**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Higher Diploma Data Analytics for Business   * Data Preparation & Visualisation * Statistical Techniques for Data Analytics * Machine Learning |
| **Assessment Title:** | Individual / Practical |
| **Lecturer Name:** | Dr. Muhammad Iqbal  David McQuaid  Marina Soledad Iantorno |
| **Student Full Name:** | Miroslava Slavikova |
| **Student Number:** | sba22382 |
| **Assessment Due Date:** | 05 Jan 2024 |
| **Date of Submission:** | 02 Jan 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. |

Attrition – title……

**Subject area:**

Table of Contents

**Introduction3**

**Business Analysis and Project Plan3**

**Machine Learning Models3**

**Libraries 4**

**Dataset and Data Understanding4**

**Data Preparation4**

**Data Visualization5**

**Hotel Revenue Management7**

**Analysis of Results14**

**Future recommendations** 14

**Challenges15**

**Milestones**15

**Conclusion15**

**GitHub link16**

**Timeline link16**

**Bibliography & References16**

**Introduction**

Employee turnover peaked in 2021 at 18%, up from 8% in 2020. This was the highest percentage recorded, since the HR Barometer Report was established in 2017. It was projected that the turnover will reach 18.2% for the year 2022. As reported by 56% employers, the main cause of the increase is an attractive renumeration, followed by the key factors, like better work-life balance and career advancement.

Commenting on increasing salaries, Sarah Fagan, Managing Director of Adare Human Resource Management said: “While the market is seeing increases in salary, total reward is not just about the financial benefits. It should incorporate the intrinsic value brought through your business culture. It recognises and shares successes of employees, it balances paid benefits with those of a non-financial value. And, most importantly, it considers the employee experience as an asset that requires the right investment.” (Fagan, 2022)

At the same time, the expense of hiring has significantly increased to as high as €7,491 per employee, going up from €4,215 in 2021 and €6,895 in 2020. For large companies, with more than 250 workers, the cost of new hire soars to €14,690 per person.

Over the past year, 39% of businesses have considered flexing working arrangements, that includes hybrid or work from home models, to be crucial indicative that will lead to higher employee retention.

**Business Analysis and Project Plan**

A company has gathered data on its employees with intention to uncover patterns and trends that will assist in improving employee satisfaction, increase productivity and gain an understanding of how to keep a healthy employee retention rate.

With the provided dataset that includes wide range of information about the employees, we are going to look at the variety factors impacting employee attrition. By understanding the reasons why employees leave, the company can revise or develop new strategies to improve talent retention and in the long term reduce their recruitment budget.

**Machine Learning Models**

Three different supervised machine learning models have been used:

* Decision Tree with accuracy score of
* KNN with accuracy score of
* Logistic Regression with accuracy score of

The results provided reasonable values, albeit not outstanding. Therefore, I have tried to perform PCA that not only compresses data, but aims to improve the accuracy score. Running the KNN model thereafter improved the accuracy score to

**Libraries**

I have loaded all the necessary libraries, panda for data manipulation and analysis library, matplotlib.pyplot and seaborn for data visualisation, numpy for numerical computing, sklearn libraries for future scaling for ML models, PCA, encoding, training and testing sets, linear regression model for predictive modelling. Libraries have been imported and assigned the abbreviated formats. The abbreviated format makes recalling and use of these libraries more efficient. Lastly, I have uploaded a dataset "Employee\_Attrition.csv".

**Dataset and Data Understanding**

**Data Preparation**

By using command df.head(), we get a quick overview, it shows first 5 rows of the DataFrame and we can look at the structure of the data, which is particularly useful when working with large datasets. When further inspecting the dataset, we have 1470 rows (observations) and 35 columns (features or variables). Additionally, the DataFrame contains following data types: 26 columns “float64(26)” representing numerical values with a decimal point and 9 columns “object(9)” representing string values.

