A picture containing icon

Description automatically generated

**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Data Preparation & Visualisation  Statistical Techniques for Data Analytics  Machine Learning |
| **Assessment Title:** | Continuous Assessment |
| **Lecturer Name:** | Muhammad Iqbal  David McQuaid  Marina Soledad Iantorno |
| **Student Full Name:** | [Riccardo Possieri](https://moodle.cct.ie/user/view.php?id=13536&course=2625) |
| **Student Number:** | sba23439 |
| **Assessment Due Date:** | 7th Jan 2024 |
| **Date of Submission:** | 05/01/2024 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

Table of contents

1. Introduction

2. Goal

3. Import Datset & libraries

4. Data Preparation

5. Overview

6. EDA

7. Data Pre-processing

8. Statistical Techniques

9. Descriptive Statistical Analyses

10. Hypothesis Formulation and Testing

11. Jupyter Notebook Analysis

12. Machine Leaning

13. Splitting

14. Modelling

15. Oversampling

16. Conclusions

17. References

18. GitHub repo link

1. Introduction

This work is divided into 3 principal parts: Data Preparation, Statistical Techniques and Machine Learning. We are going to work on each section separately, even though the project is consequential, proceeding step by step.

What is Employee Attrition?

Employee attrition is a natural process where an employee leaves the workforce. Some of them are unavoidables, such as retirement or involuntary, others can be voluntary, internal, or demographic based. Attrition is an inevitable part of any business, and it occurs when the workforce diminishes over time for personal or professional reasons of the employee. It is important also to mark the difference between the employee attrition and the turnover. Employee attrition create vacancies that are not immediately filled up, while turnover is a more short-term metric and it can be addressed faster.

2. Goal

The purpose of this project is to understand the reason why of the *Attrition* and which are the most relevant features of our dataset that contribute to our target variable. In addition to identifying the reasons behind attrition, we must focus on distinguishing between unavoidable and voluntary attrition. Factors such as *Age* will be examined in relation to inevitable attrition, providing insights into the natural progression of workforce changes. However, most of the project will focus on voluntary attrition because of the mean and the mode of *Age* respectively:

|  |  |
| --- | --- |
| Mean | 36.6 |
| Median | 34.9 |

As part of the overall goal, the project will delve into statistical analyses to uncover patterns and relationships among various features in our dataset. This will facilitate us the identification of the most relevant contributors to attrition, guiding subsequent machine learning efforts for predictive modelling.

3. Import Datset & libraries

We are going to use the most common libraries like pandas, numpy & seaborn, and some specific libraries such as:

*from sklearn.preprocessing import StandardScaler, LabelEncoder  
from sklearn.discriminant\_analysis import LinearDiscriminantAnalysis  
from scipy.stats import f\_oneway*

4. Data Preparation

5. Overview

The dataset, which we are going to rename *df*, is composed by (1470, 35), where 1470 are the number of observations, and 35 are the features. Each feature has been explained by the professors with the Word document *HR\_Data\_Dictionary*. Looking at the DataFrame, most of the features need to be handled in terms of rounding, filling missing values, data cleaning and preparation. The features we have are divided into 26 numerical and 9 categorical features with 10% of missing values (5145).

Plotting our numerical features, most of them look normal distributed.

A group of blue bars

Description automatically generatedFig 1

Using the in-built Pandas function for data visualization, we check if they really are normal distributed:

A group of blue lines

Description automatically generatedFig 2

Considering the normal distribution of a feature, some of them look normal distributed.

A diagram of a normal distribution

Description automatically generatedFig 3. Bhandari, P. (2020). Normal Distribution | Examples, Formulas, & Uses. [online] Scribbr.

