CAPSTONE PROJECT PRESENTATION

Analytics Systems

A Final Project on Analytics Systems

- Analytics Systems (ALY6140)

- **Guided by:** Vivian Clements Edwin

- **Submitted by:** Group – 2

- **Date of submission :** 28th March’2023

Northeastern University

AGENDA

1

INTRODUCTION

- Aim
- Dataset

2

DATA CLEANING

- Initial Data Analysis
- Business Questions with visualizations

3

MACHINE LEARNING MODELS

- Techniques
- Models with results

4

CONCLUSION

- Comparison

INTRODUCTION

- The organization's largest challenge is choosing the right candidates for promotion and getting them ready on time.
- Choosing set of employees based on recommendations or prior performance and then training them and based on several factors deciding if he/she should be promoted is time consuming.
- Therefore, machine learning techniques are essential to overcome this.

Aim *: In this project, we will use predictive analytics to identify the individuals who are most likely to be promoted based on a variety of factors.*

DATASET

- The information about the recruitment channel can help the company **identify** the most effective channels **for hiring new employees**, while information about employee education and training history can help the company identify areas where **additional training is needed**.

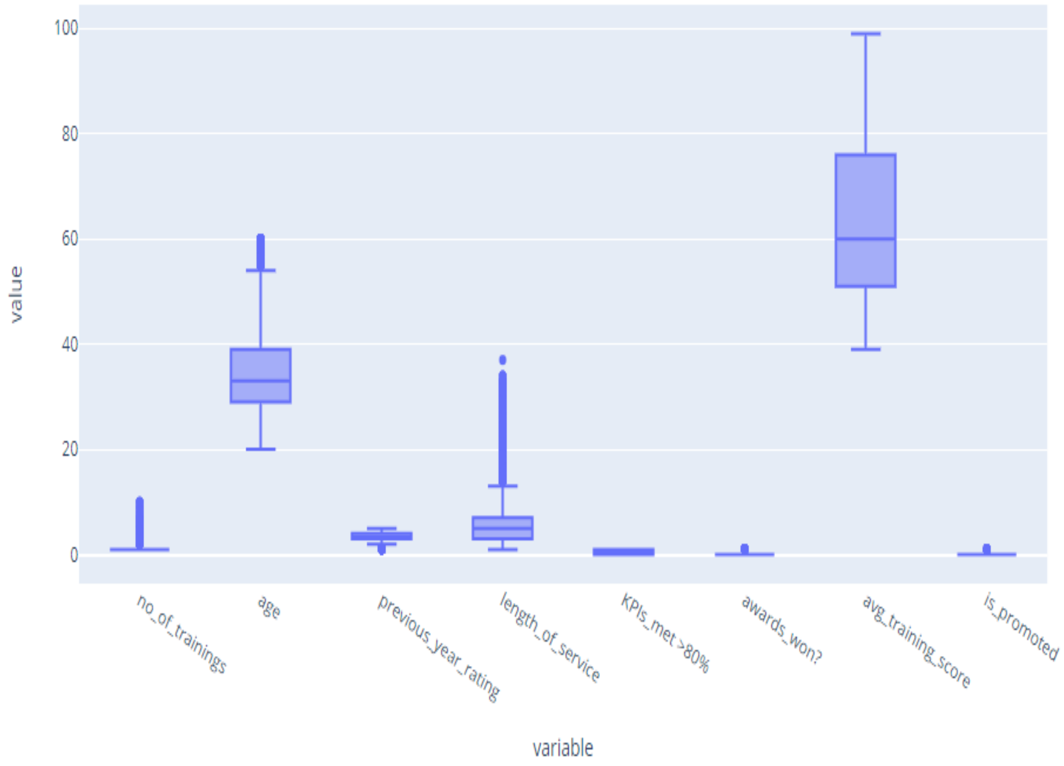
- Moreover, the dataset can also be used to **develop strategies** for improving employee retention and performance as well as departments or regions where the employees are not performing well.

- The data **can be used to analyse** various aspects of employee **performance** and **behaviour**, such as their training history, **KPIs met**, and **awards won**, which can help identify the best-performing employees.

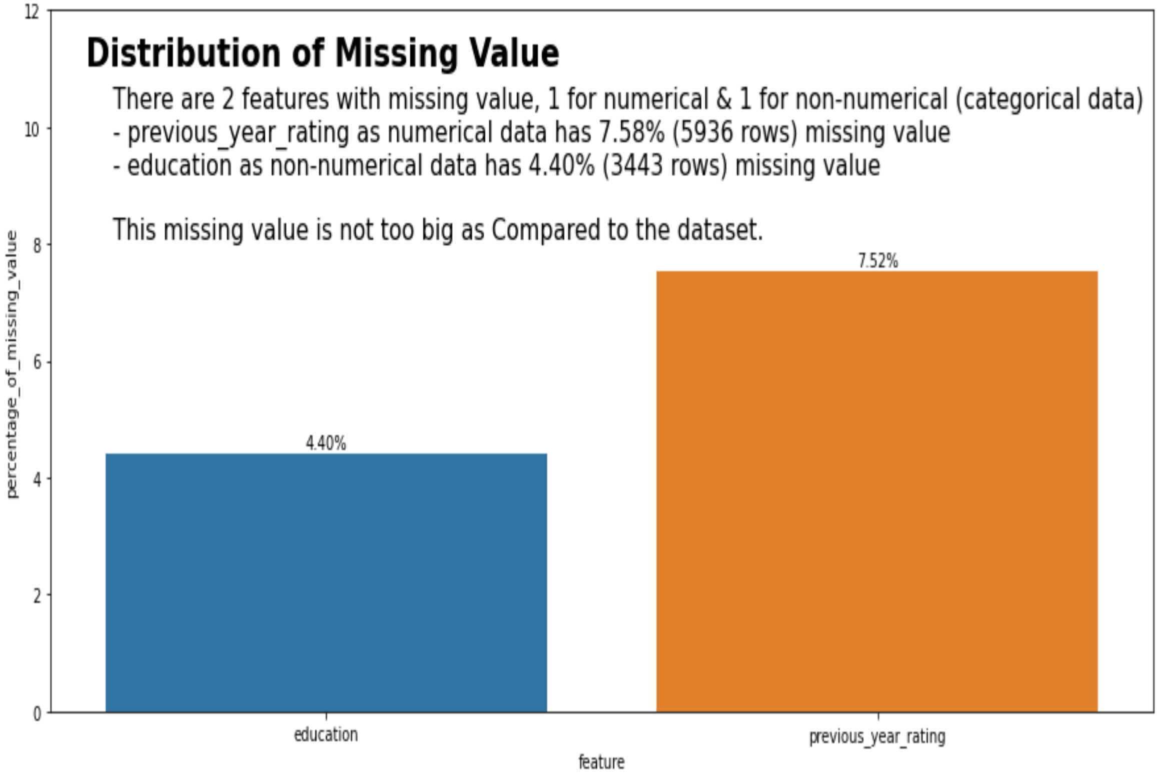
- The given dataset provides information about the **employees of a company**, including various demographic and job-related attributes. The dataset consists of **14 columns**, with each row representing a single **employee's information**.

- The Dataset contains information about the employees of a company, including their **department, region, education, gender, recruitment channel, number of trainings attended, age, previous year rating, length of service, KPIs met greater than 80%, awards won, average training score**, and whether or not they were **promoted**.

DATA CLEANING



- Checking for outliers is an important step to ensure that the data we have is accurate and does not contain any unusual values.
- Upon analysing the dataset, it was found that there **were some outliers present.**
- However, upon further investigation, it was determined that **these outliers were valid and did not have any impact on the dataset or its analysis**.



Distribution of Missing Value

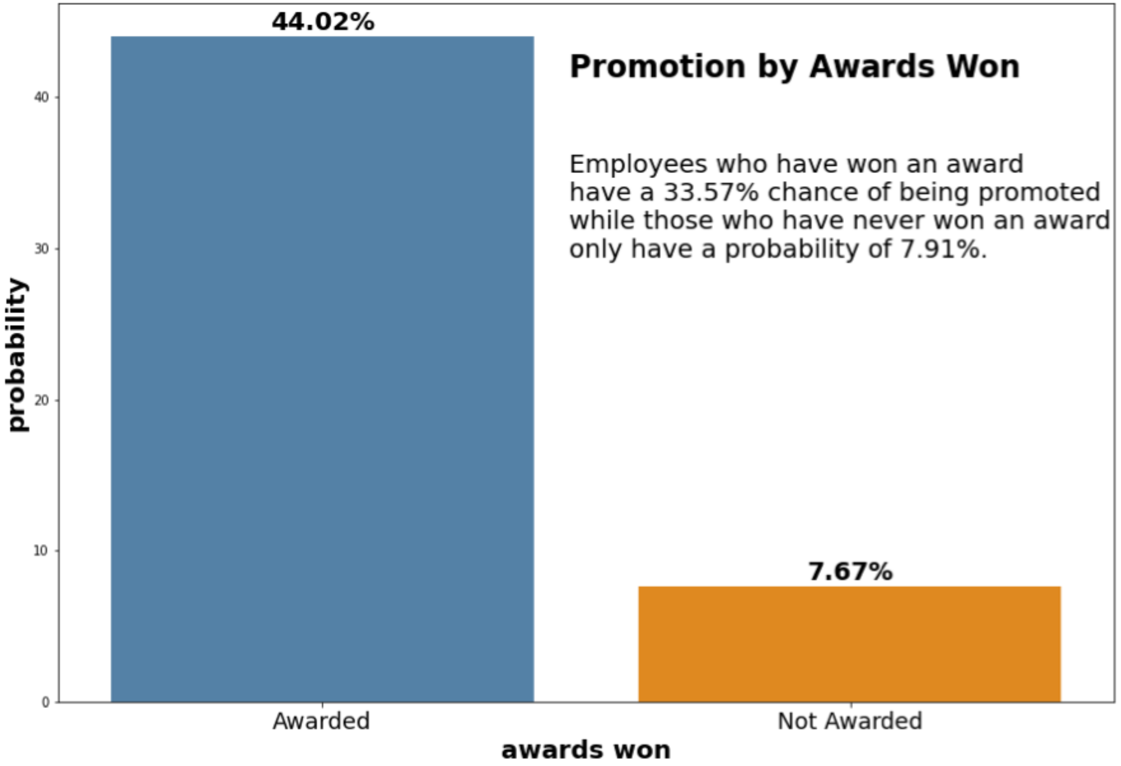
There are 2 features with missing value, 1 for numerical & 1 for non-numerical (categorical data)
- previous_year_rating as numerical data has 7.58% (5936 rows) missing value
- education as non-numerical data has 4.40% (3443 rows) missing value

This missing value is not too big as Compared to the dataset.

- The bar chart shows us how many missing values we have in the dataset. We found that only **two columns have missing values: education and previous year rating.**
- For the education column, we **filled in the missing** values with the **modes** value in the column because only a small percentage of the data was missing.
- For the previous year rating column, which has numerical values, we **filled in the missing** values with the **median value** of the column, since only a small percentage of data was missing in that column too.

