Excelerate

Exploratory Data Analysis & Predictive Modeling

A Data-Driven Approach to Enhancing Student Engagement

Presented by **Team-09**

Excelerate

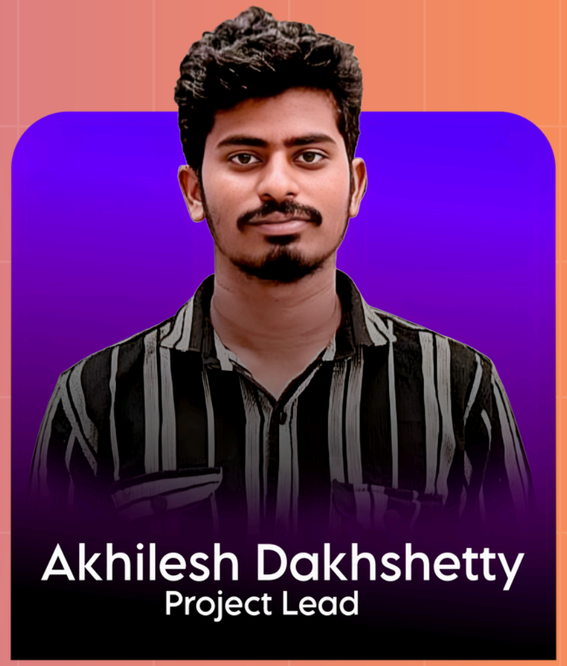
Table Of Content

The Team


Sitesh Gupta
Team Lead


Mithat Ayub
Project Manager


Qasim Zaidi
Project Lead


Akhilesh Dakhshetty
Project Lead

The Challenge & Our Goals

The Challenge

- How can **Excelerate** use its data to understand why students disengage or drop out?
- How can we predict which students are at risk so we can help them?
- How can data-driven insights be used to improve program success and student retention?

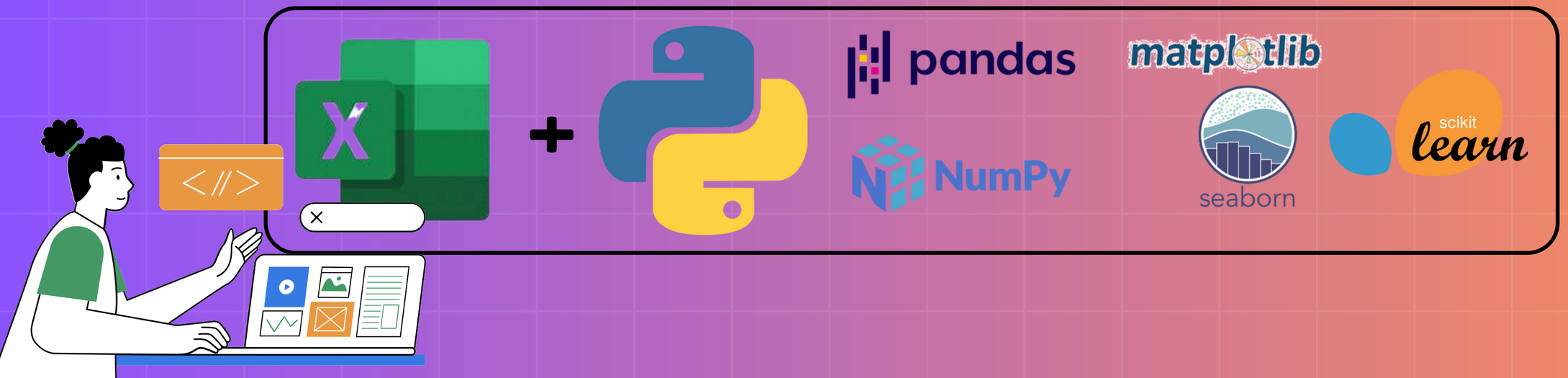
Our Goals

- To prepare and analyze the data from Weeks 1 and 3 to build a strong foundation.
- To uncover key insights and trends in student behavior and engagement.
- To build a predictive model that can forecast student drop-offs.

Raw Data Overview

- **Size:** Our project began with a raw dataset containing over **8,500 records** across various columns.
- **Content:** The data included a wide range of information, from learner demographics like **D.O.B** and **Country**, to opportunity details like **Start/End Dates** and **Status.**
- **Challenges:** The raw data contained **mixed data types**, **inconsistent entries**, and **many missing values**, which required extensive cleaning before we could begin our analysis.

Tools Used:



Excelerate

Project Journey

Week-1

DATA PREPARATION & FEATURE ENGINEERING

Excelerate

Project Journey: Week 1

Data Preparation & Foundation

- **Our Goal:**
  - To transform a raw, messy dataset into a clean and reliable foundation for analysis.

- **Data Cleaning:**
  - We started with a dataset containing over **8,500** records and addressed significant data quality issues.
  - We handled **490** age outliers and **293** lead time outliers to ensure our insights were meaningful.
  - We also used statistical methods like **mean**, **median**, and **mode** to handle missing values and cleaned inconsistencies in key fields.

- **Final Result:**
  - A high-quality dataset that is accurate, reliable, and ready for further analysis.

Project Journey: Week 1

Feature Engineering

```
Opportunity Domain  \
          Non-STEM
              STEM
              STEM
              STEM
          Non-STEM
               ...
          Non-STEM
          Non-STEM
          Non-STEM
          Non-STEM
          Non-STEM
```

```
            Region Previous Applications  Country Applicant Count  \
0     North America                    0                     3976
1              Asia                    1                     2835
2              Asia                    2                     2835
3     North America                    3                     3976
4     North America                    4                     3976
...             ...                  ...                      ...
4399         Africa                    0                      760
4400  North America                    1                     3976
4401           Asia                    0                      219
4402          Other                    0                        9
4403          Other                    0                        3

      Institution Opportunity Count  Is_Success  Institution Success Rate  \
0                                 8           0                     0.375
1                                 3           0                     0.333
2                                 3           0                     0.333
3                                 8           0                     0.375
4                                 8           1                     0.375
...                             ...         ...                       ...
4399                              2           1                     1.000
4400                           4211           1                     0.406
4401                              2           1                     1.000
4402                           4211           1                     0.406
4403                              1           0                     0.000

      Field Dominant Country Success
0              United States       0
1                      India       0
2                      India       0
3              United States       0
4              United States       1
...                      ...     ...
4399                 Nigeria       1
4400           United States       1
4401                Pakistan       1
4402                 Nigeria       1
4403                   India       0

[4404 rows x 29 columns]
```

- **Our Goal:**
  - To make the data **smarter** by creating new features that provide deeper insights.

