cct College Dublin
Computing • IT • Business

2024

Data Survey Analysis

INTEGRATED CA2

EDUARDO J. MATOS ROMERO

CCT COLLEGE 2023/24

MACHINE LEARNING

DATA PREPARATION

STATISTICAL TECHNIQUES

cct College Dublin
Computing • IT • Business

Index

1. Exploratory Data Analysis (379 words)

1.1 Data Interpretation

How many categorical and numerical columns we have?

As we can see: We have 1470 entries in total and 147 with missing values (1470-1323). We have 9 categorical columns and 26 numerical ones. Our datatypes are float and objects.

Columns to perform the study are "JobSatisfaction","PerformanceRating" and 'Attrition'.

1.1.1 Statistical description of numerical columns

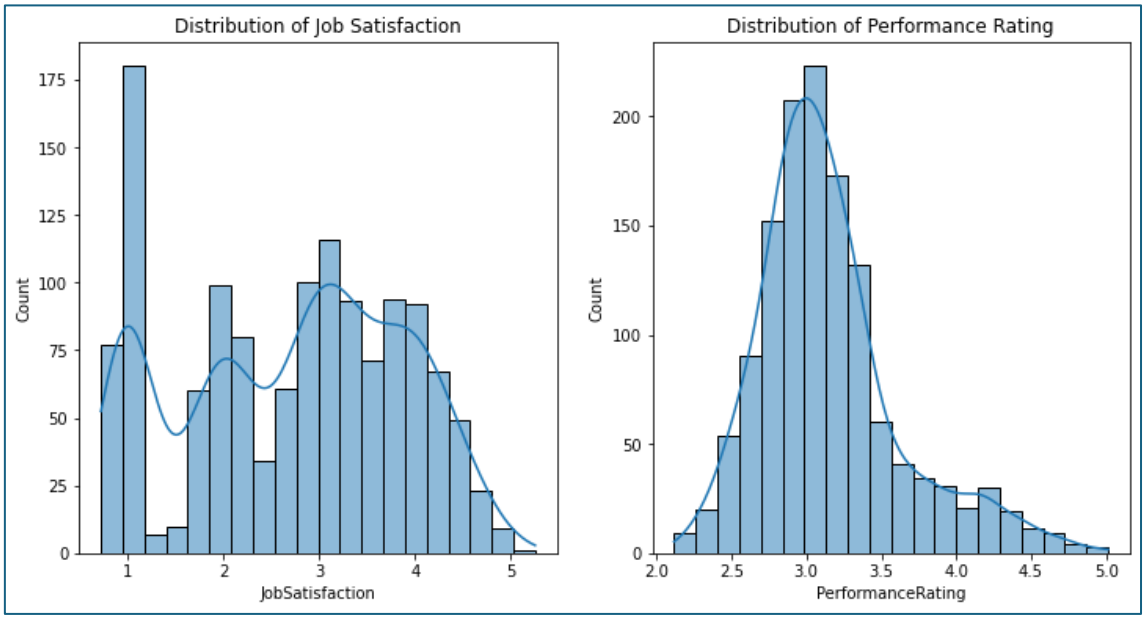

Illustration 1. Distribution of Job Satisfaction and Performance Rating.

- Age: Employees' ages vary moderately, averaging around 35.7 years.
- Daily Rate: Average £801.45, but with wide individual differences.
- Commute Distance: On average, employees live 8.06 km from work, with significant variation.
- Education: Most employees have some college education (average level 2.7/5).
- Hourly Rate: Average £64.29, showing notable variation among employees.
- Job Level: Primarily lower to mid-level roles (average level 1.85).
- Work Experience: Employees average 9.7 years in total.
- Company Tenure: Average 5.7 years but ranges widely.
- Job Satisfaction: Varies widely from very dissatisfied to very satisfied.
- Performance Ratings: Less varied, mostly around the average.
- Data Distribution: Most numerical data follow a normal distribution.

Note

Employee performance ratings follow a normal distribution around the average, while job satisfaction varies widely, indicating diverse levels of satisfaction among employees.

