Optimizing Lead Conversion: Insights and Predictions for ExtraaLearn

Classification and Hypothesis Testing Potential Customers Prediction

Jan 16

Contents

- Business Problem Overview and Solution Approach
- Data Overview
- EDA Results - Univariate and Multivariate
- Data Preprocessing
- Model Performance Summary
- Conclusion and Recommendations

Business Problem Overview and Solution Approach

- Business Problem
  - The EdTech industry has experienced rapid growth, especially during the COVID-19 pandemic, leading to a surge in online education and a significant increase in customer leads. ExtraaLearn, an early-stage EdTech startup offering upskilling programs, is generating a large number of leads through digital channels. However, the company struggles to identify which leads are more likely to convert into paying customers, resulting in inefficient resource allocation. A data-driven approach is needed to predict lead conversion, optimize resources, and improve overall conversion rates.
- Solution Approach
  1. Exploratory Data Analysis (EDA): Perform a comprehensive analysis to uncover trends and identify key features influencing lead conversion, such as time spent on the website, first interaction, and profile completion levels.
  2. Model Development: Develop classification models (Decision Tree and Random Forest) to predict the likelihood of lead conversion with a focus on maximizing recall to reduce false negatives and ensure potential customers are prioritized.
  3. Hyperparameter Optimization: Leverage GridSerachCV to fine-tune model parameters and mitigate class imbalance by applying weighted class adjustments.
  4. Feature Importance Analysis: Extract insights by ranking influential features, enabling data-backed recommendations for resource prioritization.
  5. Evaluation and Deployment: Evaluate model performance on unseen data for generalization, and deploy the optimized model to operationalize lead prioritization for better efficiency and conversion outcomes.

Data Overview

- The dataset contains 4,612 records with 15 features representing customer demographics, interactions, and advertisement exposure.

- The data has no missing values or duplicates, and behavioral metrics like time spent on the website and profile completion are key drivers of lead conversion.

- Key features include:

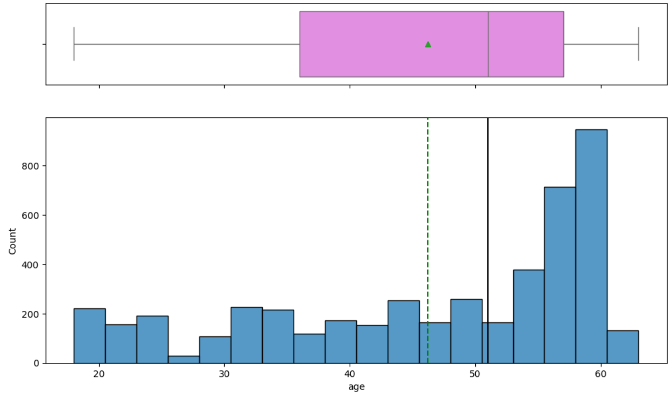
1. Demographics: Age and occupation (Professional, Unemployed, or Student)

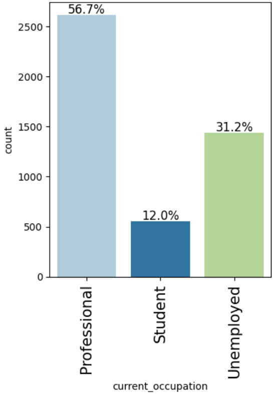
2. Interactions: First interaction channel, profile completion level, website visits, time spent on the website, and last activity (Email, Phone, or Website)

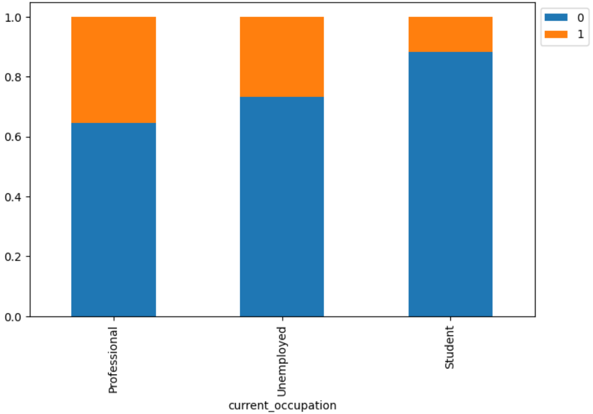
3. Advertisement Exposure: Exposure to print, digital media, educational channels, and referrals.

