Your video should be of 10 mins total (strict maximum). Please plan and prepare it accordingly.

This is just a guideline -- feel free to choose your own style, depending on your own data-story.

First 2 minutes (approx.) of your presentation -- *Your Motivation*

- You MUST mention which dataset(s) you used, and what EXACTLY is your problem definition.
- This is your chance to explain your MOTIVATION. Don't get carried away; be precise and clear.

Next 3 minutes (approx.) of your presentation -- *Set The Stage*

- Present your Exploratory Data Analysis and some initial data-driven Insights from the dataset.
- You MAY also mention how you are planning to set up the Analysis / ML problem for this case.
- You MUST mention how you collected / curated / cleaned / prepared the data for this problem.
- Did you only use tools and techniques learned in this course? What ELSE did you learn / try?

Next 3 minutes (approx.) of your presentation -- *Core Analysis*

- If you used ML (regression, classification, or something else); mention mainly WHICH one(s).
- You may now briefly CLARIFY why and how the ML problem(s) aim(s) to solve your objective.
- How did you apply ML technique(s) to SOLVE your problem? Which model(s), how and why?
- Did you only use tools and techniques learned in this course? What ELSE did you learn / try?

Last 2 minutes (approx.) of your presentation -- *Finish Strong*

- What is the OUTCOME of your project? Did it solve your original problem? Anything interesting?
- What are your data-driven INSIGHTS and recommendations / views towards the target problem?

Everything you plan to showcase should be presented within the 10 mins. Plan carefully and smartly.

Present the main aspects of your project and data-story in the 10 mins; extra items may be on GitHub.

No need to mention who in your team worked on which part of the project -- this should be on GitHub.

No need to cite references to other related works in your presentation -- this should be on GitHub too.

First 2 minutes (approx.) of your presentation -- *Your Motivation*

- You MUST mention which dataset(s) you used, and what EXACTLY is your problem definition.
- This is your chance to explain your MOTIVATION. Don't get carried away; be precise and clear.

CS : Change Slide
(https://www.bls.gov/news.release/archives/jolts_03092022.pdf)
Hi, my name is Danish and together with my group mate Andrea, we would be breaking down and tackling the plague that has haunted companies around the world: employee retention and attrition. [CS]In 2021, the US Bureau of Labor Statistics reported a total employee voluntary attrition rate of 25%[CS] for just the United States alone. For reference, [CS]it is recommended for companies to have only a 10% turnover rate. [CS]Therefore, our group's goal is to minimise this casualty rate by leveraging data analytics and machine learning, in order to predict potential cases of attrition. By analysing trends in the data and building predictive models, the HR department can take proactive steps to retain valuable employees and mitigate the negative impact of attrition on organisational performance. Predictive models analyse historical data to identify employees likely to leave, allowing HR to intervene early with retention measures. Understanding attrition factors helps tailor strategies like addressing overtime, job satisfaction, and career growth.

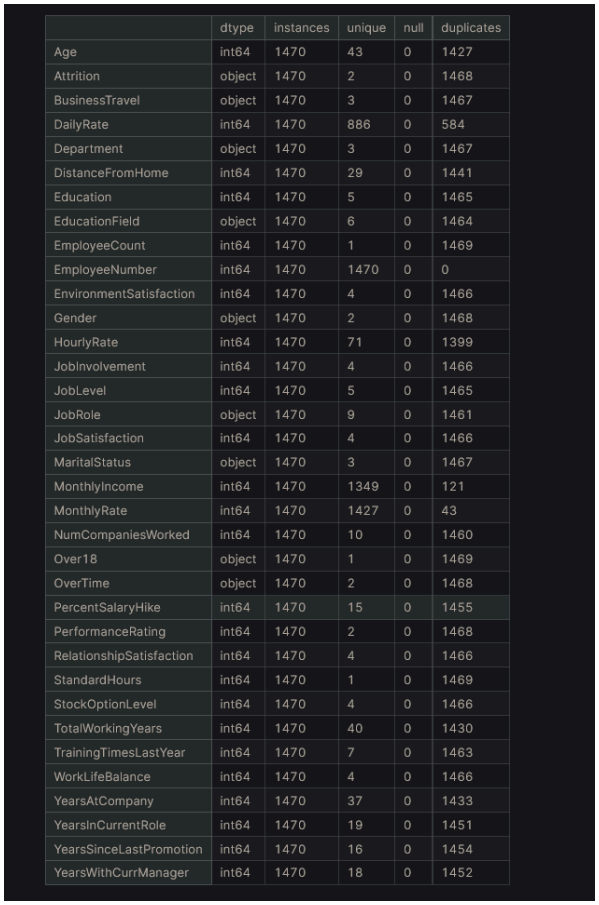
[CS]The dataset we will be utilising is from Kaggle, titled "HR Analytics : Employee Attrition Prediction" and we chose this because it contains organisational data relevant to employee attrition, including various attributes such as age, job role, marital status, education level, etc.