1.1.2 Statistical description of categorical columns

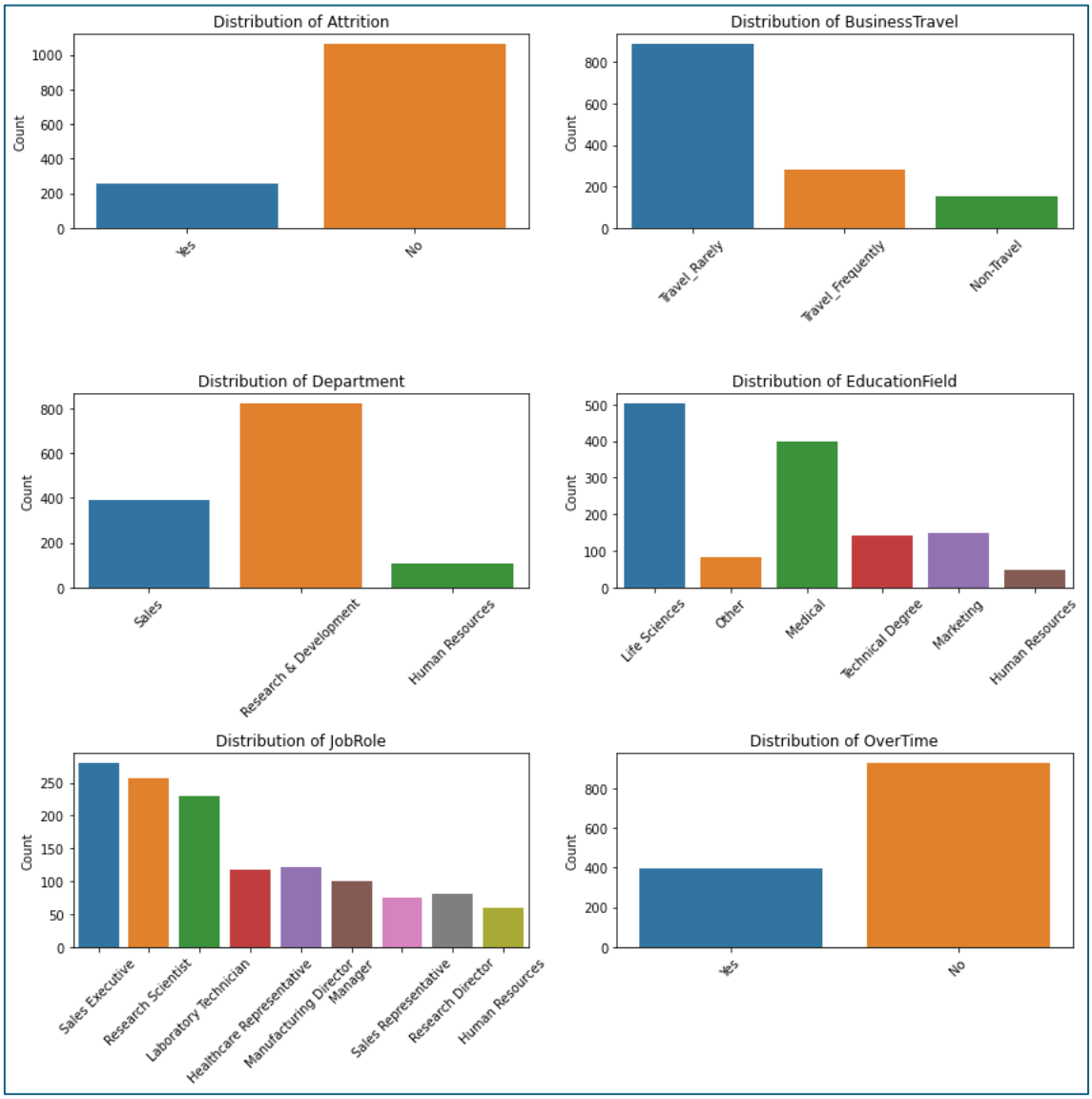

Illustration 2. Distribution of Categorical features.

- The profile of the average employee is the next one:
  - BusinessTravel - Travel_Rarely
  - Department - Research & Development
  - EducationField - Life Sciences
  - JobRole - Sales Executive
  - OverTime - No
- The dataset is imbalanced, with certain categories occurring much more frequently than others.

1.3 Data processing

1.3.1 Labelling missing values

After confirming the presence and distribution of null values, we labeled NaNs in categorical features as a separate category.

