WEEK – 3 REPORT

NAME : Syed Hur ABBAS Naqvi

DATE :7-july-2025

INTRODUCTION

Objective : The purpose of churn analysis is to find out why students stop using the platform or leave before completing a program. By looking at
how students use the website, which programs they join, and when they become inactive, we can spot patterns and reasons for student drop-off.

For example, some students may leave due to a lack of engaging content, difficulty navigating the site, or not finding programs that match their interests. Others might drop off because of technical issues, a lack of timely support, or feeling disconnected from the community. By collecting feedback and analyzing user behavior, we can better understand the
specific points where students lose interest or face challenges. This helps in making changes, like improving content, sending reminders, or offering more support. It can also lead to the introduction of new features, such as personalized recommendations, interactive elements, or community-building activities.

Churn analysis is important because it helps keep more students engaged, and increases program completion rates, building a stronger platform over time. By continuously monitoring and addressing the reasons behind student churn, the platform can create a more positive experience, encourage long-term participation, and support the success of its users.

DATA PREPARATION

- Data Cleaning

Removed duplicate rows using drop_duplicates().

Converted 'Date of Birth' to datetime format using pd.to_datetime().

Dropped rows with missing or invalid 'Date of Birth' values using dropna().

Calculated a new feature called Age from 'Date of Birth' using today's date (2025-07-05).

Removed outliers: excluded records where age was less than 13 or greater than 60.

- Feature Selection

The following features were selected for analysis and further use:

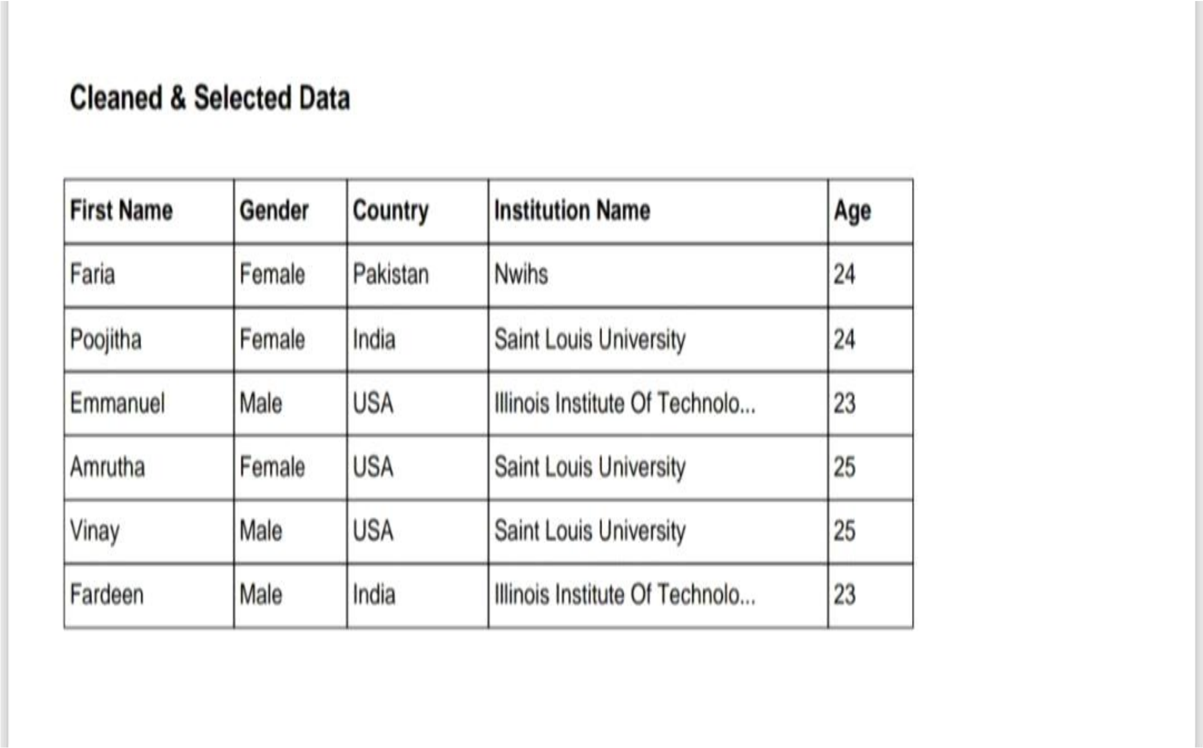
First Name

Gender

Country

Institution Name

Age (newly created from DOB)



Cleaned & Selected Data

First Name	Gender	Country	Institution Name	Age
Faria	Female	Pakistan	Nwihs	24
Poojitha	Female	India	Saint Louis University	24
Emmanuel	Male	USA	Illinois Institute Of Technolo...	23
Amrutha	Female	USA	Saint Louis University	25
Vinay	Male	USA	Saint Louis University	25
Fardeen	Male	India	Illinois Institute Of Technolo...	23

EXPLORATORY DATA ANALYSIS

Descriptive characteristics

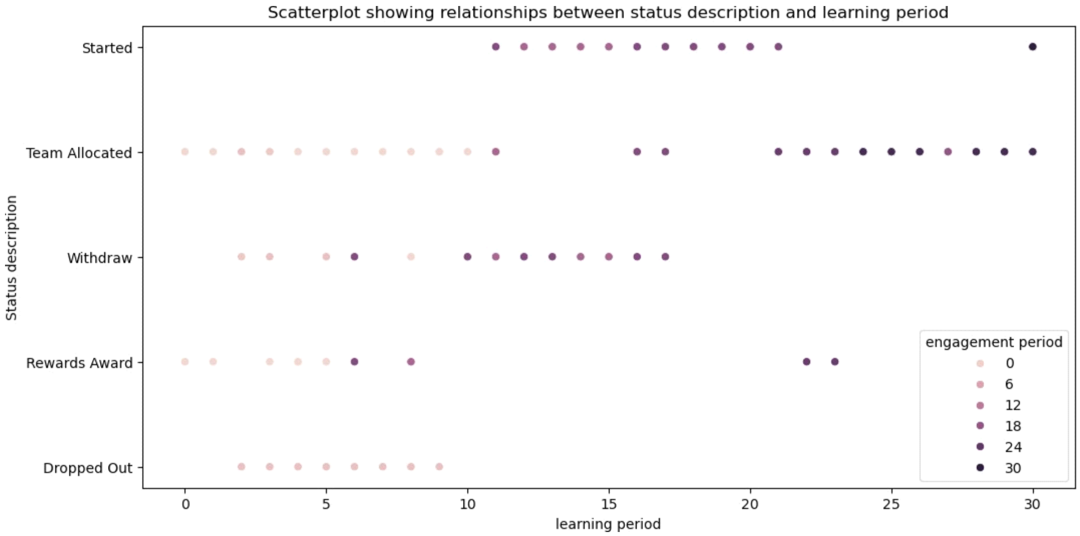
- The dataset includes demographic and academic variables such as age, gender, country, and course completion.

- Age is a continuous variable with a mean of 25.2 years.

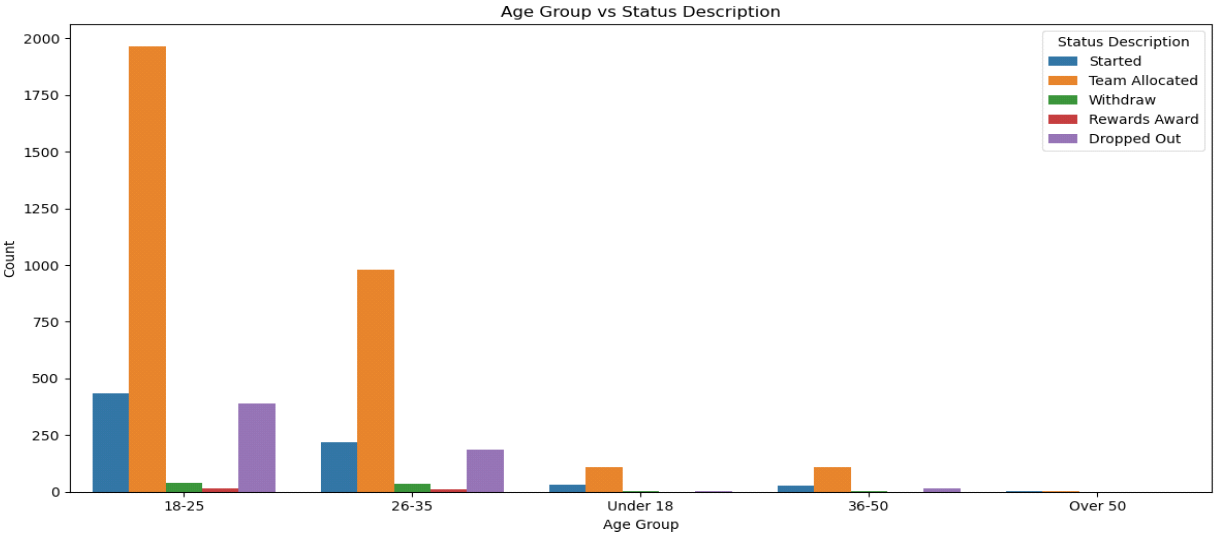
- Gender is a binary variable with more males than females.

- Country is a categorical value with 66 different countries the most represented country is USA with 1919 learners and India is the second most represented country with 1500 learners.
- Status description describes the status of the students through the program.
  with Team Allocated having the highest number of students of 3200, and Dropped Out has around 600 students.

