

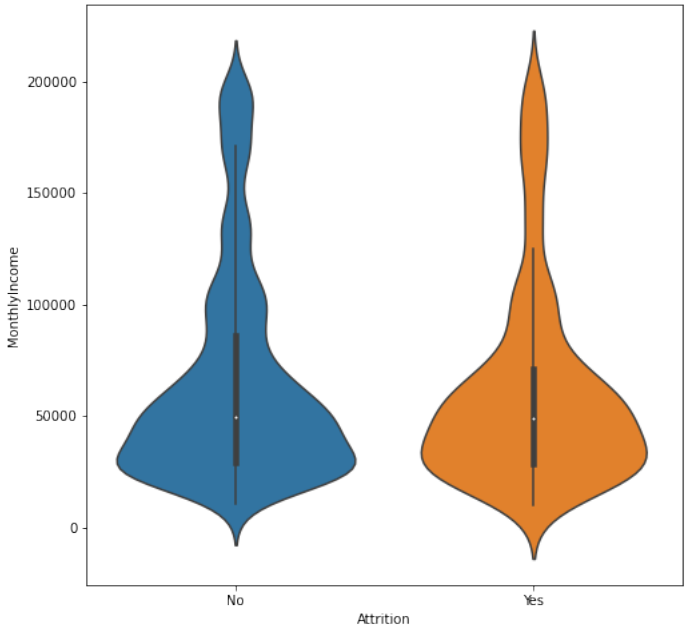


Fig.10:Violin Plot of Attrition versus MonthlyIncome

Insights: The probability of attrition is higher among the people with lower income than that of higher income.

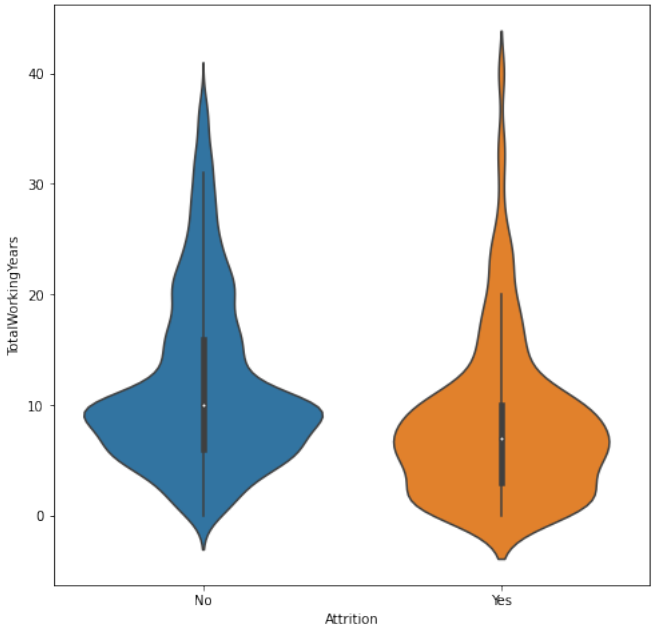


Fig.11:Violin Plot of Attrition versus totalWorkingYears

Insights: The violin plot indicates that lesser the number of total working years, higher the chance of attrition.

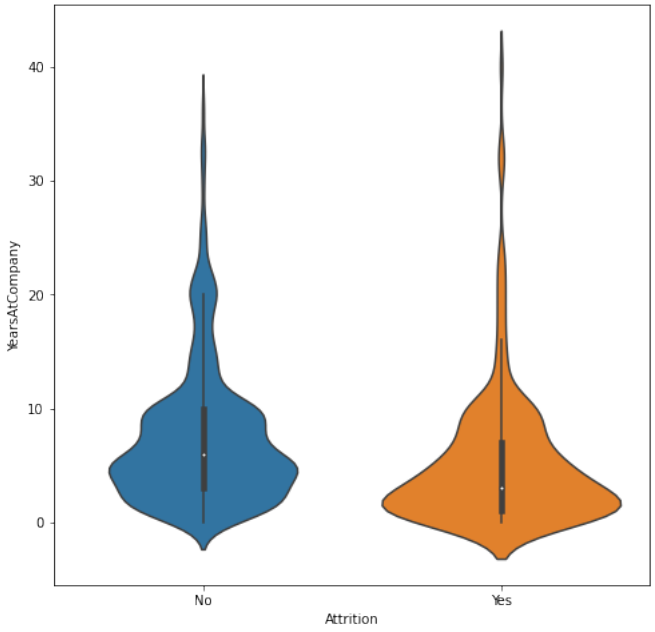


Fig.12:Violin Plot of Attrition versus YearsAtCompany

Insights: The violin plot indicates that the employee with lesser number of years at company has higher chance of attrtion.