- **We have decided to label them as "Missing".**

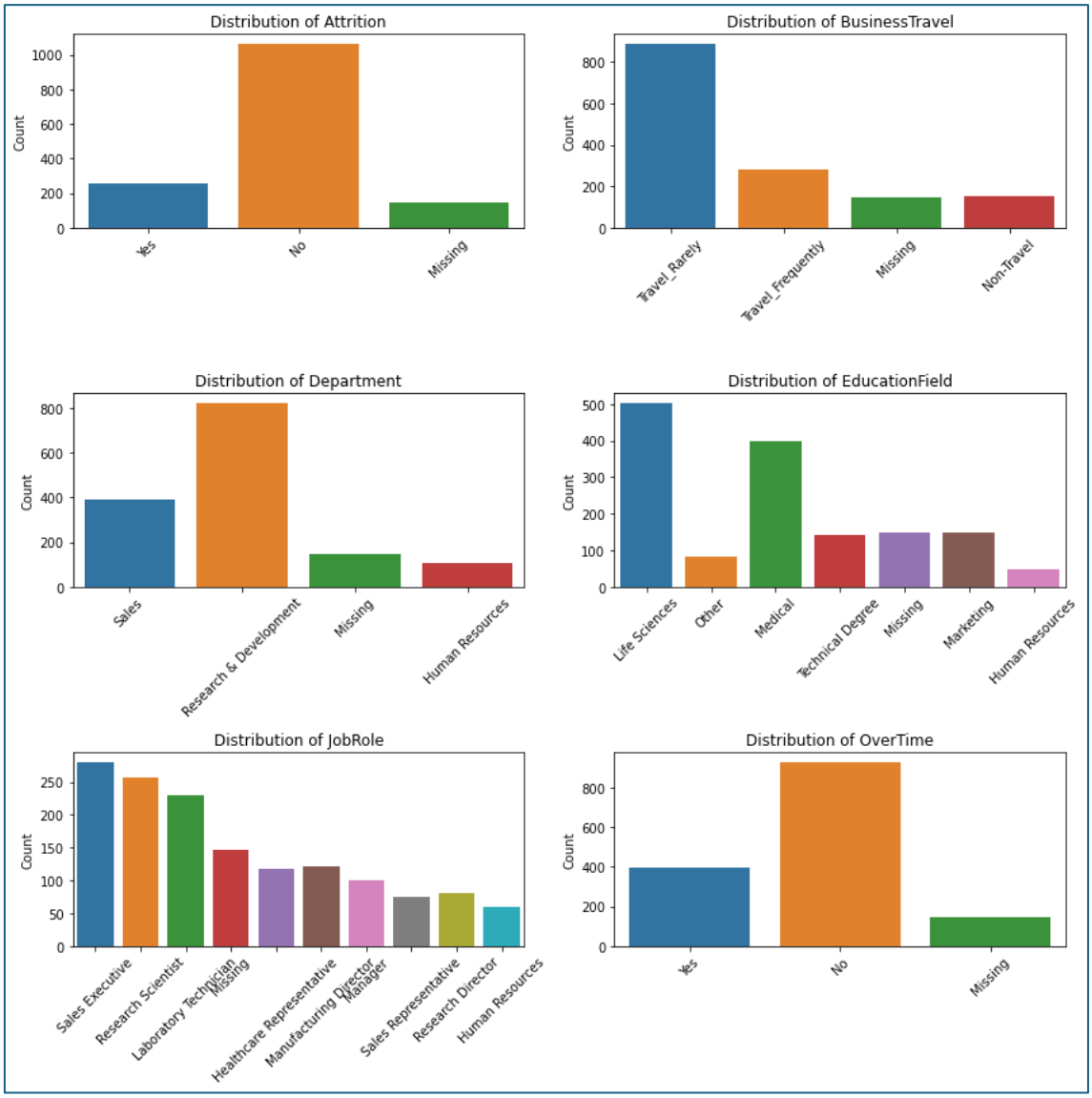

Illustration 3. Distribution of Categorical features after labelling.

Given the random distribution and low number of missing values, we are using median imputation via sklearn.impute's SimpleImputer. Outliers are not a concern as all data responses are from a predefined question set.

1.3.2 Handling missing values in categorical features

1.3.2.1 Iterative imputation

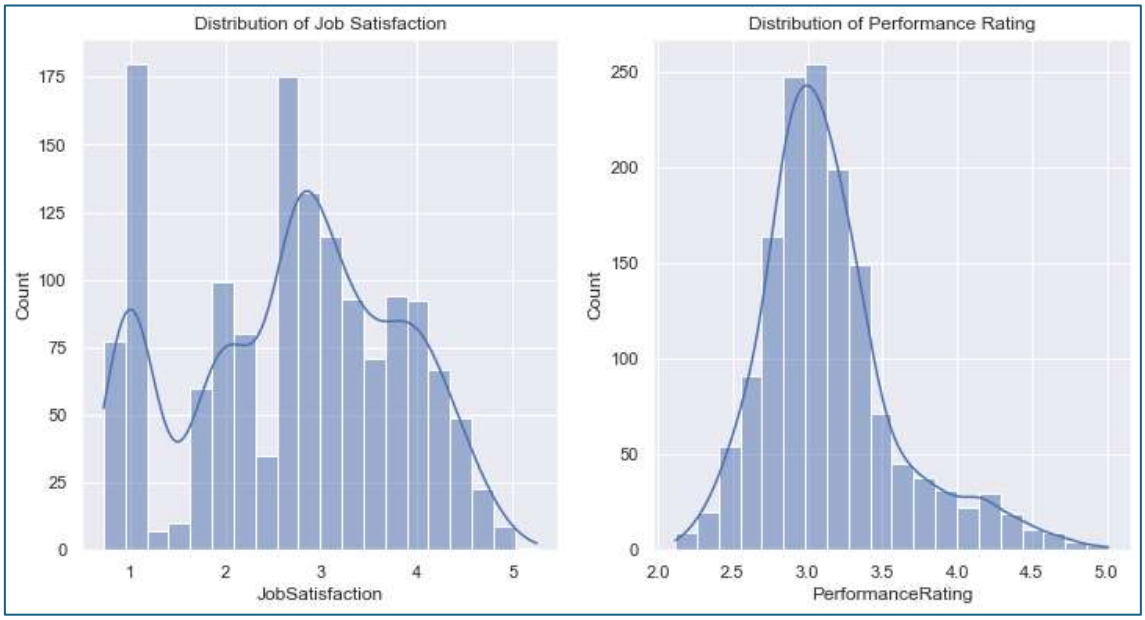

Illustration 4. Dataset after handling null values with iterative imputation.

Multiple Imputation is used to handle numerical nulls while preserving data distribution, and to label nulls in categorical features for effective data preprocessing.

1.4 Data Overview - Visual analysis

1.4.1 Do overtime stresses employees and makes them have to leave?

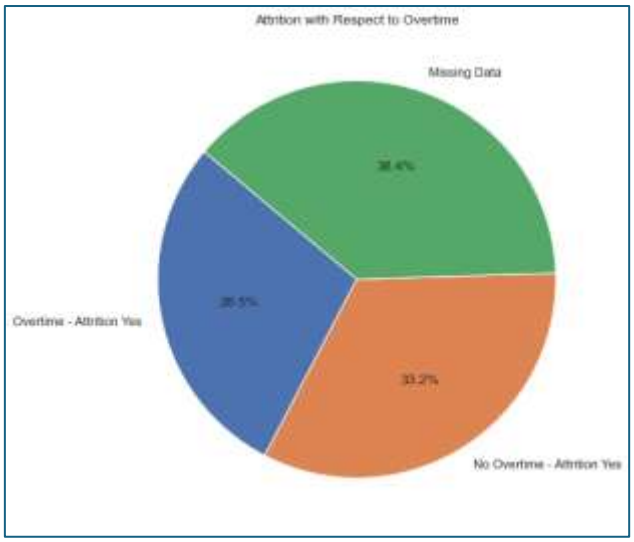

Illustration 5. Attrition with respect to Overtime.

A considerable number of employees left the company without working overtime, suggesting that overtime may not have a direct impact on the decision to leave.