Visualisation

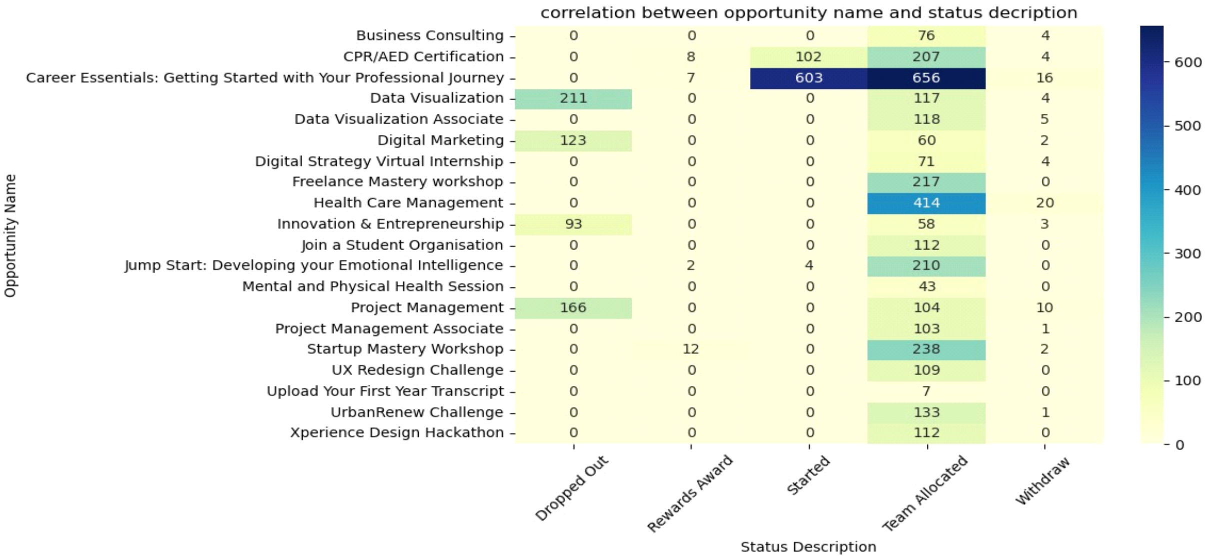


- Longer learning and engagement periods are associated with more positive statuses (Started, Team Allocated, Rewards Award), while negative statuses (Withdraw, Dropped Out) occur across all periods. There is a visible trend that higher engagement is linked to better outcomes.
- ‘Withdraw’ and ‘dropped out’ statuses are spread across various learning periods suggesting that dropping out can occur or withdrawing can occur at any stage.
- ‘Rewards Award’ status appears at higher learning and engagement periods indicating that rewards are more likely for those with longer involvement
- The engagement period tends to be higher for ‘Started’ ‘Team Allocated’,and ‘Rewards Award’ statuses and lower for ‘Withdraw’ and ‘Dropped Out’



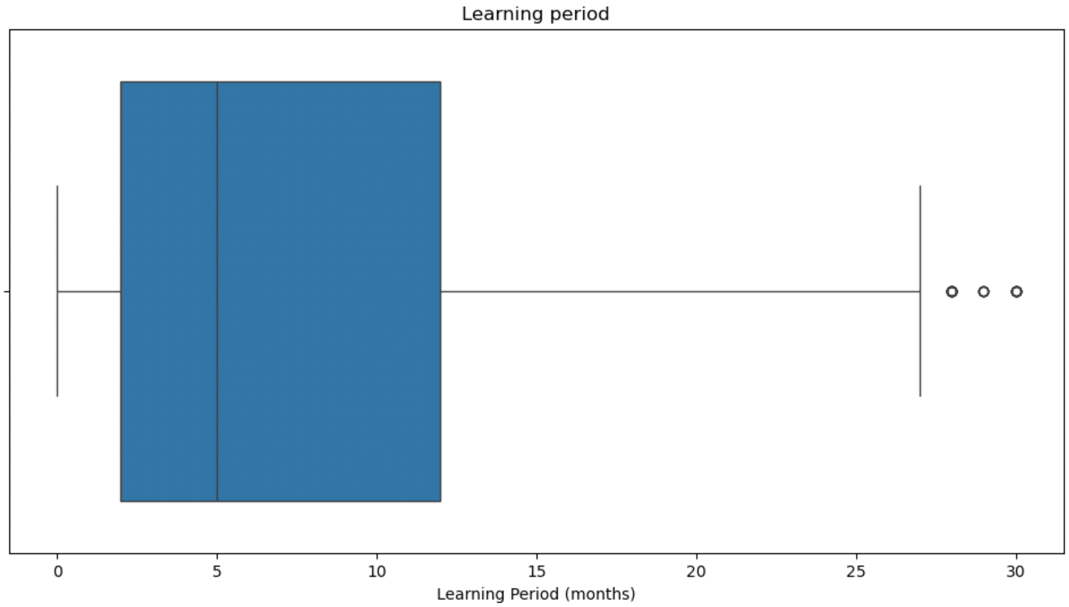
- Most of the learners are young adults aged 18-25.
- ‘Team Allocated ‘ is the most common statuses.

- Dropouts and withdrawals occur in all age groups but are much less. frequent than team allocations.



Some opportunities like ‘Data Visualisation Digital Marketing and project management had a number of notable drop outs.

Opportunities like free mastery workshop and join a student organisation and upload your first year transcripts have a very low participation or status counts.



- Most learners have a learning period of 0-12 months with a median of around five months.
- Some learners take much longer to complete their learning period which are considered as outliers.
- The distribution is right skewed.

PREDICTIVE MODELING

Preparation:

- Removed irrelevant columns like names, dates, and IDs that don't help predict outcomes.
- Created new features from dates: student age and time between signup and application
- Converted categorical data to numbers: gender (0/1), tech major classification, and opportunity categories
- Created target variable: binary drop-off indicator based on withdrawal/dropout status
- Cleaned up missing values by filling with column averages

- Prepared data for modeling by separating features from the target variable

Training:

- Trained three machine learning models with specific configurations: Logistic Regression (1000 max iterations, balanced class weights), Decision Tree (random state 42), and Random Forest (100 estimators, random state 42)
- Uses stratified train-test split (70/30) to preserve the original class distribution between dropout and non-dropout students in both training and testing sets
- Applies balanced class weights and random states to handle class imbalance issues and ensure reproducible results across multiple runs

Evaluation:

- Iteration 0:
  - Evaluates model performance with comprehensive metrics (accuracy, precision, recall, F1) and generates visualizations including confusion matrices, and ROC curves
  - Analyzes feature importance for tree-based models to identify which student characteristics are most predictive of dropout risk
- Iteration 1:

  - Evaluates model performance with comprehensive metrics (accuracy, precision, recall, F1) and generates visualizations including confusion matrices, and ROC curves
  - Analyzes feature importance for tree-based models to identify which student characteristics are most predictive of dropout risk
  - Uses stratified train-test split (70/30) to preserve the original class distribution between dropout and non-dropout students in both training and testing sets
  - Applies balanced class weights and random states to handle class imbalance issues and ensure reproducible results across multiple runs
- Iteration 2:
  - Optimizes prediction thresholds by finding the threshold that maximizes F1 score instead of using the default 0.5, which is crucial for imbalanced datasets like dropout prediction
  - Evaluates model performance with comprehensive metrics (accuracy, precision, recall, F1) and generates visualizations

    including confusion matrices, ROC curves, and threshold tuning curves
  - Analyzes feature importance for tree-based models to identify

which student characteristics are most predictive of dropout risk

Results:

Run 1:

Logistic
Accuracy : 0.8673469387755102
Precision: 0.0
Recall : 0.0 F1 Score : 0.0

Logistic

	precision	recall	f1-score	support
0	0.87	1.00	0.93	1191
1	0.00	0.00	0.00	181
accuracy			0.87	1372
macro avg	0.43	0.50	0.46	1372
weighted avg	0.75	0.87	0.81	1372

Tree
Accuracy : 0.8534985422740525
Precision: 0.4315068493150685
Recall :
0.34806629834254144 F1
Score :
0.3853211009174312

Tree

precision recall f1-score support

0	0.90	0.93	0.92	1191	
1	0.43	0.35	0.39	181	
accuracy			0.85	1372	
macro avg		0.67	0.64	0.65	1372
weighted avg		0.84	0.85	0.85	1372

Forest

Accuracy : 0.8644314868804664

Precision: 0.48299319727891155

Recall :

0.39226519337016574 F1

Score :