Next 3 minutes (approx.) of your presentation -- *Set The Stage*

- Present your Exploratory Data Analysis and some initial data-driven Insights from the dataset.
- You MAY also mention how you are planning to set up the Analysis / ML problem for this case.
- You MUST mention how you collected / curated / cleaned / prepared the data for this problem.
- Did you only use tools and techniques learned in this course? What ELSE did you learn / try?

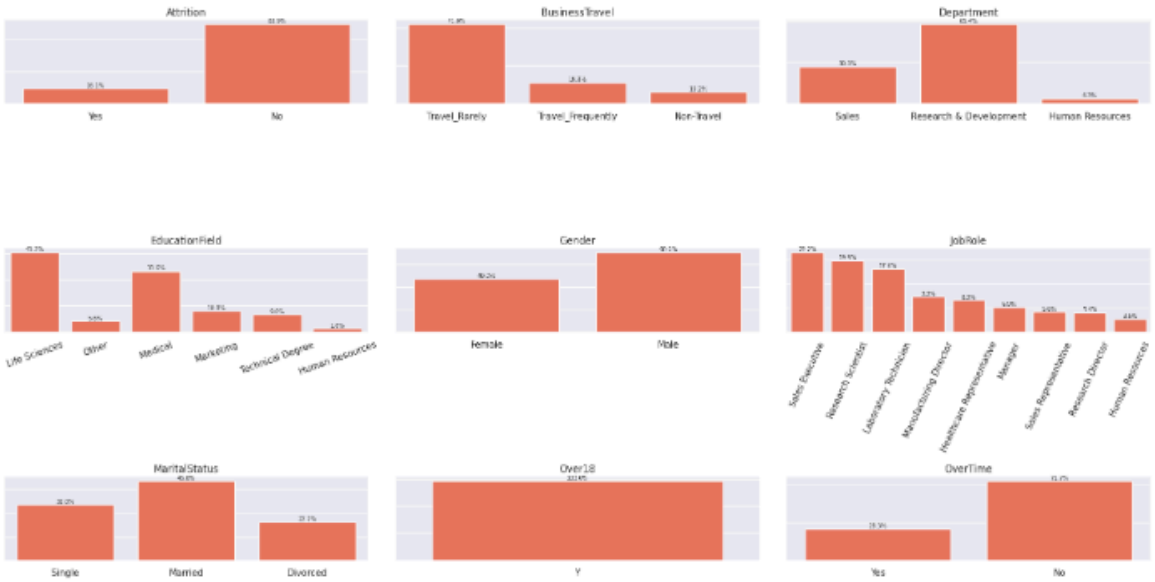
Data Cleaning

[CS]Initially, we tried to do some data cleaning by looking for null values, to which there are none. We then decided to clean our data even more by eliminating redundant columns from our dataset which we will combine with our [CS]exploratory data analysis.