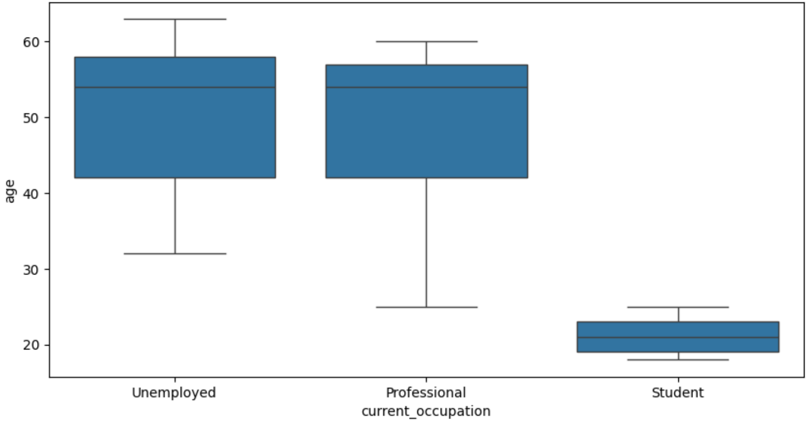
4. Target Variable: status (1 = Converted, 0 = Not Converted)

EDA Results: Observations on Demographics



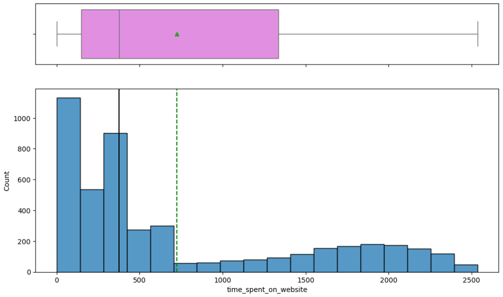


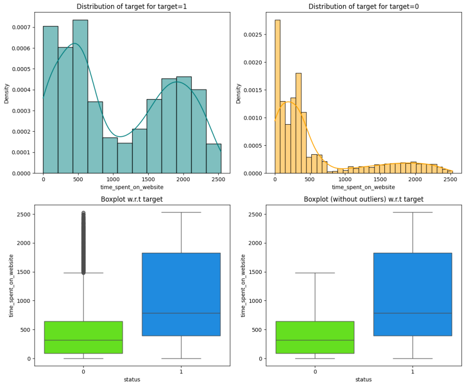


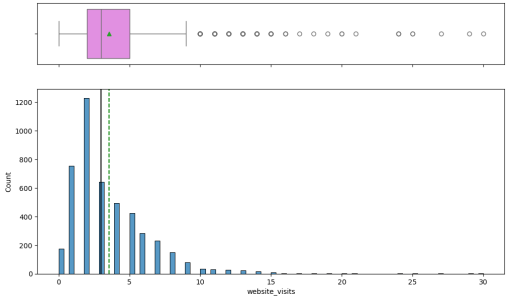


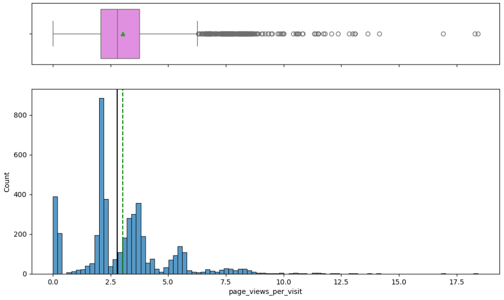
- The majority of leads are professionals (56.7%) and middle-aged (40-60 years), with professionals showing the highest conversion rates.
- Students represent the smallest group (12%) and have the lowest conversion rates, indicating lower engagement or financial constraints.