**Handling missing values**

Missing data in dataset is a very common occurrence, therefore we need to understand, why is it missing and what exactly is missing. If we ignore the missing data, our results could lead into a bias result and reduce statistical power ultimately leading to invalid conclusions.

There are two possible options to deal with this, if the percentage of missing date is relatively low, we can perform imputation to replace the missing data. The second option is to remove data but it’s not recommended as it’ll lead into fewer observations resulting in inaccurate analysis.

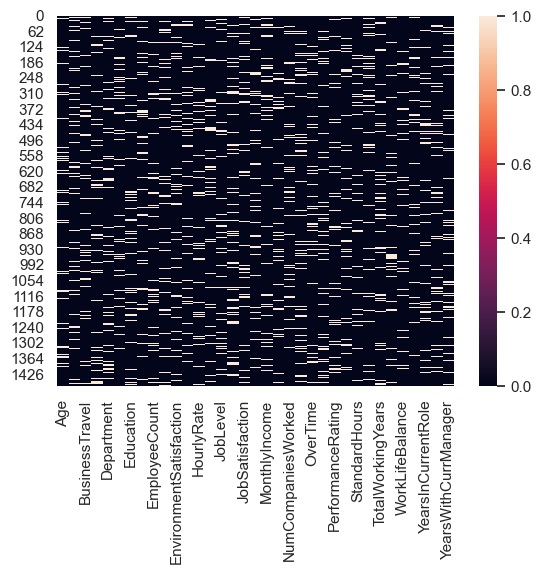
From the heatmap below we can see that missing values are spread all across the dataset and when checking for Null Values, it’s apparent that 147 values are missing for all the variables.

Imputing missing values with the Mean involves replacing NaN entries with the average value. Numeric with Mean and categorical with Mode.

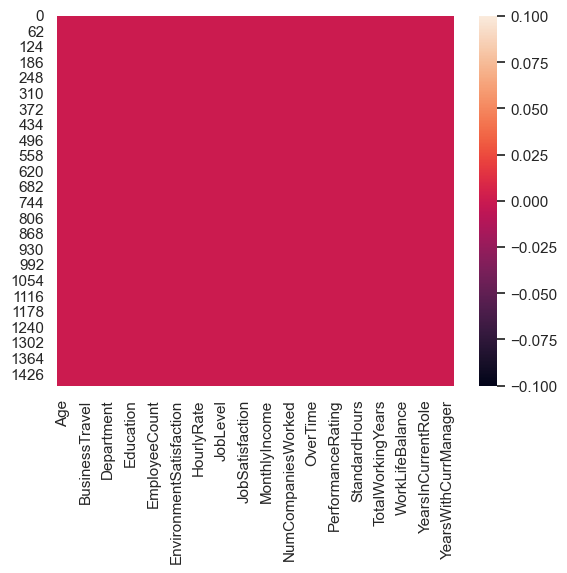
Numeric with Mean and categorical with Mode.

Mean is for numerical values (mean of each column

Mode for categorical (cannot do average)



We have no missing values now, 2 heatmaps to compare



**Data Visualisation**

performed several visualizations to review the data in order to have better understanding of the dataset.