1.4.2 Is salary one of the main reasons?

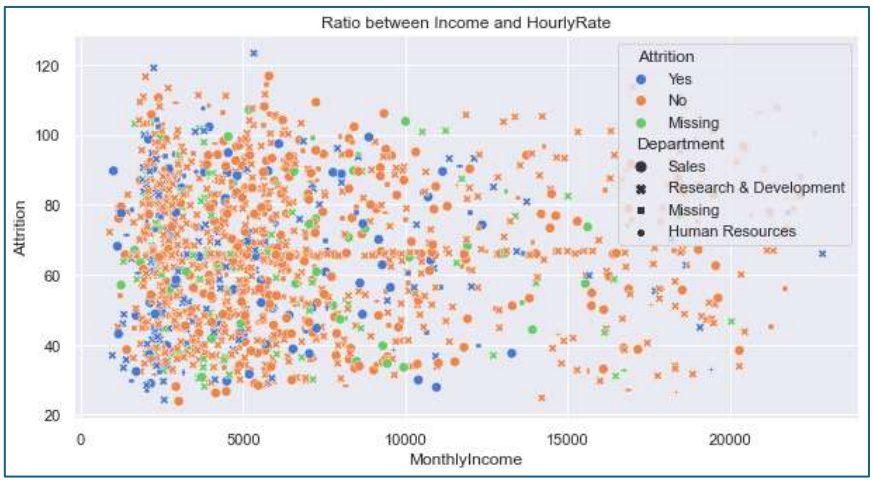

Illustration 6. Ratio between income and hourly Rate.

Most employees who leave the company have average incomes, which reduces the importance of the correlation between attrition and monthly income.

1.4.3 What is the effect of age on attrition?

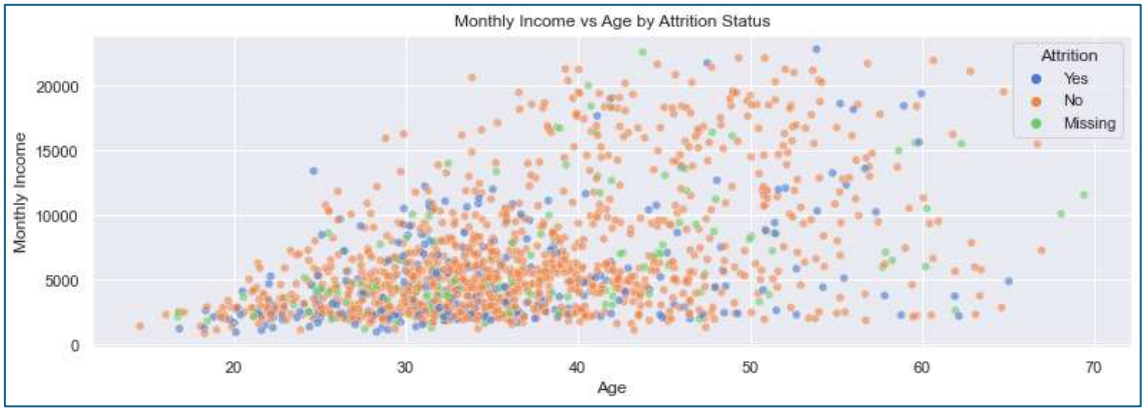

Illustration 7. Monthly income vs Age by Attrition Status.

Apparently, it is mainly young people who are leaving the company.

1.4.4 What is the distribution of departments?

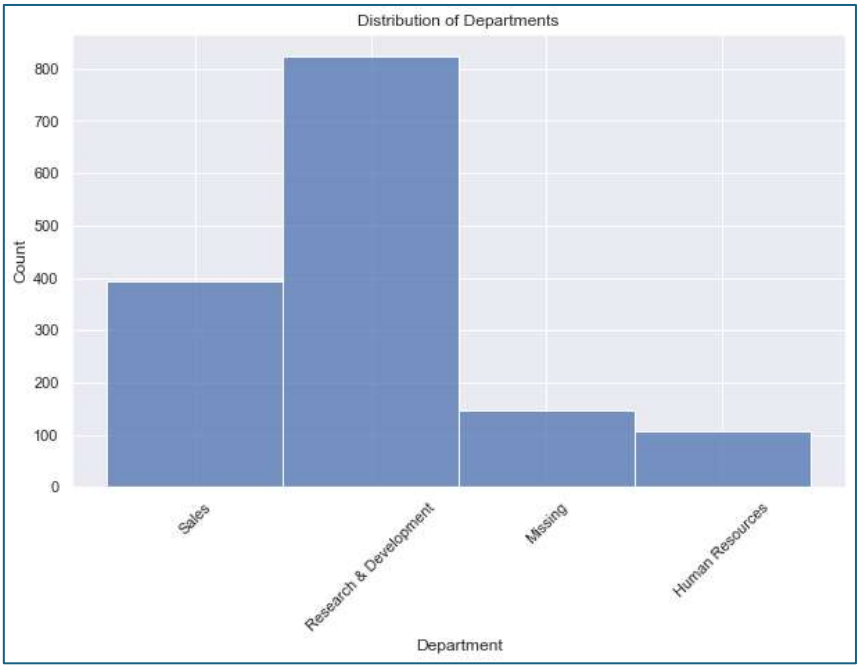

Illustration 8. Distribution of departments.

The department with more employees is Research & Development.

1.4.5 How does work experience affect attrition?

Illustration 9. Effect of Total Working Years on Employee Attrition.

Employees typically leave the company within the first 10 years.