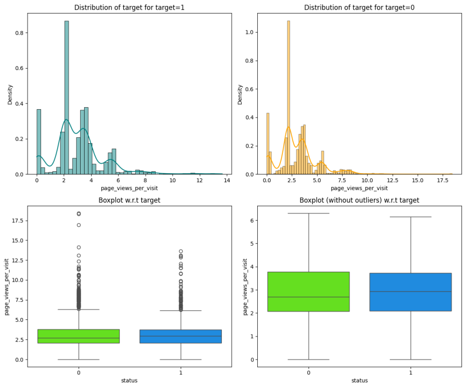
EDA Results: Observations on Interactions 1

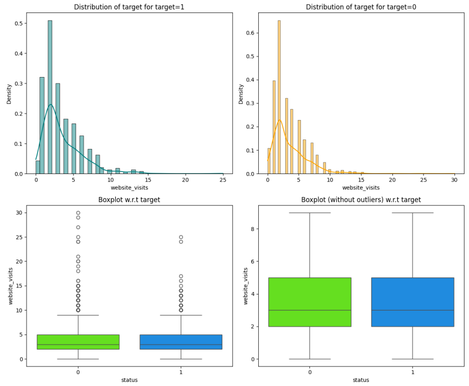






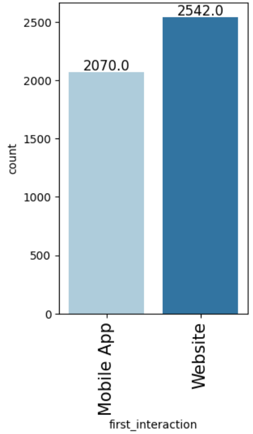


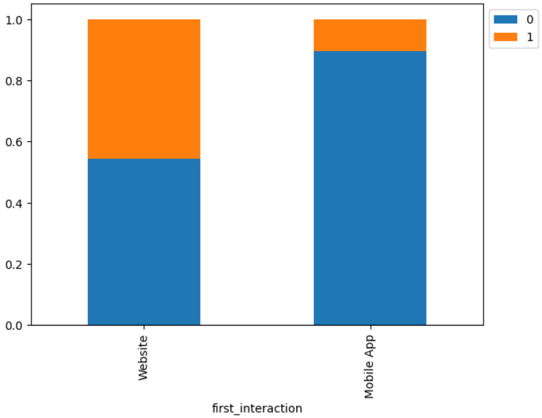


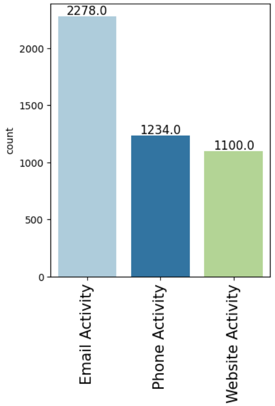


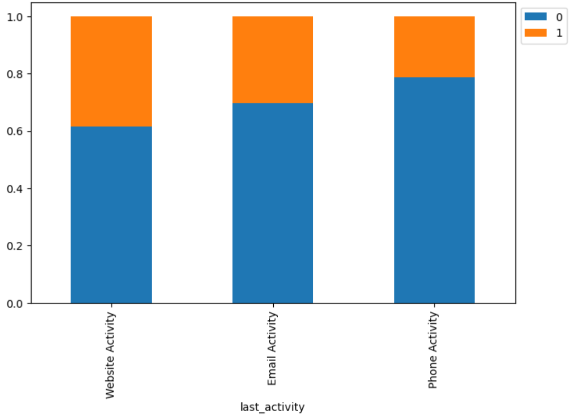
- Time spent on the website strongly correlates with lead conversion, as leads spending more time are more likely to convert.
- In contrast, page_views_per_visit and website_visits show little to no direct impact on conversion, highlighting the importance of engagement quality over quantity.

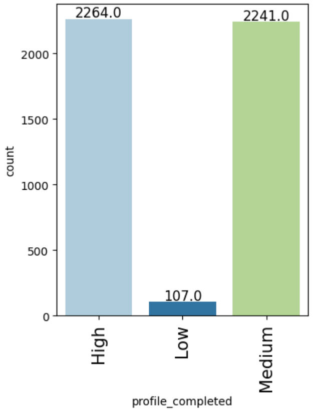
EDA Results: Observations on Interactions 2