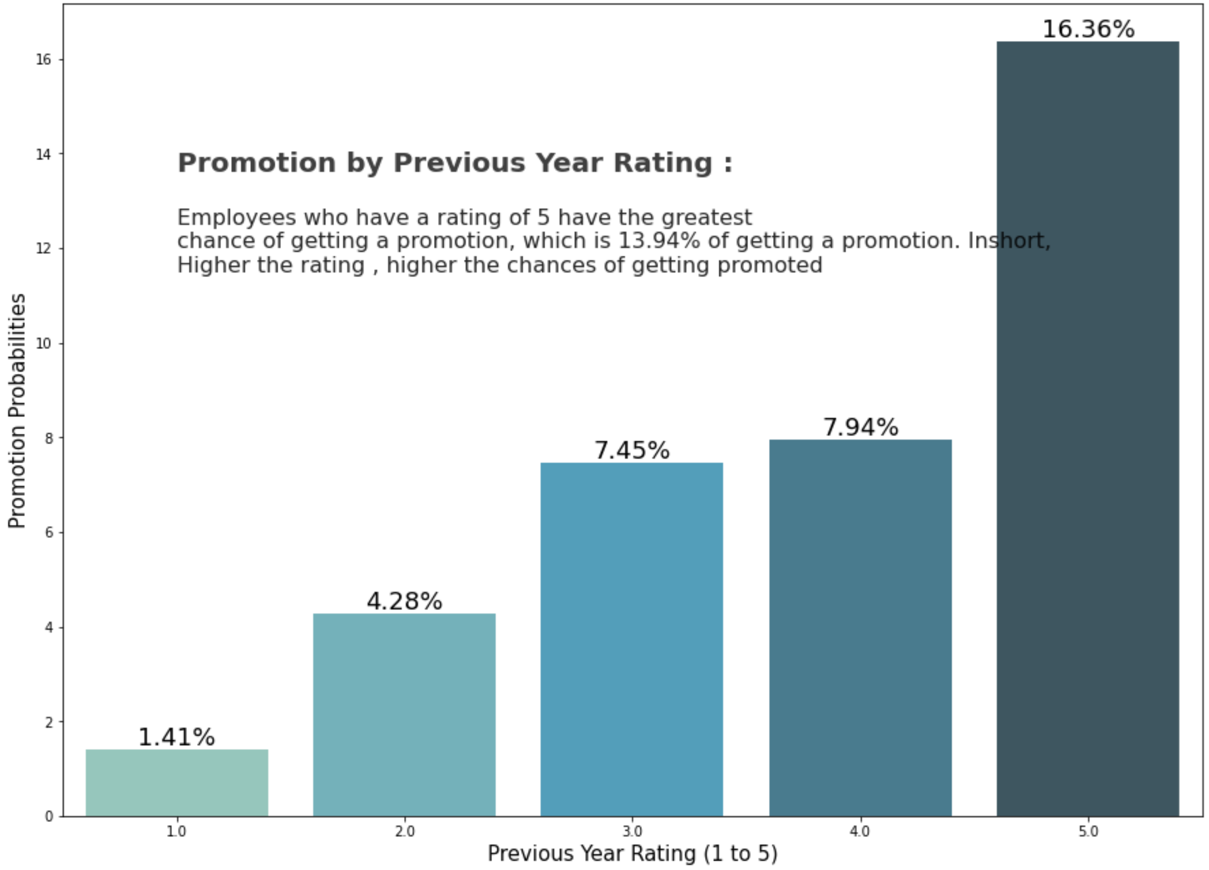
VISUALIZATIONS

Does winning awards for the company have an added advantage for getting promoted?



- The plot we saw earlier shows that employees who have received an award are more likely to be promoted.
- The chance of being promoted for those **who have received an award is 44.02%,** while it is **only 7.67% for those who have not received an award**.
- It's important to remember that receiving an award means the employee has done something exceptional that has benefited the company. It takes a **lot of hard work to receive recognition** from the company.

1. Does previous years rating affects in getting promoted?

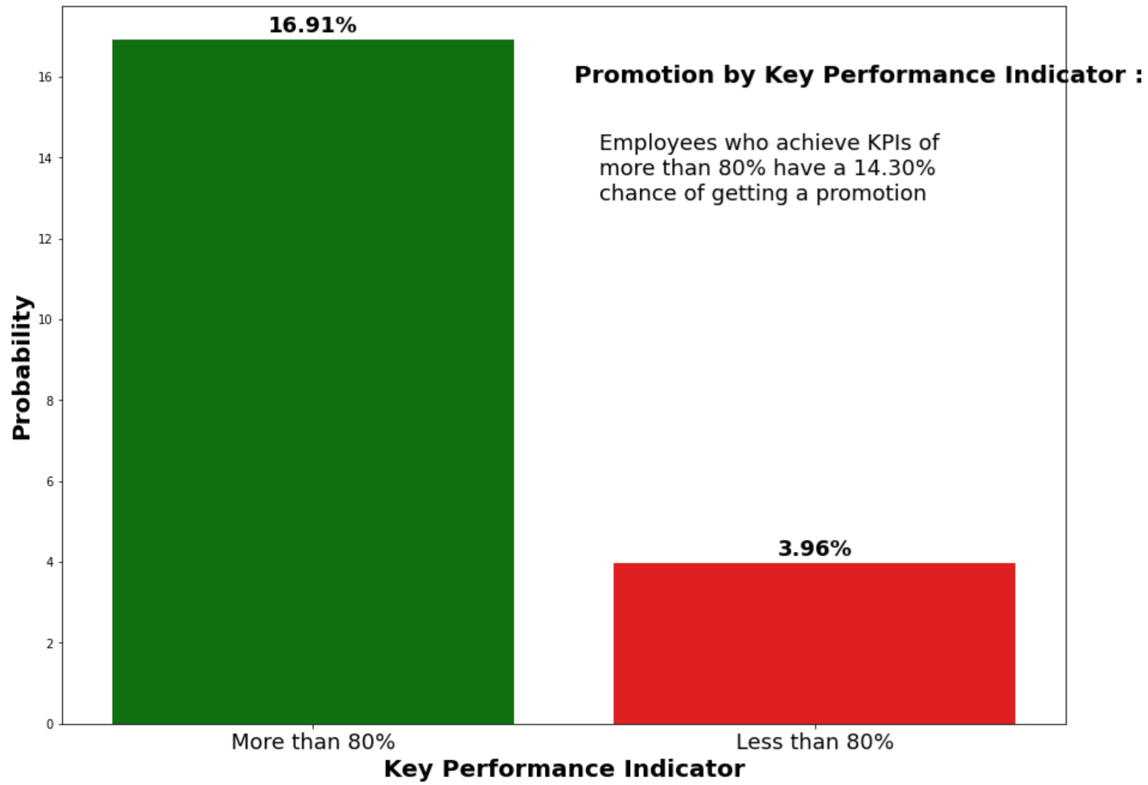


Promotion by Previous Year Rating :

Employees who have a rating of 5 have the greatest chance of getting a promotion, which is 13.94% of getting a promotion. Inshort, Higher the rating , higher the chances of getting promoted

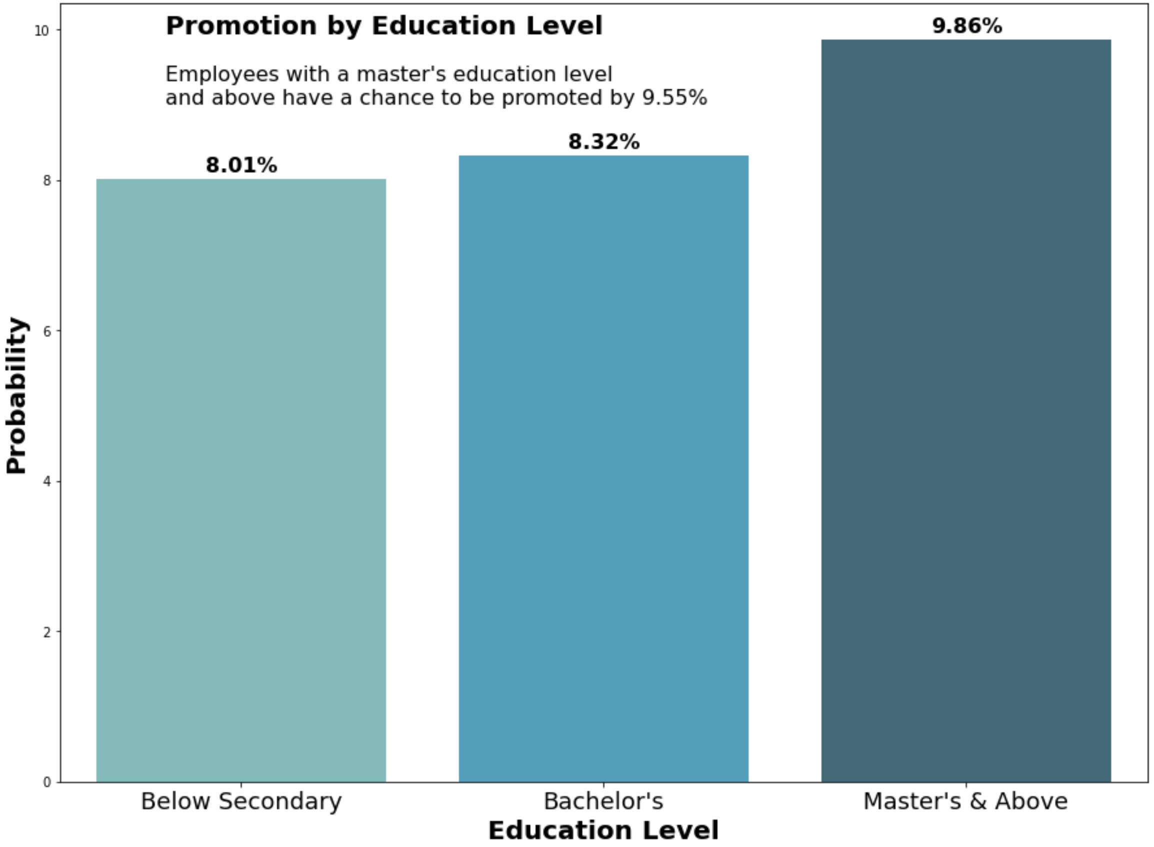
- The graph shows how well employees were rated in the past. **Ratings range from 1 to 5**, with 5 being the best. If an **employee got a 5, there is a 16.36% chance they'll get promoted**, but it's not guaranteed.
- Sometimes even the **best-rated employees don't get promoted** because of factors like **long tenure**, **management positions, or previous promotions**. A rating of 3 means that around 45% of employees do their job adequately but don't go above and beyond.