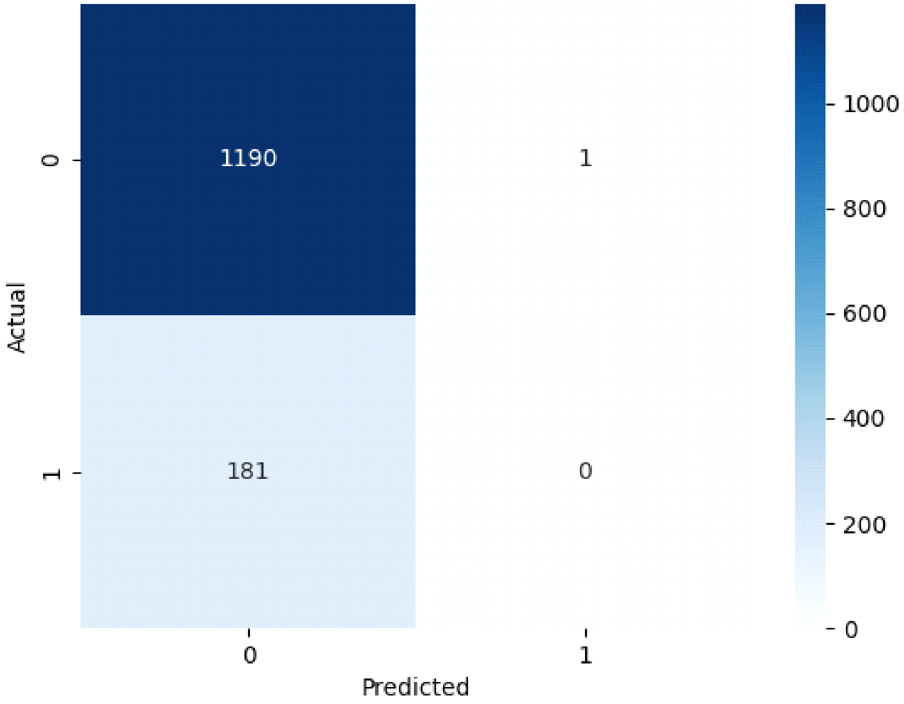
0.4329268292682927

Forest

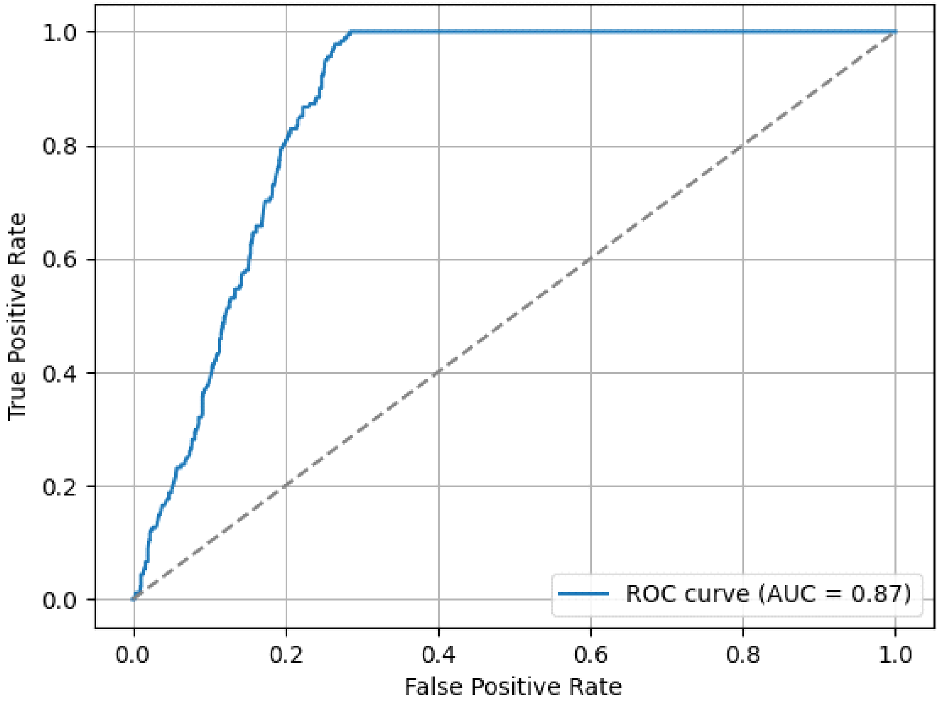
precision recall f1-score support

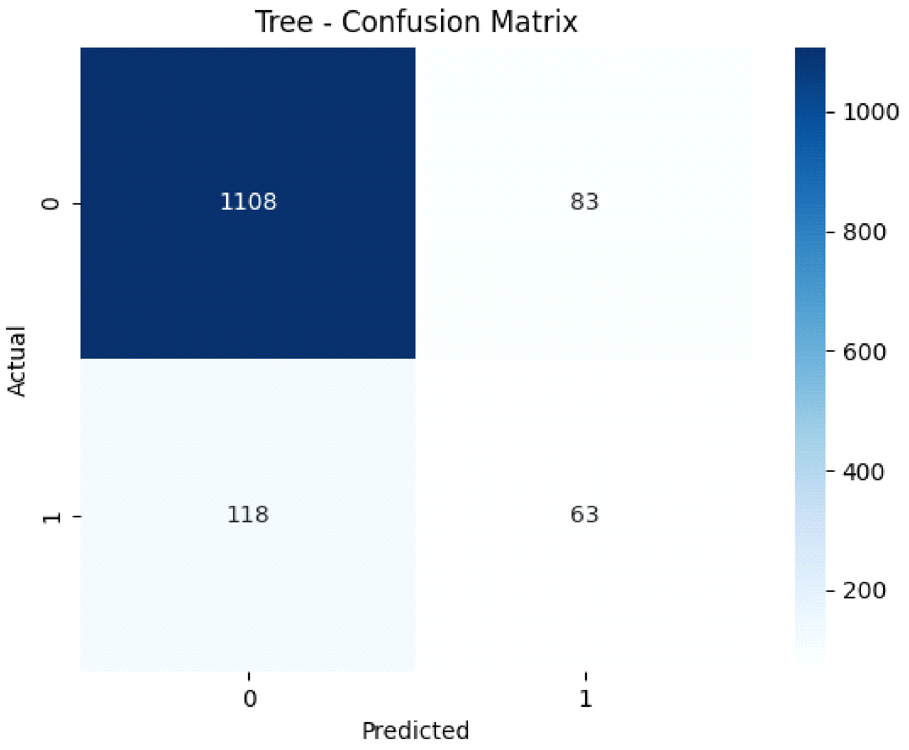
0	0.91	0.94	0.92	1191
1	0.48	0.39	0.43	181
accuracy			0.86	1372
macro avg 0.70		0.66	0.68	1372
weighted avg 0.85		0.86	0.86	1372

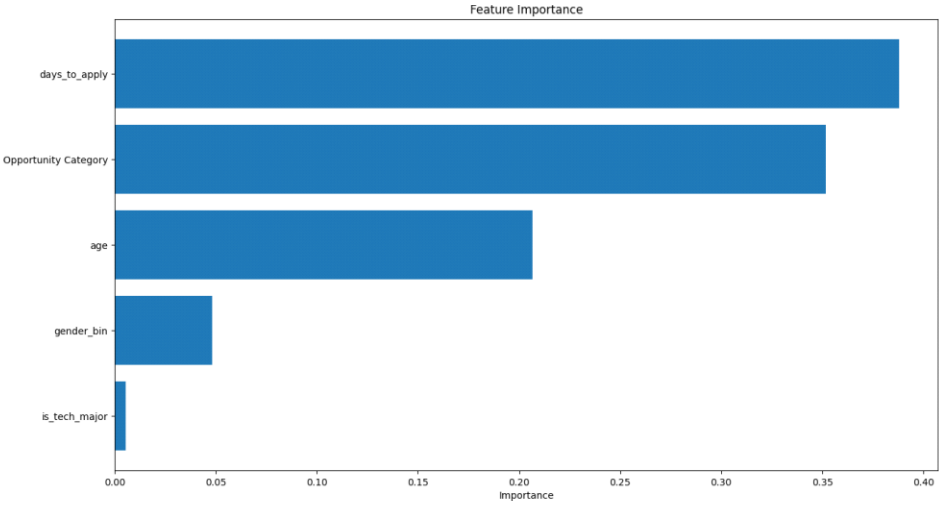
Logistic - Confusion Matrix

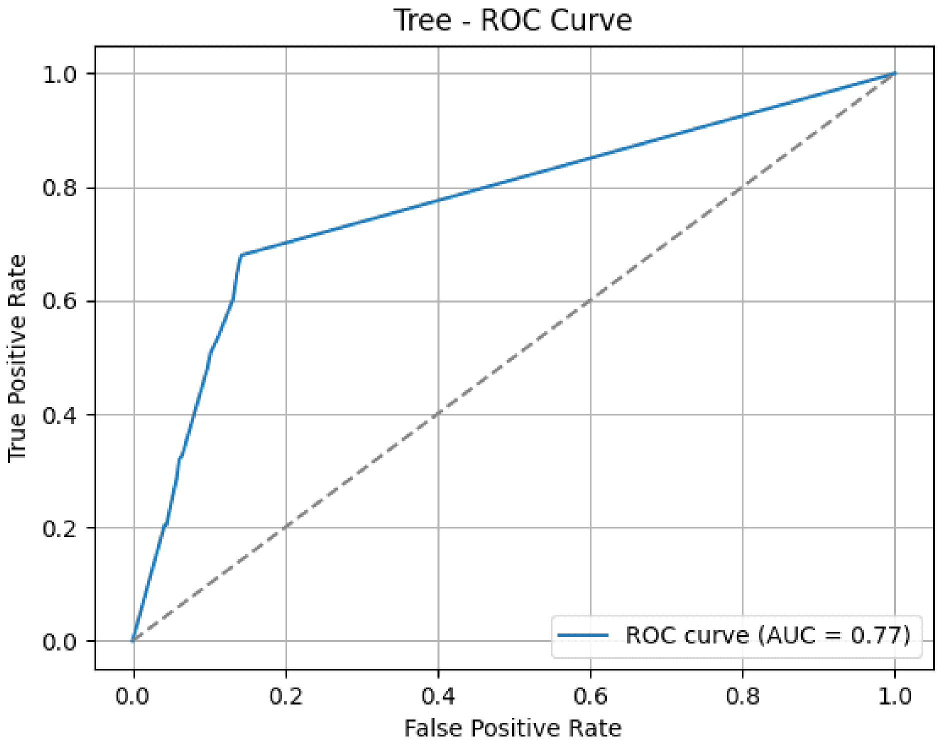


Logistic - ROC Curve

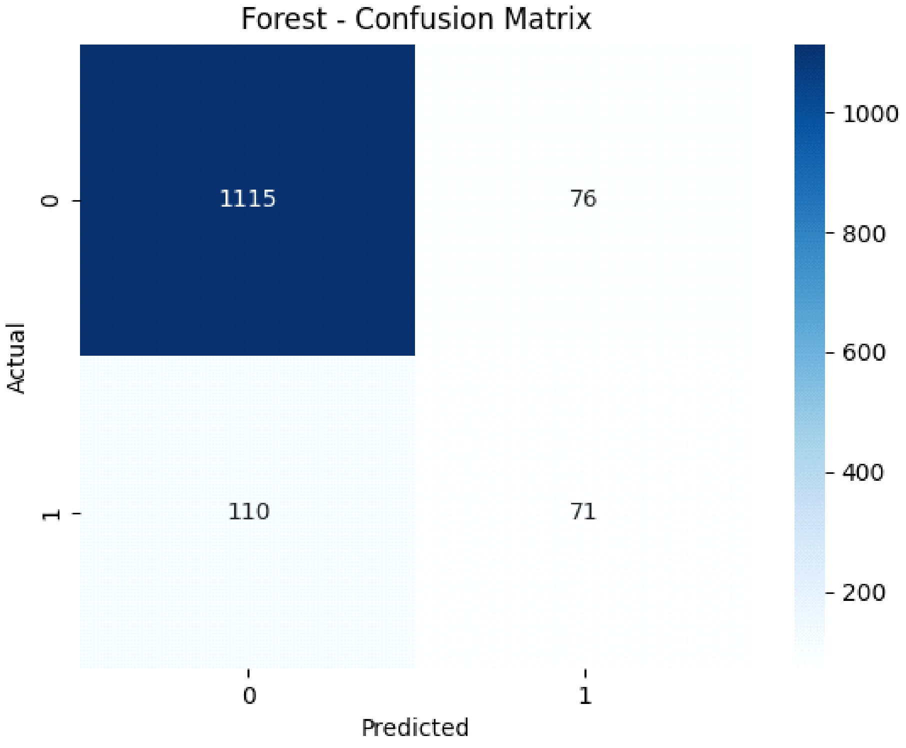



Tree - Confusion Matrix
Actual
0
1
1108
83
118
63
Predicted
0
1
1000
800
600
400
200

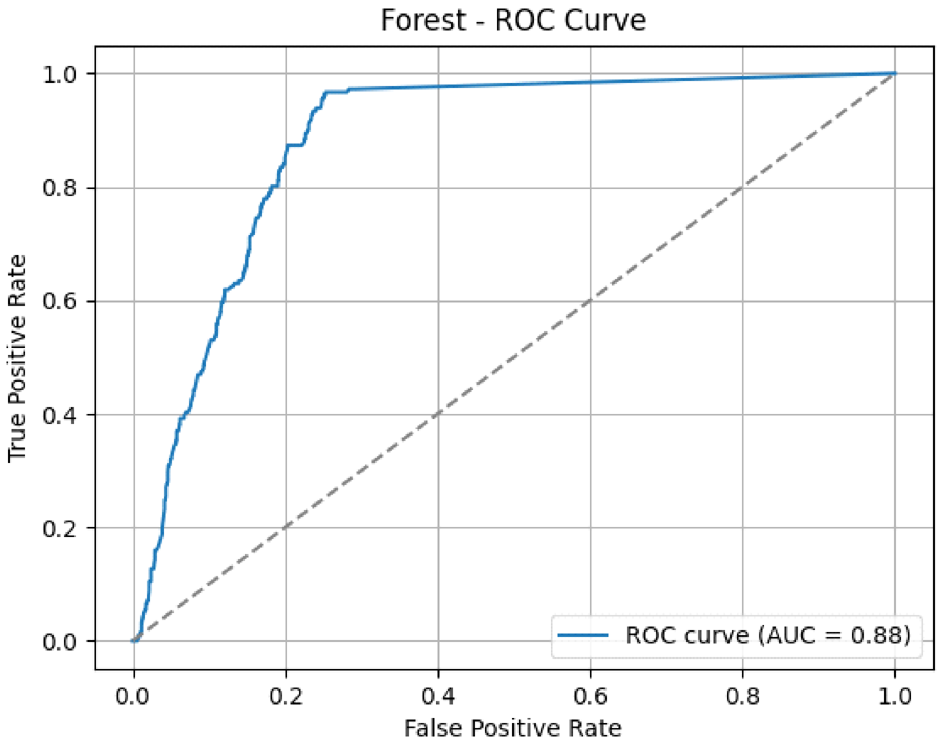

Feature Importance
days_to_apply
Opportunity Category
age
gender_bin
is_tech_major
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Importance