Data Visualization Using Pair Plot:

Pair Plots are a really simple way to visualize relationships between each variable where the variables can be continuous or categorical.It produces a matrix of relationships between each variable for an instant visual inspection of our data.It can also be a great jumping off point for determining types of regression analysis to use.Here the below pair plot shows the extent of relationship between the response variable with each of the explanatory variable.

Fig.13:Pair Plot of the dataset

4 Train And Test Data:

To fit a model on an data the procedure involves taking that dataset and dividing it into two subsets.
1.Training set:a subset to train a model.
2.Test Set:a subset to test the trained model.Performance of the model is measured using the test data
We have started with 4410 observations,out of those 4410 observations we have sample 80% observation for the training set i.e we have total 3528 observations in the training set.
First we fit the logistic model on the training dataset and obtained the following result–

	coef	std err	z	P>|z|	[0.025	0.975]
const	2.4782	0.111	22.252	0.000	2.260	2.696
hrs	0.6321	0.035	18.250	0.000	0.564	0.700
Age	-0.4134	0.047	-8.844	0.000	-0.505	-0.322
NumCompaniesWorked	0.2300	0.036	6.406	0.000	0.160	0.300
StockOptionLevel	-0.1761	0.034	-5.208	0.000	-0.242	-0.110
TotalWorkingYears	-0.3458	0.058	-6.010	0.000	-0.459	-0.233
TrainingTimesLastYear	-0.2837	0.034	-8.345	0.000	-0.350	-0.217
YearsSinceLastPromotion	0.5206	0.044	11.943	0.000	0.435	0.606
YearsWithCurrManager	-0.6057	0.048	-12.591	0.000	-0.700	-0.511
JobInvolvement_Low	-0.1909	0.155	-1.230	0.219	-0.495	0.113
JobInvolvement_Medium	-0.2453	0.081	-3.015	0.003	-0.405	-0.086
JobInvolvement_Very High	-0.1562	0.122	-1.280	0.201	-0.396	0.083
EnvironmentSatisfaction_Low	0.4359	0.088	4.952	0.000	0.263	0.608
EnvironmentSatisfaction_Very High	-0.8644	0.084	-10.287	0.000	-1.029	-0.700
JobSatisfaction_Low	-0.0440	0.086	-0.510	0.610	-0.213	0.125
JobSatisfaction_Very High	-1.1652	0.086	-13.482	0.000	-1.335	-0.996
WorkLifeBalance_Best	-2.4500	0.153	-15.998	0.000	-2.750	-2.150
WorkLifeBalance_Better	-2.4805	0.115	-21.488	0.000	-2.707	-2.254
WorkLifeBalance_Good	-2.4740	0.130	-19.088	0.000	-2.728	-2.220

Table.3:Summary of fitting Logistic model on the training dataset

5 Multicollinearity:

Next we proceed to another important step of Model Accuracy Checking that is checking the presense of multicollinearity in our model.A basic assumption is multiple linear regression model is that the rank of the matrix of observations on explanatory variables is the same as the number of explanatory variables. In other words, such a matrix is of full column rank. **Multicollinearity** refers to a situation in which more than two explanatory variables in a multiple regression model are highly linearly related. There can be more than one reason behind multicollinearity, such as:

1. The data collection method employed

2.Model specification using too many regressors

3.An over-defined model etc.

As a consequence the design matrix becomes ill-conditioned producing regression coefficients with large standard errors which can potentially damage the prediction capability of the model. In order to deal with multicollinearity problem of our data, we have plotted **Correlation Heat Map** for the Mean Imputed Dataset.In the Correlation Heat Map darker shade implies higher correlation among the variables whereas the lighter shade indictes the less correlation.