2. Statistics Analysis (601 words)

2.1 Statistics Analysis (I)

2.1.1 Correlation Overview

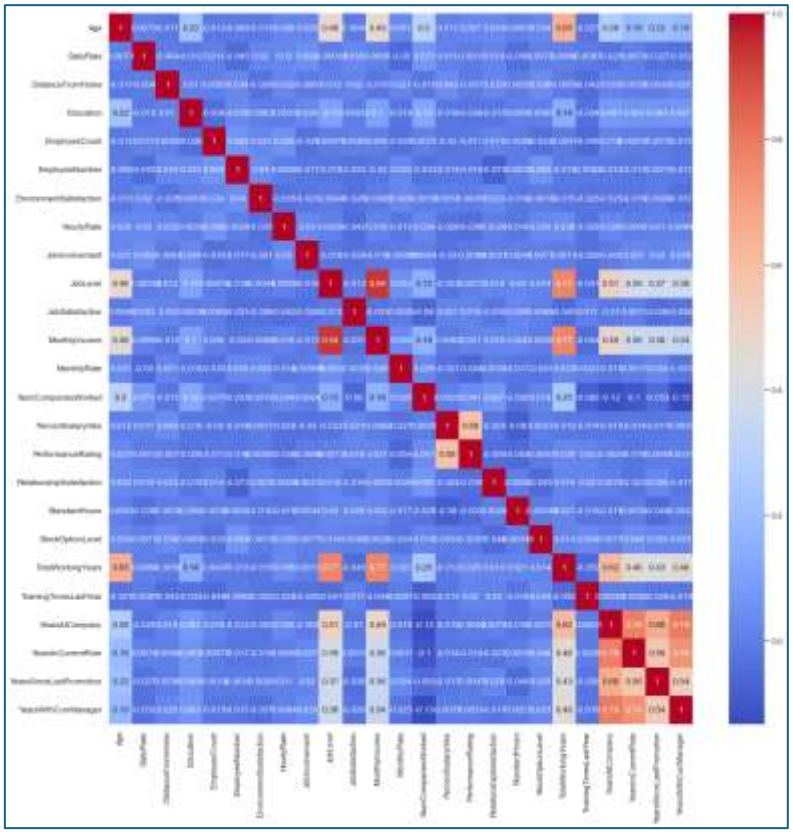

Illustration 10. Correlation overview.

Upon initial review, the correlation encountered does not provide any insights to our questions.

2.1.2 Correlation - Years In Current Role

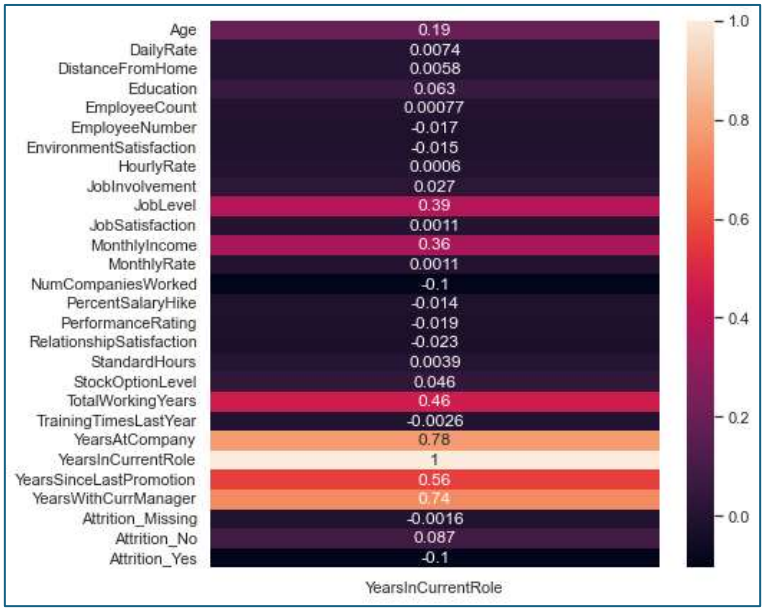

Illustration 11. Correlation Years in Current Role.

Employees who have worked under the same manager for a long time and have been in stable roles often receive more promotions and are less likely to leave. However, spending an extended period in the same position can lead to complacency, despite the benefits of job security and long-term company tenure.

This situation highlights the need for further analysis and potential investment in enhanced career development opportunities to prevent stagnation and encourage continuous professional growth within the company.

2.1.3 Correlation - Years With Current Manager

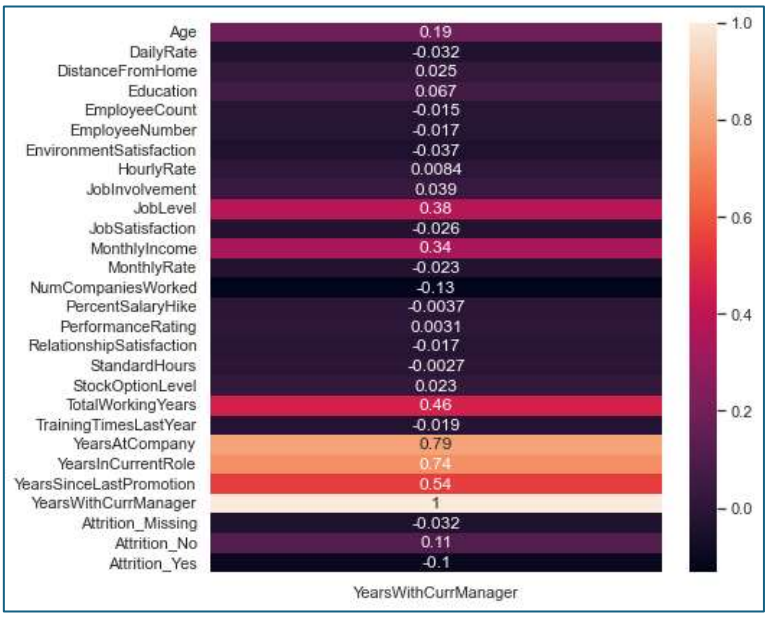

Illustration 12. Correlation Years With Current Manager.

Employee tenure with the same manager and role duration correlate with promotion frequency, indicating that career growth is tied to managerial stability.

The company should focus on implementing structured career development strategies beyond manager-employee relationships to provide more professional advancement opportunities.