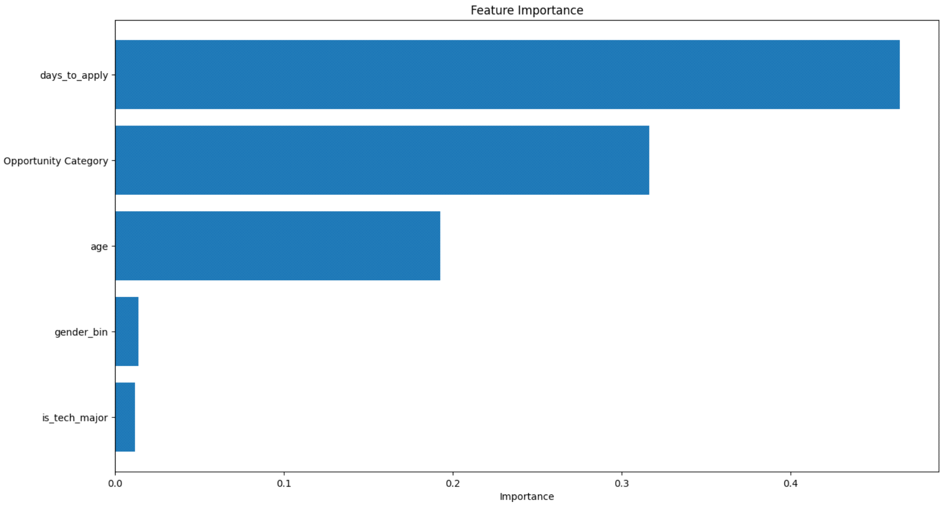


Tree - ROC Curve



Forest - Confusion Matrix





Run 2:

Logistic

Accuracy : 0.7325072886297376

Precision: 0.326605504587156

Recall   : 1.0

F1 Score : 0.49239280774550487

Logistic

precision recall f1-score support

0	1.00	0.69	0.82	1194	
1	0.33	1.00	0.49	178	
accuracy			0.73	1372	
macro avg		0.66	0.85	0.66	1372
weighted avg		0.91	0.73	0.78	1372

Tree

Accuracy : 0.8527696793002916

Precision: 0.40625

Recall :

0.29213483146067415 F1

Score :

0.33986928104575165

Tree

precision recall f1-score support

0	0.90	0.94	0.92	1194	
1	0.41	0.29	0.34	178	
accuracy			0.85	1372	
macro avg		0.65	0.61	0.63	1372
weighted avg		0.83	0.85	0.84	1372

Forest

Accuracy : 0.8520408163265306

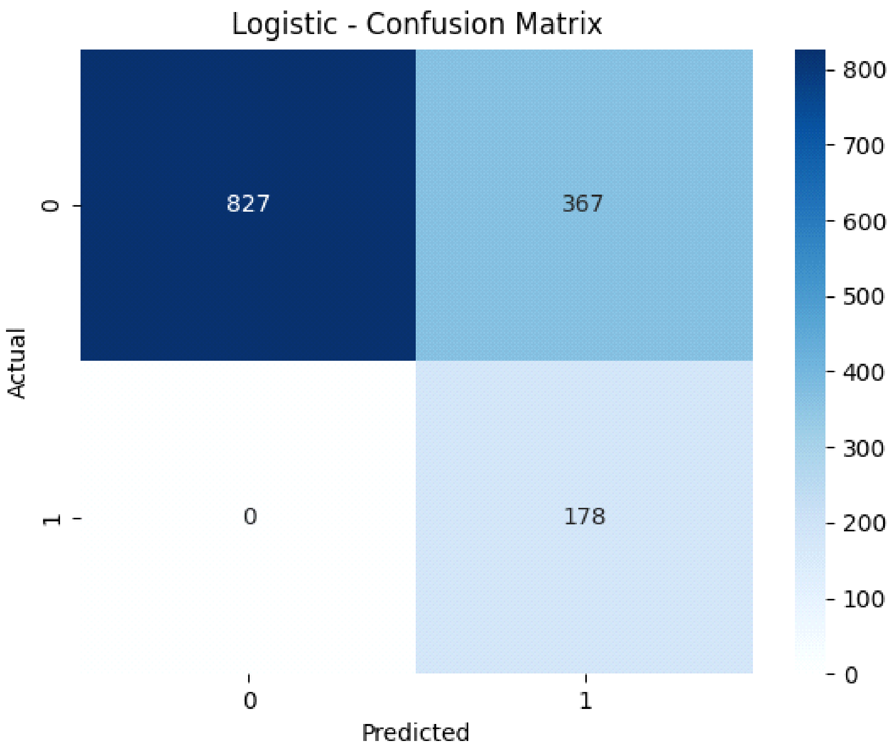
Precision: 0.40601503759398494

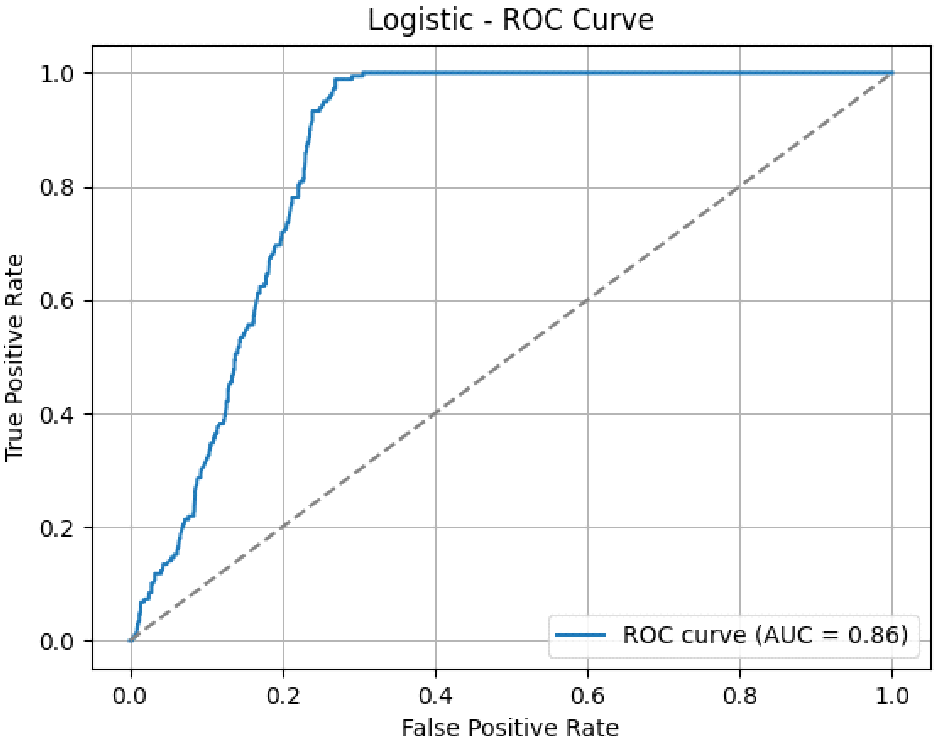
Recall :
0.30337078651685395 F1
Score :
0.34726688102893893