2. What is the probability of getting promoted based on KPI?



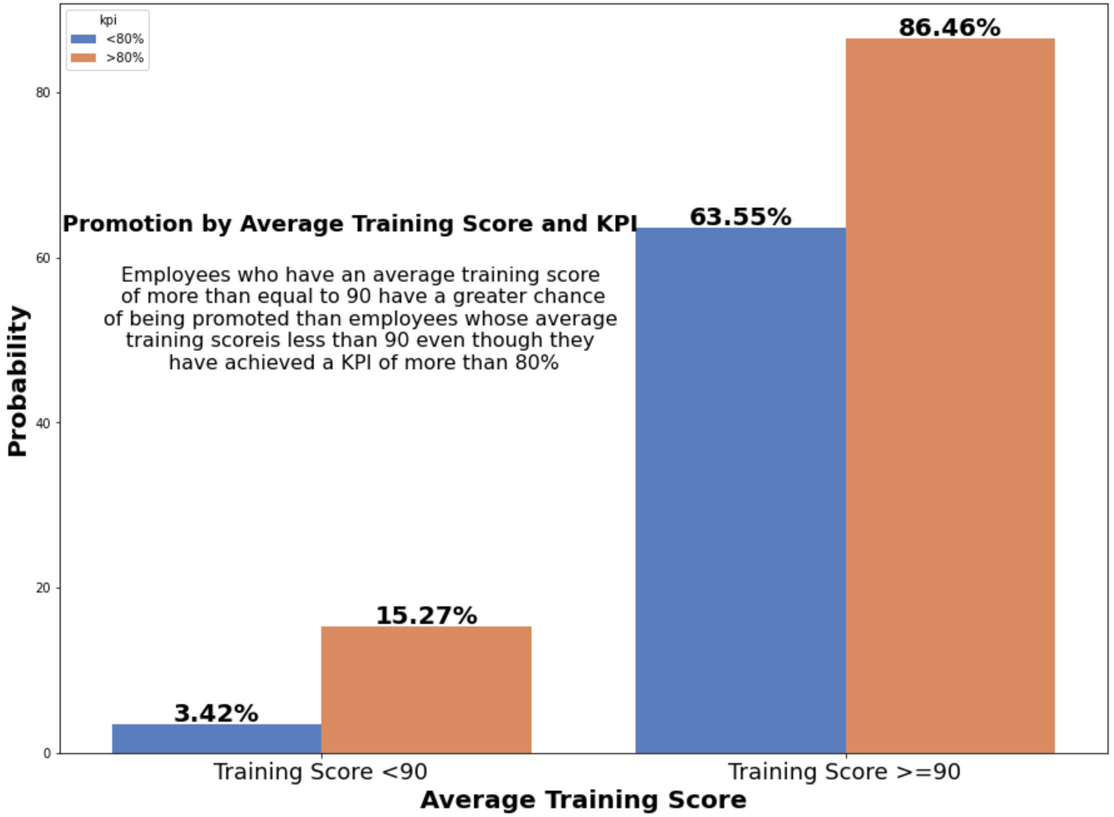
- The bar graph shows a performance measure called **KPI and how it affects promotion chances.** If KPI is **above 80%,** the **chance of promotion is 16.91%**, but if it's **below 80%, the chance drops to 3.96%.**
- However, some people with a KPI **above 80% were not promoted**, maybe because of their high position or low training score.
- Similarly, some with a KPI **below 80% were promoted**, possibly because of their **good training score.**

3. Can a person's education affect their chances of promotion?



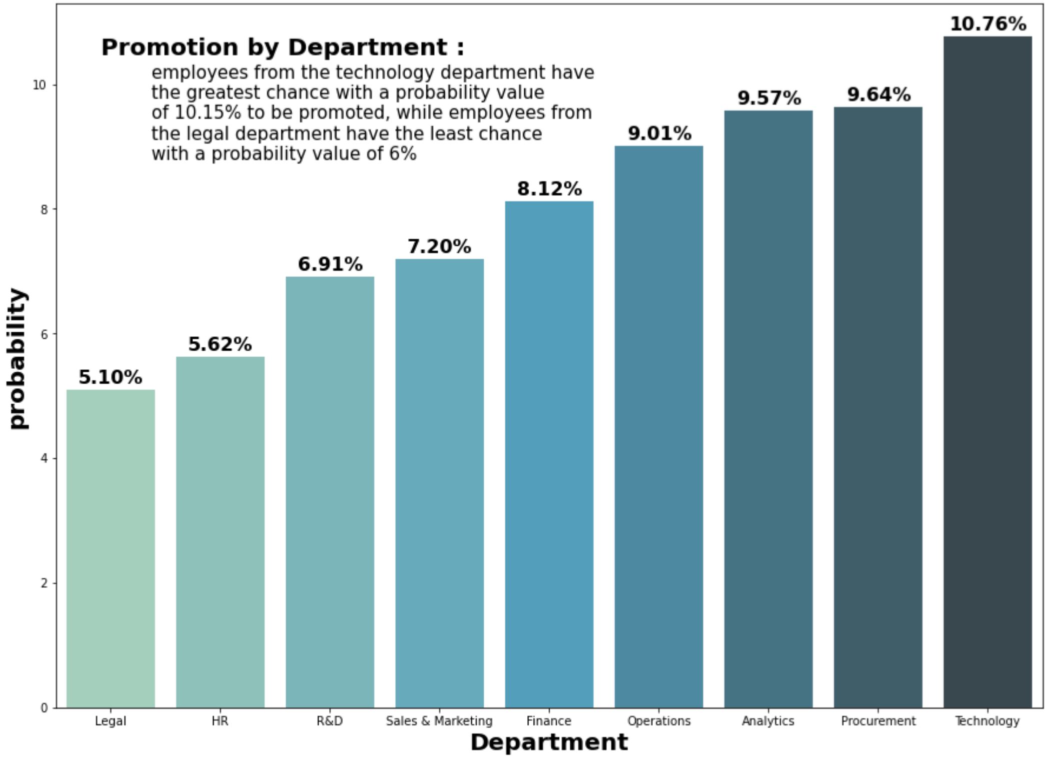
- The bar graph shows the education level of employees. About **75%** of them had a **bachelor's degree** and were **hired through campus placements** or other processes.
- Those who have a **master's degree** or higher have a **9.86% chance** of being **promoted**, but not everyone can afford to pursue higher education.
- Employees with qualifications lower than secondary level usually get lower-level positions like sales and marketing and have a lower chance of being promoted due to their limited exceptional skills compared to their educational background.

4. How can we enhance career growth and use data analytics to identify potential candidates for promotion within our company?



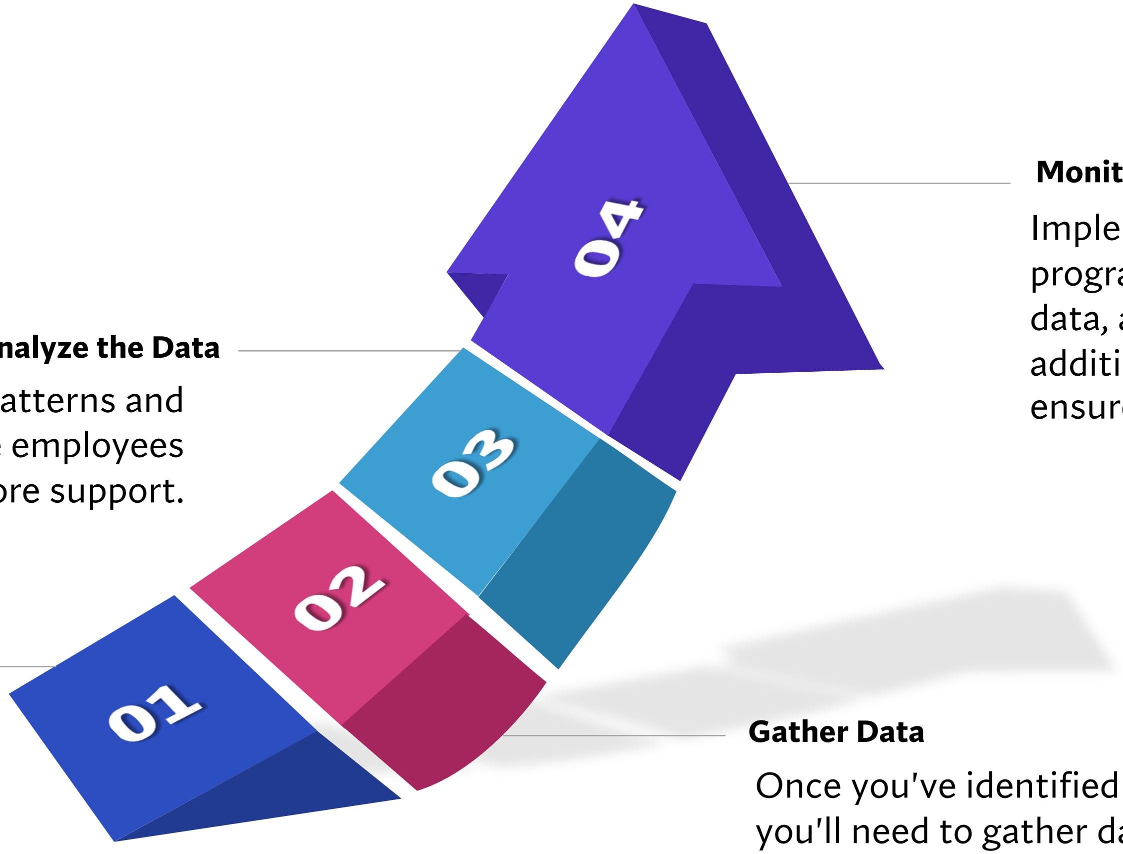
- If an employee has a high key **performance score above 80%,** but their **training score is below 90**, they still have a **good chance** of being **promoted at 13%.**
- However, employees who **score well** in **both domains** have the **highest probability** of being promoted, around 63%.
- It is **important** to have a **high KPI score for promotion**. Data analytics can help in the promotion process for employees.

5. Does a department's importance in promotion play a part?



- The bar plot shows the likelihood of **employees being promoted based on their department**, with the legal department having the **lowest chance at 5.10%, technology** having the **highest at 10.76%, and operations and sales/marketing** having a strong chance due to their larger number of employees.

6. How can we use data to assess and improve the skills and abilities of our employers that are essential for higher-level role?



Monitor Progress

Implement training and development programs, monitor progress with data, and identify areas where additional support may be needed to ensure the desired impact.

Analyze the Data

Analyse the collected data to find patterns and trends, including both areas where employees excel and where they may need more support.

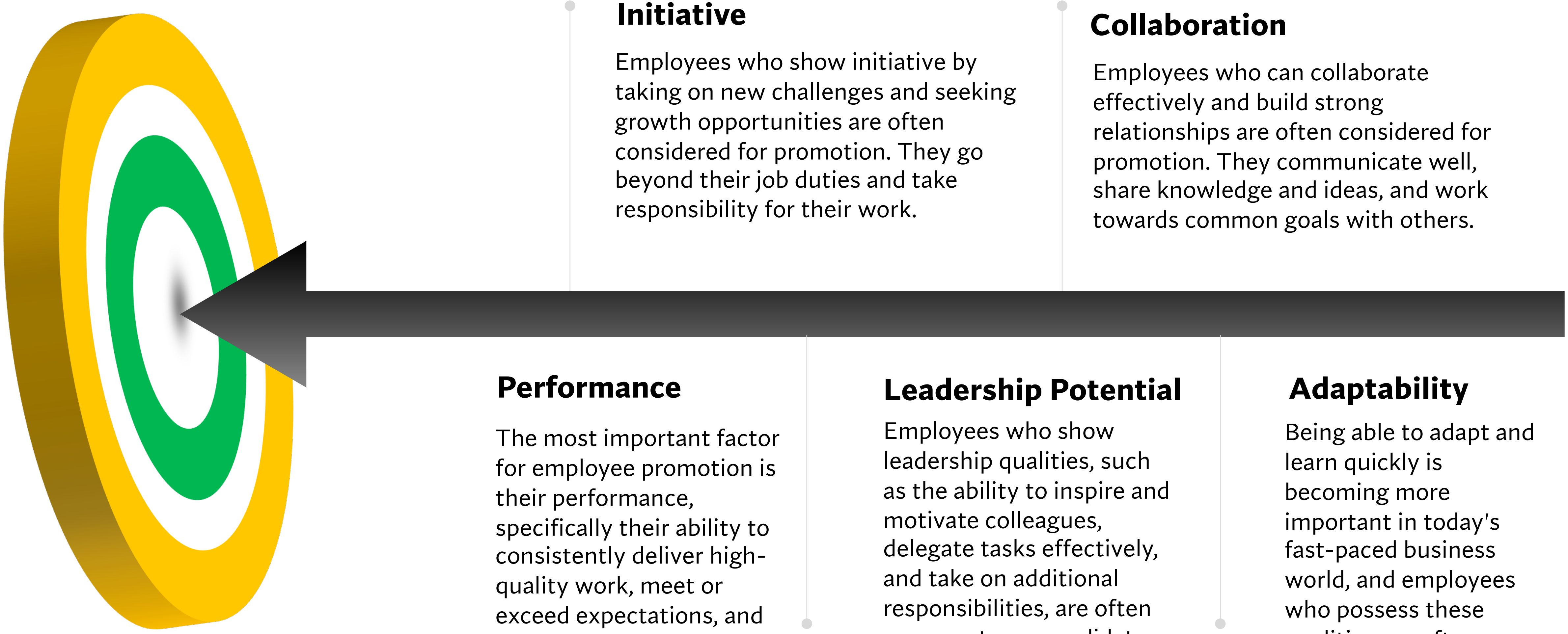
Identify the Essential skills

Start by identifying the key skills and abilities that are essential for success in higher-level roles in your organization. This could include skills like leadership, strategic thinking, communication, and problem-solving.