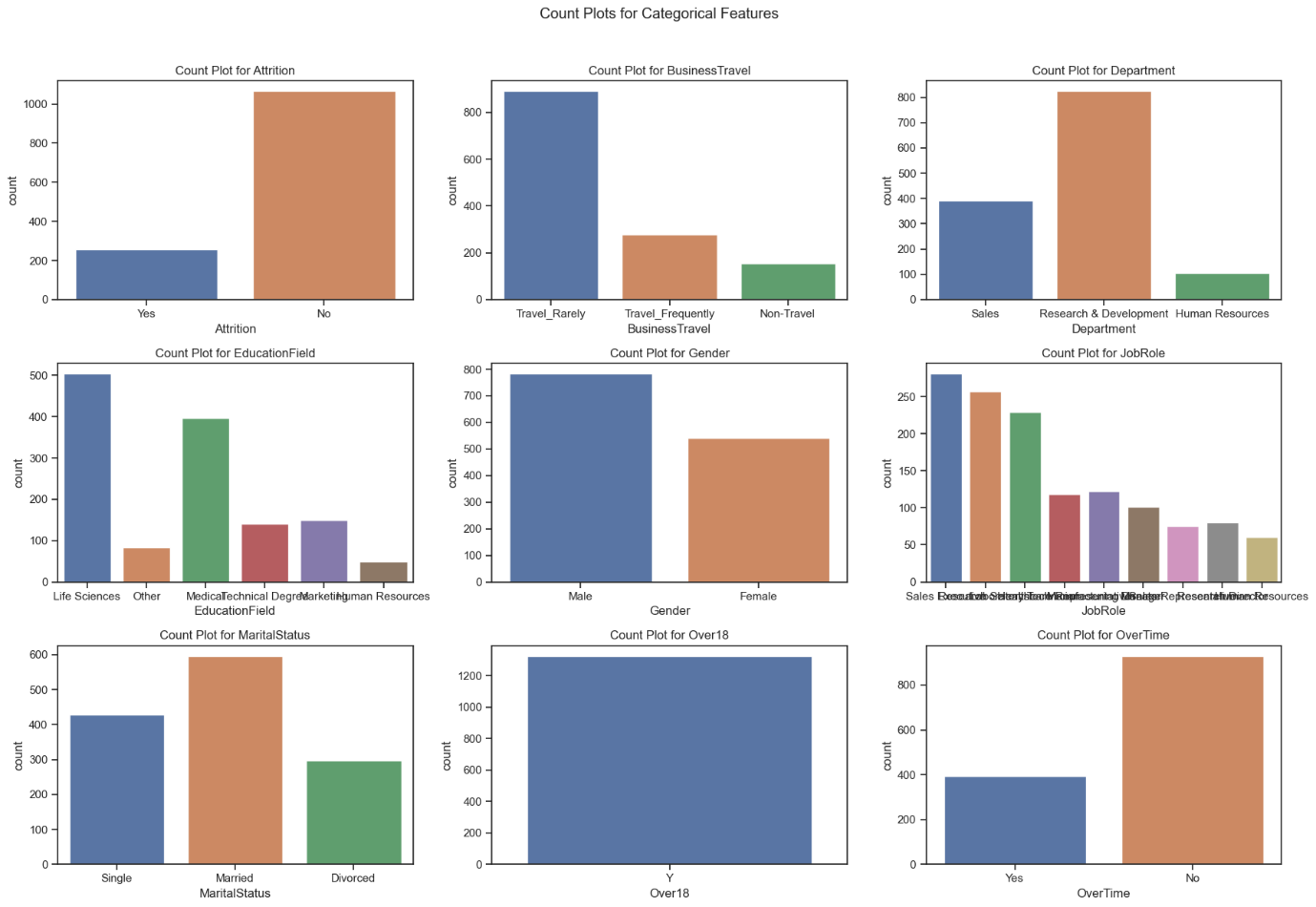
We then plot the categorical features. Looking at our target variable *Attrition*, it looks very umbalanced. Consequently, in machine learning models, we are going to run some models with balanced and unbalanced target variable. It is always a good idea to balance our target variable before running the models, but it is important to evaluate how balance it due to not lose important information.   
  