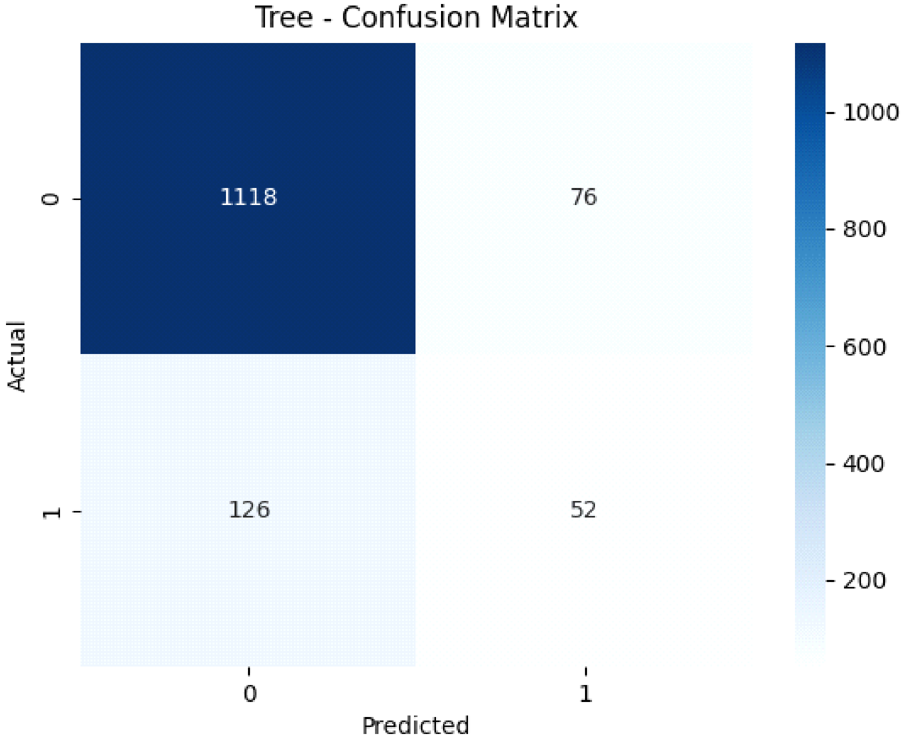
Forest

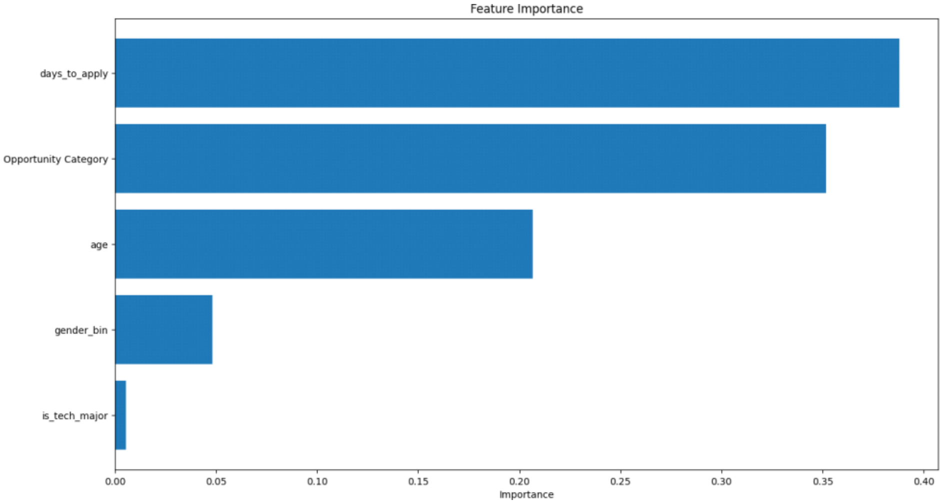
precision recall f1-score support

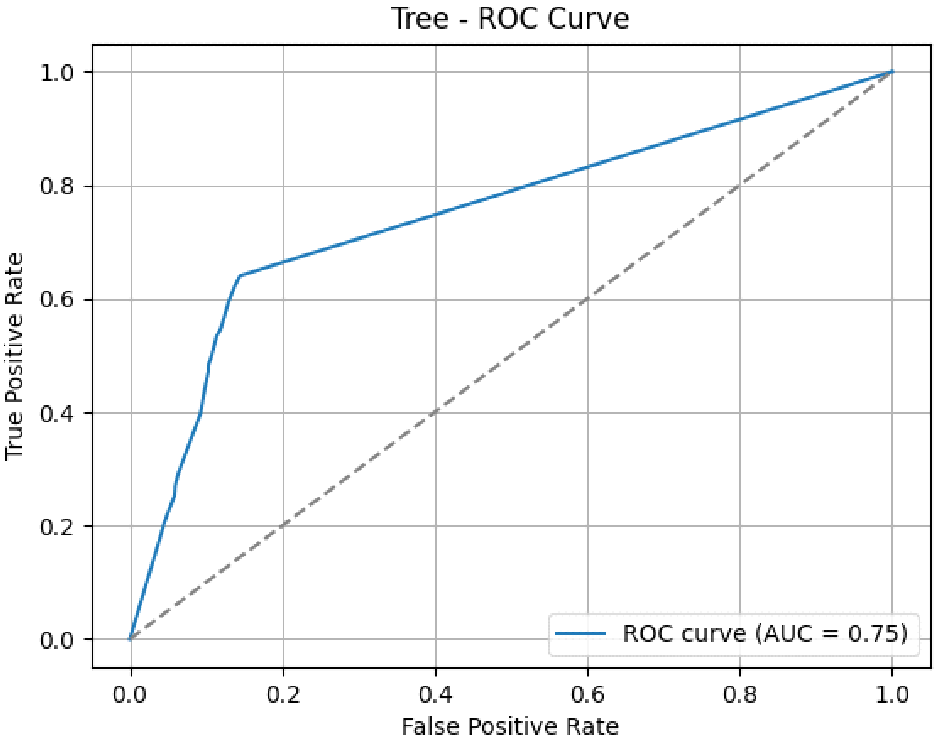
0	0.90	0.93	0.92	1194
1	0.41	0.30	0.35	178
accuracy		0.85	1372	
macro avg	0.65	0.62	0.63	1372
weighted avg	0.84	0.85	0.84	1372

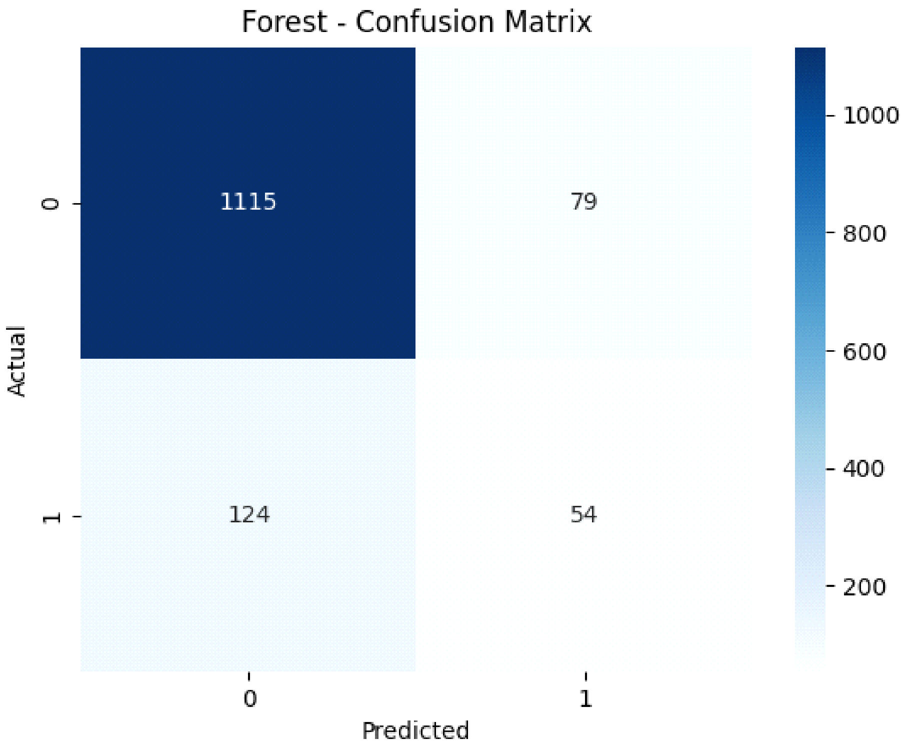


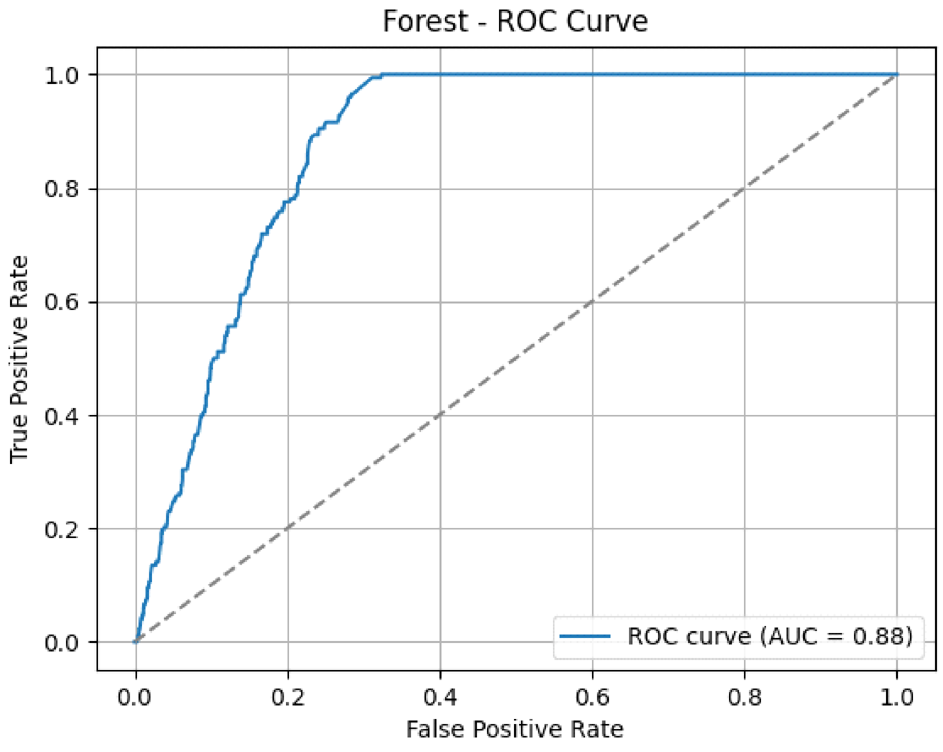

Logistic - ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (AUC = 0.86)

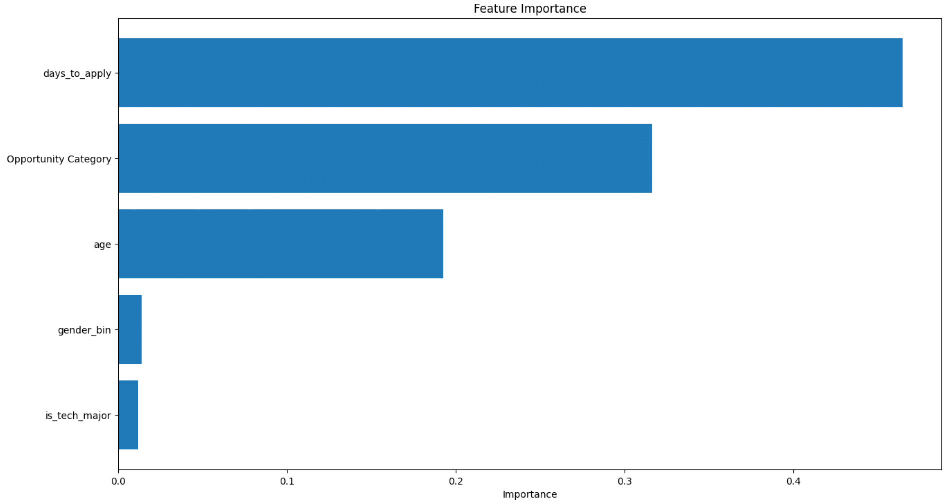

Tree - Confusion Matrix
Actual
Predicted
0
1
0
1
1118
76
126
52
1000
800
600
400
200


Feature Importance
days_to_apply
Opportunity Category
age
gender_bin
is_tech_major
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Importance


Tree - ROC Curve
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
ROC curve (AUC = 0.75)


Forest - Confusion Matrix
Actual
Predicted
0
1
0
1
1115
79
124
54
1000
800
600
400
200


Forest - ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (AUC = 0.88)



Run 3:

Logistic - Threshold Tuning Best Threshold: 0.74 (F1: 0.527)

Accuracy : 0.7827988338192422

Precision: 0.3672566371681416

Recall : 0.9325842696629213 F1 Score : 0.526984126984127

precision recall f1-score support

0	0.99	0.76	0.86	1194
1	0.37	0.93	0.53	178
accuracy		0.78	1372	
macro avg	0.68	0.85	0.69	1372
weighted avg	0.91	0.78	0.82	1372

Tree - Threshold Tuning

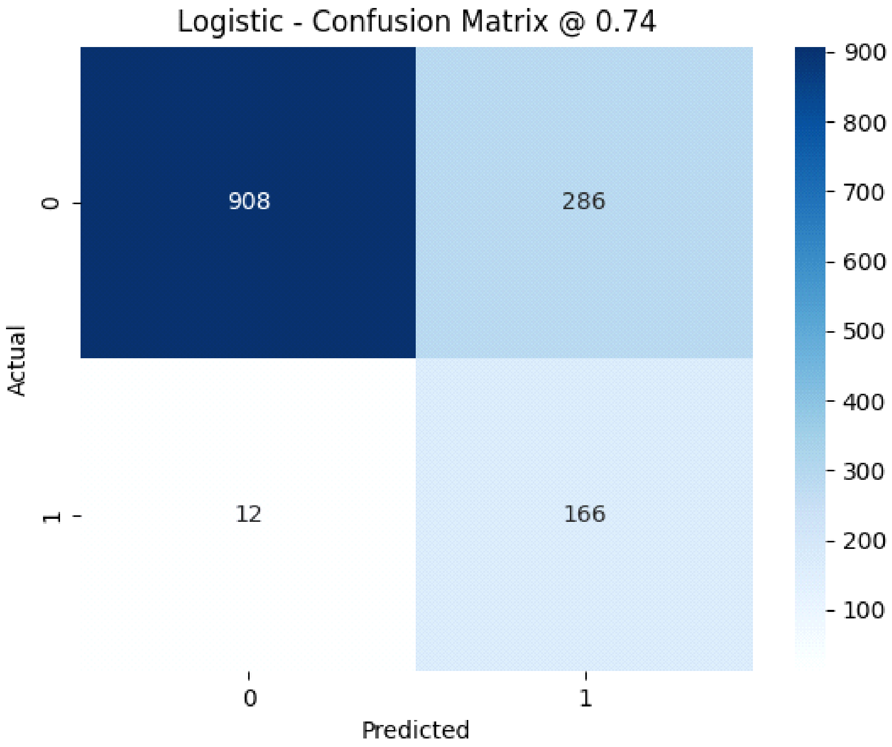
Best Threshold: 0.08 (F1: 0.490)