Gather Data

Once you've identified the essential skills, you'll need to gather data to assess how well your employees are performing in each area. This could involve conducting surveys, collecting performance reviews, or using other tools to gather feedback.

7. What are the key elements that tend to have a significant impact on employee promotion, despite variations based on industry, company culture, and job demands?



Initiative

Employees who show initiative by taking on new challenges and seeking growth opportunities are often considered for promotion. They go beyond their job duties and take responsibility for their work.

Collaboration

Employees who can collaborate effectively and build strong relationships are often considered for promotion. They communicate well, share knowledge and ideas, and work towards common goals with others.

Performance

The most important factor for employee promotion is their performance, specifically their ability to consistently deliver high-quality work, meet or exceed expectations, and achieve their goals

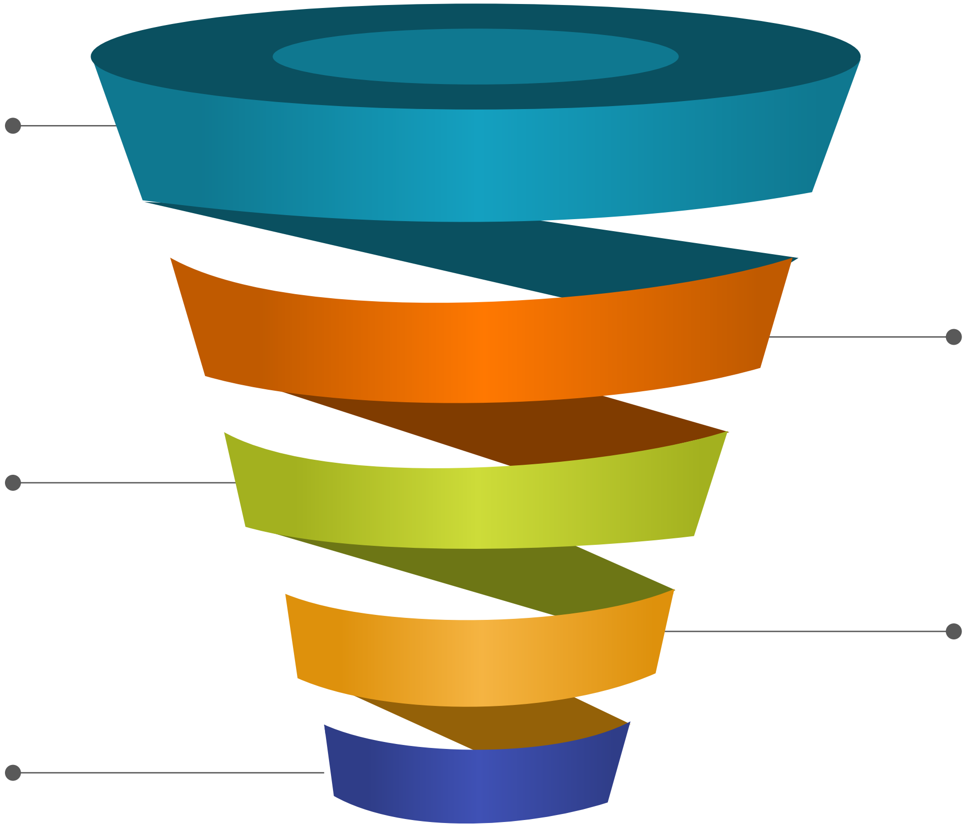
Leadership Potential

Employees who show leadership qualities, such as the ability to inspire and motivate colleagues, delegate tasks effectively, and take on additional responsibilities, are often seen as strong candidates for promotion.

Adaptability

Being able to adapt and learn quickly is becoming more important in today's fast-paced business world, and employees who possess these qualities are often considered for promotion.

8. What are the common performance patterns of employees who have succeeded in? High-level positions?



Strong Work Ethic

Successful employees have a strong work ethic, which means they are hardworking, focused, and determined to succeed. They are dedicated to their work and are willing to put in extra effort to reach their goals.

Adaptability

Being adaptable and flexible is crucial for success in higher-level positions, as successful employees are able to adjust their strategies and approaches to achieve their objectives.

Leadership Skills

Successful employees have strong leadership skills, inspiring, motivating and guiding others, while working collaboratively with their team and delegating tasks effectively.

Strategic Thinking

Successful employees think strategically, seeing the big picture and identifying opportunities for growth, and can develop and execute plans aligned with the organization's goals.

Continuous Learning

Successful employees are committed to ongoing learning and development, actively seeking out new knowledge and experiences to improve their skills and achieve success.

MACHINE LEARNING MODELS

Logistic Regression

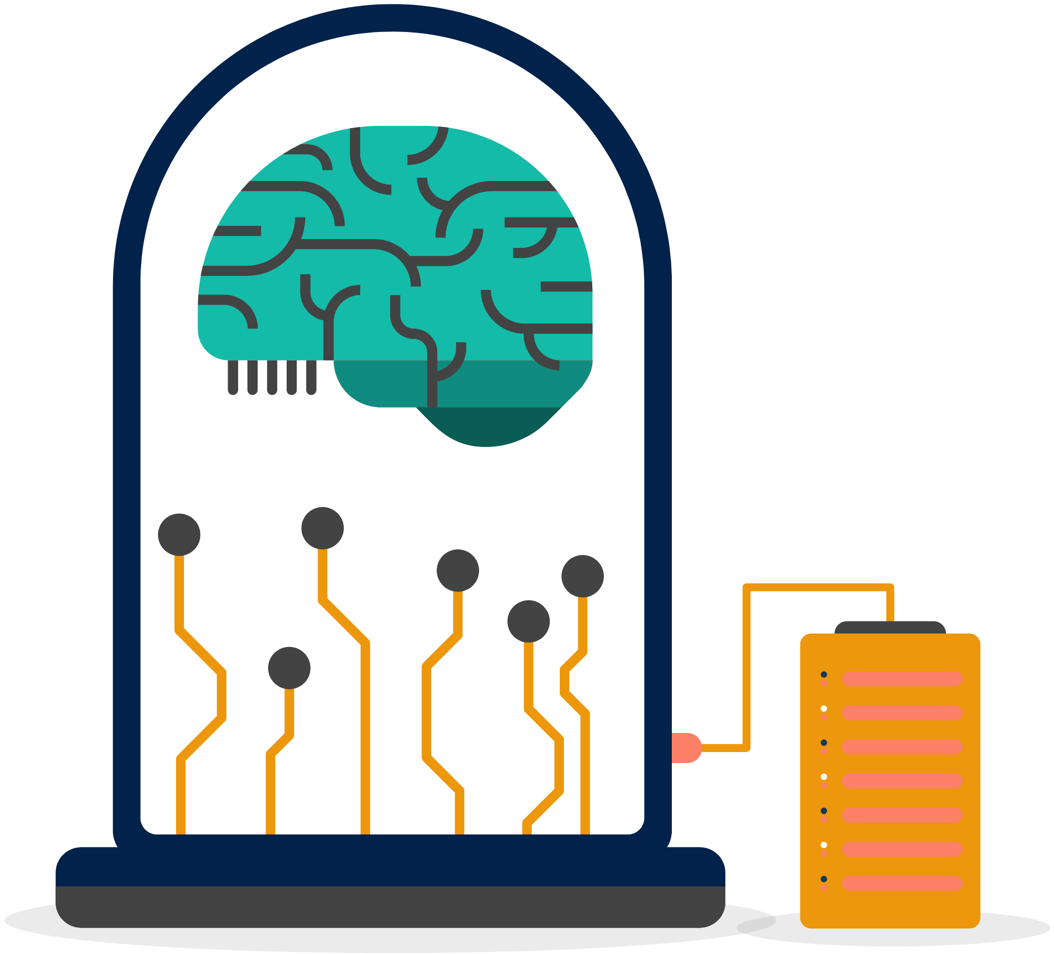
Logistic regression is a statistical method used to predict the likelihood of a binary outcome (such as yes/no or true/false) based on input features.

Random Forest

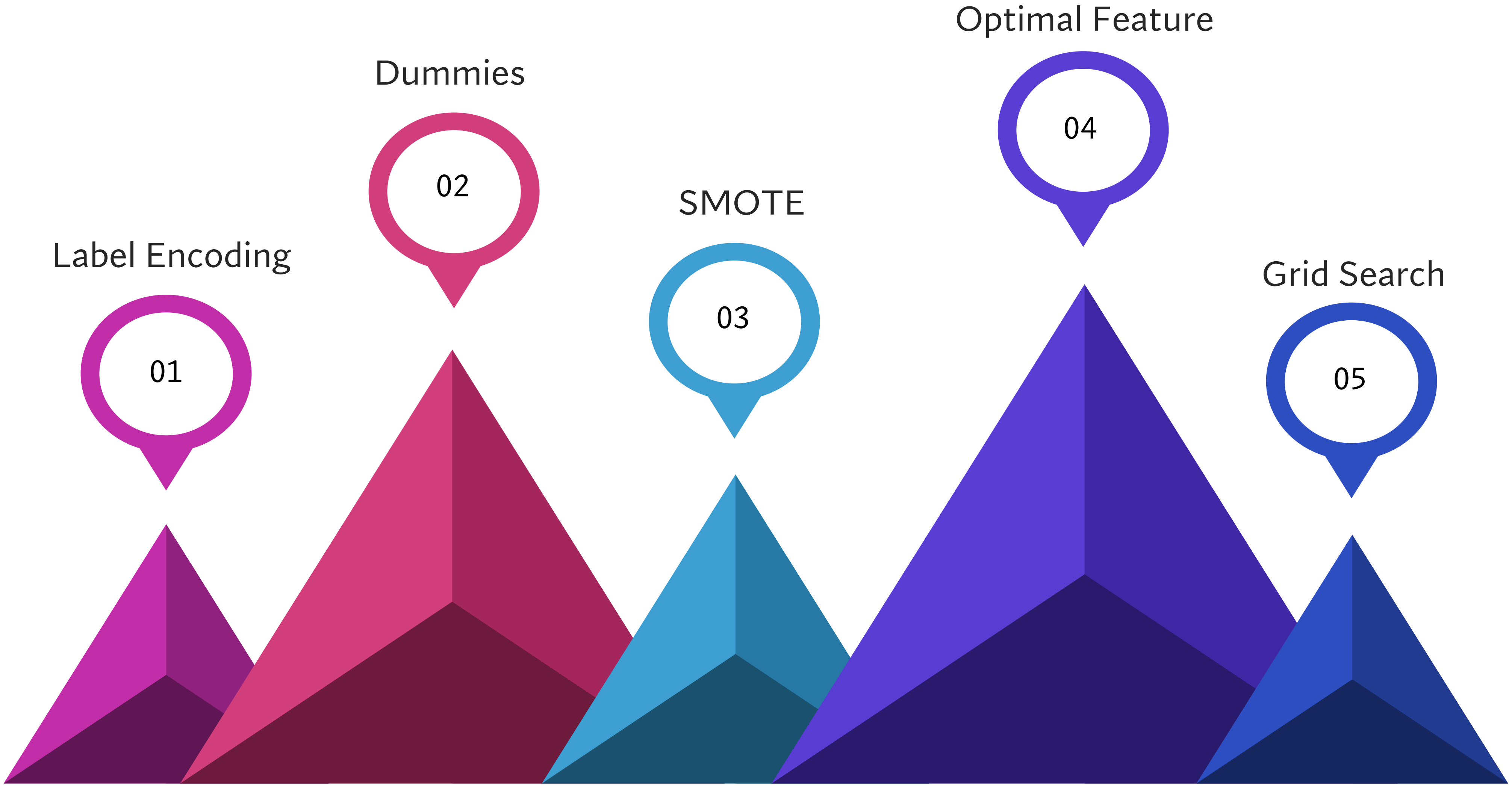
Random Forest is a type of machine learning algorithm that creates multiple decision trees and combines their predictions to make a more accurate and stable prediction.