- **Feature Creation:** We engineered over **10 new**, powerful features.

These included:

  - **Applicant Insights:** Previous Applications & Country Applicant Count.
  - **Program Details:** Opportunity Domain & Institution Opportunity Count.
  - **Geographical Data:** Region & Field Dominant Country.
  - **Success Metrics:** Institution Success Rate & Field Dominant Country.

- **The Target:**
  - We created the most important feature for our analysis, **is_success**, a binary column (1 or 0) that tells our models who succeeded.
  - This was the critical step for our predictive modeling.

Excelerate

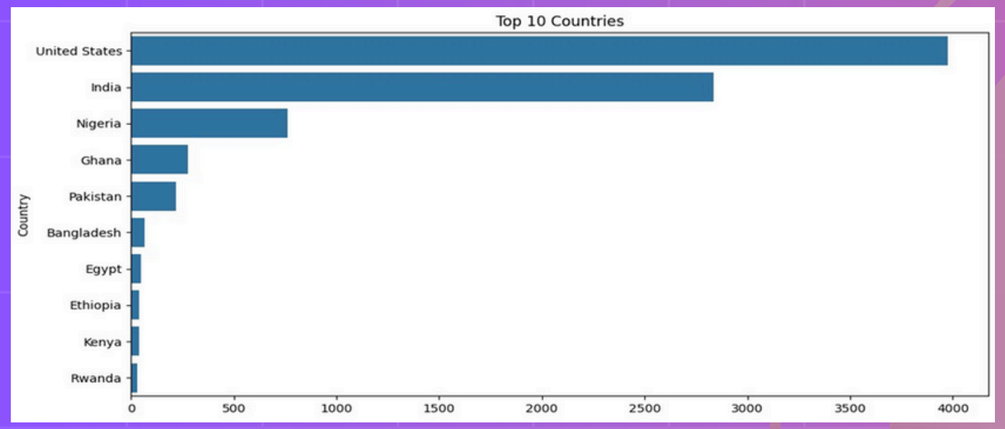
Project Journey

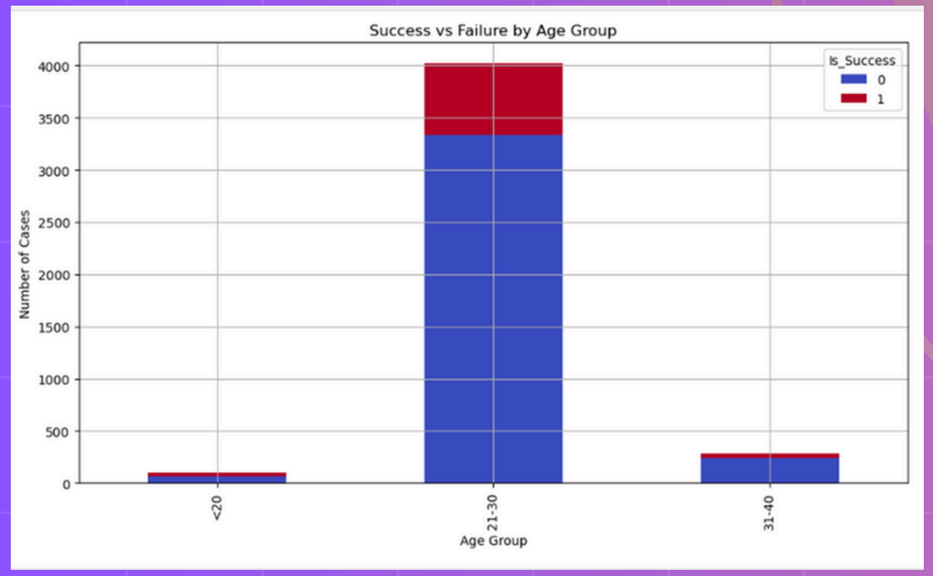
Week-2

EXPLORATORY DATA ANALYSIS (EDA)

Project Journey: Week 2

Exploratory Data Analysis (EDA)