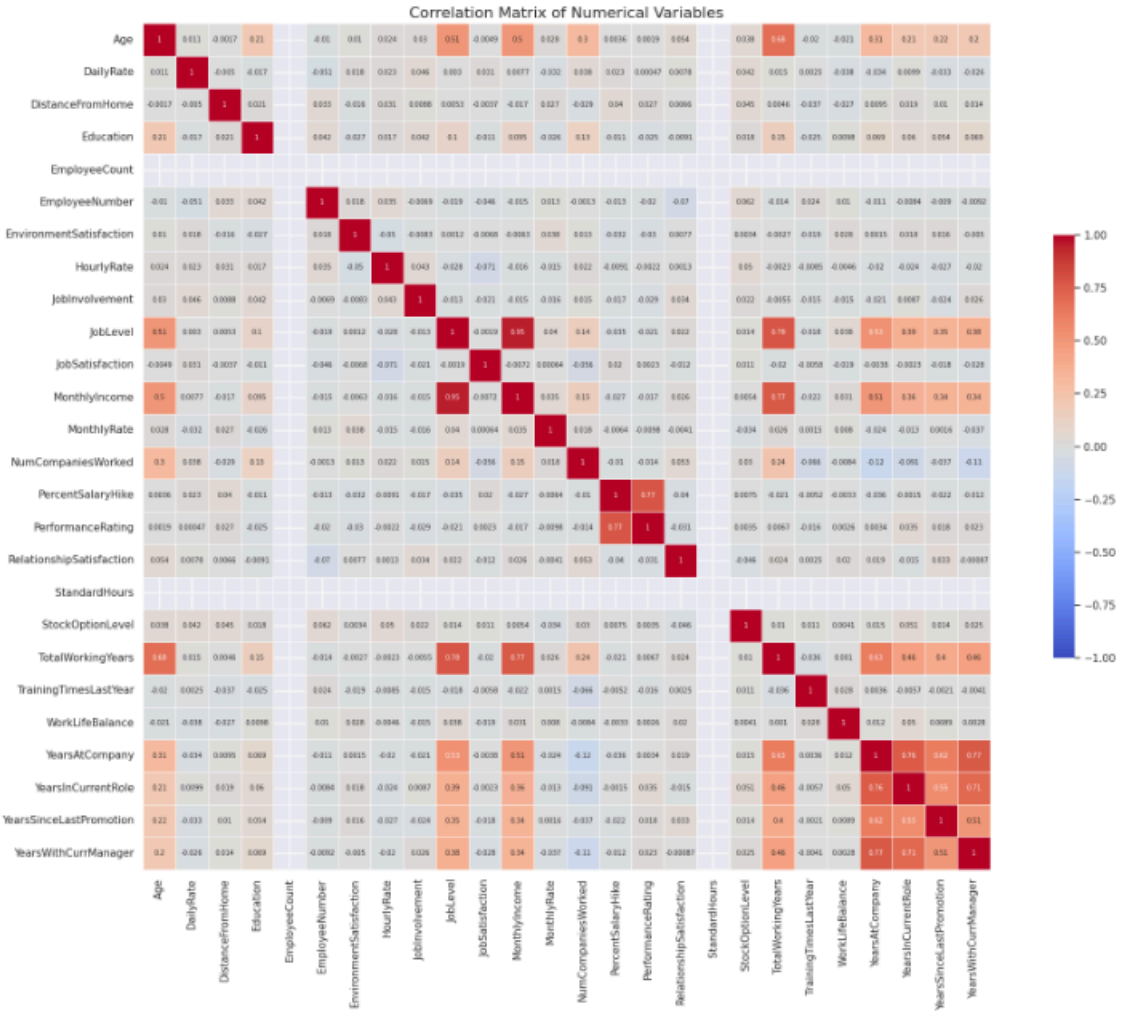
	dtype	instances	unique	null	duplicates
Age	int64	1470	43	0	1427
Attrition	object	1470	2	0	1468
BusinessTravel	object	1470	3	0	1467
DailyRate	int64	1470	886	0	584
Department	object	1470	3	0	1467
DistanceFromHome	int64	1470	29	0	1441
Education	int64	1470	5	0	1465
EducationField	object	1470	6	0	1464
EmployeeCount	int64	1470	1	0	1469
EmployeeNumber	int64	1470	1470	0	0
EnvironmentSatisfaction	int64	1470	4	0	1466
Gender	object	1470	2	0	1468
HourlyRate	int64	1470	71	0	1399
JobInvolvement	int64	1470	4	0	1466
JobLevel	int64	1470	5	0	1465
JobRole	object	1470	9	0	1461
JobSatisfaction	int64	1470	4	0	1466
MaritalStatus	object	1470	3	0	1467
MonthlyIncome	int64	1470	1349	0	121
MonthlyRate	int64	1470	1427	0	43
NumCompaniesWorked	int64	1470	10	0	1460
Over18	object	1470	1	0	1469
OverTime	object	1470	2	0	1468
PercentSalaryHike	int64	1470	15	0	1455
PerformanceRating	int64	1470	2	0	1468
RelationshipSatisfaction	int64	1470	4	0	1466
StandardHours	int64	1470	1	0	1469
StockOptionLevel	int64	1470	4	0	1466
TotalWorkingYears	int64	1470	40	0	1430
TrainingTimesLastYear	int64	1470	7	0	1463
WorkLifeBalance	int64	1470	4	0	1466
YearsAtCompany	int64	1470	37	0	1433
YearsInCurrentRole	int64	1470	19	0	1451
YearsSinceLastPromotion	int64	1470	16	0	1454
YearsWithCurrManager	int64	1470	18	0	1452

Categorical Attributes Analysis:



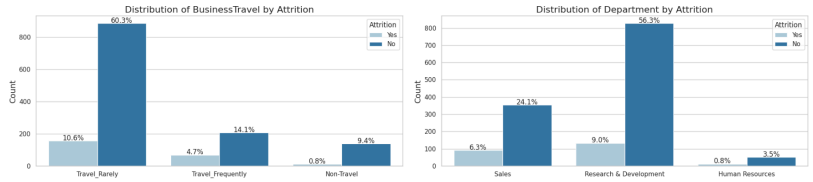
[CS]We first split the dataset into categorical and numerical attributes and produced box plots for each categorical column, and by doing so we are able to notice certain patterns and details without even going through machine learning. We can see that [CS]most employees are from R&D, [CS]they do minimal overtime and [CS]travel rarely.

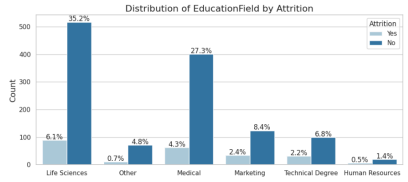
[CS]As the ‘Over18’ column has only one value, it is redundant and will be the first out of many columns that we will begin to remove. [CS] We also notice more non-attrition cases than attrition, that we will have to rectify later on.

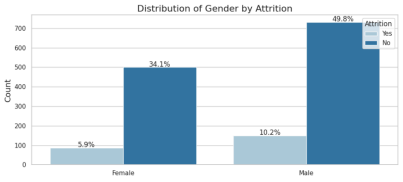


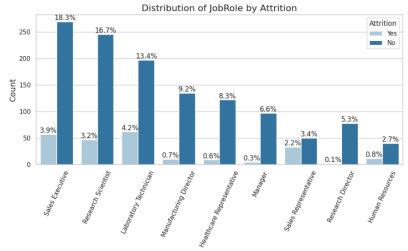
[CS]During our exploration, we conducted a correlation analysis to uncover significant insights[CS] into the relationships between the[CS] numerical variables. From the colours, our analysis revealed positive[CS] correlations amongst ‘Age’,[CS] 'JobLevel', [CS] 'MonthlyIncome.', [CS] and 'TotalWorkingYears' This indicates that individuals with higher job levels tend to command higher monthly incomes. This correlation[CS] underscores the importance of experience and tenure in driving career advancement and financial rewards within the organisation.[CS] Furthermore, our exploration uncovered strong positive correlations among 'YearsInCurrentRole,' 'YearsWithCurrentManager,' and 'YearsAtCompany.' These correlations provide valuable insights into employee tenure and career progression dynamics. [CS]Specifically, prolonged tenure in one's current role appears to align with extended periods working under the same manager and remaining with the company. This observation suggests a symbiotic relationship between employee satisfaction, managerial stability, and organisational loyalty.

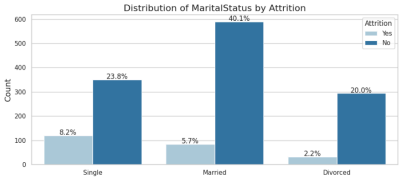
[CS]Next, we created various count plots for each categorical variable in the dataset so we can visualise the distribution with respect to employee attrition.