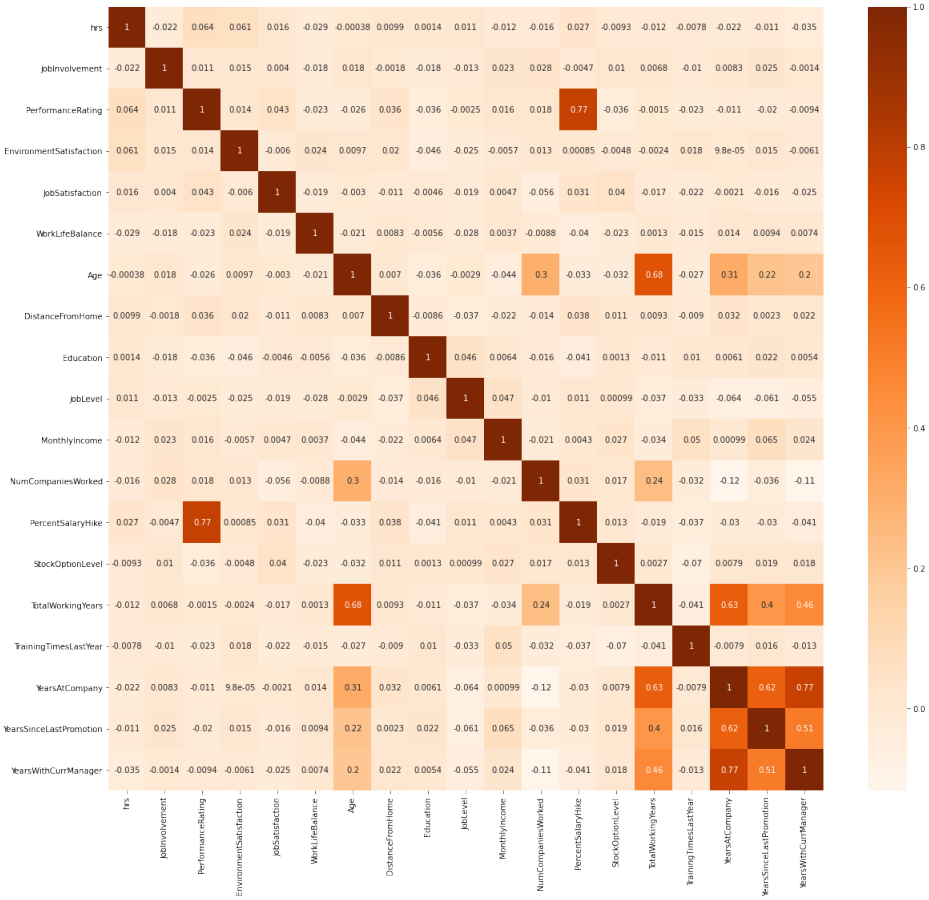


Fig.14:Correlation Heatmap for Mean Imputed dataset

Dealing With Multicollinearity:

There are several diagonostics measure for multicollinearity and each of them are based on a particular approach.Here we have used **Variance Inflation Method** to deal with the multicollinearity problem.

VIF or Variance Inflation Factor for the j-th explanatory variable is defined as:

$VIF_j = \frac{1}{1-R_j^2}$

R_j denotes the coefficient of determination obtained whenX_j is regressed on the remaining (k-1) variables excluding X_J.

In practice, usually, if R_j^2 >o.8 or VIF $>$ 5 indicates that mulicollinearity leading to the poor estimates of the regression coefficients Hence, in order to deal with it,we have adopted an iterative algorithm that drops variable with highest VIF and then checks

VIF again and then drop until VIF of all variables is less than 5.Here the below table shows the VIF values of each of the important explanatory variables–

Index	feature	VIF
1	TotalWorkingYears	2.68
2	Age	2.09
3	WorkLifeBalance_Better	2.06
4	YearsWithCurrManager	1.79
5	YearSinceLastPromotion	1.52
6	JobSatisfaction_Very High	1.43
7	WorkLifeBalance_Good	1.42
8	EnvironmentSatisfaction_Very High	1.42
9	JobInvolvement_Medium	1.35
10	JobSatisfaction_Low	1.31
11	EnvironmentSatisfaction_Low	1.28
12	WorkLifeBalance_Best	1.22
13	NumcompaniesWorked	1.19
14	JobInvolvement_Very High	1.14
15	JobInvolvement_Low	1.10

Table.4

From The above table it is clear that our data is not affected by multicollinearity ,so there is no need to proceed any further step to remove the multicollinearity issue.