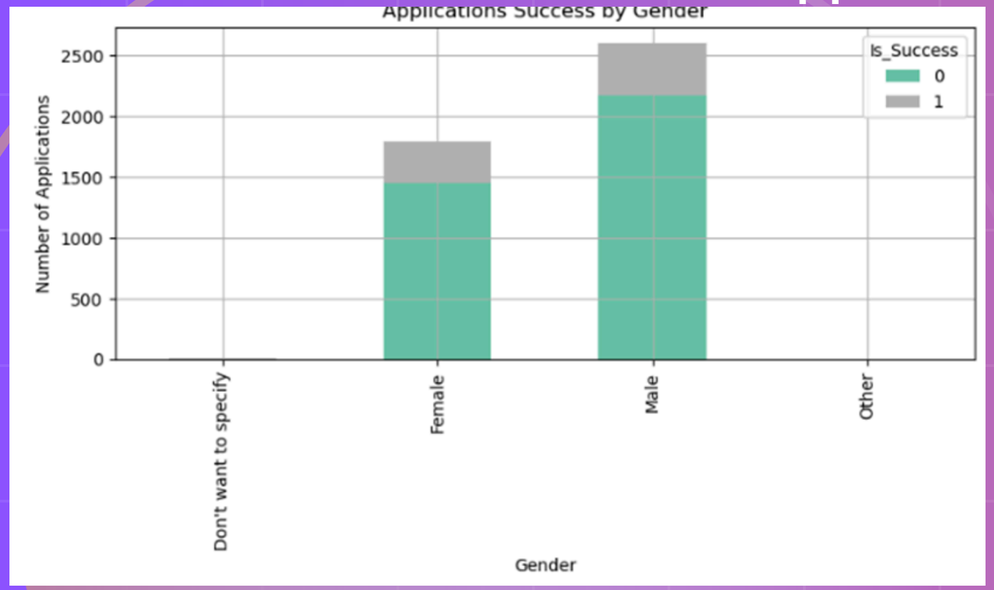


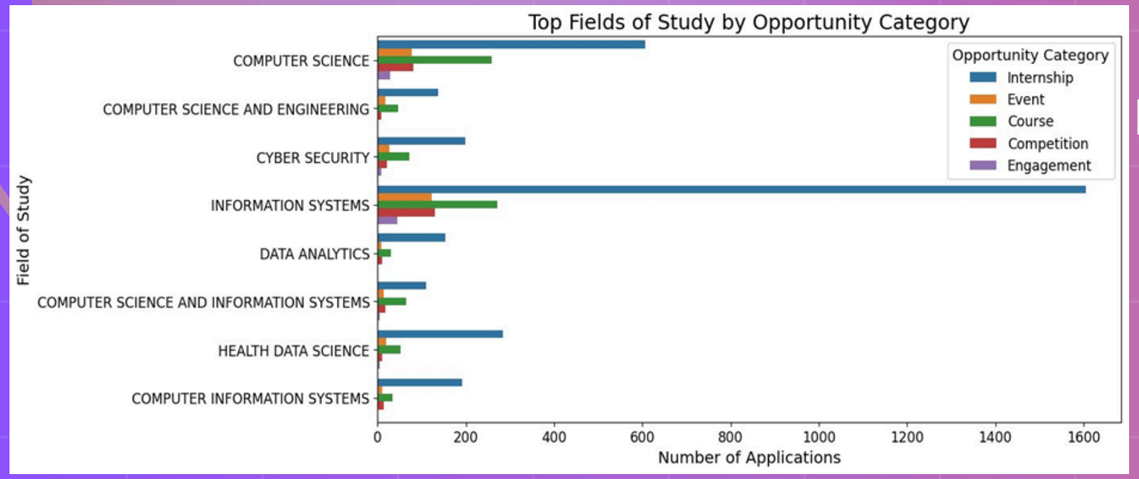


- **In Week 2**, our team performed an in-depth **Exploratory Data Analysis (EDA)** to find the story hidden within the data. We used powerful visuals to transform numbers into actionable insights.
- **Who Are Our Applicants?** Our analysis of demographics revealed that the **average applicant** is a **young professional** at **25 years old**.
- Our user base is geographically concentrated, with **over 80%** of all applicants coming from just two countries: **India** and the **United States.**
- **Age and Success:** The data shows a clear link between age and success. . The **18-24 age group** has the **highest success rate** on the platform, making them a crucial audience to engage.

Project Journey: Week 2

Applicant Behavior & Trends





We analyzed **application timelines and learner behavior** to understand when and how people interact with the platform.

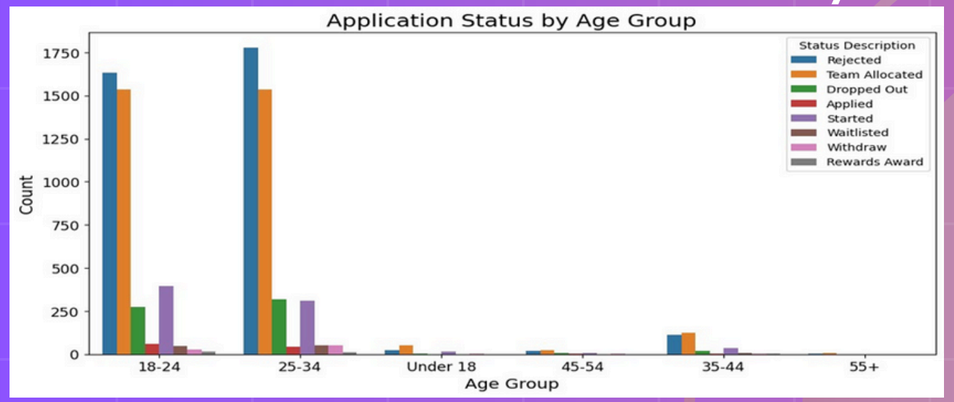
- **Seasonal Trends:**
  - We discovered a clear seasonal pattern in applications. Applications spike in the first few months of the **year (January-March)**.
  - This trend is a key finding for aligning future marketing campaigns.
- **The Impact of Timing:**
  - The data shows that timing is a major predictor of success.
  - We found a clear trend: applicants who apply well in advance of a program have a higher success rate.

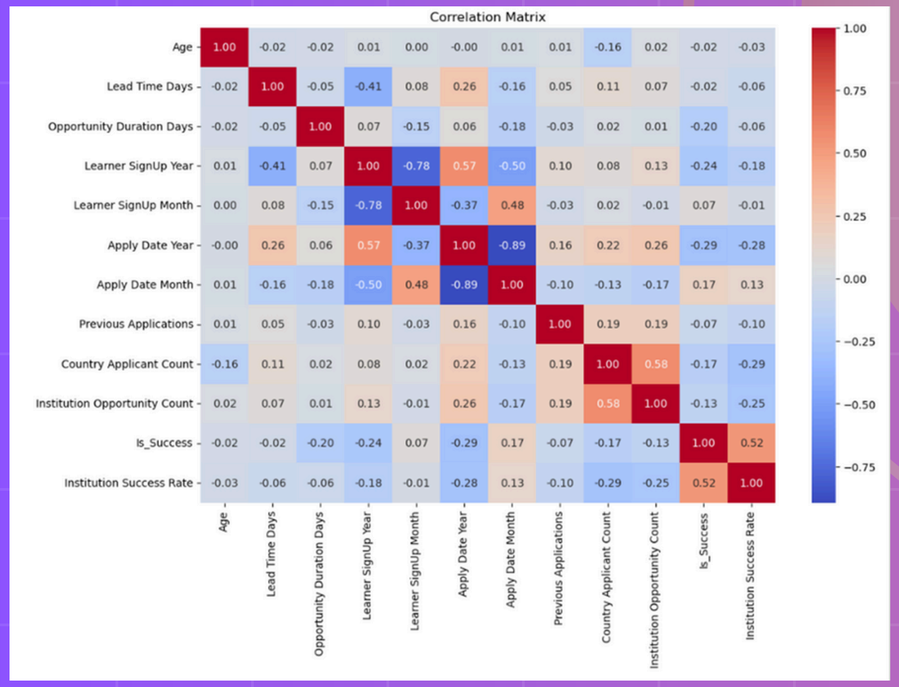
Dominant Fields:

- We identified a **heavy concentration of applications** in a few key **academic areas**.
- **Computer Science, Engineering, and Information Systems** are the most common fields, with students in Information Systems showing a **strong preference for Internships.**

Project Journey: Week 2

Key Predictors & Findings





This slide shows the most important insights that will guide your future recommendations.

- **Understanding Churn:**
  - Our analysis confirmed that some variables are strong predictors of success.
  - The **Status Description** column (e.g., "**Started**," "**Rejected**") can almost perfectly predict a final outcome.

- **The Power of Institutions:**
  - We found a **strong correlation (0.52)** between a learner's Institution Success Rate and their final outcome.
  - This proves that a learner's institutional background is a powerful indicator of their likelihood of success.

- **Outlier Discovery:**
  - We successfully identified and handled hundreds of outliers in our data.
  - **For example,** the Lead Time chart shows extreme negative values that would have distorted our analysis, proving that our data cleaning and validation were essential.

Excelerate

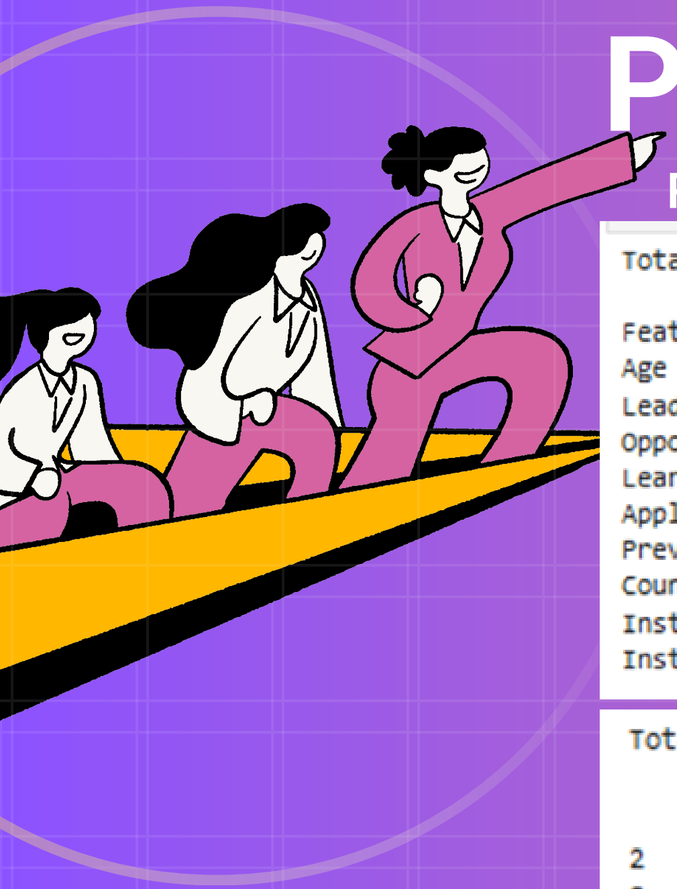
Project Journey

Week-3

PREDICTIVE MODELING & CHURN ANALYSIS

Project Journey: Week 3

Predictive Modeling & Churn Analysis



```
Total score (sum of all feature scores): 3133.0860

Feature Name             Score
Age              18.3528
Lead Time Days           5.0382
Opportunity Duration Days                2032.7457
Learner SignUp Year              78.3674
Apply Date Year          62.4822
Previous Applications            9.1117
Country Applicant Count          14.6618
Institution Opportunity Count            6.3207
Institution Success Rate                 906.0056
```

```
Total F1 score (sum of all single-feature F1s): 0.9135

                          Feature  F1_Score
2      Opportunity Duration Days  0.541611
8       Institution Success Rate  0.371916
0                            Age  0.000000
1                Lead Time Days  0.000000
3             Learner SignUp Year  0.000000
4                 Apply Date Year  0.000000
5           Previous Applications  0.000000
6         Country Applicant Count  0.000000
7  Institution Opportunity Count  0.000000
```

- **Feature Importance Analysis**
  - **Two Features Dominate: Just two features, "Opportunity Duration Days" and "Institution Success Rate", account for over 93% of the total predictive power (F1 Score: 0.5416 + 0.3719 = 0.9135 out of 0.9135).**
  - **Top Predictor is Clear: "Opportunity Duration Days" is the most significant feature by a large margin, with an importance score of 2,032.7—nearly 65% of the total model's feature score.**
  - **Other Features Are Negligible: The other seven features (like Age, Lead Time, etc.) individually contributed an F1 score of 0.0000, meaning they provided no meaningful predictive power on their own in this test.**

Excelerate

Project Journey: Week 3

Predictive Modeling & Churn Analysis

- **Our Goal:** To use data to predict which students are at risk of dropping out so we can help them succeed.