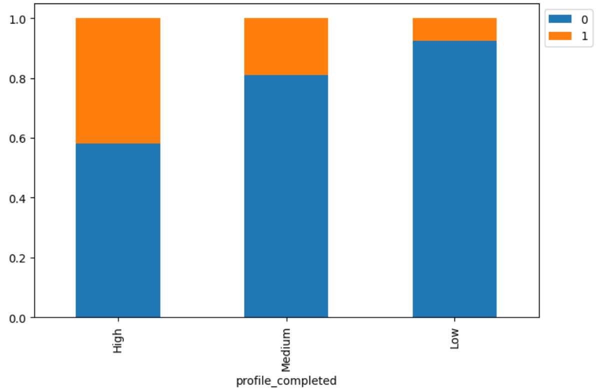






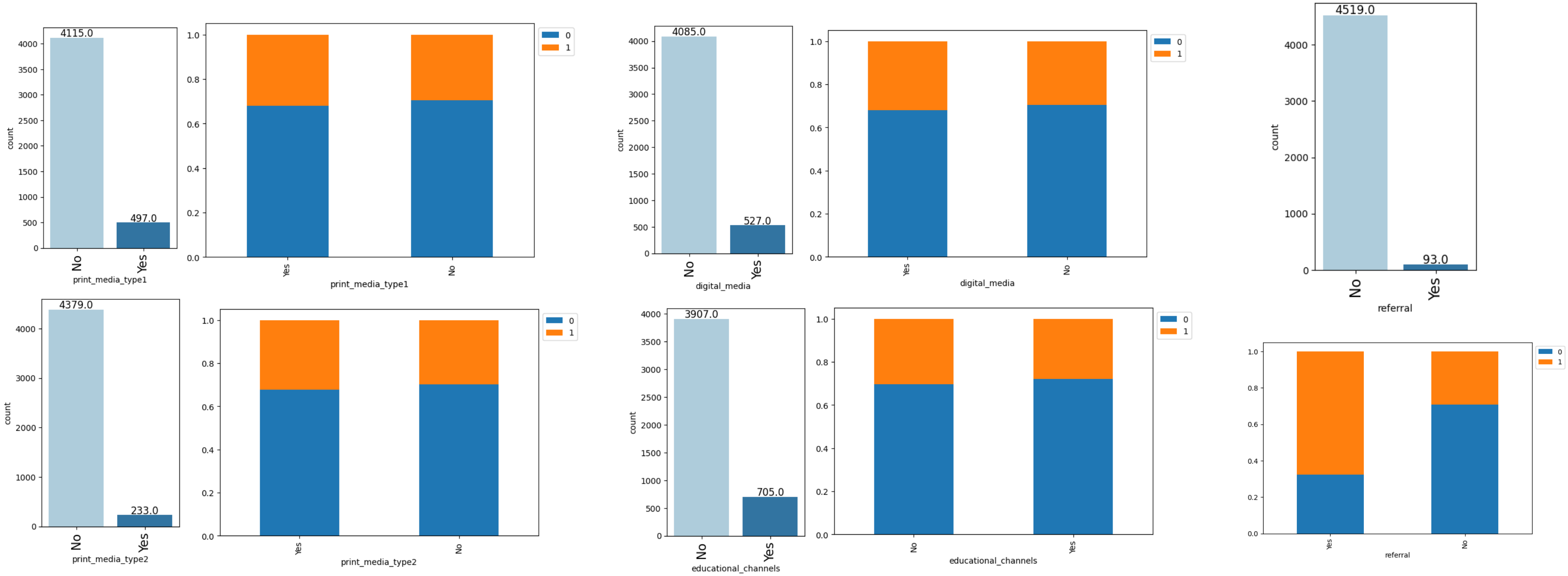






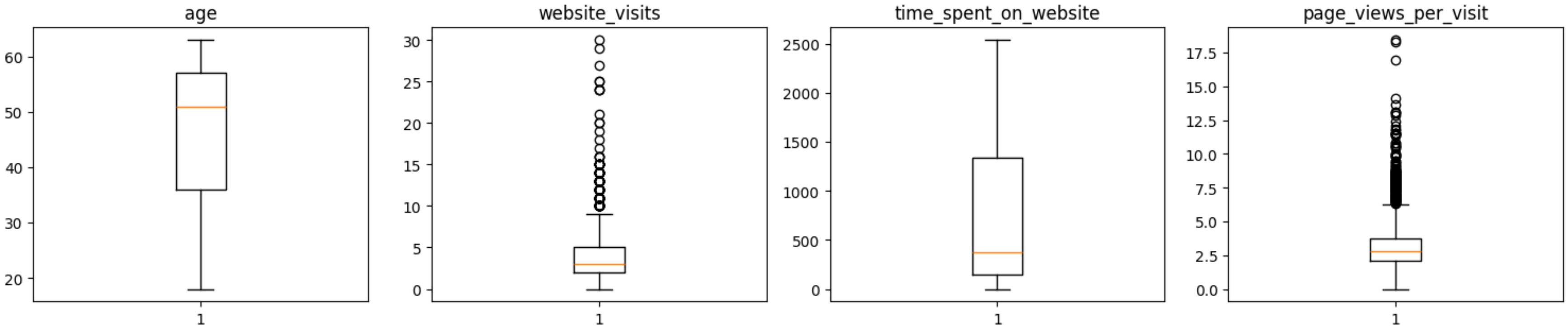
- The company’s initial interactions with leads tend to be more effective on the website than on the mobile app, resulting in a higher percentage of paid customers.
- There is a clear correlation between profile completion and conversion rates: leads who complete more of their profile are more likely to become paying customers.
- Most leads enroll directly through the website, followed by email and phone activity.

EDA Results: Observations on Advertisement Exposure



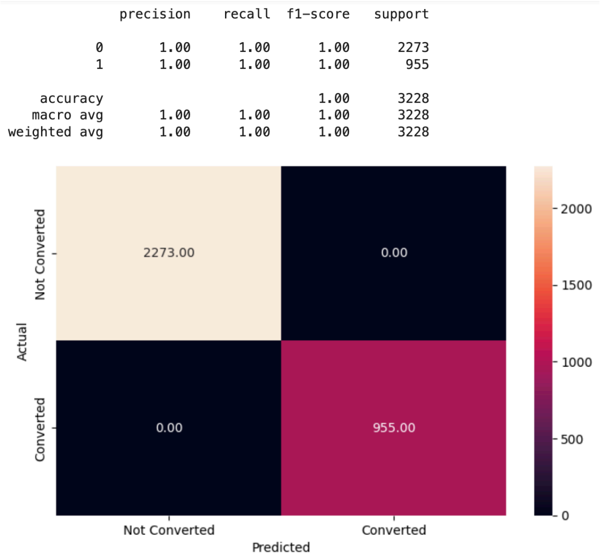
- Referrals and educational channels demonstrate the highest conversion rates, with referrals being especially effective despite generating the fewest leads.