2.1.4 Correlation - JobRole - Sales Executive

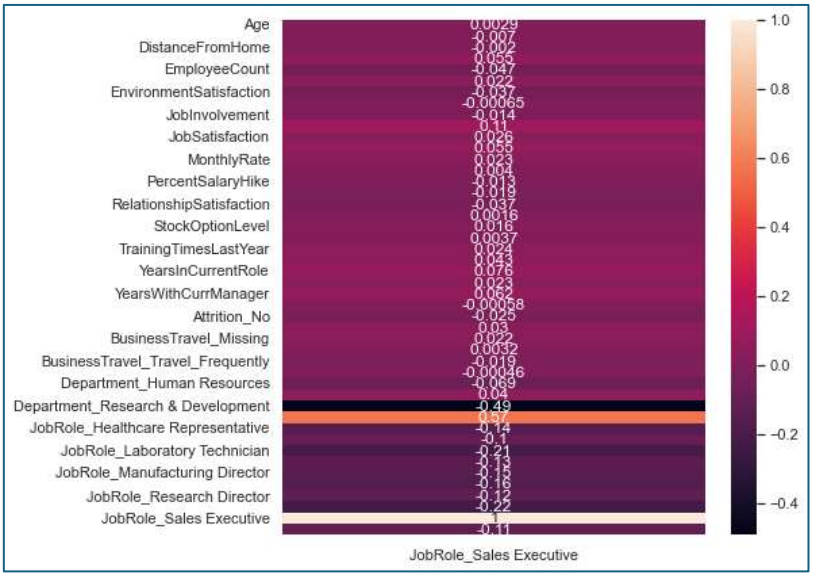

Illustration 13. Correlation Sales Executive.

Sales executives in the hierarchy experience job stability and higher income. However, there is a disconnect with core company sectors. Integrating sales roles into career plans could foster interdepartmental growth.

Structured career development strategies should extend beyond manager-employee dynamics to enhance professional advancement.

2.2 Statistics Analysis (II)

2.2.1 T-test

We are using a t-test to compare the average job satisfaction between two groups, such as male and female employees.

2.2.1.1 T-test | Age – Attrition

Hypothesis:

Null Hypothesis (H0): There is no significant difference in the average age of employees between the two attrition groups.

Alternative Hypothesis (H1): There is a significant difference in the average age of employees between the two attrition groups.

T-Statistic for Age: -3.3514082845782953, P-Value: 0.0008267367692839238

Given the low P-Value, we reject the null hypothesis (H0) which stated there is no significant difference in the average age of employees between the two attrition groups.

The negative T-Statistic value (-3.35) indicates that the average age of employees who have left the company ('Yes' to attrition) is lower than that of the employees who haven't left ('No' to attrition).

The P-Value of 0.00084 is significantly less than the common alpha level of 0.05. This low P-Value indicates that the observed difference in ages between the two groups is statistically significant.

- **Possible solutions**
  - **Career Development Opportunities.**
  - **Recognition and Feedback.**

2.2.1.2 T-test | Training time last year – Attrition

Hypothesis:

Null hypothesis (H0): There is no significant difference in the amount of training received by employees who left and those who stayed.

Alternative hypothesis (H1): There is a significant difference in the amount of training received in the last year between employees who left and those who stayed.

T-Statistic for Distance: -2.134648835520212, P-Value: 0.03297332495134566

The t-statistic's negative value indicates that, on average, employees who left the company had fewer training sessions in the last year than those who stayed.

The p-value is slightly less than the typical alpha level of 0.05, suggesting that the result is statistically significant, albeit close to the threshold.

This could indicate that training opportunities, or lack thereof, are a factor in employee retention. Employees who receive more training may feel more engaged, valued, and equipped to perform their roles, thereby reducing their likelihood of leaving.

2.3 Insights and Recommendations Summary Based on the Analysis

- Employees who remain in the same role for an extended period of time tend to have longer company service and managerial relationships. However, they may experience delays in promotion, which could indicate potential complacency. This highlights the need for enhanced career development initiatives.
- In the sales hierarchy, job level is positively correlated with stability and income, unlike in the technical, manufacturing, and healthcare sectors. Integrating sales into career plans could bridge departmental gaps and increase growth opportunities.
- Employees who work under the same manager for an extended period may experience professional stagnation.
- Younger employees tend to have higher attrition rates, which may be due to fewer training opportunities. Therefore, increasing training could improve retention and job satisfaction.

3. Machine Learning algorithms – Classification (660 words)

As it was analysed previously:

- The majority of numerical features demonstrate a normal distribution, with data concentrated around the mean.
- The dataset is imbalanced, with certain categories occurring much more frequently than others in categorical features.

The purpose of our machine learning classification model is to predict whether or not employees are likely to leave the company.

3.1 Data Pre-processing

- To address data imbalance, resampling techniques like SMOTE are used, generating synthetic samples for the minority class to enhance model performance.
- Certain non-determinant features like 'EducationField' and 'DistanceFromHome' are removed for better model focus.
- Binary labels like attrition and multi-level categories are not normalized due to class imbalance concerns.
- One-Hot Encoding is applied for categorical variables, suitable for nominal data.
- The chosen machine learning models are Random Forest, Gradient Boosting, and Decision Tree, suitable for mixed data types.
- To mitigate overfitting, a validation strategy like cross-validation is employed, ensuring reliable model performance assessment on unseen data.
- Entries with missing attrition responses are removed.