Fig 4

JobRole is a feature with many groups. This is the reason why we plot a pie plot and look at the percentage of each of them. SalesExecutive is the JobRole with more than 20% of frecuency. We are going to analyze it in the statistic techniques part.

A pie chart with numbers and text

Description automatically generatedFig 5

According to the correlation matrix we are going to show later, one important feature that can highlight factors associated with our target variable is: *TotalWorkingYears*. Let’s compare it with *PercentSalaryHike.*

A graph showing a blue circle

Description automatically generated with medium confidenceFig 6

Keep analysing the dataFrame, we notice we have 147 missing values for each feature. The Heatmap can show us each of them.

A purple and yellow background with white text

Description automatically generatedFig 7

We also notice the DataFrame has not duplicate values. Afterwards, printing out our target variable, we have the following information:

*array(['Yes', 'No', nan], dtype=object)  
No 1065  
Yes 258  
NaN 147*

A graph with a blue and brown rectangular bar

Description automatically generated with medium confidenceFig 8

The Dataset looks an interesting dataset to analyse. We would have used many features as target variable such as *JobSatisfaction, MonthlyRate, PerformanceRating*, though we decided to focus on *Attrition* to get the best performance of the models because of this binary categorical variable.

6. EDA

The first interesting point to analyse in the EDA part are the missing values in *Over18*. We realize *Over18* can be associated with Age feature and fill the nan values with the corresponding *Age* value. For obvious reasons, we can’t do vice versa. Once we have done it, we get the following information:

Over18  
Y 1456  
UNKNOWN 14

14 of 147 missing *Over18* values don't have the age information and they are also relatively small and these rows don't carry critical information for our analysis so we choose to drop them.

A graph of different age groups

Description automatically generatedFig 9

We then use the mean for the normal distributed features, the median for all the floating variables which have a skewed distribution or are concerned about the influence of outliers as we could have seen in the Overview step, and the mode for the categorical features.

In order to decide how can we Handle missing values of our target variable, is important to see the realtionship between *Attrition* variable's missing values and other features. To do that, we create a Flag for Missing values in *Attrition.*

A graph of a graph of a missing person

Description automatically generated with medium confidenceFig 10

A graph showing different colored squares

Description automatically generatedFig 11

A graph of different colored boxes

Description automatically generated with medium confidenceFig 12

A graph of a bar chart

Description automatically generated with medium confidenceFig 13

A graph of a bar graph

Description automatically generated with medium confidenceFig 14

A graph of missing and department

Description automatically generatedFig 15

By comparing Attrition with the other variables, the trend of the values of our target variable remains unchanged. Therefore, we can consider the choice of dropping the missing values of that. Also, looking at *Age*, we need to round it into integer values:

[42 52 33 35 25 30 62 27 37 28 21 60 34 24 31 32 44 48 41 40 29 46 22 39

54 47 63 56 59 53 45 51 55 36 20 26 38 43 18 61 19 50 57 49 23 58 67 17

65 69 68 15]

Having a look to all the features which have more than 10% of 0 values, we decide to leave it as it is because these 0 values have an imporant meaning for our goal.

After checking the second attrition File we have generated, we notice that in this situation, some features won't offer any discriminatory power, and their inclusion in analysis or modelling might not be necessary. We also decide to round some specified columns to integers and to round others specified columns to 2 decimals in our Dataset. Our features are now handled and the *df* is ready to be manipulated with the pre-processing step.

7. Data Pre-processing

Total number of features: 31

After dropping the Target Variable, we proceed to encode our dataFrame. We decide to apply the StandardScaler for the numerical variables, the LabelEncoder for the categorical ones, and the One-Hot encoding technique for the binary categorical ones. This is the best option we assume it is because we don’t want to lose important information throughout the encoding process. After these preprocessing steps, our data in the DataFrame is ready for PCA. We can now proceed to apply PCA to reduce the dimensionality of our data, but before of it, we split the data into training and testing sets 10% because of the dimension of our DataFrame. The features are stored in X, and our target variable 'Attrition' is extracted from the *df*.