XGBoost

XGBoost is a machine learning algorithm that combines multiple decision trees to make accurate predictions for classification and regression tasks, known for its speed and scalability.



POWERFUL DATA PREP TECHNIQES FOR ML MODELS



01 Label Encoding

02 Dummies

03 SMOTE

04 Optimal Feature

05 Grid Search

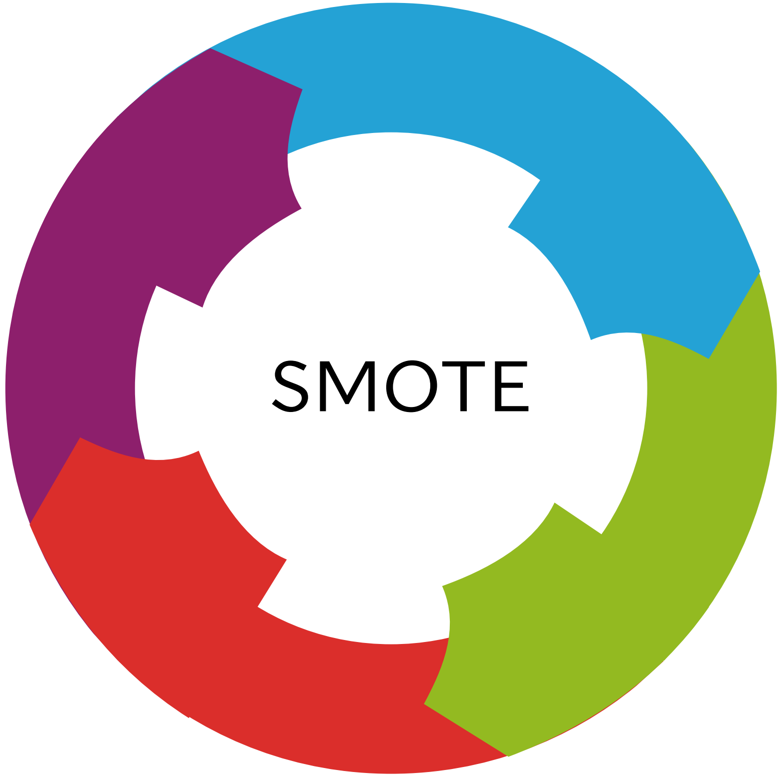
Outcome

We can **balance** the number of **samples in each class** and improve the **performance** of our machine learning model.

Where it can be use ?

Tuning hyperparameters using grid search can help **improve the accuracy, precision, or recall of our model,** leading to better overall performance.

To optimize model performance



About

SMOTE (Synthetic Minority Over-sampling Technique) is a data augmentation method used in machine learning to address the problem of **imbalanced datasets**.

How does it work?

It works by **generating synthetic** examples of the minority class by **interpolating new points between existing ones.**

About

Grid search is a technique used in machine learning to **find the best combination of hyperparameters** for a **model**.

How does it work?

It works by

- **Defining a grid** of possible hyperparameter values,
- **Training and evaluating** the model for each combination of values and
- **Selecting the combination** that performs the best.

Where it can be use ?

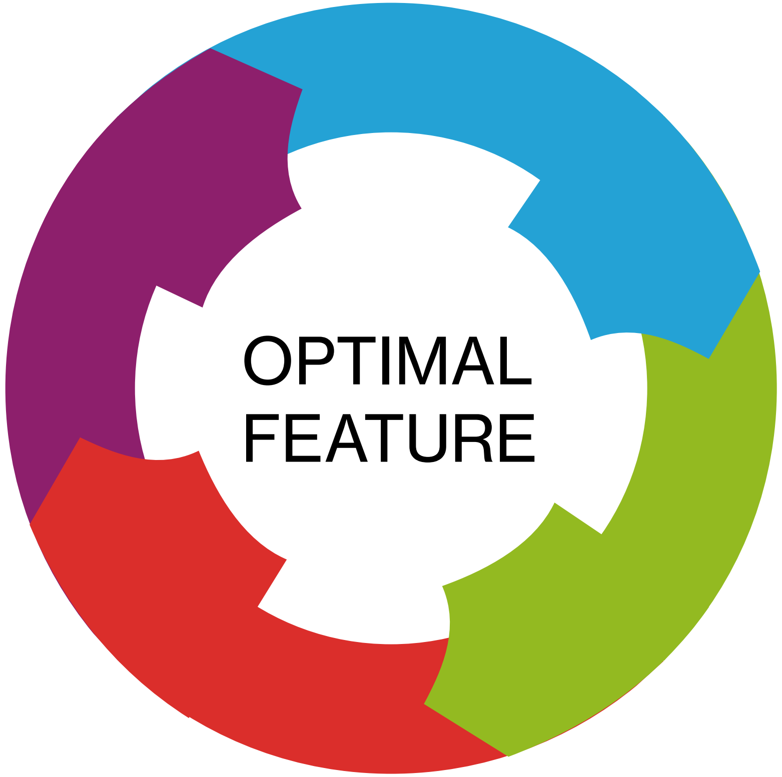
When the **number of samples in one class is significantly lower** than the number of samples in the other class

Outcome

Grid search helps **to identify the best set of hyperparameters** for a given model. This optimal combination of hyperparameters can then be **used to train a final model.**

Outcome

It **filter** out **irrelevant or redundant** features and **ranking or scoring** the remaining features based on their **ability** to improve model performance.



About

In machine learning, an optimal feature is a feature or **set of features** that can best **represent the underlying patterns and relationships in the data.**

Where it can be use ?

A large number of features are **generated based on the available data**. These can be **domain-specific**, **manually crafted** or can be generated automatically using feature engineering techniques

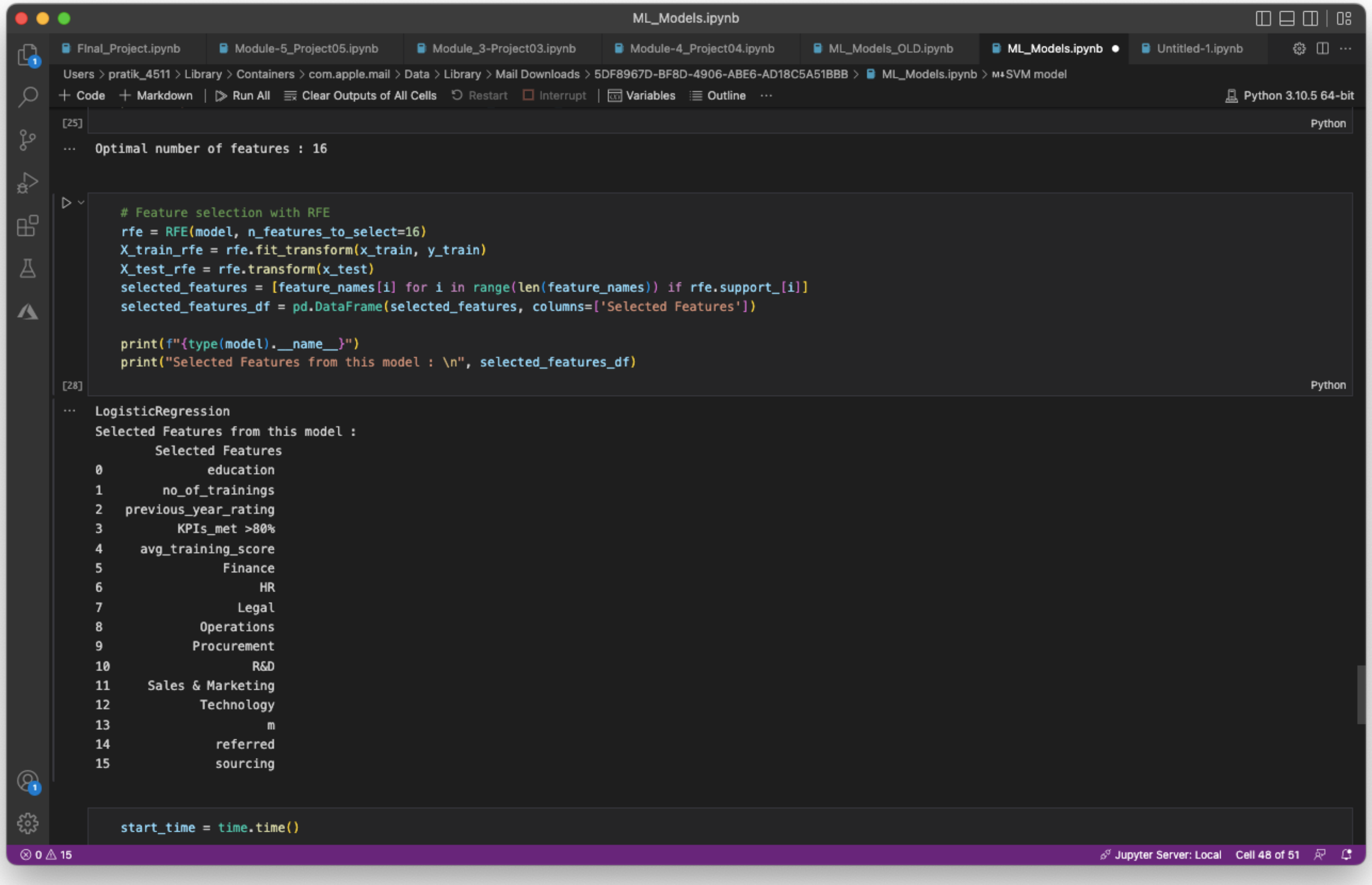
How does it work?