Accuracy : 0.827259475218659
Precision: 0.397212543554007
Recall :
0.6404494382022472 F1
Score :
0.49032258064516127

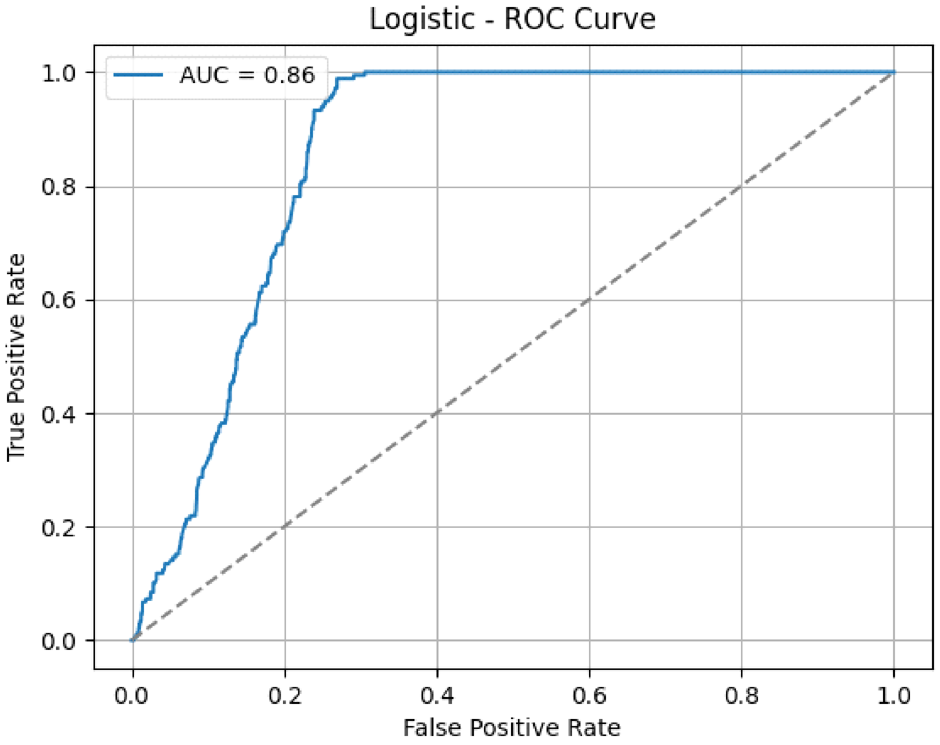
	precision	recall	f1-score	support
0	0.94	0.86	0.90	1194
1	0.40	0.64	0.49	178
accuracy		0.83	1372	
macro avg	0.67	0.75	0.69	1372
weighted avg	0.87	0.83	0.84	1372

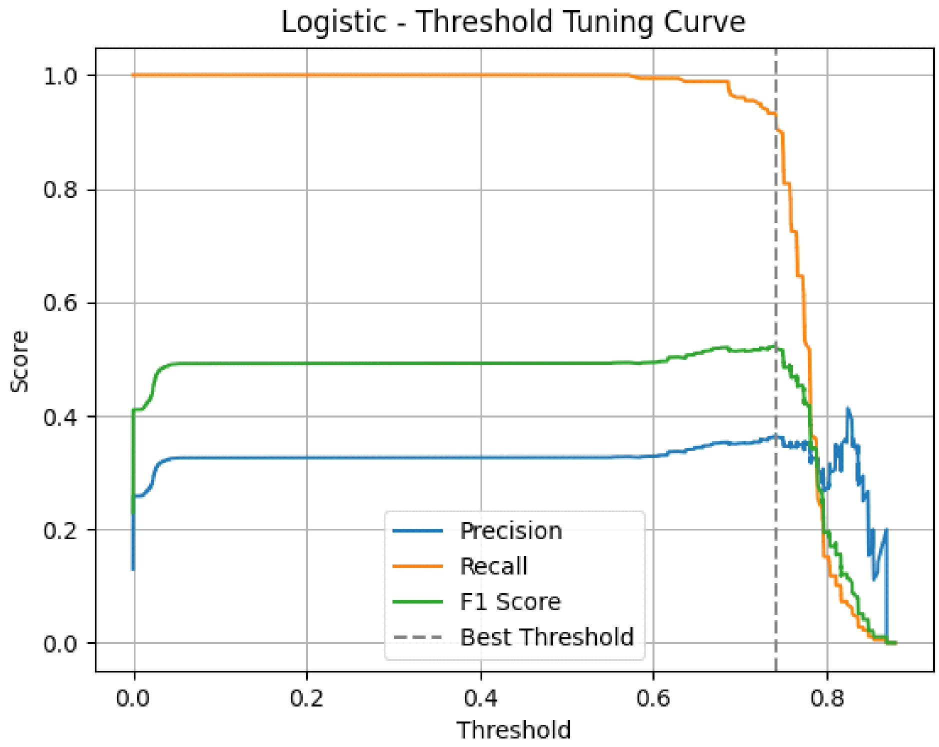
Forest - Threshold Tuning
Best Threshold: 0.06 (F1: 0.516)
Accuracy : 0.7842565597667639
Precision: 0.3640552995391705
Recall :
0.8876404494382022 F1
Score :
0.5163398692810458

	precision	recall	f1-score	support
0	0.98	0.77	0.86	1194
1	0.36	0.89	0.52	178
accuracy		0.78	1372	
macro avg	0.67	0.83	0.69	1372
weighted avg	0.90	0.78	0.82	1372