6 Variable Selection:

The complete regression analysis depends on the explanatory variables present in the model.Variable selection is intended to select the "best" subset of regressors from the pool of regressors.Unnecessary explanatory variables will add noise to the estimation of other quantities that we are interested in and also the efficiency of the model will also be compromised.Now it is understood that in the regression analysis that only correct and important explanatory variables appear in the model.Generally, all such candidate variables are not used in the regression modelling, but a subset of explanatory variables is chosen from this pool.In our project, we have used Lasso Regression to perform the variable selection method.

Lasso Regression:
Lasso regression is a type of Regularization that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection,like variable selection/parameter elimination. The acronym "LASSO" stands for Least Absolute Shrinkage and Selection Operator.

Lasso regression performs L1 regularization technique, which adds a penalty equal to the absolute value of the magnitude of coefficients. Larger penalties result in coefficient values closer to zero, which is the ideal for producing simpler models. On the other hand, L2 regularization (e.g. Ridge regression) doesn't result in elimination of

coefficients or sparse models. This makes the Lasso far easier to interpret than the Ridge.
List of existing variables in the model after performing variable selection using Lasso Regression:

index	Variable Name
1	'hrs'
2	'Age'
3	'NumCompaniesWorked'
4	'StockOptionLevel'
5	'TotalWorkingYears'
6	'TrainingTimesLastYear'
7	'YearsSinceLastPromotion'
8	'YearsWithCurrManager'
9	'JobInvolvement_Low'
10	'JobInvolvement_Medium'
11	'JobInvolvement_Very High'
12	'EnvironmentSatisfaction_Low'
13	'EnvironmentSatisfaction_Very High'
14	'JobSatisfaction_Low'
15	'JobSatisfaction_Very High'
16	'WorkLifeBalance_Best'
17	'WorkLifeBalance_Better'
18	'WorkLifeBalance_Good

Table.5

7 Model:Logistic Regression

After cleaning and preprocessing our data set,now we are all set to define our model.Here our response variable **'Y'** is a categorical variable with two nominal categories.If **Y** gives the value **0** it indicates that there is no attrition in the company,and**1** indicates that there is attrition in the company.Logistic regression makes use of the canonical link function $\ln(\frac{p}{(1-p)})$.
The logistic regression model is given as:

$$Y_i \sim Binomial(\mathrm{N}_i, p_i)$$
$$\ln(\tfrac{p_i}{1-p_i}) = \beta_0 + \textstyle\sum_{j=1}^{k} \beta_j x_{i_j} = \mathrm{X}\beta$$
$$\text{for i=1,2....n}$$

Where x_{i_j} is the i'th row and j'th column of the design matrix **X**.In order to check the accuracy of the logistic model,we use a confusion matrix here.

7.1 Confusion Matrix:

Confusion matrix is a technique for summarizing the performance of the model and as the name suggests it gives us a matrix as a output.It gives information about errors made by the classifier and the types of errors that are being made.It reflects how a classification model is disorganized and confused while making predictions.

Confusion matrix is a very popular measure used while solving classification problems.It visualizes and summarizes the performance of a classification algorithm.In general classification accuracy fails on classification problems with a skewed class distribution because of the intuitions developed by practitioners on data sets with an equal class distribution.
The 4 important terms are mentioned below–
1.**True Positives(TP)**:In this case the model correctly predicts the positive class.

2.**True Negatives(TN)**:In this case the model correctly predicts the negative class.

3.**False Positives(FP)**:In this case the model incorrectly predicts the positive class.

4.**False Negatives(FN)**:In this case the model incorrectly predicts the negative class.

7.2 Metrics Of Confusion Matrix:

7.2.1 Accuracy Score:

Accuracy is how close or far off a given set of measurements are to their true value.So,as a rule of thumb accuracy can tell us immediately whether a model is being trained correctly and how it may perform generally. Generally accuracy is most suited when we just care about single individuals instead of multiple classes,in case of multiclass problem it computes subset accuracy.
The main problem behind using accuracy as the main performance metric is that it does not perform well in presense of severe class imbalance.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

7.2.2 Precision Score:

Precision indicates that how many of the actual positives our model is able to capture labelling it as positive.It is a fraction of true postive elements divided by total no. of positively predicted elements.Precision is a good measure to determine if the cost of False Postive id high.

$$\text{Precision} = \frac{TP}{TP+FP}$$

7.2.3 Recall Score:

The Recall measures the model's predictive accuracy for the positive class.It is computed as a fraction of true postive classes divided by total no. of positively classified classes.This is also termed as sensitivity.