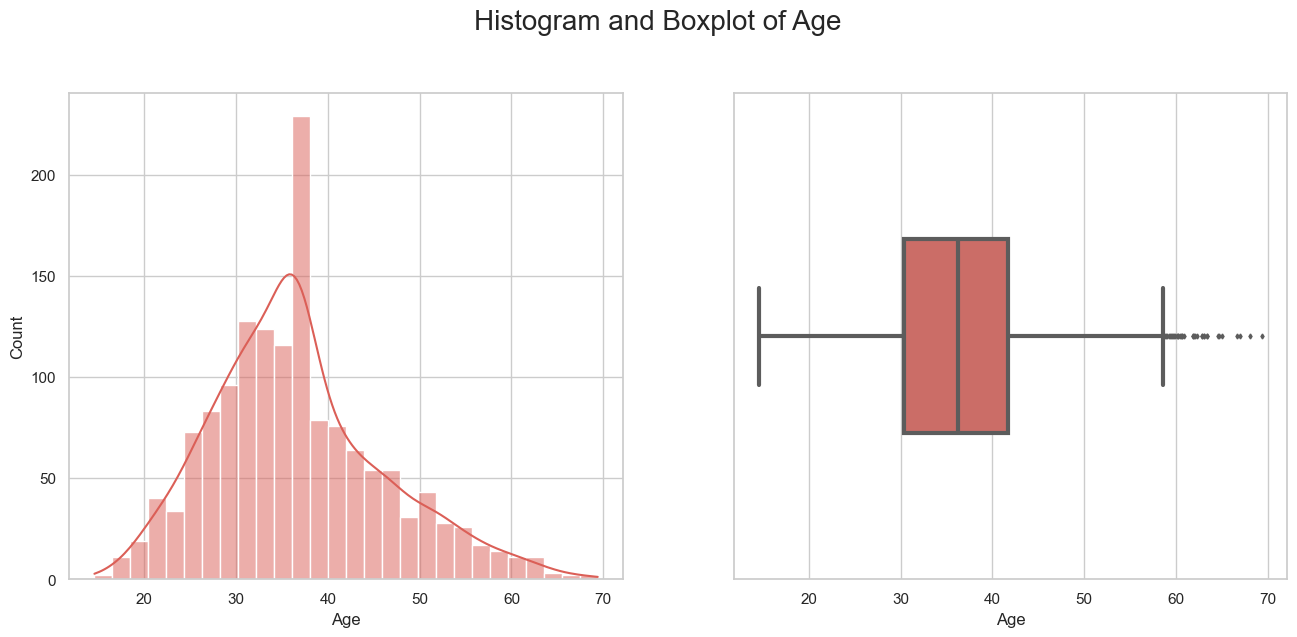
Describe – we display statistical values to identify outliers

Correlations shows us relationships between variables and how they are correlated.

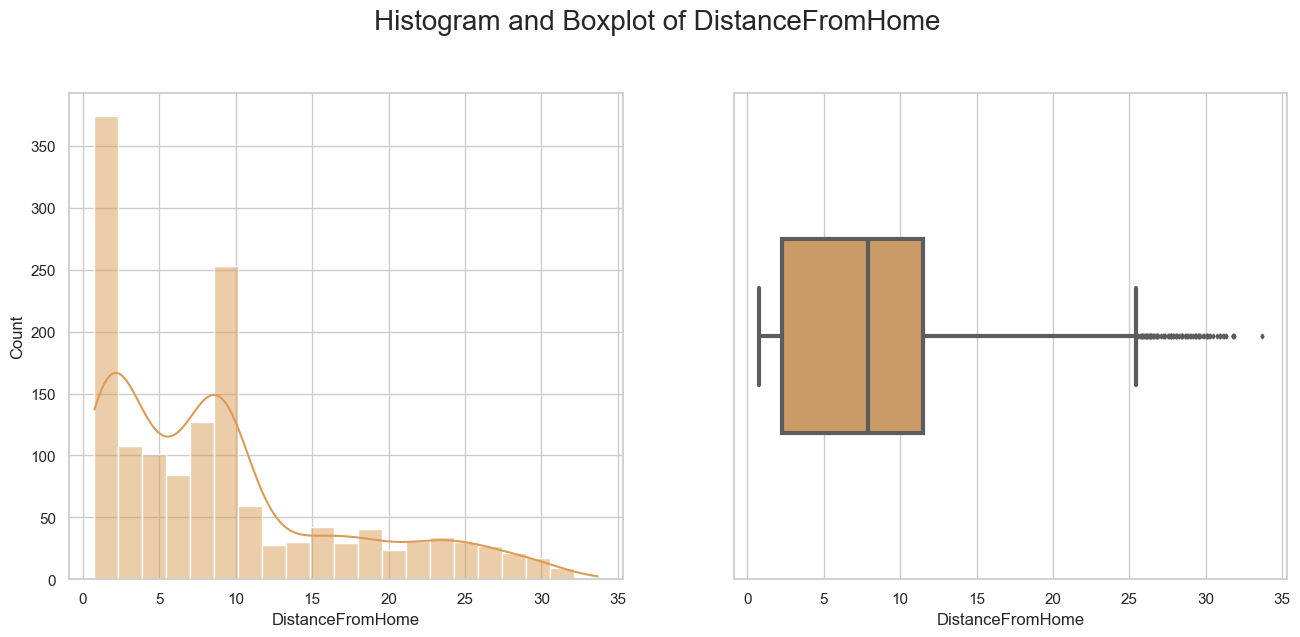
this is library that helps to perform EDA more effectively source:

REF: <https://pypi.org/project/fasteda/>

Checked outliers, that majority employees are 30-40 yo Mean age is 36



There is an outlier on the distance from home, but I did not drop it, because based on our correlation results, I found out it won’t have any effect the data.



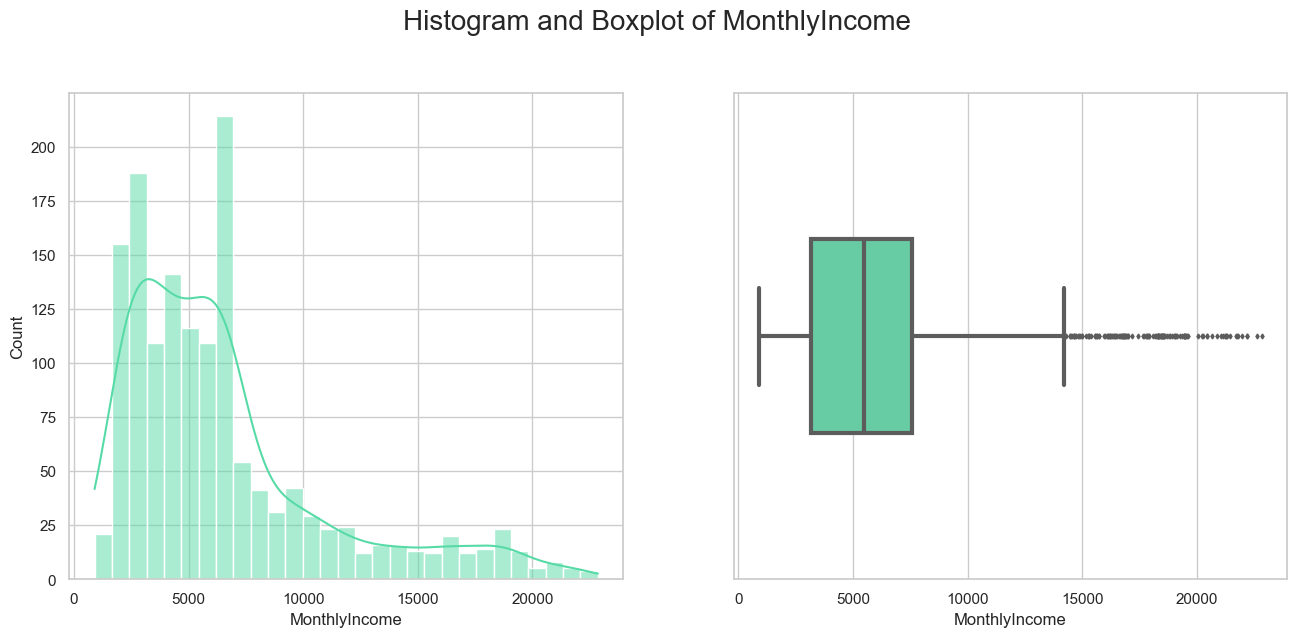
**There’re no outliers in the monthly income variable, the range of monthly income is following:**

MonthlyIncome

Mean 6544.72

Median 5495.90

Mode 6544.72



Encoding, Scaling, Feature Engineering

PCA/ KNN

We use StandardScaler to encode categorical values into numerical

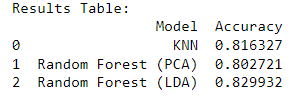
Results

LDA/ KNN



We reduced the noise …….