It involves generating a

- large set of candidate features,
- filtering out irrelevant or redundant features,
- ranking or scoring the remaining features based on their ability to improve model performance.

LOGISTIC REGRESSION

Que : Which features are given highest importance when the employer promotion is considered?



```
[25]
Optimal number of features : 16
```

```python
# Feature selection with RFE
rfe = RFE(model, n_features_to_select=16)
X_train_rfe = rfe.fit_transform(x_train, y_train)
X_test_rfe = rfe.transform(x_test)
selected_features = [feature_names[i] for i in range(len(feature_names)) if rfe.support_[i]]
selected_features_df = pd.DataFrame(selected_features, columns=['Selected Features'])

print(f"{type(model).__name__}")
print("Selected Features from this model : \n", selected_features_df)
```

```
[28]
LogisticRegression
Selected Features from this model :
        Selected Features
0               education
1         no_of_trainings
2    previous_year_rating
3             KPIs_met >80%
4      avg_training_score
5                 Finance
6                      HR
7                   Legal
8              Operations
9             Procurement
10                    R&D
11      Sales & Marketing
12             Technology
13                      m
14               referred
15               sourcing
```

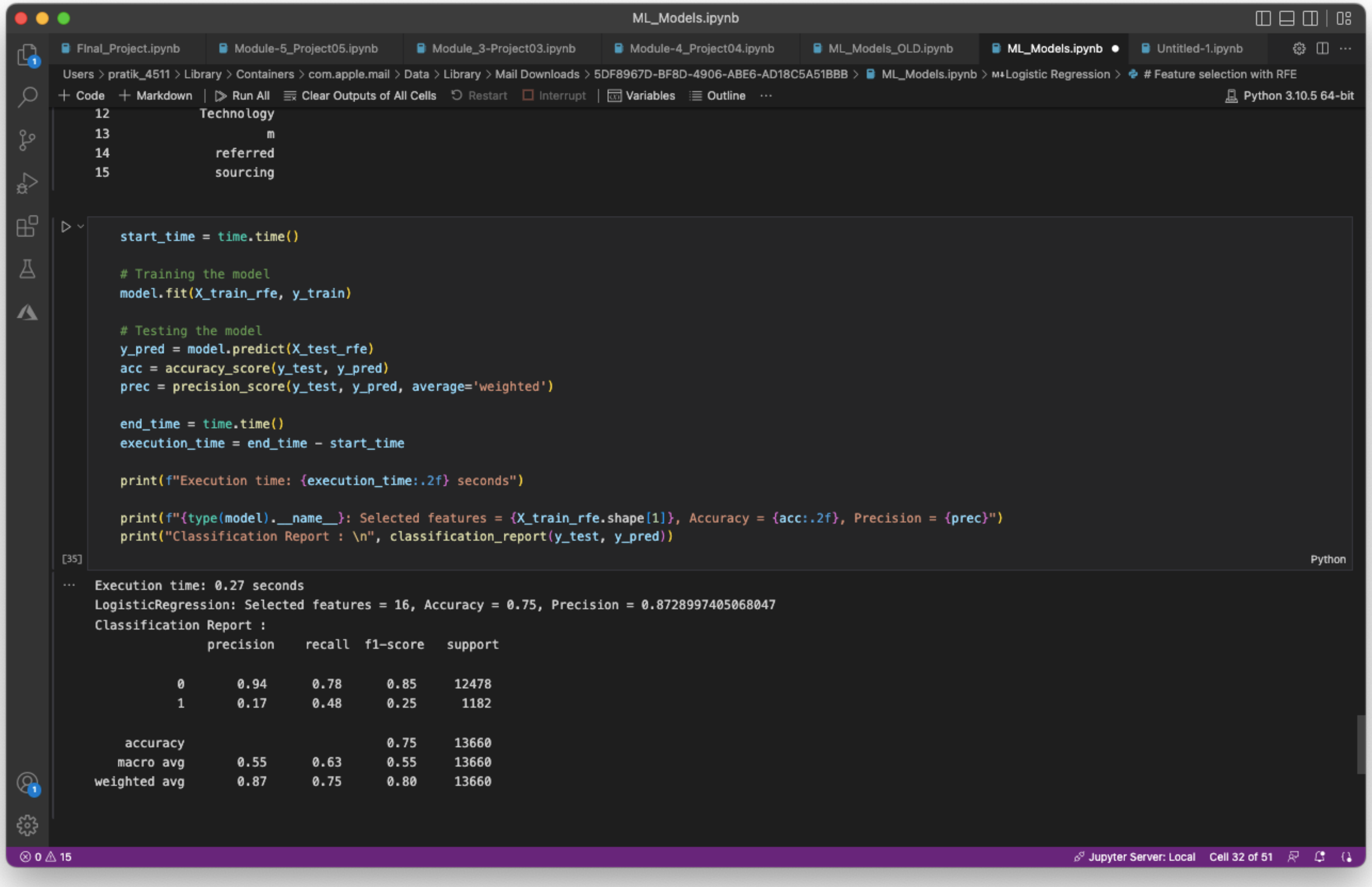
```python
start_time = time.time()
```

```
⊗ 0 ⚠ 15
```

After performing Recursive Feature Elimination **(RFE)** analysis on logistic regression, we found that certain features had the **highest impact** on predicting **employee promotion.**

- Education
- No_of_training
- Previous_year_rating
- KPI_met>80
- Average_training_score

Que : Is it possible to determine from the available data whether or not the company can promote an employee?



```
12          Technology
13                   m
14            referred
15            sourcing
```

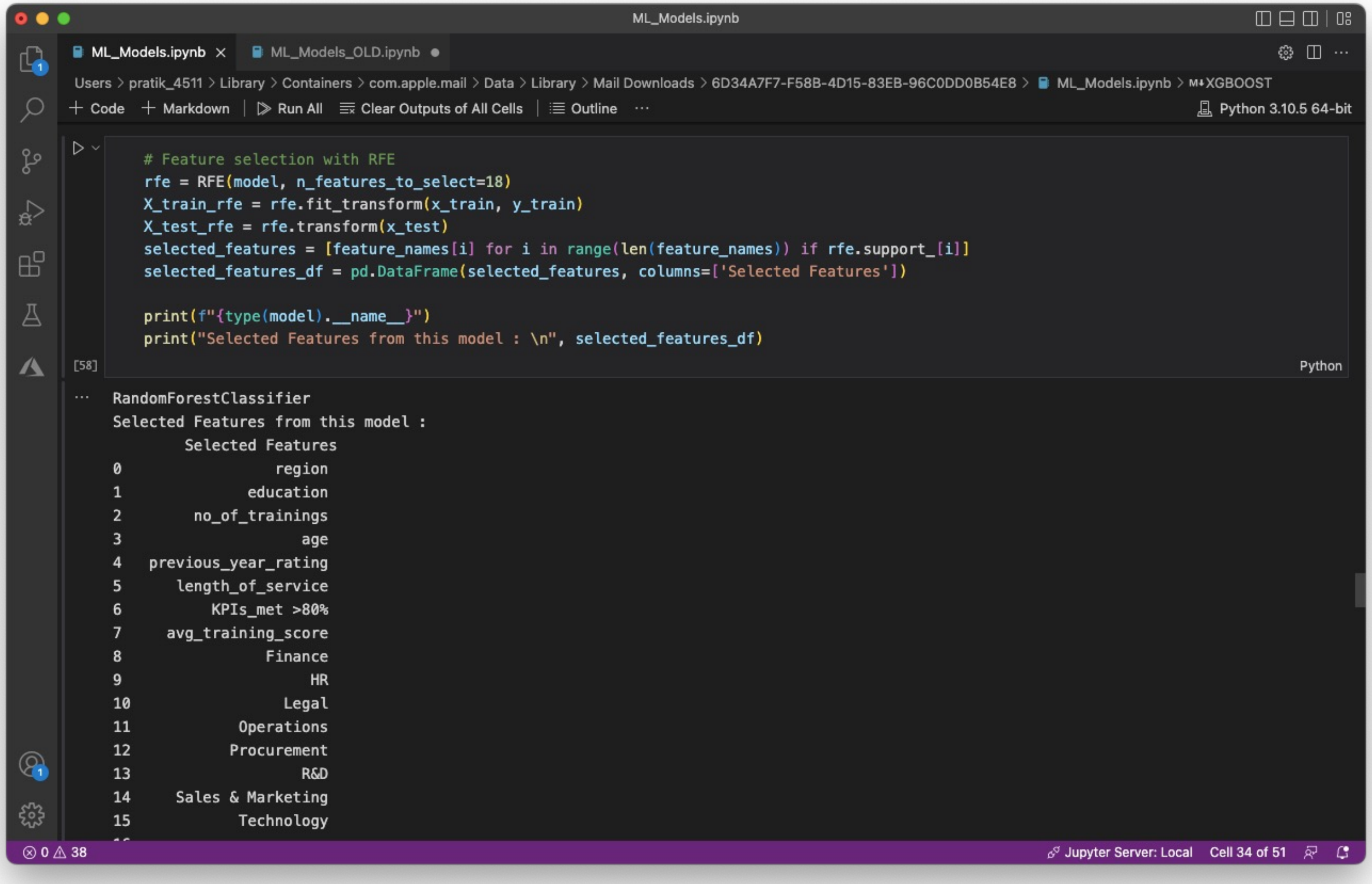
```python
start_time = time.time()

# Training the model
model.fit(X_train_rfe, y_train)

# Testing the model
y_pred = model.predict(X_test_rfe)
acc = accuracy_score(y_test, y_pred)
prec = precision_score(y_test, y_pred, average='weighted')

end_time = time.time()
execution_time = end_time - start_time

print(f"Execution time: {execution_time:.2f} seconds")

print(f"{type(model).__name__}: Selected features = {X_train_rfe.shape[1]}, Accuracy = {acc:.2f}, Precision = {prec}")
print("Classification Report : \n", classification_report(y_test, y_pred))
```

```
Execution time: 0.27 seconds
LogisticRegression: Selected features = 16, Accuracy = 0.75, Precision = 0.8728997405068047
Classification Report :
               precision    recall  f1-score   support

           0       0.94      0.78      0.85     12478
           1       0.17      0.48      0.25      1182

    accuracy                           0.75     13660
   macro avg       0.55      0.63      0.55     13660
weighted avg       0.87      0.75      0.80     13660
```

- **Yes**, a logistic regression model using **16 crucial features** can be used to predict whether a company can promote an employee. The model achieved an **accuracy score of 75%** and a **precision score of 87%**
- However, the model predicted the **negative class (not promoted) better than the positive class (promoted),** with a recall score of 0.78 for the negative class and 0.48 for the positive class.

RANDOM FOREST

Que : Which features are given highest importance when the employer promotion is considered?