3.2 70-30 % Data split

```
Gradient Boosting Model:
Accuracy: 0.7808564231738035
Classification Report:
              precision    recall  f1-score   support

          No       0.83      0.91      0.87       316
         Yes       0.44      0.26      0.33        81

    accuracy                           0.78       397
   macro avg       0.63      0.59      0.60       397
weighted avg       0.75      0.78      0.76       397
```

```
Decision Tree Model:
Accuracy: 0.6750629722921915
Classification Report:
              precision    recall  f1-score   support

          No       0.82      0.75      0.79       316
         Yes       0.28      0.37      0.32        81

    accuracy                           0.68       397
   macro avg       0.55      0.56      0.55       397
weighted avg       0.71      0.68      0.69       397
```

```
Random Forest Model:
Accuracy: 0.7934508816120907
Classification Report:
              precision    recall  f1-score   support

          No       0.82      0.95      0.88       316
         Yes       0.48      0.17      0.25        81

    accuracy                           0.79       397
   macro avg       0.65      0.56      0.57       397
weighted avg       0.75      0.79      0.75       397
```

Illustration 14. Accuracy metrics.

- Gradient Boosting Model:
  - The Gradient Boosting Model is effective in predicting the "No" class but struggles significantly with the "Yes" class, both in terms of precision and recall.
- Decision Tree Model:
  - The Decision Tree Model shows a more balanced performance between the two classes than Gradient Boosting, though it still struggles with the "Yes" class. The model is more consistent across precision and recall for each class.
- Random Forest Model:
  - The Random Forest Model shows the highest accuracy and is very effective in predicting the "No" class but has a significant issue with the "Yes" class, missing the vast majority of actual "Yes" instances.

3.2.1 Cross Validation 70-30 % - Random Forest

[0.69 - 0.91333333 - 0.92 - 0.95317726 - 0.94983278]

The accuracy of the model varies significantly, ranging from 69% to 95%. This suggests that the model's performance is sensitive to data splits and potential data peculiarities. High scores in certain folds raise concerns about overfitting, particularly if they exceed domain knowledge expectations or previous performance benchmarks.

3.3 80-20 % Data split

Gradient Boosting Model:
Accuracy: 0.769811320754717
Classification Report:

	precision	recall	f1-score	support
No	0.83	0.89	0.86	213
Yes	0.38	0.27	0.31	52
accuracy			0.77	265
macro avg	0.61	0.58	0.59	265
weighted avg	0.74	0.77	0.75	265

Decision Tree Model:
Accuracy: 0.6641509433962264
Classification Report:

	precision	recall	f1-score	support
No	0.83	0.74	0.78	213
Yes	0.25	0.37	0.30	52
accuracy			0.66	265
macro avg	0.54	0.55	0.54	265
weighted avg	0.71	0.66	0.68	265

Random Forest Model:
Accuracy: 0.8
Classification Report:

	precision	recall	f1-score	support
No	0.83	0.95	0.88	213
Yes	0.47	0.17	0.25	52
accuracy			0.80	265
macro avg	0.65	0.56	0.57	265
weighted avg	0.76	0.80	0.76	265

Illustration 15. Accuracy metrics.

- Gradient Boosting Model:
  - The Gradient Boosting model performs well in predicting the "No" class but struggles with the "Yes" class, missing many true "Yes" cases.
- Decision Tree Model:
  - The Decision Tree model has a lower overall accuracy compared to Gradient Boosting. It shows similar trends with better performance on the "No" class but weaker performance on the "Yes" class, with slightly better recall for "Yes" than Gradient Boosting.
- Random Forest Model:
  - The Random Forest model shows the highest overall accuracy and better precision for the "Yes" class compared to the other models. However, it has a very low recall for the "Yes" class, indicating a significant number of false negatives.

3.3.1 Cross Validation 80-20 % - Random Forest

[0.68035191 - 0.91202346 - 0.94428152 - 0.93548387 - 0.93823529]

The model demonstrates high accuracy, exceeding 90%, except for the first fold. This suggests effective prediction capabilities. Investigating the lower performance of the first fold may reveal the model's specific shortcomings or data complexities.

3.5 Hyperparameter Tuning 70-30 % (I)

The 70-30 split appears to be the best choice, as it consistently achieved higher scores across the cross-validation iterations, suggesting better stability and generalization for the Random Forest model.

```
Best Parameters: {'max_depth': 10, 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 100}
Accuracy with Best Parameters: 0.7984886649874056
Classification Report:
              precision    recall  f1-score   support

          No       0.80      1.00      0.89       316
         Yes       0.67      0.02      0.05        81

    accuracy                           0.80       397
   macro avg       0.73      0.51      0.47       397
weighted avg       0.77      0.80      0.72       397
```

Illustration 16. Accuracy metrics.

3.5.1 Cross Validation 70-30 % - Random Forest Hyperparameter tunned

[0.69 - 0.91333333 - 0.92 - 0.95317726 - 0.94983278]

The model's precision in predicting attrition ('Yes') is decent at 67%, but its recall is extremely low at 2%, missing many actual cases. The F1-score for 'Yes' is just 0.05, reflecting a significant imbalance between precision and recall.

3.6 Hyperparameter Tuning 70-30 % (II) – Final Model

3.6.1 Adjusting Class Weights in Random Forest - Automatic Balancing

Using Random Forest, it possible to adjust the class weights to give more importance to the "No" class. This can help the model pay more attention to the minority class.

```
Random Forest Model:
Accuracy: 0.7934508816120907
Classification Report:
              precision    recall  f1-score   support

          No       0.82      0.95      0.88       316
         Yes       0.48      0.17      0.25        81

    accuracy                           0.79       397
   macro avg       0.65      0.56      0.57       397
weighted avg       0.75      0.79      0.75       397
```

Illustration 17. Accuracy metrics.