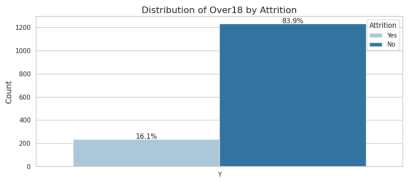


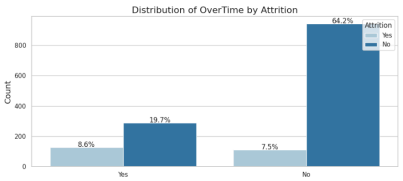






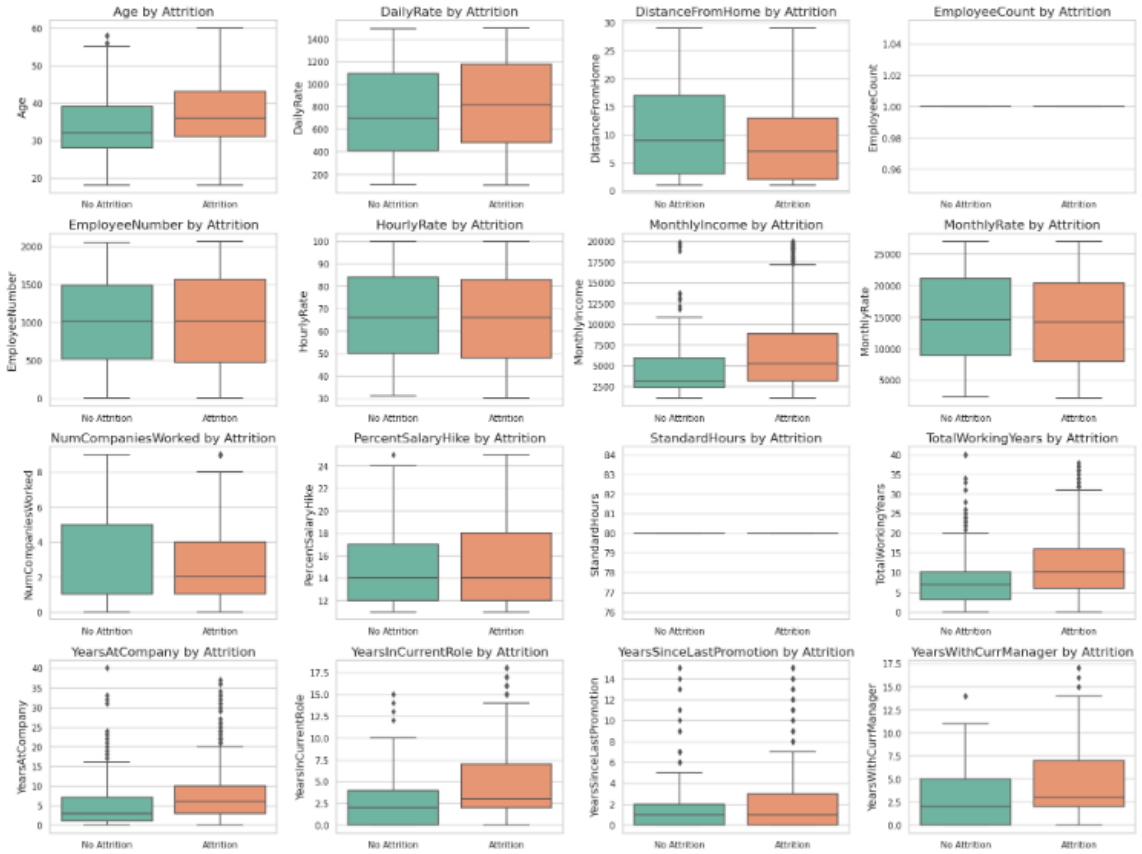






Delving deeper, we can notice that employees who travel less have a lower attrition rate than those who travel frequently, suggesting that frequent travel may contribute to stress, work-life imbalance, and higher attrition rates. Remote working could be a preferred option for many employees. Among departments, Research & Development has the lowest attrition rate, while the sales department has a higher rate. Human Resources has the smallest attrition rate but also the smallest proportion of employees. [CS]In terms of education fields, employees from Life Sciences and Medical backgrounds are the most represented and show relatively low attrition rates. Male employees have slightly higher attrition rates compared to female employees, but the difference is not significant enough to suggest gender discrimination. [CS]Job role analysis reveals higher attrition rates among Research Scientists and Laboratory Technicians, while Research Directors and Managers have the lowest rates. Sales Representatives have the highest attrition rate in sales-related roles. [CS]Single employees show higher attrition rates compared to married and divorced employees, which could be influenced by factors such as work-life balance, job flexibility, or career growth opportunities. [CS]Employees working overtime have a slightly higher attrition rate, implying that long working hours may contribute to employee dissatisfaction. In essence, from these observations, we can see that job roles, presence of overtime and

business travelling schedules are the major important factors in examining attrition and developing retention strategies.



[CS]Upon analysing the box plots of numerical variables and excluding categorical ones with a data type of integer 64, it becomes evident as per mentioned that certain attributes[CS] like EmployeeCount, StandardHours, HourlyRate, EmployeeNumber, and MonthlyRate lack substantial variations that could provide meaningful insights into attrition patterns or contribute effectively to predictive modelling efforts.

In total, the following attributes are being removed from consideration due to their limited variability, relevance, and having only 1-2 unique values:

- PerformanceRating
- StandardHours
- EmployeeCount
- Over18
- EmployeeNumber
- HourlyRate
- MonthlyRate

By dropping attributes with limited variability, we have streamlined our analysis to focus on key factors that offer meaningful insights into employee behaviour. This optimization process enhances the predictive power and interpretability of our models, enabling us to develop more accurate retention strategies to address attrition challenges effectively.