- In contrast, print and digital media show low effectiveness, contributing minimally to both lead generation and conversions, suggesting these channels may require strategic reassessment.

EDA Results: Observations on Outliers



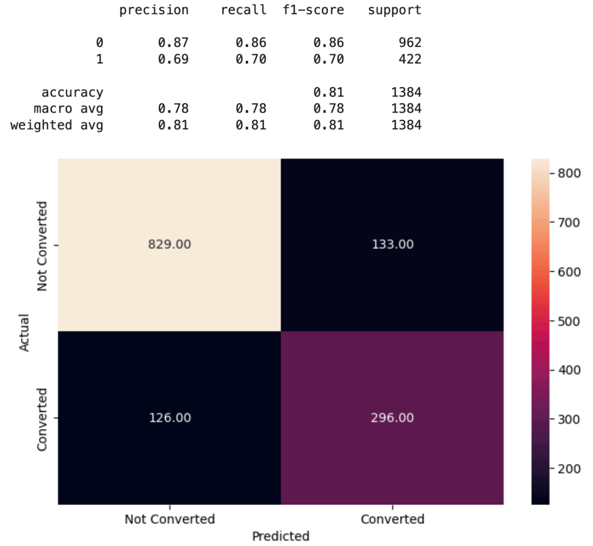
- Leads who convert are predominantly middle-aged (40-60 years) and spend significantly more time on the website, highlighting engagement as a key factor in conversion.
- Website visits and page views per visit show a wide distribution with many outliers. These outliers suggest that while high engagement (frequent visits and page views) may exist, it does not necessarily guarantee conversion and could point to unproductive or repetitive behavior by certain users.