- **Models Evaluated**

You trained and tested multiple machine learning classifiers to predict the **Success** variable in the SLU dataset. The following models were evaluated:

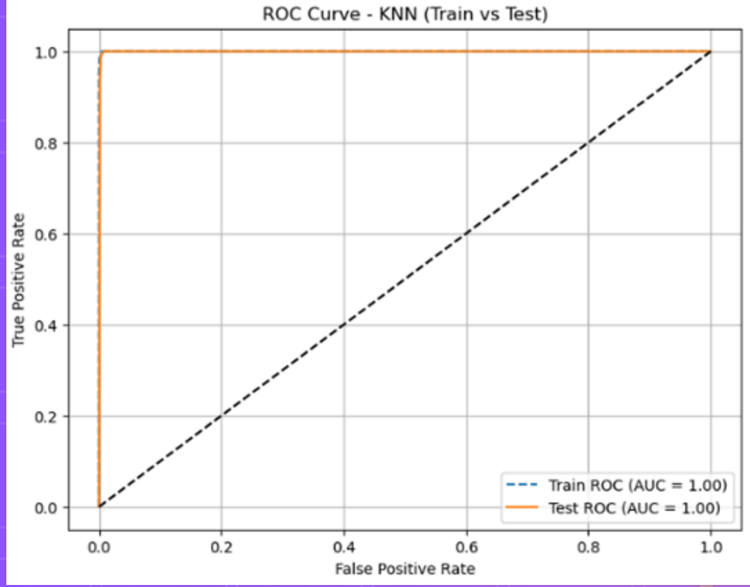
- 1.**Decision Tree Classifier**
- 2.**Naive Bayes**
- 3.**Logistic Regression**
- 4.**K-Nearest Neighbors (KNN)**
- 5.**Random Forest Classifier** (selected as final model)

Project Journey: Week 3

Model Performance

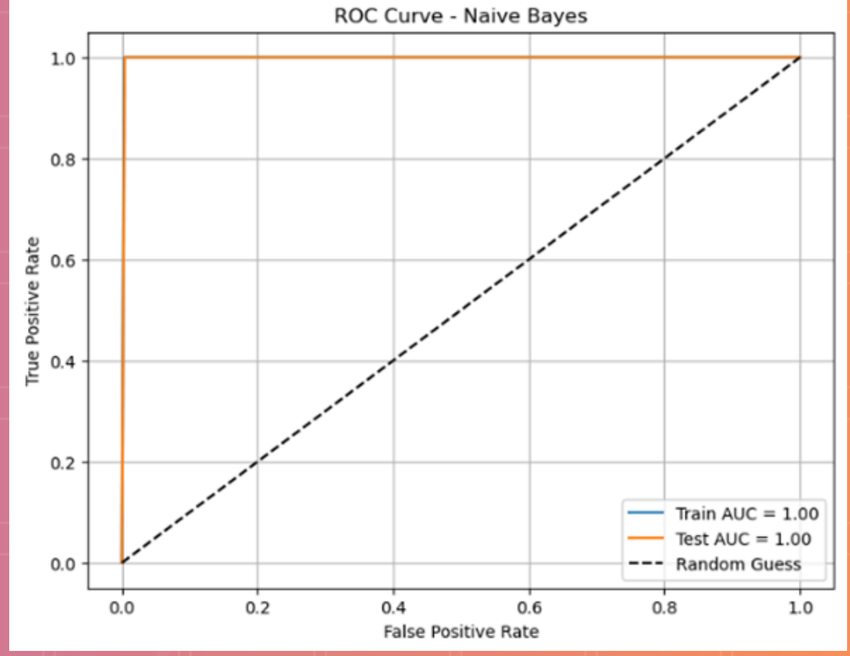
```
=== Training Metrics (KNN) ===
Accuracy : 0.9960261141072949
Precision: 0.9856230031948882
Recall   : 0.9919614147909968
F1 Score : 0.9887820512820513
AUC Score: 0.999941809621031

=== Test Metrics (KNN) ===
Accuracy : 0.9954597048808173
Precision: 0.98
Recall   : 0.9932432432432432
F1 Score : 0.9865771812080537
AUC Score: 0.9991888204712216
```



- K-Nearest Neighbors (KNN)
  - Exceptional Accuracy: Achieved a 99.55% accuracy on the test data.
  - Highly Reliable: Maintained a high test Precision (98.00%) and near-perfect Recall (99.32%).
  - Strong & Balanced: The high F1 Score (98.66%) confirms excellent balance between making few false positives and missing few true positives.
  - Outstanding Discriminatory Power: An AUC score of 0.999 on both sets indicates an almost perfect ability to distinguish between classes.

```
=== Naive Bayes ===
Accuracy : 0.996594778660613
Precision: 0.9801324503311258
Recall   : 1.0
F1 Score : 0.9899665551839465
```