We are going to try to preserve as much variance as possible and we should pick the eigenvectors with the biggest eigenvalues, because they are going to capture the most data.

A graph with numbers and a line

Description automatically generated  
Fig 16

Eigenvectors:

array([[ -3242.58961429],

[ 12748.77045982],

[ -29.4087693 ],

...,

[ -8341.7471347 ],

[-10128.98249087],

[ -669.77582222]])

As per Fig 16., choosing just 1 component we can reach 99.8% of variance, which is a great result considering that between 90-95% is already good. However, we apply the LDA, and we compare it with the PCA. Is important to highlight that one of LDA parameters must be (n-1) where ‘n’ is the number of unique classes of our target variable. Having a binary categorical variable as target variable, the number of components in the LDA is limited by 'n-1’. We decide then to plot PCA & LDA in 1D and comparing them.

A chart with numbers and lines

Description automatically generated with medium confidence  
Fig 17

A chart with a line and numbers

Description automatically generated with medium confidence  
Fig 18

The goal of PCA is to maximize the variance, while the observation that in the LDA plot, points of different classes (yellow and purple) are more clustered and separated along the line, informs us that LDA is successful to find a direction that maximizes the class separation.

In summary, our results can be resumed as good results if the goal is having high percentage of variance with as less principal components as possible. However, our goal is to build predictive models, and using a reduced set of principal components may impact the model performance.We will see the results in the Machine Learning model step.

8. Statistical Techniques

9. Descriptive Statistical Analyses

First at all we compute the descriptive statistics for all the numerical variables. Here, we can see measures of central tendency such as Mean, Median, and Mode, as well as measures of dispersion like Standard Deviation. Measures of dispersion quantify how spread out or dispersed the values in a dataset are. Standard Deviation specifically measures the average deviation of each data point from the mean of the dataset.

As we described in our goal, the Mean & Median of Age is rounding 35 years old, which is the reason why we focus on the voluntary Attrition, and not on involuntary one. Following the descriptive Statistical Analysis, we calculate the relative frequency for each of the features we consider crucial in identifying the reasons for employees voluntarily leaving the company.

Frequency Distributions for categorical variables

A group of pie charts

Description automatically generated  
Fig 19

Before calculating the relative frequency of each variable, we must determine which are the most important variables for us. To do that we encode our features and then we run the correlation matrix.

A screenshot of a computer screen

Description automatically generatedFig 20

A graph of a graph with different colored bars

Description automatically generated with medium confidenceFig 21

While negative correlations can highlight factors associated with lower attrition, they may not directly indicate the reasons behind attrition. This is why we are going to choose positive & negative values as well of the correlation matrix we just did. We also add *JobSatisfaction* as *reason\_of\_attrition\_chosen*:

*Correlation with Attrition for Chosen Features:*

*OverTime: 0.17890804531309443*

*MaritalStatus: 0.11145201213995101*

*JobRole: 0.05800251080888211*

*DistanceFromHome: 0.051686682533385035*

*NumCompaniesWorked: 0.03070241504329511*

*TotalWorkingYears: -0.13926623444084954*

*MonthlyIncome: -0.13804998724929374*

*JobSatisfaction: -0.08154831921179398*

The most frequent OverTime is No, with a relative frequency of 73.6%

The most frequent MaritalStatus is Married, with a relative frequency of 50.1%

The most frequent JobRole is Sales Executive, with a relative frequency of 28.9%

The most frequent DistanceFromHome is 7km, with a relative frequency of 10%

The most frequent NumCompaniesWorked is 1, with a relative frequency of 32.4%

The most frequent TotalWorkingYears is 9, with a relative frequency of 17.2%

The most frequent MonthlyIncome is 4954.75, with a relative frequency of 10%

The most frequent JobSatisfaction is 2.90, with a relative frequency of 10.6%

A pie chart with numbers and text

Description automatically generated

Fig 22

Even though this pie chart shows the highest relative frequencies, it doesn’t mean that those are the classes with highest Attrition. This is why we are going to apply t-test and ANOVA for these features in the following steps.