The model accurately predicts the 'No' class but struggles with the 'Yes' class due to the imbalanced dataset. The high accuracy is misleading as the model shows a strong bias towards the 'No' class, as evidenced by the high recall and F1-score for 'No' and very low recall and F1-score for 'Yes'.

3.6.2 Cross Validation 70-30 % - Random Forest Hyperparameter tunned- Class Weight adjusted with Automatic Balancing

[0.69 - 0.91333333 - 0.92 - 0.94983278 - 0.94314381]

The Random Forest model exhibits varying accuracy across folds, ranging from 0.69 to 0.9498, indicating performance dependence on data subsets. With the exception of the first fold, accuracies exceed 90%, indicating overall effectiveness. The model's final analysis considers the complexity of capturing voluntary attrition factors.

4. References

Shukla, P. (2022). Handling Missing Data with SimpleImputer. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2022/10/handling-missing-data-with-simpleimputer/.

scikit-learn. (n.d.). 6.4. Imputation of missing values. [online] Available at: https://scikit-learn.org/stable/modules/impute.html#:~:text=The%20SimpleImputer%20class%20provides%20basic [Accessed 28 Dec. 2023].

Stack Overflow. (n.d.). Pandas not recognizing NaN as null. [online] Available at: https://stackoverflow.com/questions/30604893/pandas-not-recognizing-nan-as-null [Accessed 28 Dec. 2023].

pandas.pydata.org. (n.d.). pandas.isnull — pandas 1.4.2 documentation. [online] Available at: https://pandas.pydata.org/docs/reference/api/pandas.isnull.html.

note.nkmk.me. (2023). pandas: Find rows/columns with NaN (missing values) | note.nkmk.me. [online] Available at: https://note.nkmk.me/en/python-pandas-nan-extract/ [Accessed 28 Dec. 2023].

saturncloud.io. (2023). How to Check if a Single Cell Value is NaN in Pandas | Saturn Cloud Blog. [online] Available at: https://saturncloud.io/blog/how-to-check-if-a-single-cell-value-is-nan-in-pandas/#:~:text=One%20of%20the%20most%20common [Accessed 28 Dec. 2023].

Vanawat, N. (2021). How To Perform Exploratory Data Analysis -A Guide for Beginners. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/08/how-to-perform-exploratory-data-analysis-a-guide-for-beginners/.

Stack Overflow. (n.d.). ConvergenceWarning: Liblinear failed to converge, increase the number of iterations. [online] Available at: https://stackoverflow.com/questions/52670012/convergencewarning-liblinear-failed-to-converge-increase-the-number-of-iterati [Accessed 29 Dec. 2023].

Li, P., Stuart, E.A. and Allison, D.B. (2015). Multiple Imputation. JAMA, 314(18), p.1966. doi:https://doi.org/10.1001/jama.2015.15281.

scikit-learn.org. (n.d.). 6.4. Imputation of missing values — scikit-learn 0.22.2 documentation. [online] Available at: https://scikit-learn.org/stable/modules/impute.html.

www.linkedin.com. (n.d.). What are the pros and cons of different imputation methods in python? [online] Available at: https://www.linkedin.com/advice/1/what-pros-cons-different-imputation-methods-python [Accessed 1 Jan. 2024].`

builtin.com. (n.d.). How to Do a T-Test in Python | Built In. [online] Available at: https://builtin.com/data-science/t-test-python.

Indeed Career Guide. (n.d.). 20 Career Development Opportunities. [online] Available at: https://www.indeed.com/career-advice/career-development/career-development-opportunities.

pandas.pydata.org. (n.d.). pandas.DataFrame.merge — pandas 1.3.4 documentation. [online] Available at: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.merge.html.

Stack Overflow. (n.d.). Difference between Standard scaler and MinMaxScaler. [online] Available at: https://stackoverflow.com/questions/51237635/difference-between-standard-scaler-and-minmaxscaler [Accessed 1 Jan. 2024].

kaggle.com. (n.d.). Employee Attrition EDA. [online] Available at: https://www.kaggle.com/code/alaaelnakeeb/employee-attrition-eda [Accessed 5 Jan. 2024].

kaggle.com. (n.d.). Employee Attrition Classifications. [online] Available at: https://www.kaggle.com/code/hassanahmed093/employee-attrition-classifications [Accessed 1 Jan. 2024].

Cross Validated. (n.d.). machine learning - Do we normalise the dataset before or after performing one hot encoding? [online] Available at: https://stats.stackexchange.com/questions/473401/do-we-normalise-the-dataset-before-or-after-performing-one-hot-encoding.

GeeksforGeeks. (2020). StandardScaler, MinMaxScaler and RobustScaler techniques - ML. [online] Available at: https://www.geeksforgeeks.org/standardscaler-minmaxscaler-and-robustscaler-techniques-ml/.

Amy @GrabNGoInfo (2023). Balanced Weights For Imbalanced Classification. [online] GrabNGoInfo. Available at: https://medium.com/grabngoinfo/balanced-weights-for-imbalanced-classification-465f0e13c5ad#:~:text=The%20RandomForestClassifier%20in%20sklearn%20has [Accessed 1 Jan. 2024].

Stack Overflow. (n.d.). Proper use of 'class_weight' parameter in Random Forest classifier. [online] Available at: https://stackoverflow.com/questions/58275113/proper-use-of-class-weight-parameter-in-random-forest-classifier [Accessed 3 Jan. 2024].

kaggle.com. (n.d.). Employee Attrition Prediction - Classification. [online] Available at: https://www.kaggle.com/code/valeriamulina/employee-attrition-prediction-classification#Random-Forest-Classifier [Accessed 7 Jan. 2024].

imbalanced-learn.org. (n.d.). BorderlineSMOTE — Version 0.10.1. [online] Available at: https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.BorderlineSMOTE.html.