Preparing for ML:

[CS]From the graphs before, you can see that there are many more cases of non-attrition compared to attrition, which may cause the dataset to be imbalanced. Imbalanced datasets can lead machine learning algorithms to perform poorly, as the model tends to be biassed towards the majority. So, we have to first run the modified dataset under a resampling technique, using SMOTE, Synthetic Minority Over-sampling Technique, as our oversampling method. As there are 1470 employees in the dataset, it is a small number. Therefore, we have to oversample to avoid any data loss. The dataset will then bloat the attrition cases as a result. Now we move on to the machine learning part where Andrea will explain more.

Andrea:

Feature Engineering

For feature engineering, we employed one-hot encoding to handle categorical variables. This technique converts categorical data into a binary format, aiding machine learning algorithms in processing it effectively. Without one-hot encoding, numerical labels may be misinterpreted as having a meaningful order, resulting in biassed predictions. Therefore, we opted for one-hot encoding to enable the model to differentiate between different roles and their impacts on attrition.

Model Selection

Logistic Regression

Metric	Training Set	Test Set
Accuracy	0.8801020408163265	0.8673469387755102
Precision	0.7094017094017094	0.6666666666666666
Recall	0.4368421052631579	0.3404255319148936
F1-score	0.5407166123778502	0.4507042253521127

Decision Tree Classifier

Metric	Training Set	Test Set
Accuracy	0.7814625850340136	0.7517006802721088
Precision	0.3562231759656652	0.3088235294117647
Recall	0.4368421052631579	0.44680851063829785
F1-score	0.3924398817966904	0.3652173913043478

Random Forest Classifier

Metric	Training Set	Test Set
Accuracy	0.8545918367346939	0.8639455782312925
Precision	0.7111111111111111	0.8181818181818182
Recall	0.16842105263157894	0.19148936170212766
F1-score	0.2723404255319149	0.31034482758620685

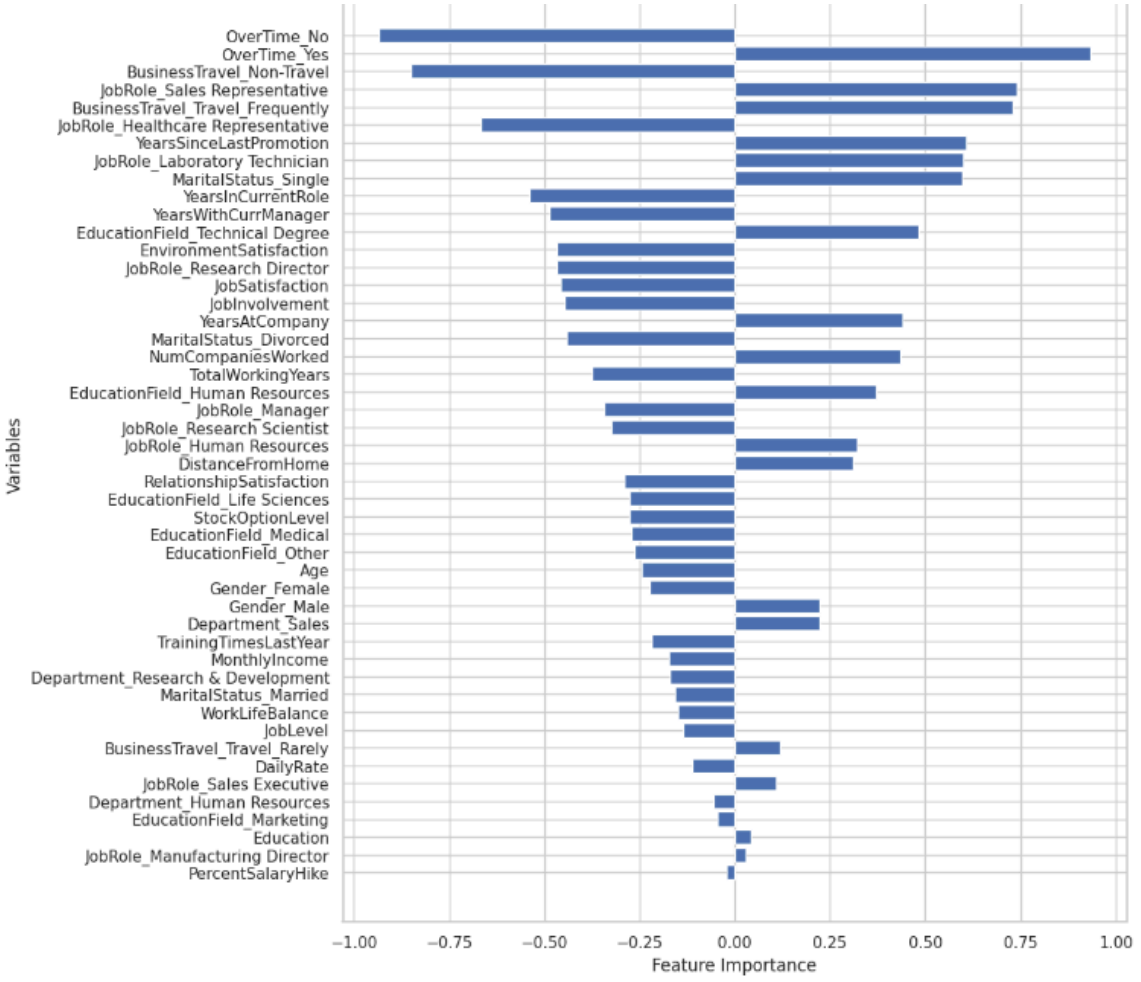
When choosing a model for predicting attrition, it's crucial to consider performance factors. After evaluating several machine learning models, we selected Logistic Regression, Decision Tree Classifier, and Random Forest Classifier as these models are well-established for classification tasks like predicting attrition.