Model Building: Decision Tree Model



```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      2273
           1       1.00      1.00      1.00       955

    accuracy                           1.00      3228
   macro avg       1.00      1.00      1.00      3228
weighted avg       1.00      1.00      1.00      3228
```

Training data

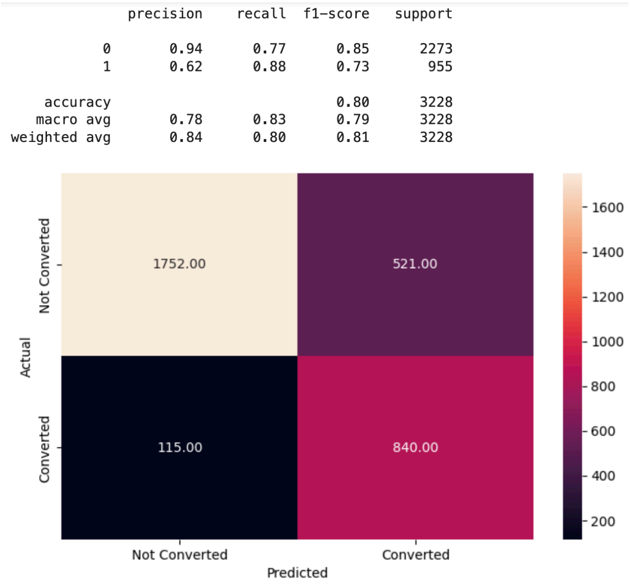


```
              precision    recall  f1-score   support

           0       0.87      0.86      0.86       962
           1       0.69      0.70      0.70       422

    accuracy                           0.81      1384
   macro avg       0.78      0.78      0.78      1384
weighted avg       0.81      0.81      0.81      1384
```

Test data

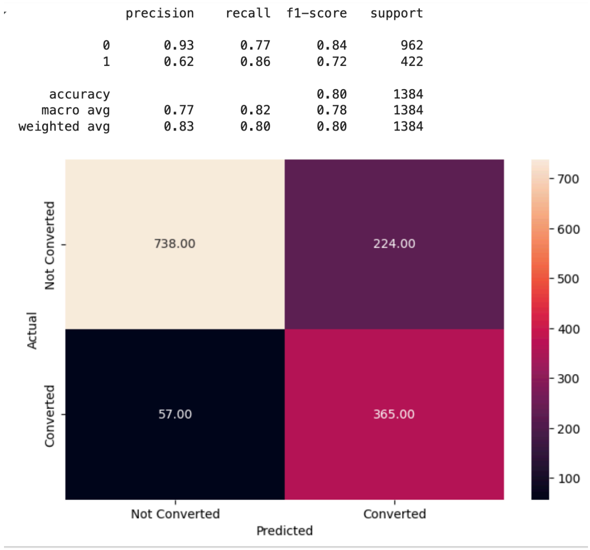
- The initial decision tree model achieved perfect accuracy (100%) on training data, but its performance dropped significantly on test data (81% accuracy), indicating overfitting.
- Precision and recall for the Converted class are relatively low, suggesting the need for regularization and better handling of potential class imbalance to improve generalization.

Decision Tree - Hyperparameter Tuning (Using GridSearchCV)



```
              precision    recall  f1-score   support

           0       0.94      0.77      0.85      2273
           1       0.62      0.88      0.73       955

    accuracy                           0.80      3228
   macro avg       0.78      0.83      0.79      3228
weighted avg       0.84      0.80      0.81      3228
```