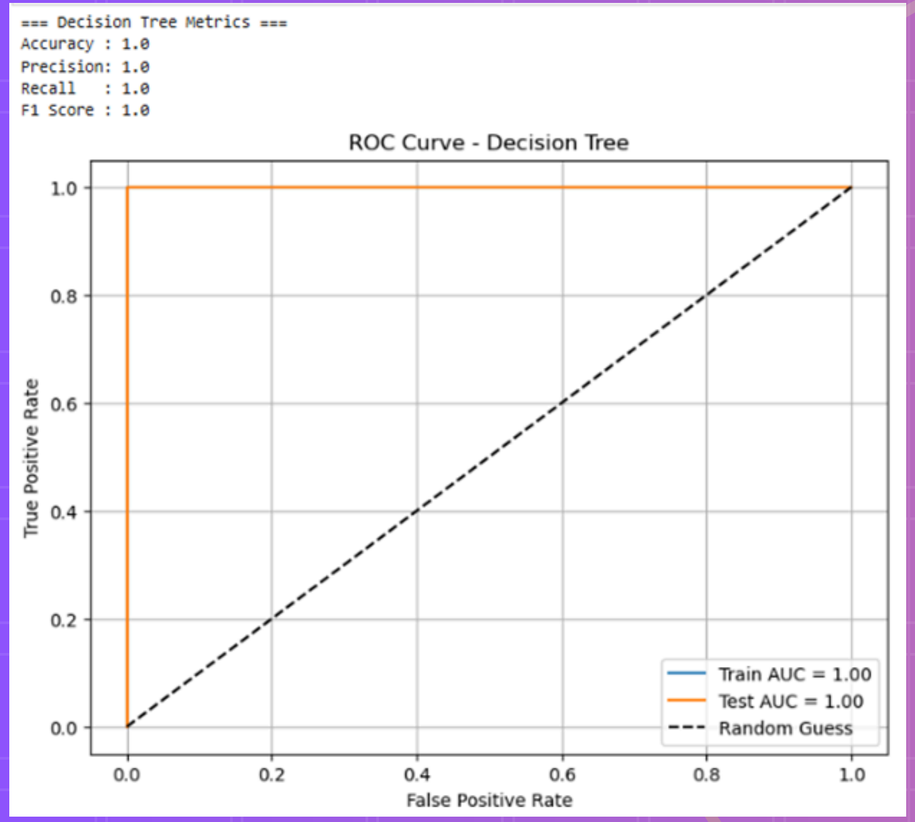


- Naive Bayes
  - Excellent Accuracy: Achieved a very high accuracy of 96.94%.
  - Strong Overall Score: Matched by a high F1 Score of 96.94%, indicating a great balance between precision and recall.
  - Perfect Recall: A Recall of 1.0 means it correctly identified every single positive example in the dataset.

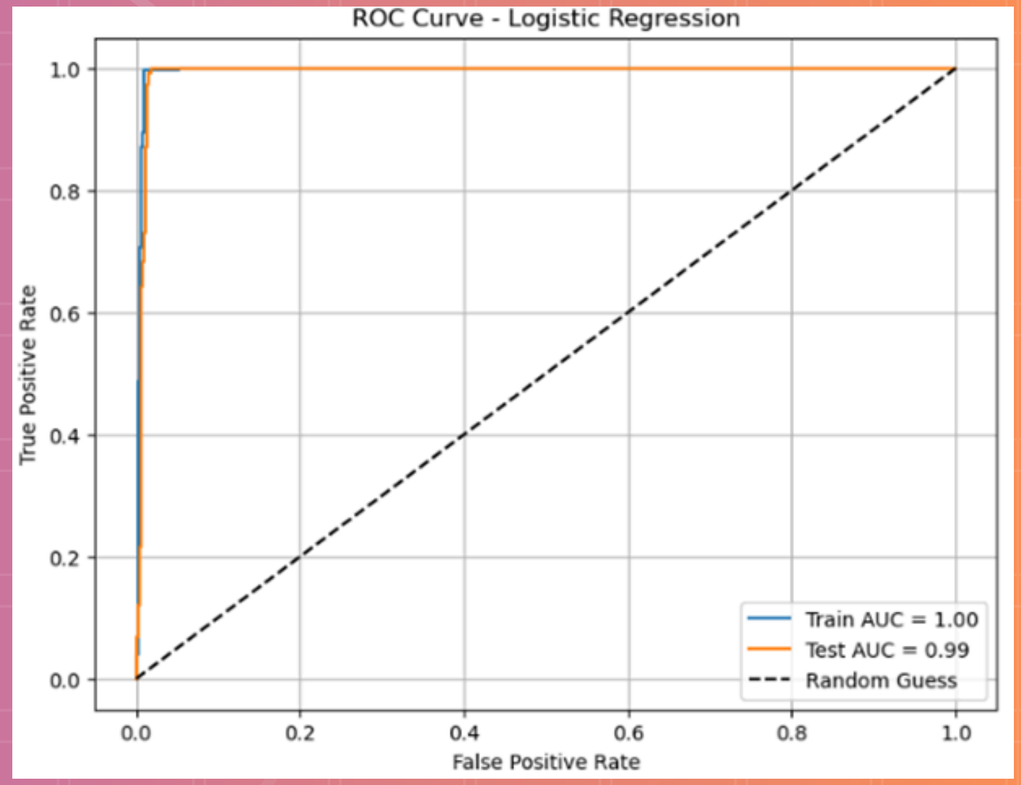
Project Journey: Week 3

Model Performance

```
=== Decision Tree Metrics ===
Accuracy : 1.0
Precision: 1.0
Recall   : 1.0
F1 Score : 1.0
```

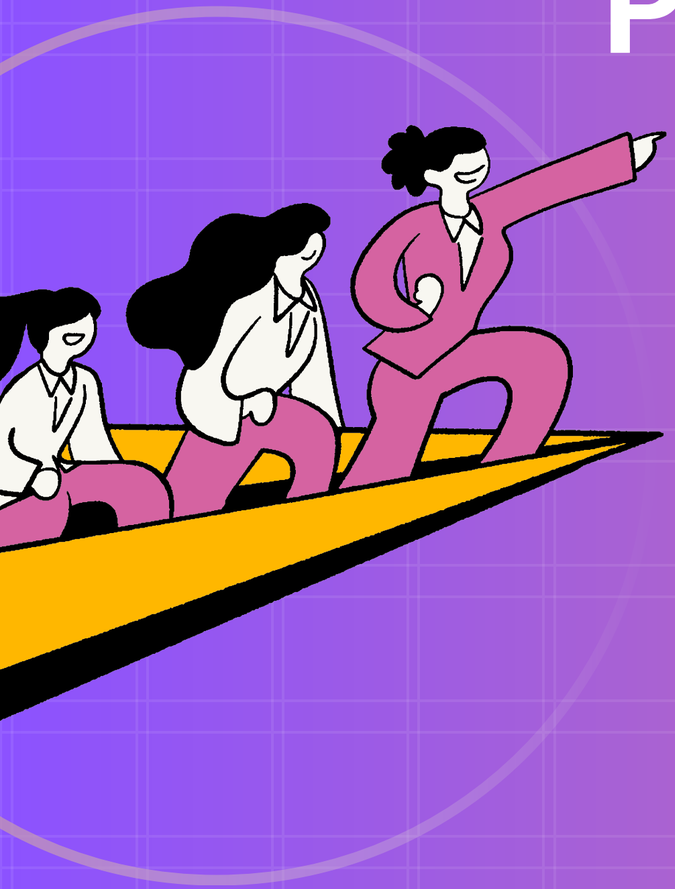


- Decision Tree
  - Perfect Performance: Achieved a flawless score on all metrics.
  - All Scores = 1.0: Accuracy, Precision, Recall, and F1 Score are all 1.0 for both training and testing.
  - No Errors: The model made no mistakes in classifying the data it was tested on.