10. Hypothesis Formulation and Testing

We apply a t-test when comparing means between two groups or levels, suitable for binary or two-level categorical variables. We Use ANOVA for comparing means across multiple groups or levels, especially with categorical variables featuring more than two levels. Based on our reason\_of\_attrition\_chosen, we would like to apply:

T-test:

OverTime (Binary categorical variable)  
MaritalStatus (Categorical variable with multiple levels, but we can perform pairwise t-tests for each level against the others)

ANOVA:

JobRole (Categorical variable with multiple levels)

DistanceFromHome (Numerical variable, but if we categorize it into groups, we can use ANOVA)

NumCompaniesWorked (Numerical variable, but if we categorize it into groups, we can use ANOVA)

TotalWorkingYears (Numerical variable, but if we categorize it into groups, we can use ANOVA)

MonthlyIncome (Numerical variable, but if ywe categorize it into groups, we can use ANOVA)

JobSatisfaction (Categorical variable with multiple levels)

The T-test suggests that there is a significant difference in some aspect (the means) between employees who work overtime ("Yes") and those who don't ("No"). The small P-value strengthens the evidence against the null hypothesis.

*T-test for OverTime:*

*Result: Reject H0 - There is enough evidence to say that the means are different.*

*T-test for MaritalStatus:*

*Result: Reject H0 - There is enough evidence to say that the means are different.*

*ANOVA for JobRole:*

*Result: Reject H0 - There is enough evidence to say that at least one group mean is different.*

*ANOVA for DistanceFromHome:*

*Result: Fail to reject H0 - There is not enough evidence to say that all group means are different.*

*ANOVA for NumCompaniesWorked:*

*Result: Fail to reject H0 - There is not enough evidence to say that all group means are different.*

*ANOVA for TotalWorkingYears:*

*Result: Reject H0 - There is enough evidence to say that at least one group mean is different.*

*ANOVA for MonthlyIncome:*

*Result: Reject H0 - There is enough evidence to say that at least one group mean is different.*

*ANOVA for JobSatisfaction:*

*Result: Reject H0 - There is enough evidence to say that at least one group mean is different.*

Results indicate significant differences in means for *OverTime* and *MaritalStatus* based on t-tests, while *JobRole* and *JobSatisfaction* show evidence of at least one group mean being different in ANOVA. However, no conclusive evidence is found for *DistanceFromHome* and NumCompaniesWorked in ANOVA, suggesting potential similarity in group means. We can then conclude that:

***JobRoles*,**

***TotalWorkingYears,***

***MonthlyIncome,***

***JobSatisfaction*,**

might have varying impacts on *Attrition.*

11. Jupyter Notebook Analysis

12. Machine Leaning

To predict employee attrition, the dataset provides labelled data where our target variable is known. This makes it as a supervised learning problem. Our DataFrame included examples of both classes of Attrition. Supervised machine learning is one of the most used types of machine learning and it requires our effort to build the training set. The goal here is to predict a class label because in this case it's separated into binary classification, which is distinguishing between exactly two classes. In order to make accurate predictions on new, unseen data that has the same characteristics as the training set we use, we need to build a model able to generalize from training to test set.

Features like *OverTime, MaritalStatus, JobRole, DistanceFromHome, NumCompaniesWorked, TotalWorkingYears, MonthlyIncome, and JobSatisfaction* showed significant differences in means between attrition and non-attrition groups. This is the reason why we are going to consider these *reason\_of\_attrition\_chosen* for our classification models. Four of these features, such as (*JobRole, TotalWorkingYears MonthlyIncome, JobSatisfaction*) are considered very important on impacts on Attrition based on ANOVA & t-test.