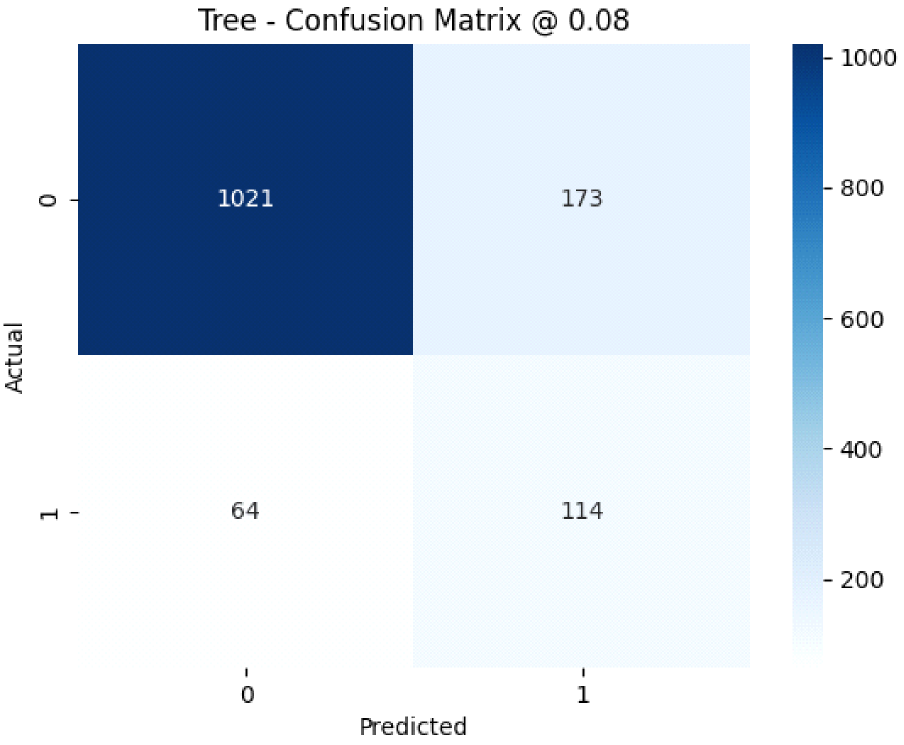

Logistic - Confusion Matrix @ 0.74

Actual \ Predicted	0	1
0	908	286
1	12	166

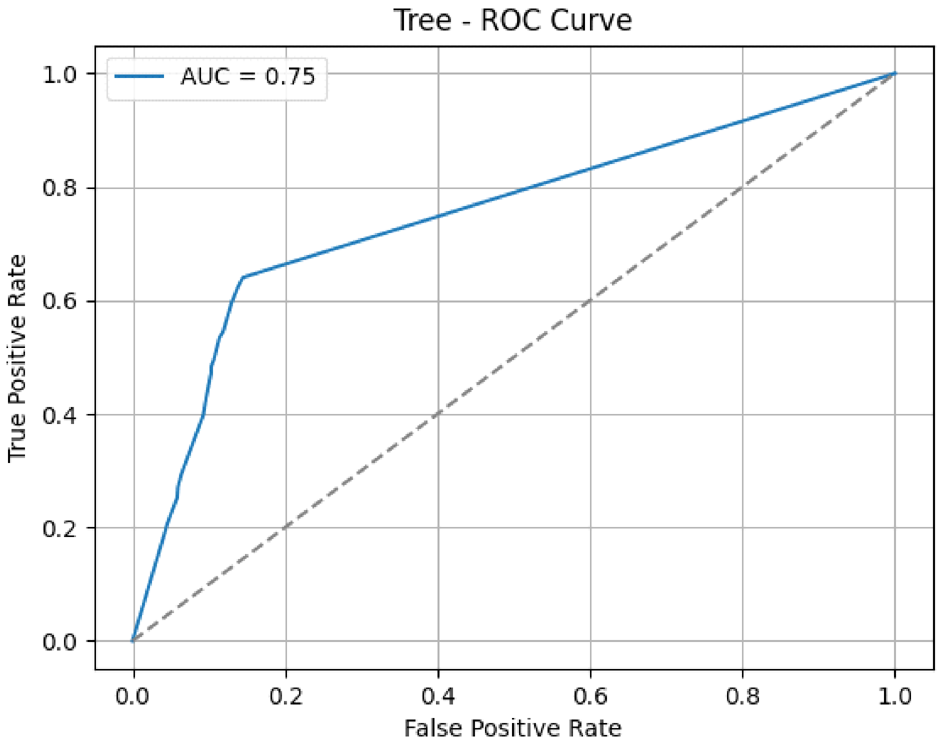

Logistic - ROC Curve

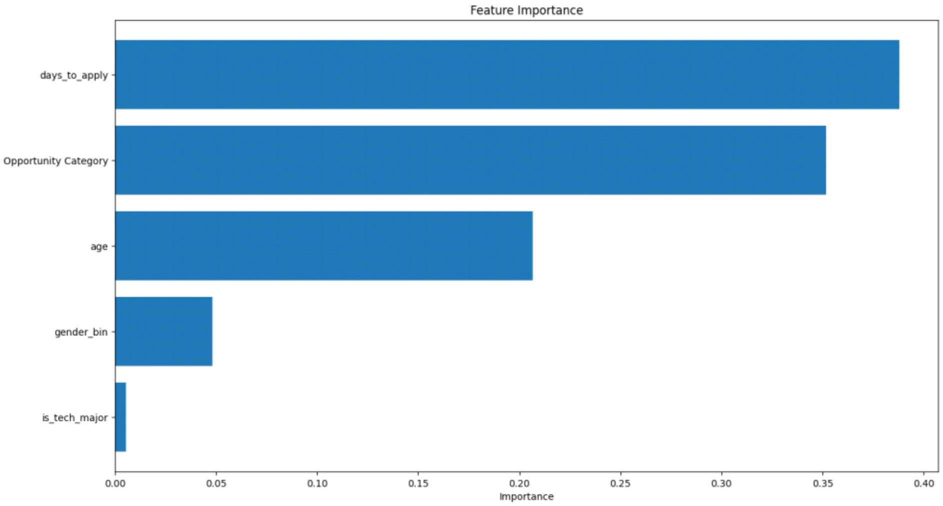


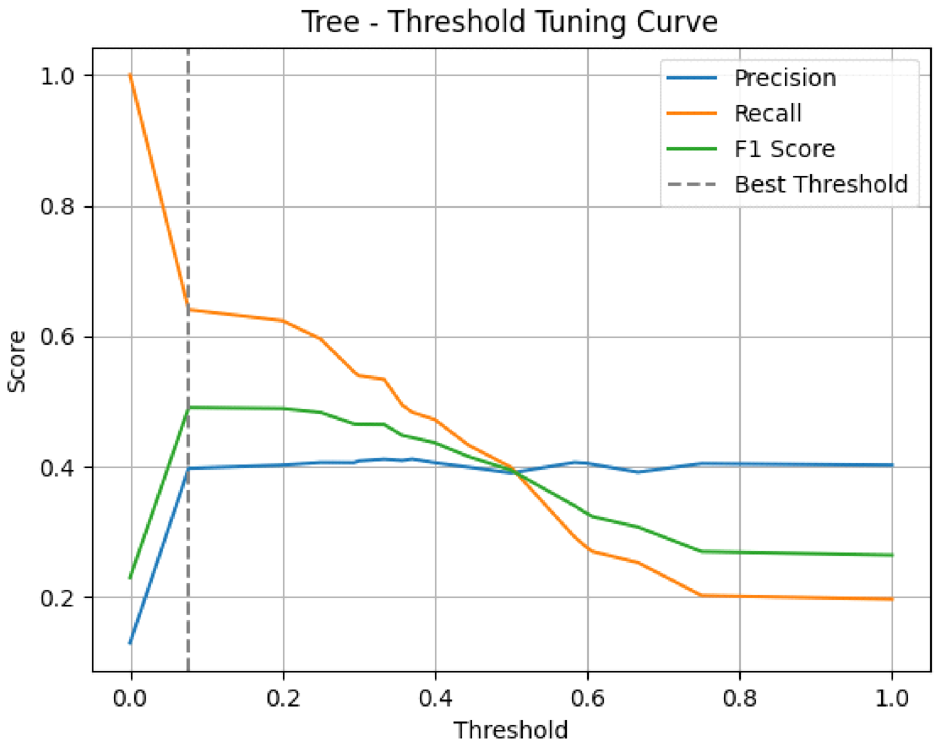
Logistic - Threshold Tuning Curve

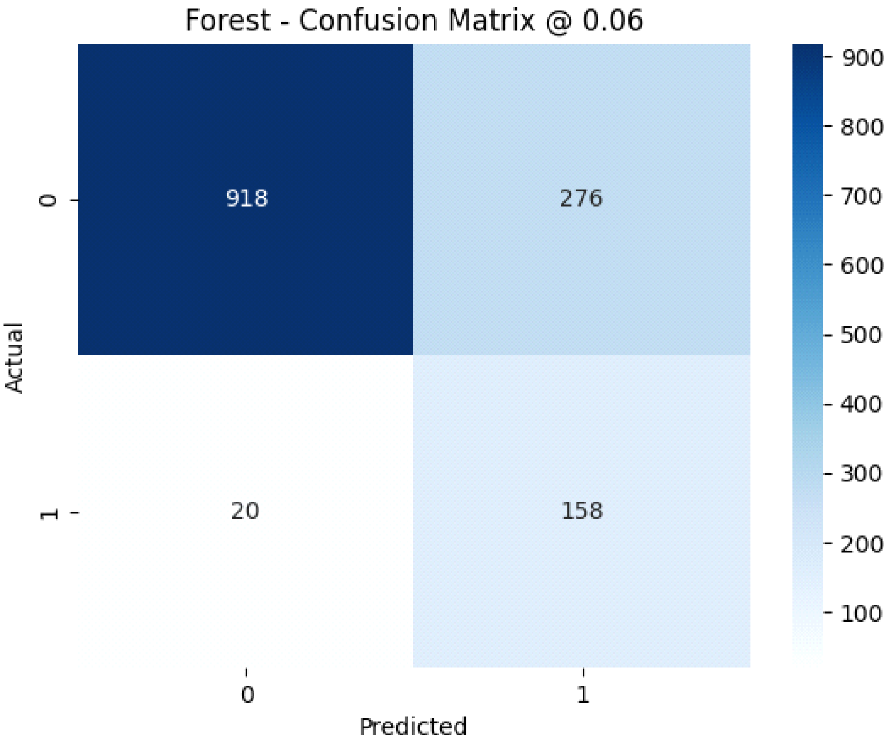


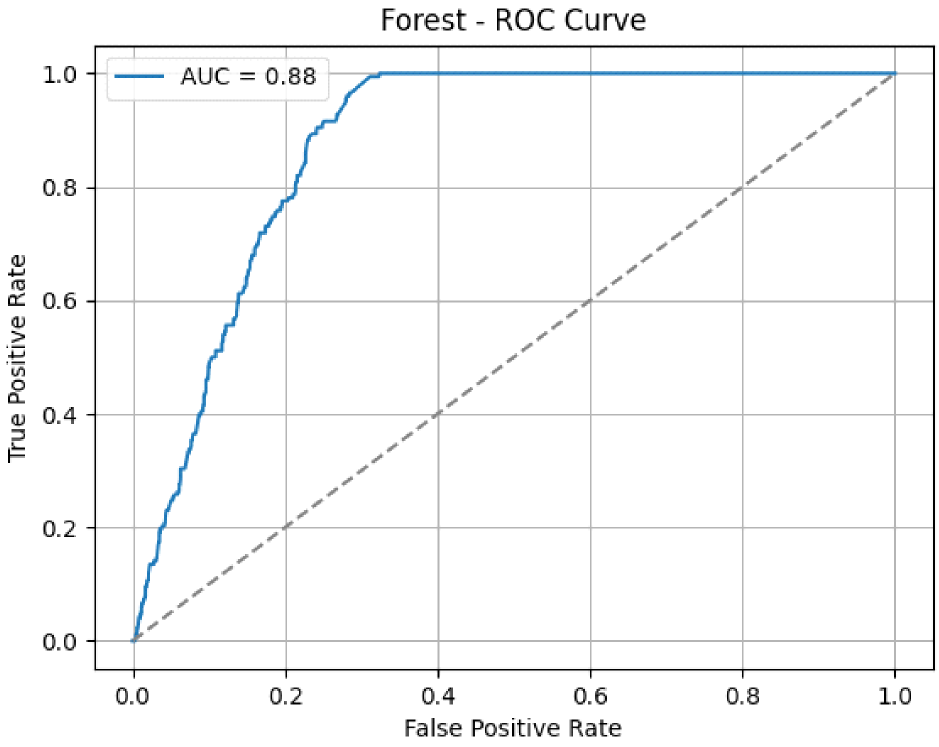
Tree - Confusion Matrix @ 0.08

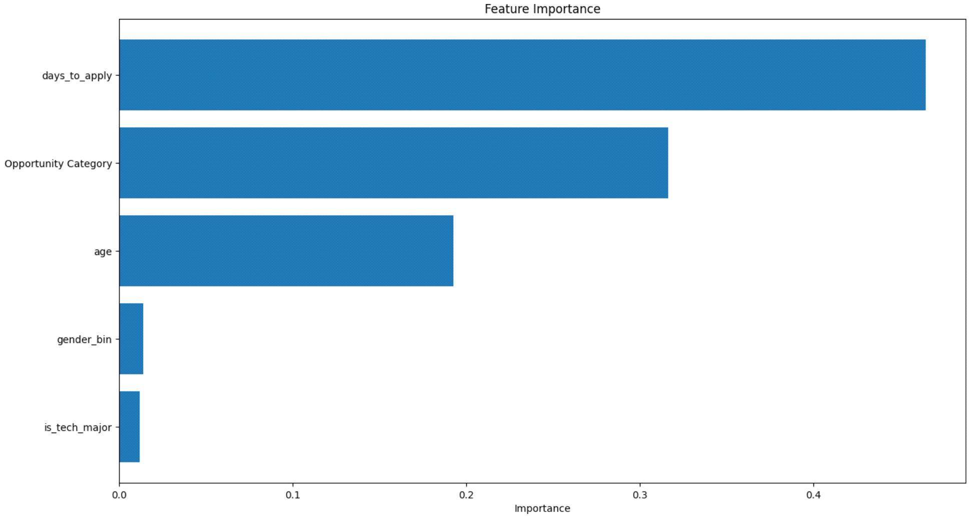

Tree - ROC Curve
AUC = 0.75
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