$$\text{Recall} = \frac{TP}{TP+FN}$$

7.3 True Negative Rate(Specificity):

It is just an opposite measure of recall score.It is a fraction of true negative elements divided by total no. of negatively presicted elements.

$$\text{True Negative Rate} = \frac{TN}{TN+FP}$$

7.3.1 F1 Score:

F1 score is a measure of robustness and preciousness of the model.It is needed when we want to seek a balance between precision and recll.It is the harmonic mean of precision and recall.It is generally used if false negatives and false positives play aa crucial role in the model.It takes maximun score of 1 and minimun score of 0.

$$\text{F1 Score} = \frac{2(PrecisionXRecall)}{Precision+Recall}$$

$$= \frac{2TruePositive}{2TruePositive+FalsePositive+FalseNegative}$$

7.3.2 Precision Recall Curve:

A precision-recall curve is a plot of the precision and the recall for different thresholds.It is used for evaluating the performance of binary classification algorithms.In case of logistic regression the threshold value would be the predicted probability of an observation belonging to positive class.

7.3.3 ROC Curve:

A ROC (Receiver Operating Curve)curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system at its different classification thresholds.This curve is produced by plotting sensitivity as the y coordinate versus false positive rate (FPR) as the x coordinate for a single classifier at different thresholds.For the logistic regression purpose the threshold value would be the predicted probability of an observation belonging to positive class.

8 Model Building And Comparison:

In our project, we have built two models using each of the imputed datasets,which are as follows:

1. Logistic Regression using Lasso Regression.
2. Logistic Regression not using Lasso Regression.

We then compared the accuracy score of each of the model .

Accuracy Score using Lasso Regression	0.78
Accuracy Score not using Lasso Regression	0.83

Table.6:Accuracy Score comparison table

we can see that the Lasso Regression Model has accuracy 78% ,whereas the model without Lasso Regression has accuracy 83%.Let us explore the confusion matrix of the model with Lasso Regression:

True neg 2350 39.7%	False pos 608 10.27%
False neg 690 11.65%	True pos 2270 38.35%

Confusion Matrix for Training Dataset

True neg 741 50%	False pos 0 0%
False neg 738 49.8%	True pos 1 0.67%

Confusion Matrix for Testing Dataset

The summary of our model can be seen in the following table:

Out[282]:

Generalized Linear Model Regression Results

Dep. Variable:	Attrition	**No. Observations:**	5918
Model:	GLM	**Df Residuals:**	5899
Model Family:	Binomial	**Df Model:**	18
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2770.1
Date:	Fri, 15 Apr 2022	**Deviance:**	5540.2
Time:	21:33:41	**Pearson chi2:**	6.49e+03
No. Iterations:	6		
Covariance Type:	nonrobust		

Fig.15:Classification report for Lasso Regression for Mean Imputed dataset

The classification report for Lasso Regression for mean imputed dataset is shown below:

Data	Precision	Recall	F1-score	Sensitivity	Specificity
Training Data	0.78	0.76	0.76	0.76	0.79
Testing Data	1.0	0.0013	0.0025	0.0013	1

Table.7

Now the following table is the table of estimates of the coefficients , standard errors ,Z-Scores , Probability $Z > |Z|$ and the confidence intervals for each of the variables which are included in the model.

	coef	std err	z	P>|z|	[0.025	0.975]
const	2.4782	0.111	22.252	0.000	2.260	2.696
hrs	0.6321	0.035	18.250	0.000	0.564	0.700
Age	-0.4134	0.047	-8.844	0.000	-0.505	-0.322
NumCompaniesWorked	0.2300	0.036	6.406	0.000	0.160	0.300
StockOptionLevel	-0.1761	0.034	-5.208	0.000	-0.242	-0.110
TotalWorkingYears	-0.3458	0.058	-6.010	0.000	-0.459	-0.233
TrainingTimesLastYear	-0.2837	0.034	-8.345	0.000	-0.350	-0.217
YearsSinceLastPromotion	0.5206	0.044	11.943	0.000	0.435	0.606
YearsWithCurrManager	-0.6057	0.048	-12.591	0.000	-0.700	-0.511
JobInvolvement_Low	-0.1909	0.155	-1.230	0.219	-0.495	0.113
JobInvolvement_Medium	-0.2453	0.081	-3.015	0.003	-0.405	-0.086
JobInvolvement_Very High	-0.1562	0.122	-1.280	0.201	-0.396	0.083
JobInvolvement_Low	-0.1909	0.155	-1.230	0.219	-0.495	0.113
JobInvolvement_Medium	-0.2453	0.081	-3.015	0.003	-0.405	-0.086
JobInvolvement_Very High	-0.1562	0.122	-1.280	0.201	-0.396	0.083
EnvironmentSatisfaction_Low	0.4359	0.088	4.952	0.000	0.263	0.608
EnvironmentSatisfaction_Very High	-0.8644	0.084	-10.287	0.000	-1.029	-0.700
JobSatisfaction_Low	-0.0440	0.086	-0.510	0.610	-0.213	0.125
JobSatisfaction_Very High	-1.1652	0.086	-13.482	0.000	-1.335	-0.996
WorkLifeBalance_Best	-2.4500	0.153	-15.998	0.000	-2.750	-2.150
WorkLifeBalance_Better	-2.4805	0.115	-21.488	0.000	-2.707	-2.254
WorkLifeBalance_Good	-2.4740	0.130	-19.088	0.000	-2.728	-2.220