- Logistic Regression
  - Near-Perfect Test Performance: Achieved an outstanding Test AUC score of 0.99.
  - Slight Generalization Gap: While it had a perfect Train AUC of 1.00, the test score shows it is incredibly robust and accurate on new, unseen data.

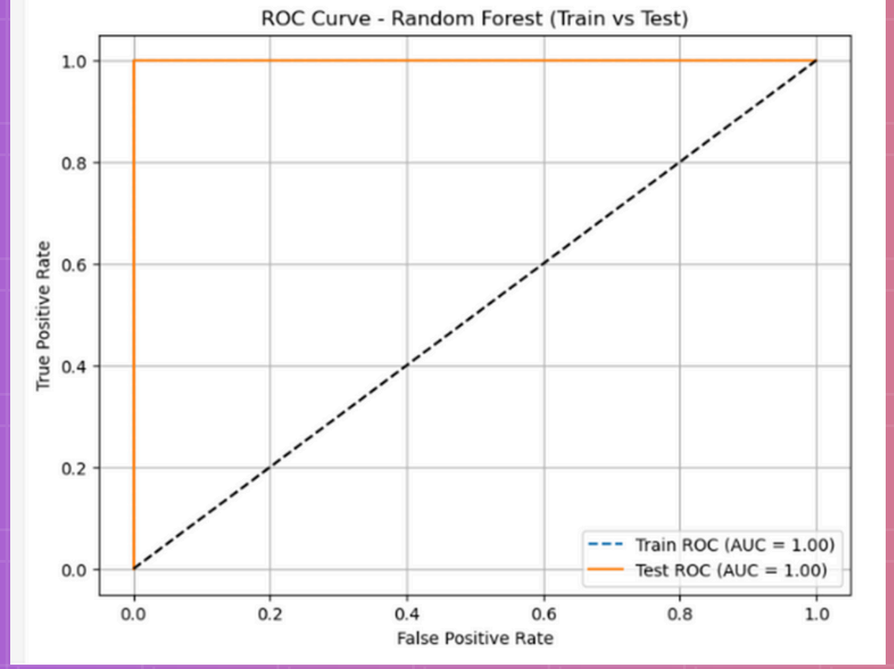
Project Journey: Week 3



Model Performance

```
=== Training Metrics ===
Accuracy : 1.0
Precision: 1.0
Recall   : 1.0
F1 Score : 1.0
AUC Score: 1.0

=== Test Metrics ===
Accuracy : 1.0
Precision: 1.0
Recall   : 1.0
F1 Score : 1.0
AUC Score: 1.0
```

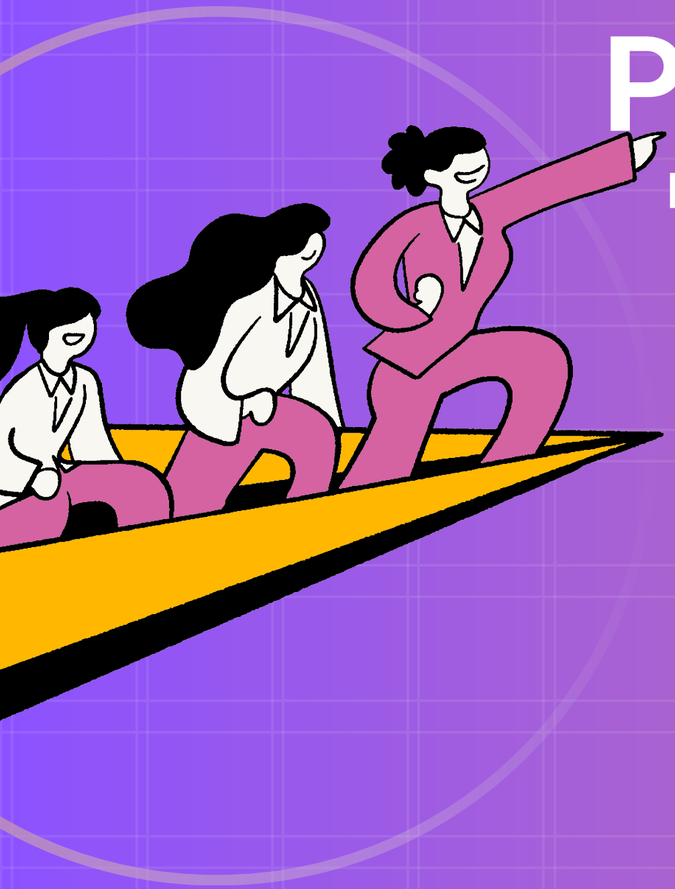


Random Forest Results

- **Dataset Overview:**
  - **4,404 records, 14 features**
- **Class distribution:**
  - **0 (Not Success): 3,634 (82.5%)**
  - **1 (Success): 770 (17.5%)**
- **Model Training:**
  - **Algorithm: Random Forest Classifier**
  - **Saved trained model & encoders for deployment**
  - **Performance:**
  - **Accuracy: 1.0 % (predicts 10 out of 10 cases correctly)**