LDA/ Random Forest

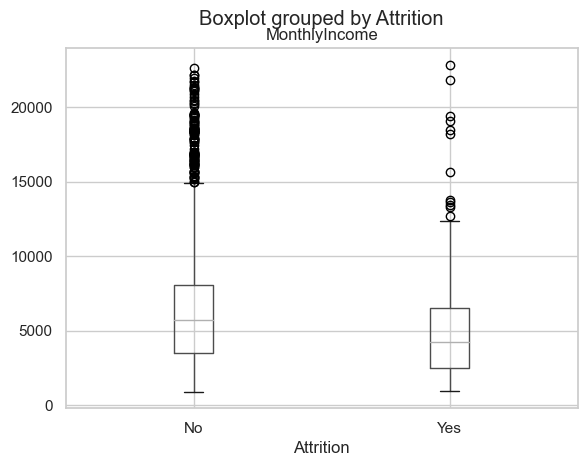


Why is PCA is working better with KNN and LDA with RandomForest …

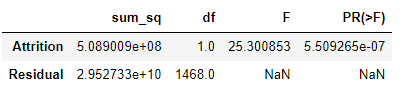
KNN works on the probabilities

Hypothesis Testing

We use 2 variables to identify how the monthly income affects the attrition and ,..

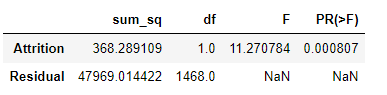


Annova



The p-value for 'Attrition' is very low (5.509265e-07), suggesting that there is a significant effect of 'Attrition' on 'MonthlyIncome.'

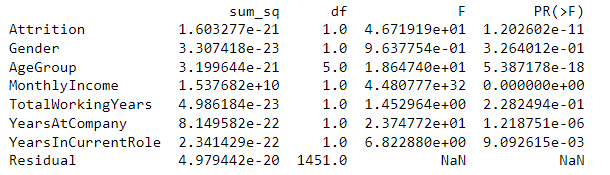
Year is the company does not affect



Hypothesis test where

Dep. Variable: MonthlyIncome

Anova table where obtain good p-values, which is higher than 0.05, anything below BAD



Selection of Hyperparameters

I used the n estimators

'n\_estimators': [50, 100, 200]

'max\_depth': [None, 10, 20],

Models decided:

Best Hyperparameters: {'max\_depth': 10, 'min\_samples\_leaf': 4, 'min\_samples\_split': 2, 'n\_estimators': 100}

Grid search cv

This means that the model correctly predicted the target variable 'Attrition' for about 80.27% of the instances in the test data. (NOTES: overfitting ……70 is bad, 95 is overfitted) too many variables, with one or 2 is easier

**Machine Learning Model Comparison using Best Hyperparameters**

KNN

Random Forest PCA

Randon Forest LDA

K-fold using split 10 (least time to run and most accurate)

, 20, 30 (30 will take more optimization)

**Data Preparation**

● Characterisation of the data set: size; number of attributes; has/does not have missing values,

number of observations etc.**[0-10]**

● Application of Data preparation/evaluation methods (Cleaning, renaming, etc) and EDA (Exploratory

Data Analysis) visualizations (plural), including a clear and concise explanation of your rationale for

what you are doing with the data and why you are doing it.**[0-20]**

● Apply encoding, scaling and feature engineering as and if required, detailing how and why you used

these techniques and the rationale for your decisions.**[0-30]**

● Explore the possibility of using dimensional reduction on the dataset. Employ both LDA (Linear