In this report we decided to work with Logistic Regression, Random Forest Classifier, and Support Vector Machine (SVM) because of the categorical variable. Very important point of this final step is that we decided to not balance the taget variable due to the academic purpose though in real life we recommend to balance the target variable for example with the oversampling of the minority class.

A black and white screen with numbers

Description automatically generatedFig 23

13. Splitting

As we can see in the Fig 23, we run the models for 2 differents splits: 10% and 20% of testing. We also applied cv=10 using ten k-folds. In this study, testing the models on 10% and 20% subsets, we observed consistent challenges in accurately predicting the minority class across all models.

14. Modelling

A group of blue and orange bars

Description automatically generatedFig 24

While Logistic Regression and Random Forest struggled with lower precision, recall, and F1-score for the minority class, SVM demonstrated higher overall accuracy but faced difficulties in handling the class imbalance. These findings underscore the complexities associated with imbalanced datasets and suggest the need for tailored strategies to enhance model performance in real-world.

16. Conclusions

The analysis of the employee attrition dataset suggests that there are several factors that can influence employee turnover. The most important of these factors are JobSatisfaction, NumCompaniesWorked, DistanceFromHome, JobRole, and TotalWorkingYears. Organizations can take steps to improve these factors to reduce employee attrition and improve their overall workforce stability.

Personal considerations

This assessment has been a challenge to find the best way to reach the goal. Once I decided to study Attrition as Target variable, I noticed that it could come from various reasons. I decided to work to the Attrition as a choice and not for an inevitable part which every business has. Data preparation part has been the longest part where I worked on different scenarios and once,I got the best one, I pre-processed the data and got PCA & LDA. For instance, as I explained before, I used just one dimension, so just one component for each of them because I wanted to show a good visual comparison. LDA shows it better because of the classification problem where we could see better the two classes. The statistical study has been fundamental to find the features I wanted to use to run the models. I selected positive and negative relationship because, for obvious reasons, even if some features had a negative relation with Attrition, it didn’t mean it was unnecessary. Monthly Income, for example is inverse proportional to Attrition: the less you earn, the more probabilities you must have an attrition. Thanks to the ANOVA study, we had 4 features which had a big impact on our target variable. Finally, in Machine Learning part, I decided not to balance the target variable because of academic reason and get the results and explain them but is important to highlight that in a real-world problem, it could be necessary to balance them and getting better results from the models. We used the GridSearchCV to find the best parameters to use in our models and then I decided to run them, for 10% and 20% of testing data in the splitting section.

17. References

Vijetha (2022) Plot Distribution of Column Values in Pandas, Spark By {Examples}. Available at: https://sparkbyexamples.com/pandas/plot-distribution-of-column-values-in-pandas/#google\_vignett

‌Frost, J. (2018) Assessing Normality: Histograms vs. Normal Probability Plots, Statistics By Jim. Available at: https://statisticsbyjim.com/basics/assessing-normality-histograms-probability-plots/

‌What Is Employee Attrition? Definition, Attrition Rate, Factors, and Reduction Best Practices (no date) Spiceworks. Available at: https://www.spiceworks.com/hr/engagement-retention/articles/what-is-attrition-complete-guide/#\_001

‌Kumar, A. (2022) PCA vs LDA Differences, Plots, Examples, Data Analytics. Available at: <https://vitalflux.com/pca-vs-lda-differences-plots-examples/>

Bhandari, P. (2020). Normal Distribution | Examples, Formulas, & Uses. [online] Scribbr. Available at: https://www.scribbr.com/statistics/normal-distribution/#:~:text=Normal%20distributions%20have%20key%20characteristics

18. GitHub repo link

https://github.com/riccardopossier/Continuous-Assessment.git

‌

‌