Fig.16:Summary of Lasso Regression Model for Mean Imputed dataset

The Precision - Recall Curve and ROC Curve are as follows:

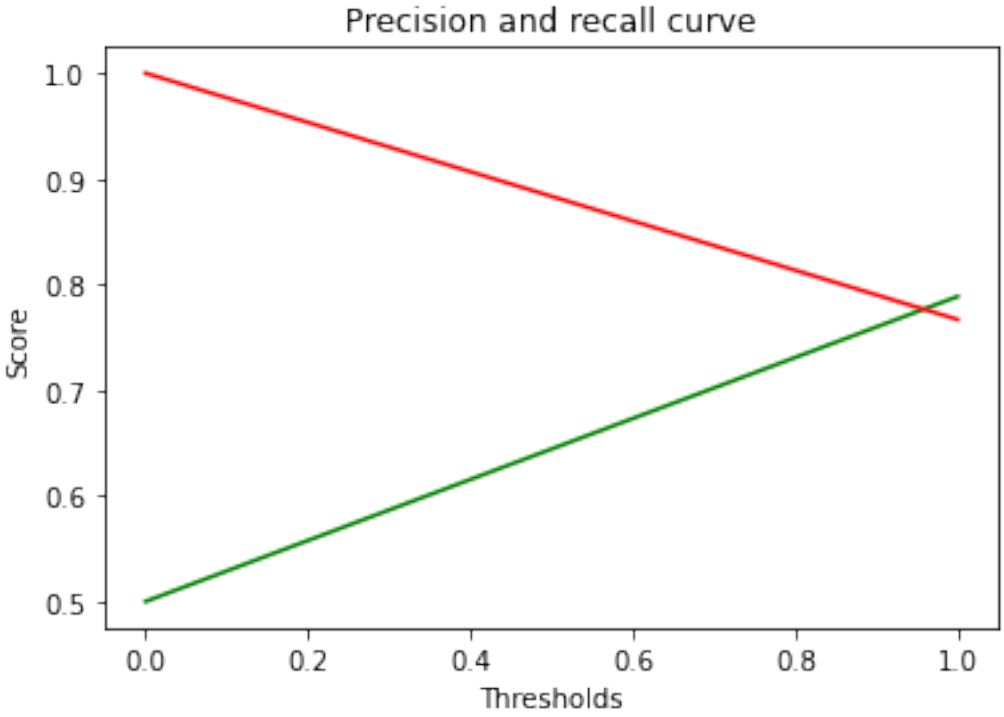


Fig.17: Precision and Recall Curve

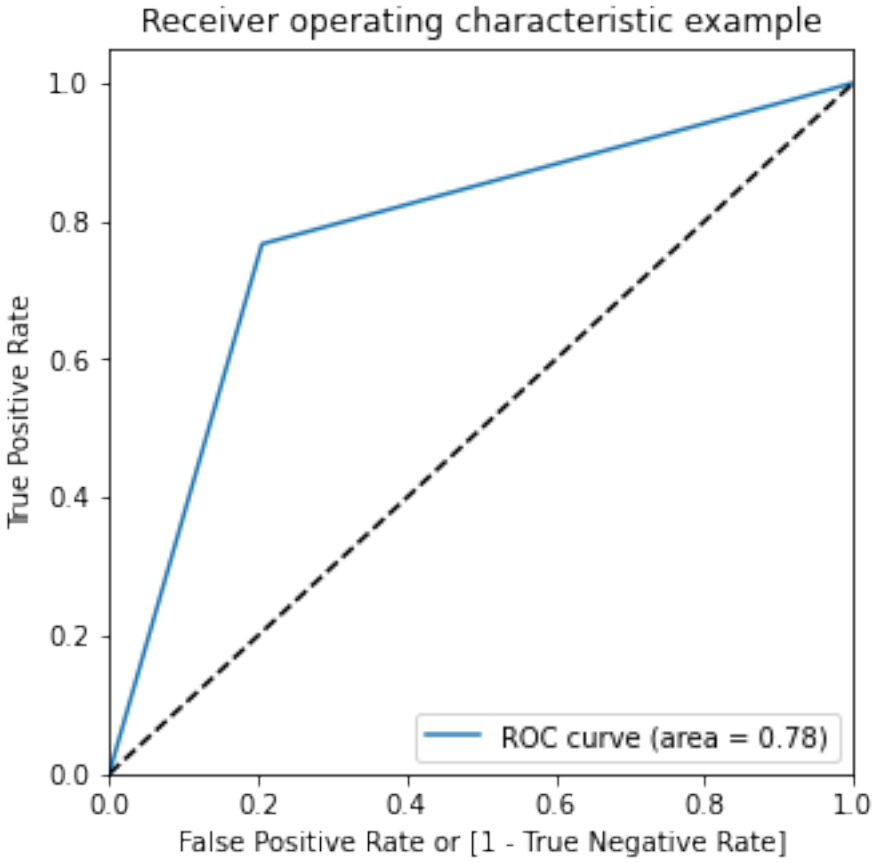


Fig.18:ROC CURVE

9 Model Diagnostics:

Now, to check the overall performance of our fitted model, we considered the ϕ Coefficient.

ϕ **Coefficient** is a measure of association for two binary variables. Higher the value of ϕ, stronger is the association. It is given by:

$\phi = \frac{f_{11}f_{00} - f_{10}f_{01}}{\sqrt{f_{1.}f_{0.}f_{.0}f_{.1}}}$

In our case, we get $\phi = 0.623$.

10 Results:

In the beginning, we first dealt with the problem of the missing values by mean imputation technique and data imbalance in our dataset. Then we plotted the bar plots and violin plot to see the distribution of attrition affected by various factors. Then we moved on to deal with the problem of multicollinearity and to tackle this, we used VIF iterative algorithm for eliminating variables from our model. We also used Lasso regression , we selected the suitable variables in our model. Then, on comparison, we got that for Lasso Regression Model with the mean imputed dataset, the overall accuracy is approximately 78% , which is lesser than the model without Lasso Regression. Finally, we used ϕ coefficient to check how good our model is.

11 Conclusions:

In our study, we have successfully modelled the problem of attrition in a company using different variants of Logistic Regression.