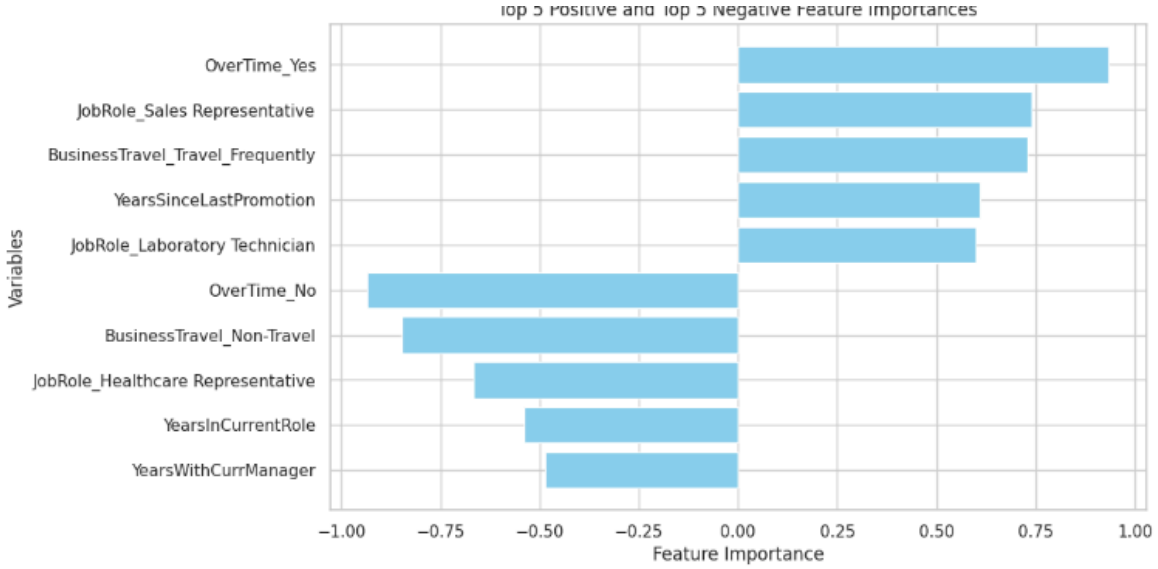
To identify the best model, we will assess metrics such as accuracy, precision, recall, and F1-score to gauge their predictive strengths. Given our focus on predicting attrition, we placed particular emphasis on maximising True Positives while minimising False Negatives. Therefore, we carefully examined the Recall metric, which measures the ability to correctly identify actual positives in the dataset. Among the three algorithms,

the Decision Tree Classifier demonstrates the highest Recall rates for both training and test sets. Although the Decision Tree Classifier seems promising due to its high Recall rates...

its low Precision raised concerns due to the significant number of False Positives.

Consequently, we opted for Logistic Regression, which exhibited higher F1-scores, indicating a balance between Recall and Precision, making it the optimal choice for predicting employee attrition.





Now, we will train a logistic regression model and then evaluate the importance of each feature in predicting attrition. The graph here displays the coefficient of each feature. By examining the coefficients assigned to each feature, it identifies which features have the greatest impact on predicting attrition.

Following that, here displays a graph of the Top 5 positive and negative feature importances.