Discriminant Analysis) and PCA (Principal Component Analysis) and compare the separation of

classes through visualization. Explain the difference between both techniques in your own words

and discuss in detail how your results may affect your analysis of classifying or clustering the normalas compared to anomalous biddings.**[0-40]**

**Statistical Techniques:**

● Use descriptive statistical analyses to explore and evaluate the data set, including measures of

central tendency and dispersion and frequency distributions. Correlation matrices are also

accepted. Provide a summary of your findings. (0-30 marks)

● Formulate and test hypotheses within a business context using appropriate statistical techniques

like t-tests or ANOVA to identify significant relationships between variables. Provide a summary of

your findings. Use at least two statistical tests. (0-40 marks)

● Use a Jupyter notebook to produce result sets from the provided dataset, such as scatter plots or

regression models. Provide a summary of your findings. (0-10 marks)

● Write the results of the analysis of your findings to stakeholders using clear and concise

explanations, visualisations, and appropriate statistical terminology. (0-20 marks)

**Machine Learning:**

● Provide a conceptual understanding and logical justification based on the reasoning for the specific

choice of machine learning approach (supervised/ Unsupervised) for the provided data set. You can

discuss the pros and cons of both approaches based on your understanding. (0-20 marks)

● Machine Learning models can be used for Prediction, Classification, and Clustering. You can choose

suitable features for the machine learning models based on feature selection methods, such as random

forest or any other method. The selection of hyperparameters for the ML models should be performed

by using hyperparameter tuning, such as GridSearchCV. Obtain the best accuracy using optimal values

of the hyperparameters. (0-30 marks)

● You should train and test the Machine learning models in the case of supervised learning for different

splits (at least 2 splits) and use appropriate metrics for unsupervised learning. Use k-fold (10 or 20 or

30) cross-validation to provide authenticity of the modelling outcomes. (0-30 marks)

● Exhibit a comparison of ML modelling outcomes using a Table or graph visualisation. Identify the

possible similarities and contrast of the Machine Learning modelling outcomes based on chosen metricand discuss their statistical understanding. (0-20 marks)

**Analysis of Results**

**Future Recommendations**

## **Challenges encountered**

I have encountered numerous challenges while working on this dataset. Larger volume of data provides more information for the training and might lead to better performance.

Additionally, I’ve identified a lot of missing vales, and identifying the best performing model with better accuracy score.

Admittedly, the project overall was a challenging task, but generally speaking, I really enjoyed learning and attempting to understand such a complex subject of study as Data Analysis and Machine Learning.

**Milestones**

Researching while working on this project helped me to gain better understanding of visualisation techniques and how to interpret the various graphs and charts, gradually gaining confidence in this subject. I've also developed a better understanding of GitHub, including version control and how to create repositories.

**Conclusion**

**GitHub link:**

**References & Bibliography:**

[**https://www.fingalchamber.ie/news/details/employee-turnover-more-than-doubles-in-12-months-as-employers-battle-to-keep-staff**](https://www.fingalchamber.ie/news/details/employee-turnover-more-than-doubles-in-12-months-as-employers-battle-to-keep-staff)

[**https://www.cipd.org/globalassets/media/knowledge/knowledge-hub/reports/2023-pdfs/2023-hr-practices-ireland-report-8238.pdf**](https://www.cipd.org/globalassets/media/knowledge/knowledge-hub/reports/2023-pdfs/2023-hr-practices-ireland-report-8238.pdf)

<https://www.kaggle.com/>

https://pypi.org/project/fasteda/

<https://towardsdatascience.com/how-to-clean-your-data-in-python-8f178638b98d>

<https://realpython.com/>

<https://www.geeksforgeeks.org/>

<https://pandas.pydata.org/>