Initially, we started with the data pre processing steps such as dealing with Missing Data and Imbalanced Data. We found during these steps that the Polish Dataset suffered from various serious problems and necessary actions were needed before modelling the data. We therefore used Mean Imputation Technique to deal with the missing values and created synthetic samples to deal with issues of missing and imbalance data.

Then, we explored the dataset to find outliers(if any) and a clear insights to the nature of the variables and their extent of association.

After obtaining a complete dataset, we did not find any multicollinearity in our dataset. To tackle the problem of multicollinearity, we primarily used VIF iterative algorithm for detecting multicollinearity in our model. Then ,we switched to the problem of Variable Selection and implemented Lasso Regression to reduce dimensionality of our model. At last we compared the models and found that the model before using Lasso Regression is more accurate than after using Lasso Regression. But the variables which are eliminated using Lasso Regression , are comparatively insignificant than the others ,so the company should focus on those factors to control further attrition.

12 References:

1. https://www.kaggle.com/code/gauravduttakiit/hr-analytics-with-logistic-reg-mle-method/notebook

2. https://www.kaggle.com/code/adepvenugopal/employee-attrition-prediction-using-ml/notebook

3. https://home.iitk.ac.in/ shalab/regression/Chapter9-Regression-Multicollinearity.pdf

4. https://home.iitk.ac.in/ shalab/regression/Chapter13-Regression-VariableSelectionAndModelBuilding.pdf

5. https://home.iitk.ac.in/ shalab/regression/Chapter14-Regression-LogisticRegressionModels.pdf