Based on the graph, Overtime work, frequent business travel and year since last promotion significantly increases the likelihood of attrition. Moreover, certain job roles, like sales representatives and laboratory technician, exhibit higher attrition rates.

Conversely, employees who do not work overtime or travel for business show lower attrition rates, highlighting the importance of work-life balance. Tenure within a role and with a current manager also influences attrition, with longer tenures correlating with lower attrition rates. This emphasises the importance of positive managerial relationships and growth opportunities within the organisation.

```
np.where(y_test==1)
```

```
(array([  5,   6,   7,  11,  14,  15,  20,  36,  52,  59,  61,  79,  97,
        109, 116, 128, 129, 134, 141, 146, 159, 160, 162, 175, 179, 181,
        190, 192, 198, 206, 207, 213, 214, 215, 228, 236, 242, 246, 257,
        265, 267, 270, 273, 274, 278, 284, 289]),)
```

Moving on, let's proceed to assess the probability of employee attrition.

Firstly, we found the indices in the test label where attrition is positive. This helps us identify instances where attrition has occurred. Next, we selected a random employee to make predictions. Using the trained logistic regression model, we predicted the probability of attrition for the selected employee from the test data.

```
Information of employee:

Age                                      20
Attrition                               Yes
BusinessTravel                Travel_Rarely
DailyRate                              129
Department          Research & Development
DistanceFromHome                         4
Education                                3
EducationField             Technical Degree
EnvironmentSatisfaction                  1
Gender                                Male
JobInvolvement                           3
JobLevel                                 1
JobRole                 Laboratory Technician
JobSatisfaction                          1
MaritalStatus                       Single
MonthlyIncome                         2973
NumCompaniesWorked                       1
OverTime                                No
PercentSalaryHike                       19
RelationshipSatisfaction                 2
StockOptionLevel                         0
TotalWorkingYears                        1
TrainingTimesLastYear                    2
WorkLifeBalance                          3
YearsAtCompany                           1
YearsInCurrentRole                       0
YearsSinceLastPromotion                  0
YearsWithCurrManager                     0
Name: 689, dtype: object


The employee has 85.8% chances of attrition.
```

Next, the image here displays information of the selected employee with index 97. This specific index was selected because it was one of the indices identified in which the corresponding employee in the test data has experienced attrition. Using the trained logistic regression model, we estimated their attrition likelihood at (click next) 85.8%. The consistency between the historical attrition occurrence and the model's prediction reinforces the model's reliability.

```
Information of employee:

Age                                30
Attrition                          No
BusinessTravel             Travel_Rarely
DailyRate                        1082
Department                      Sales
DistanceFromHome                   12
Education                           3
EducationField          Technical Degree
EnvironmentSatisfaction             2
Gender                         Female
JobInvolvement                      3
JobLevel                            2
JobRole                 Sales Executive
JobSatisfaction                     3
MaritalStatus                  Single
MonthlyIncome                    6577
NumCompaniesWorked                  0
OverTime                           No
PercentSalaryHike                  11
RelationshipSatisfaction            2
StockOptionLevel                    0
TotalWorkingYears                   6
TrainingTimesLastYear               6
WorkLifeBalance                     3
YearsAtCompany                      5
YearsInCurrentRole                  4
YearsSinceLastPromotion             4
YearsWithCurrManager                4
Name: 402, dtype: object


The employee has 16.3% chances of attrition.
```

In addition, we had also selected an employee who has not experienced attrition to assess their probability of attrition. Despite not having a history of attrition, the logistic regression model predicted a 16.3% chance of attrition for this employee. This prediction underscores the model's ability to identify potential attrition risks even among employees with no prior history of leaving the company.

Conclusions:

In conclusion, this analysis offers valuable insights into the factors influencing employee attrition within the organisation. Employees working overtime, frequently travelling for business, or experiencing delays in promotions are more likely to consider leaving. Conversely, employees that experience no overtime, minimal business travel, and longer tenures in roles or with managers are less likely to leave. Understanding these factors empowers HR to tailor retention strategies effectively and mitigate turnover risks.

Recommendations:

Here are some recommendations that we have for the company. Firstly, the company should evaluate the workload and overtime policies to reduce stress and improve work life balance. Next, business travel requirements and their impact on employee's job satisfaction and retention should be reviewed. Following that, reasons behind high attrition rates in Sales Representative and Laboratory Technician roles should be researched in order to develop targeted retention programs. Furthermore, programs to provide regular promotions and career growth opportunities could be implemented as well. Lately, the company should recognize the importance of supportive managers and encourage longer tenure with the same manager to improve retention.