ML_Models.ipynb
ML_Models.ipynb
ML_Models_OLD.ipynb
Users > pratik_4511 > Library > Containers > com.apple.mail > Data > Library > Mail Downloads > 6D34A7F7-F58B-4D15-83EB-96C0DD0B54E8 > ML_Models.ipynb > M↓XGBOOST
+ Code
+ Markdown
Run All
Clear Outputs of All Cells
Outline
Python 3.10.5 64-bit
[58]
Python
⊗ 0 ⚠ 38
Jupyter Server: Local
Cell 34 of 51

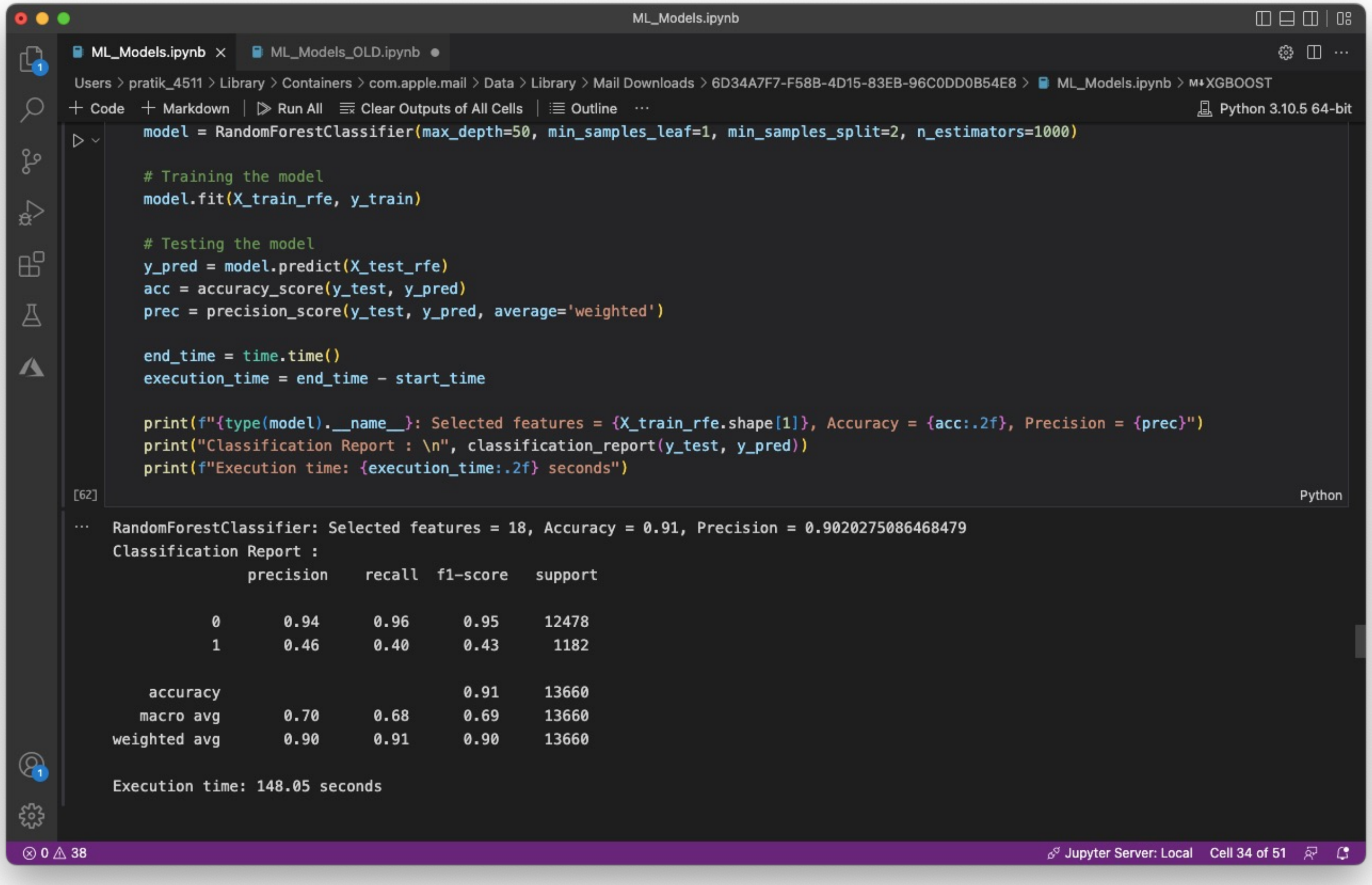
```python
# Feature selection with RFE
rfe = RFE(model, n_features_to_select=18)
X_train_rfe = rfe.fit_transform(x_train, y_train)
X_test_rfe = rfe.transform(x_test)
selected_features = [feature_names[i] for i in range(len(feature_names)) if rfe.support_[i]]
selected_features_df = pd.DataFrame(selected_features, columns=['Selected Features'])

print(f"{type(model).__name__}")
print("Selected Features from this model : \n", selected_features_df)
```

```
RandomForestClassifier
Selected Features from this model :
          Selected Features
0                    region
1                 education
2           no_of_trainings
3                       age
4      previous_year_rating
5         length_of_service
6             KPIs_met >80%
7        avg_training_score
8                   Finance
9                        HR
10                    Legal
11               Operations
12              Procurement
13                      R&D
14        Sales & Marketing
15               Technology
```

After performing Recursive Feature Elimination **(RFE)** analysis on random forest, we found that certain features had the **highest impact** on predicting **employee promotion.**

- Region
- Education
- No_of_training
- Age
- Previous_year_rating

Que : Is it possible to determine from the available data whether or not the company can promote an employee?


ML_Models.ipynb
ML_Models.ipynb
ML_Models_OLD.ipynb
Users > pratik_4511 > Library > Containers > com.apple.mail > Data > Library > Mail Downloads > 6D34A7F7-F58B-4D15-83EB-96C0DD0B54E8 > ML_Models.ipynb > M↓XGBOOST
+ Code
+ Markdown
Run All
Clear Outputs of All Cells
Outline
Python 3.10.5 64-bit
[62]
Python
Jupyter Server: Local
Cell 34 of 51
0 38

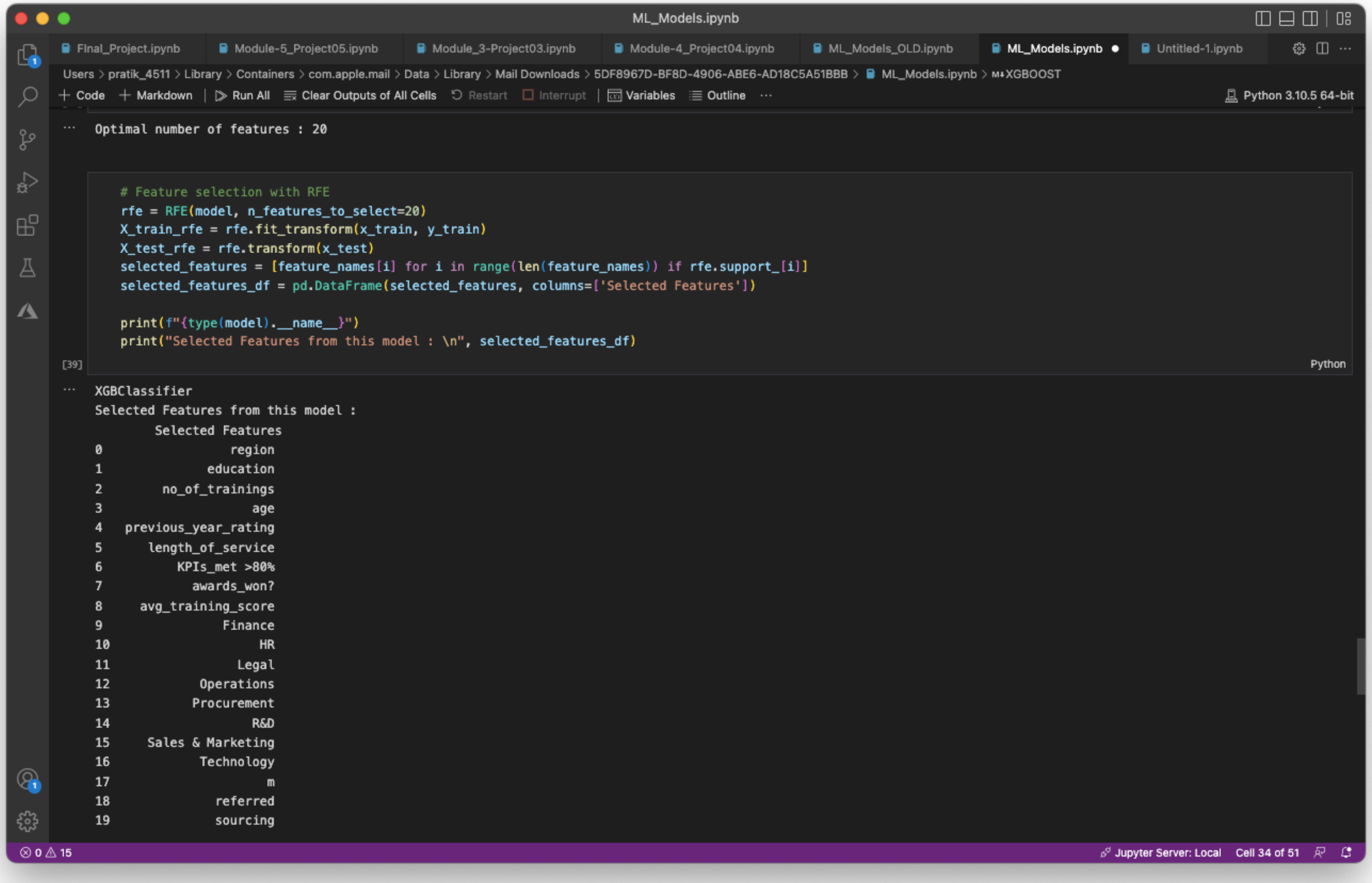
```python
    model = RandomForestClassifier(max_depth=50, min_samples_leaf=1, min_samples_split=2, n_estimators=1000)

    # Training the model
    model.fit(X_train_rfe, y_train)

    # Testing the model
    y_pred = model.predict(X_test_rfe)
    acc = accuracy_score(y_test, y_pred)
    prec = precision_score(y_test, y_pred, average='weighted')

    end_time = time.time()
    execution_time = end_time - start_time

    print(f"{type(model).__name__}: Selected features = {X_train_rfe.shape[1]}, Accuracy = {acc:.2f}, Precision = {prec}")
    print("Classification Report : \n", classification_report(y_test, y_pred))
    print(f"Execution time: {execution_time:.2f} seconds")
```

```
RandomForestClassifier: Selected features = 18, Accuracy = 0.91, Precision = 0.9020275086468479
Classification Report :
               precision    recall  f1-score   support

           0       0.94      0.96      0.95     12478
           1       0.46      0.40      0.43      1182

    accuracy                           0.91     13660
   macro avg       0.70      0.68      0.69     13660
weighted avg       0.90      0.91      0.90     13660

Execution time: 148.05 seconds
```

- **Yes**, it is possible to determine from the available data whether or not a company can promote an employee using a Random Forest Classifier model with **18 selected features**.
- The model achieved an **accuracy of 0.91** and a **precision of 0.90**, with better performance in predicting the "not promoted" class.
- The model's macro average **f1-score was 0.69**, indicating a moderate performance overall. The classification report shows that **the model had high precision and recall scores for the "not promoted" class**, but **lower scores for the "promoted" class**.

XGBOOST

Que : Which features are given highest importance when the employer promotion is considered?