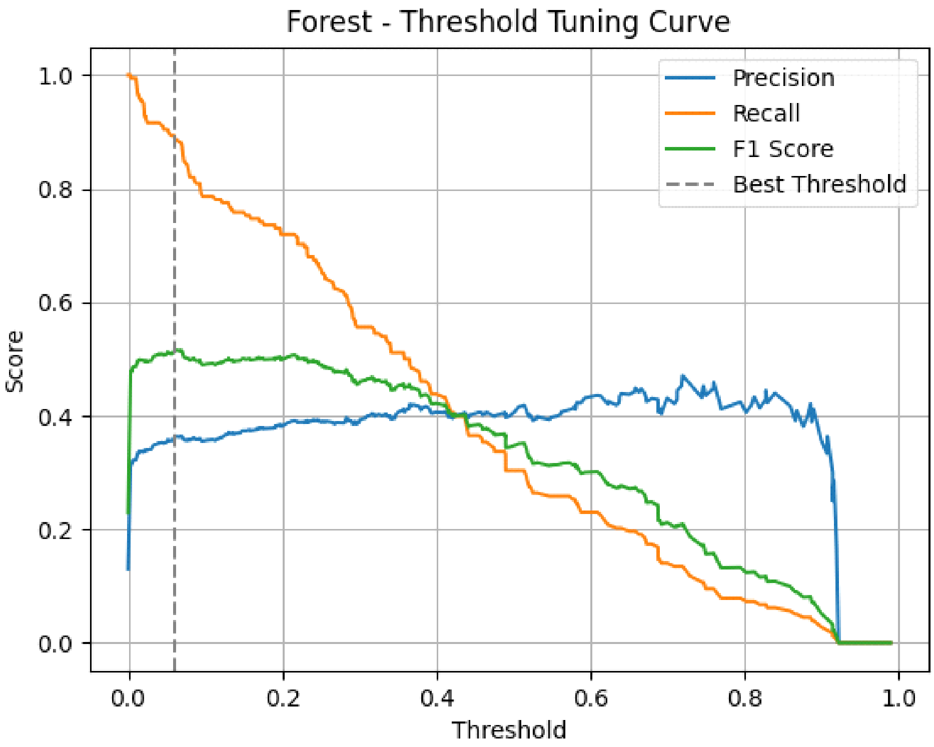

Feature Importance
days_to_apply
Opportunity Category
age
gender_bin
is_tech_major
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Importance


Tree - Threshold Tuning Curve
Precision
Recall
F1 Score
Best Threshold
Score
Threshold
1.0
0.8
0.6
0.4
0.2
0.0
0.2
0.4
0.6
0.8
1.0


Forest - Confusion Matrix @ 0.06
918
276
20
158
Actual
Predicted
0
1
0
1
900
800
700
600
500
400
300
200
100


Forest - ROC Curve
AUC = 0.88
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0


Feature Importance
days_to_apply
Opportunity Category
age
gender_bin
is_tech_major
0.0
0.1
0.2
0.3
0.4
Importance



CHURN ANALYSIS

Predictive modeling, Exploratory Data Analysis, and Data preparation revealed that the following parameters influenced churn (drop-off/withdrawal);

- **Low Duration of Engagement:**

Students are far more likely to drop out if their learning and engagement periods are short between 0 to 5 months. One important indicator is engagement. Higher levels of involvement are highly correlated with

favorable outcomes such as Team Allocated or Rewards Award statuses.

- **Opportunity Type:**

Higher dropout rates are associated with specific programs such as:

- **Data Visualization**
- **Digital Marketing**
- **Project Management**

These may have higher workload, lower alignment with student expectations, or unclear outcomes.

- **Poor Program Fit:**

Low participation rates for opportunities like Join a Student Organization or the Free Mastery Workshop could be the result of inadequate promotion or a mismatch between the programs and the interests of the students.

- **No Early Intervention:**
  Drop-offs occur at any stage of the student or interns lifecycle, not only at the start.

This suggests that without being noticed or assisted, interns or students gradually stop participating in the course or internship.

- **Influence of Demographics (Country & Age):**

Modeling indicates that age and nationality still have an impact, even though the bulk of dropouts are between the ages of 18 and 25.

To quantify this, more research may be required, but some countries or age groups might not receive enough support.

- **Absence of Tailored Support or Follow-Up:**

Students may feel disengaged or overburdened if there are no reminders, mentors, or adaptive paths available.

- **Friction between Technical and UX (Deduced from Introduction):**

Lack of interactive features, platform problems, or navigational challenges can all lead to student annoyance and eventual attrition.

Both qualitative (from exploratory Data analysis) and quantitative (from model feature importance, particularly in Decision Tree and Random Forest models) evidence supported these aspects.

Impact Analysis of Churn Factors

- **Low Duration of Engagement**

Impact: Strongest predictor. of churn.

Dropout rates are much greater for students or interns who participate for less than five months.Longer durations for example 6 months and above are associated with more dedication and a higher likelihood of achieving favorable results such as Team Allocated and Rewards Award.

Engagement time is emphasized by models like Random Forest and Decision Tree as a key factor in drop-off prediction.

- **Opportunity Type**

Impact: Certain opportunities correlate with higher churn.

Significant dropout rates are seen in programs including project management, digital marketing, and data visualization, which may indicate a misalignment with student expectations, potentially too much information or inadequate assistance. Enrolling in these programs increases the likelihood that learners or interns would leave early, which lowers retention rates overall.

- **Poor Program Fit**

Impact: Indicates missed engagement opportunities.

The low engagement rate for features like Join a Student Organization and Free Mastery Workshop may indicate that students are not aware of these opportunities.

Offerings are viewed as uninteresting or unimportant.Lack of wider participation makes the platform less "persistent," which raises the dropout risk.

- **No Early Intervention**

Impact: Churn is not limited to early stages.

Despite achieving mid-level engagement, students with statuses such as "Started" and "Team Allocated" continue to drop out.This implies that short-term involvement is insufficient.

Re-engagement tactics and continuous support are crucial otherwise even highly engaged students may churn if assistance erodes over time, according to modeling.

- **Influence of Demographics (Age and Country) Impact**: Moderate.

Age still influences dropout prediction in models, even if the dataset is dominated by people aged 18 to 25.

Infrastructure, time zone, or cultural constraints may contribute to higher turnover in some areas; however, this requires more segmentation to measure. Models show age as an important but not primary factor.

- **Absence of Tailored support or Follow-up Impact**: Subtle but important.

Lack of personalized reminders, check-ins, or flexible learning pathways might make students feel lost or abandoned.Without action, the chance of churn increases as engagement decreases.

- **Friction Between Technical**

  & UX Impact:

  Foundational.

  Silent disengagement may result from frustration with difficult platform navigation or technological issues. Even while they might not always show up in organized data, these friction points have a big impact on retention and overall experience.

RECOMMENDATIONS

- **Strategies:Actionable strategies to improve student retention.**

- **Encourage more consistent student**

engagement Reason:

Students who remain engaged for less than five months are much more likely to drop out of their programs.

Recommendation:

Encourage ongoing participation by setting clear goals, sending regular reminders, and offering small rewards such as badges or certificates when students reach key milestones.

- **Revise high-churn**

programs Reason:

Certain programs like Data Visualization, Digital Marketing, and Project Management have a higher number of students dropping out.

Recommendation:

These programs should be reviewed and improved by simplifying the content, clarifying learning outcomes, and providing additional support where needed. Collecting direct feedback from students can also help identify areas for improvement.

- **Promote underused programs more**

effectively Reason:

Some programs, such as Join a Student Organization and Free Mastery Workshop, have very low participation.

Recommendation:

Increase awareness of these programs by highlighting them during the onboarding process or through targeted email campaigns. Align the program offerings with student interests using surveys and improve how their benefits are communicated.

- **Provide early and ongoing**

interventions Reason:

Many students drop out at various stages, including after being assigned to teams, not just at the beginning.

Recommendation:

Use platform activity data to identify when students become inactive, and follow up with automated reminders, check-ins, or mentor support at
regular intervals to re-engage them.

- **Introduce personalized reminders and progress**

updates Reason:

Students often lose motivation or feel disconnected when there is no personalized support.

Recommendation:

Set up a system to send custom reminders based on individual engagement patterns and share regular progress updates. Offer easy access to a mentor or support team for questions and guidance.

- **Improve technical and user experience on the**

platform Reason:

Students may quietly disengage if the platform is difficult to use or lacks helpful features.

Recommendation:

Conduct regular usability testing to find and fix navigation issues, ensure the platform works well on mobile devices, and add features like discussion forums, progress tracking, and live support to keep students involved.

- **Interventions: Specific interventions for identified at-risk students.**

- **Low early engagement leads to higher**

dropout rates Reason:

Many students leave the platform within the first 3 to 5 months, especially if they do not stay active during the initial learning phase.

Recommendation:

Encourage regular engagement by setting short, achievable milestones. Use scheduled reminders and provide small motivational rewards such as badges or certificates to help students stay committed.

- **Specific programs have higher**

dropout rates Reason:

Programs like Data Visualization, Digital Marketing, and Project Management show more student dropouts, possibly due to complex content or unclear learning outcomes.

Recommendation:

Simplify and restructure these courses for better clarity and accessibility. Clearly communicate the learning goals and provide additional support such as FAQs, mentorship, or tutorial videos based on feedback from enrolled students.

- **Lack of personalized communication causes silent**

disengagement Reason:

Students often stop participating without warning when they feel unsupported or disconnected from the learning process.

Recommendation:

Set up a personalized reminder system that responds to each student's activity level. Send progress updates and provide easy access to mentors or support staff to answer questions and guide them through challenges.

- **Demographic factors affect student**

success Reason:

Dropout patterns vary based on a student’s age and country. Issues like time zone differences, language barriers, or regional limitations may affect their learning experience.

Recommendation:

Adapt your communication and support based on region and age group. Offer flexible learning schedules, local mentorship where possible, and culturally relevant content or examples to improve engagement and retention.

CONCLUSION

Summary

The analysis uncovered several key factors that significantly influence student churn or withdrawal across learning programs.

One of the strongest predictors of churn was a **low duration of engagement**, especially within the first 3 to 5 months. Students who failed to remain consistently active during this period were highly likely to drop out, emphasizing the need for early and sustained engagement strategies.

Additionally, certain **opportunity types**, including **Project Management**, **Digital Marketing**, and **Data Visualization**, were consistently linked to higher dropout rates. This suggests a possible mismatch between program expectations and actual delivery, potentially driven by complex content, lack of clarity, or insufficient support.

Another critical factor was **program fit**. Programs such as *Join a Student Organization* and the *Free Mastery Workshop* showed low participation, pointing to poor promotion or misalignment with student interests.
Alongside this, the absence of **early interventions** and **personalized support** was also associated with disengagement, especially for students who dropped off after reaching mid-level milestones like team allocation.

The analysis also highlighted that **demographic factors**—especially **age and country**—play a moderate role in churn. While the majority of students were aged 18–25, differences in infrastructure, time zones, or cultural alignment may contribute to varying engagement levels.

Future work

- **Deeper demographic analysis**-segment students further by region ,language,urban/rural background, or educational level. Perform country-specific churn analysis to identify localized barriers like time zones, internet access, or cultural mismatches. Use clustering techniques to group learners by demographic patterns and engagement behaviors.
- **Qualitative Feedback on High-Churn Programs**

Programs like Data Visualization, Digital Marketing, and Project Management show high churn, but the exact causes are not clearly quantified.

- Collect and analyze student feedback, surveys, and reviews on these programs.
- Use text analysis or sentiment analysis on open-ended responses to detect pain points (e.g., difficulty, confusion, poor content).
- Interview a sample of students who dropped out of these programs to gain deeper insight.
- Identify specific content or experience gaps leading to disengagement.