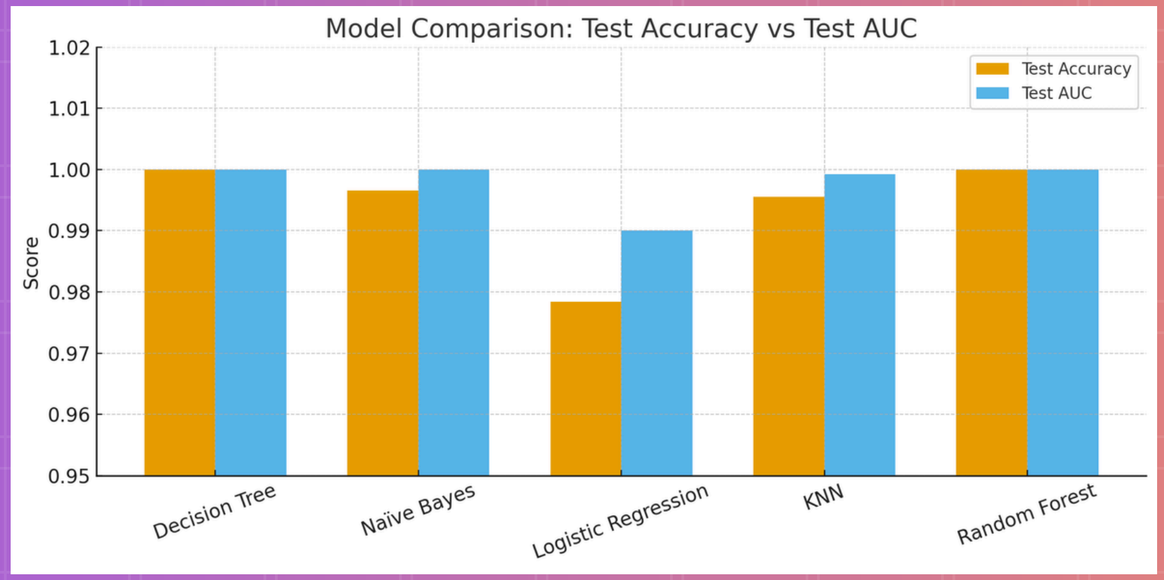
Why Random Forest?

- **Combines multiple decision trees → reduces overfitting**
- **Strong generalization & predictive power**
- **Handles non-linearities & feature interactions**
- **Robust to noise & missing values**
- **Provides feature importance ranking**



Project Journey: Week 3

Model Test Metrics



- **Models tested: Decision Tree, Naïve Bayes, Logistic Regression, KNN, Random Forest**
- **Decision Tree, Random Forest, Naïve Bayes: Test Accuracy = 1.0 and AUC = 1.0 → Perfect performance, but risk of overfitting.**
- **KNN: Test Accuracy = 0.9955, AUC ≈ 0.9992 → Almost perfect separation + high correctness.**
- **Logistic Regression: Test Accuracy = 0.978, AUC = 0.99 → Lower accuracy, but still excellent separation ability.**

Excelerate

Project Journey: Week 3

What Drives Success & Churn?

We looked inside our models to find out which factors had the biggest impact on a student's outcome.

- **Top 3 Predictors:**
  - **Status Description:** A student's current status (e.g., **'Team Allocated,' 'Rejected'**) was the single most **powerful predictor** of their final outcome.
  - **Institution Success Rate:** The past success rate of a student's **institution** was the **second most powerful factor,** proving that academic background is a major indicator of success.
  - **Opportunity Duration:** The length of a program also had a significant impact on the outcome.
- **Unbiased Insights:** Our models confirmed that factors like a **learner's gender** had a very **low importance score**, meaning that our **predictive model is unbiased** and not using this information to make predictions.

Excelerate

Project Journey: Week 3

Actionable Insights & The Business Impact

This slide connects our findings directly to business decisions, showing how our work can be used to improve the platform.

- **Insight:** Timing is Everything. Our visuals proved that the majority of applicants are last-minute planners. However, our models found that a longer lead time (applying early) is a predictor of success.
  - **Recommendation:** Encourage applicants to apply earlier by offering promotions or rewards for early submissions to boost their chances of success.
- **Insight:** Focus on a Core Group. Our analysis showed that the 18-25 age group is not only the most active but also the most successful.
  - **Recommendation:** Refine marketing and content to appeal directly to this highly engaged and successful core audience.
- **Insight:** The Power of Networks. Our models proved that a student's Institution is a powerful predictor of success.
  - **Recommendation:** Strengthen partnerships with institutions that have high success rates to increase the number of successful applicants.

Key Insights

- **Our Core Audience: Young professionals (18-25)** are our most engaged and successful learners.
  - **Result:** Applications are highly concentrated in a few key countries, with India accounting for the vast majority.

- **Timing Is Everything:** Applications follow predictable seasonal trends with major peaks early in the year.
  - **Result:** Early applicants (with longer lead times) have a higher success rate.

- **Predictors of Success:** Our models show that a learner's institution and application status are the strongest factors in predicting their final outcome.
  - **Result:** Gender was not a predictor, showing our model is unbiased.

Final Recommendations & Conclusion

- **Final Recommendations:**
  - **Broaden Outreach:** Actively engage with more academic fields and institutions to diversify your applicant pool, which is currently concentrated in a few key areas.
  - **Leverage Data:** Use the predictive model to proactively identify and help students who are at risk of dropping out, which will improve overall retention.

- **Conclusion:**
  - In this project, we successfully transformed raw data into a powerful tool.
  - We built a reliable predictive model and a recommendation system that can be used to drive business decisions and increase student success for Excelerate.

Excelerate

Future Scope

Our analysis and predictive model have created a powerful foundation. Here are the key areas for future work to maximize its value.

- **Automation:** Automate the entire analysis pipeline, from data cleaning to model updates. This would reduce the time it takes to get key insights from weeks to hours.
- **Real-time Monitoring:** Continuously monitor the predictive model to ensure its accuracy remains at over 95% as new data comes in.
- **Advanced Modeling:** Explore more advanced machine learning algorithms to potentially push the model's predictive power even higher.
- **Wider Applications:** Apply the models and insights to other areas of the business, such as predicting which marketing strategies will be most effective.

Excelerate

Thank You For Your Attention