ML_Models.ipynb
FInal_Project.ipynb
Module-5_Project05.ipynb
Module_3-Project03.ipynb
Module-4_Project04.ipynb
ML_Models_OLD.ipynb
ML_Models.ipynb
Untitled-1.ipynb
Users > pratik_4511 > Library > Containers > com.apple.mail > Data > Library > Mail Downloads > 5DF8967D-BF8D-4906-ABE6-AD18C5A51BBB > ML_Models.ipynb > M↓XGBOOST
+ Code + Markdown | Run All Clear Outputs of All Cells Restart Interrupt | Variables Outline ···
Python 3.10.5 64-bit
[39]
Python
0 15
Jupyter Server: Local Cell 34 of 51

```
Optimal number of features : 20
```

```python
# Feature selection with RFE
rfe = RFE(model, n_features_to_select=20)
X_train_rfe = rfe.fit_transform(x_train, y_train)
X_test_rfe = rfe.transform(x_test)
selected_features = [feature_names[i] for i in range(len(feature_names)) if rfe.support_[i]]
selected_features_df = pd.DataFrame(selected_features, columns=['Selected Features'])

print(f"{type(model).__name__}")
print("Selected Features from this model : \n", selected_features_df)
```

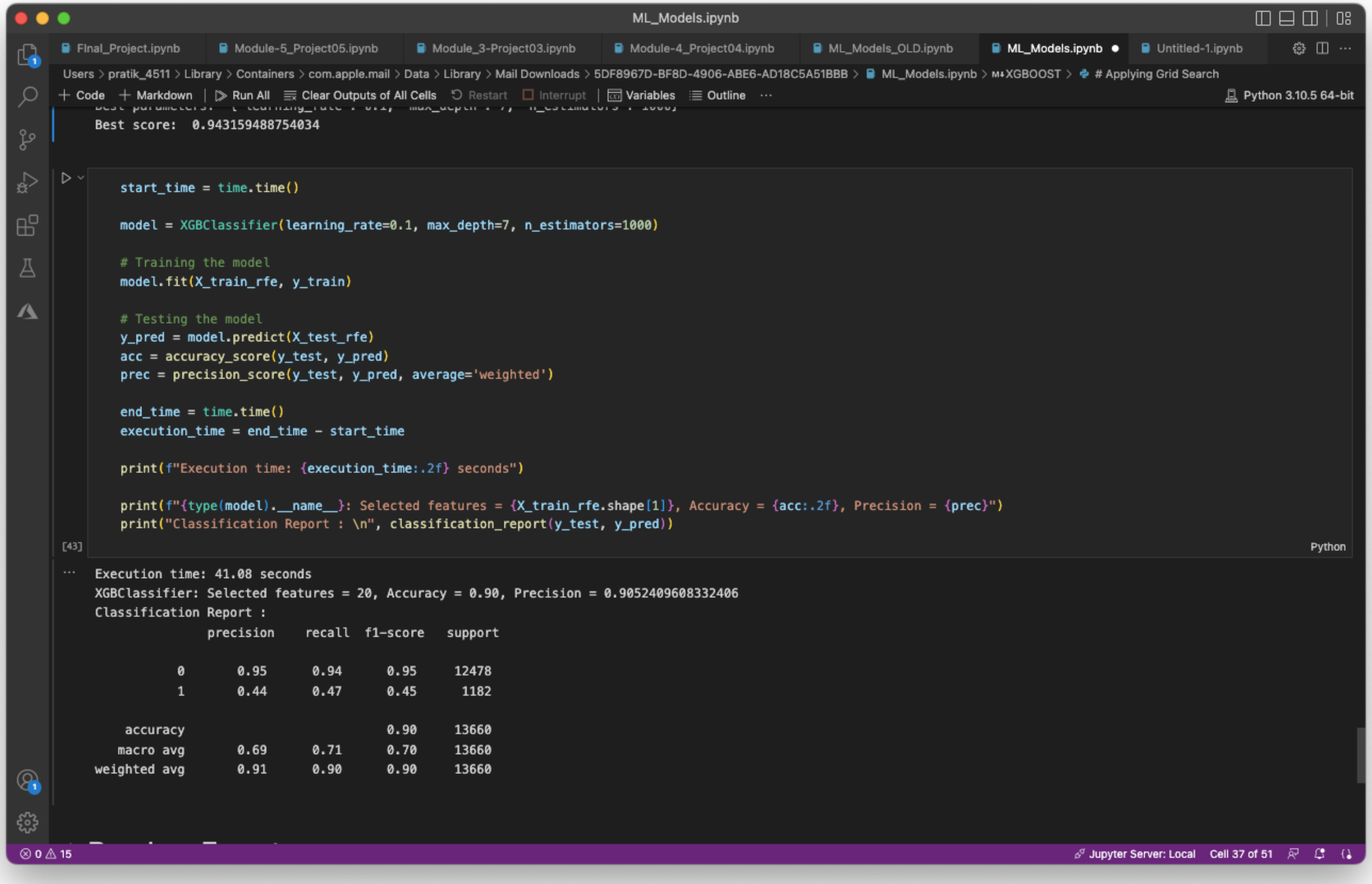
```
XGBClassifier
Selected Features from this model :
        Selected Features
0                  region
1               education
2         no_of_trainings
3                     age
4    previous_year_rating
5       length_of_service
6           KPIs_met >80%
7             awards_won?
8      avg_training_score
9                 Finance
10                     HR
11                  Legal
12             Operations
13            Procurement
14                    R&D
15      Sales & Marketing
16             Technology
17                      m
18               referred
19               sourcing
```

After performing Recursive Feature Elimination **(RFE)** analysis on XGBoost, we found that certain features had the **highest impact** on predicting **employee promotion.**

- Region
- Education
- No_of_trainings
- Age
- Average_training_score

Note : Grid Search technique used to design this ML model

Que : Is it possible to determine from the available data whether or not the company can promote an employee?



```
Best parameters:  {'learning_rate': 0.1, 'max_depth': 7, 'n_estimators': 1000}
Best score:  0.943159488754034
```

```python
start_time = time.time()

model = XGBClassifier(learning_rate=0.1, max_depth=7, n_estimators=1000)

# Training the model
model.fit(X_train_rfe, y_train)

# Testing the model
y_pred = model.predict(X_test_rfe)
acc = accuracy_score(y_test, y_pred)
prec = precision_score(y_test, y_pred, average='weighted')

end_time = time.time()
execution_time = end_time - start_time

print(f"Execution time: {execution_time:.2f} seconds")

print(f"{type(model).__name__}: Selected features = {X_train_rfe.shape[1]}, Accuracy = {acc:.2f}, Precision = {prec}")
print("Classification Report : \n", classification_report(y_test, y_pred))
```

```
Execution time: 41.08 seconds
XGBClassifier: Selected features = 20, Accuracy = 0.90, Precision = 0.9052409608332406
Classification Report :
               precision    recall  f1-score   support

           0       0.95      0.94      0.95     12478
           1       0.44      0.47      0.45      1182

    accuracy                           0.90     13660
   macro avg       0.69      0.71      0.70     13660
weighted avg       0.91      0.90      0.90     13660
```

- **Yes,** it is possible to determine from the available data whether or not a **company can promote** an employee **using an XGBoost model with 20 selected features.**
- The model achieved an **accuracy of 90%** and a **precision of 91%,** but it **struggled with predicting the "promoted" class**.
- The model's macro average **f1-score** was **70%**, indicating a **moderate performance** overall.

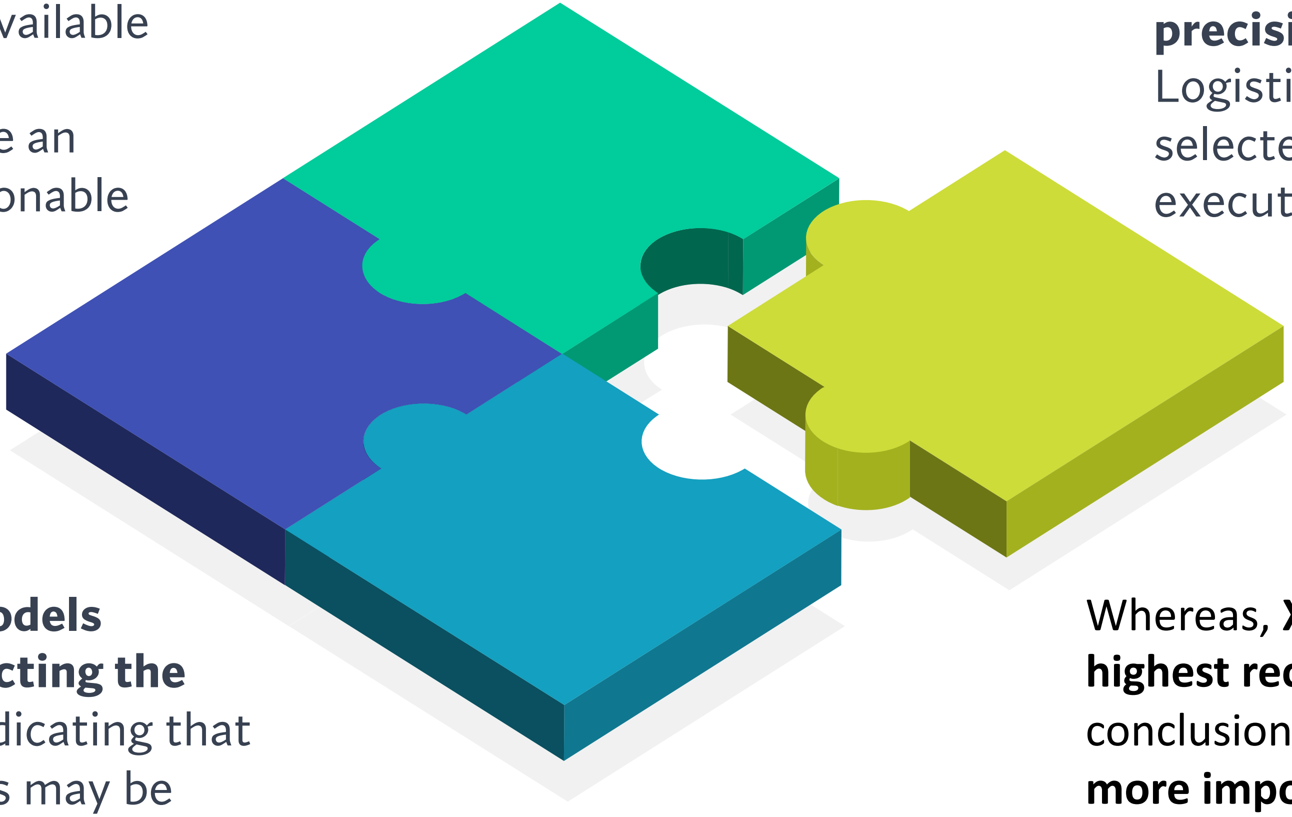
COMPARISION

Model	**Selected Features**	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**Execution Time**
Logistic Regression	16	0.75	0.87	0.47	0.25	41.08 sec
Random Forest Classifier	18	0.91	0.90	0.39	0.58	148.05 sec
XGB Classifier	20	0.90	0.90	0.46	0.53	0.27 sec

CONCLUSION

Based on the results of the three models, it is possible to determine from the available data whether or not a company can promote an employee with a reasonable degree of accuracy.

The **Random Forest** Classifier model with **18 selected features** achieved the **highest accuracy and precision scores**, while the Logistic Regression model with 16 selected features had the fastest execution time.



However, all **three models struggled with predicting the "promoted" class**, indicating that further improvements may be necessary for more accurate predictions in this area.

Whereas, **XGBClassifier has the highest recall and f1-score**. In conclusion, if **recall and f1-score are more important**, the **XGBoost model** would be the **better choice**, especially if speed is a consideration.

THANK YOU!