Training data

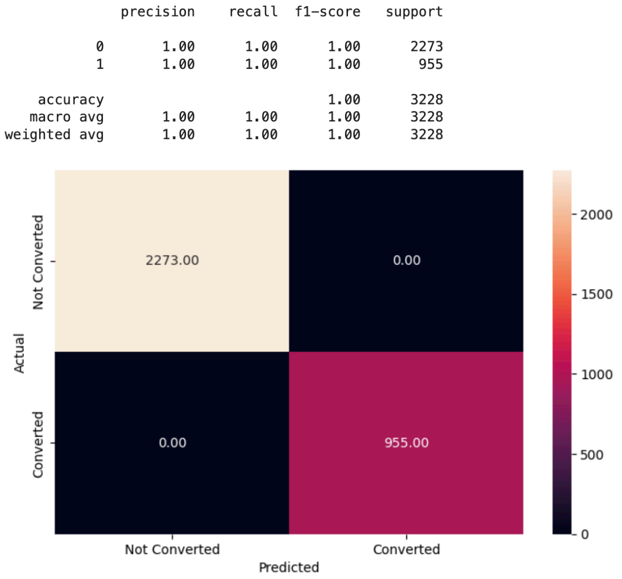


```
              precision    recall  f1-score   support

           0       0.93      0.77      0.84       962
           1       0.62      0.86      0.72       422

    accuracy                           0.80      1384
   macro avg       0.77      0.82      0.78      1384
weighted avg       0.83      0.80      0.80      1384
```

Test data

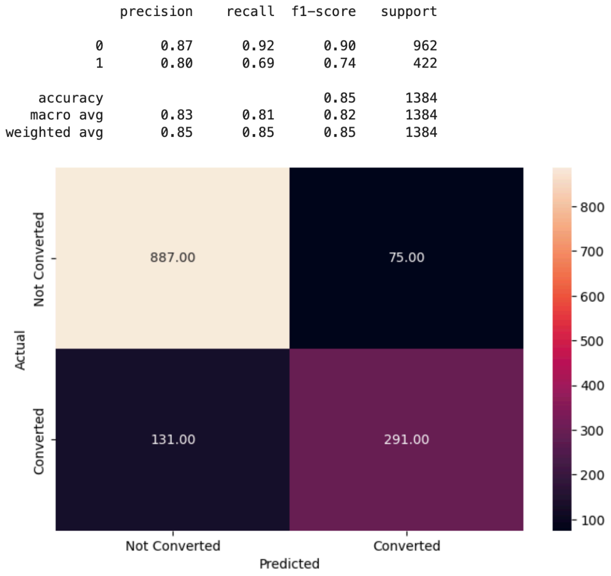
- Post-tuning, the decision tree model showed improved generalization with consistent accuracy (80%) on both training and test data.
- Recall for class 1 improved significantly (88% on training data and 86% on test data), effectively reducing false negatives.
- However, the precision for the Converted class is 62%, indicating a higher number of false positives that could be optimized further.

Model Building: Random Forest Model



```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      2273
           1       1.00      1.00      1.00       955

    accuracy                           1.00      3228
   macro avg       1.00      1.00      1.00      3228
weighted avg       1.00      1.00      1.00      3228
```

Training data

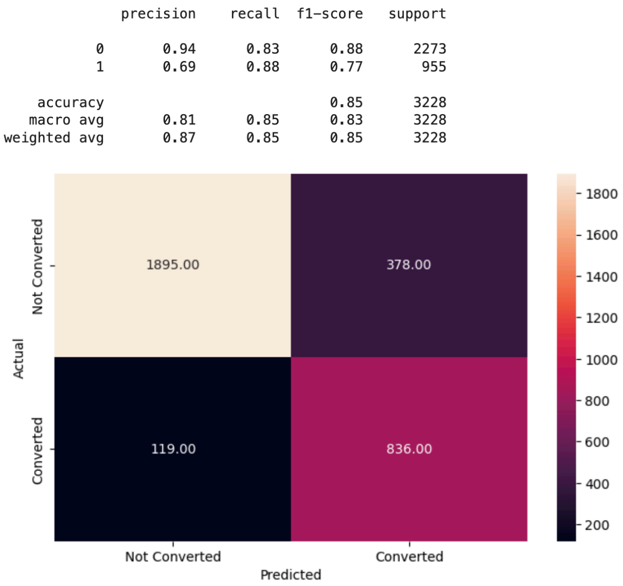


```
              precision    recall  f1-score   support

           0       0.87      0.92      0.90       962
           1       0.80      0.69      0.74       422

    accuracy                           0.85      1384
   macro avg       0.83      0.81      0.82      1384
weighted avg       0.85      0.85      0.85      1384
```

Test data

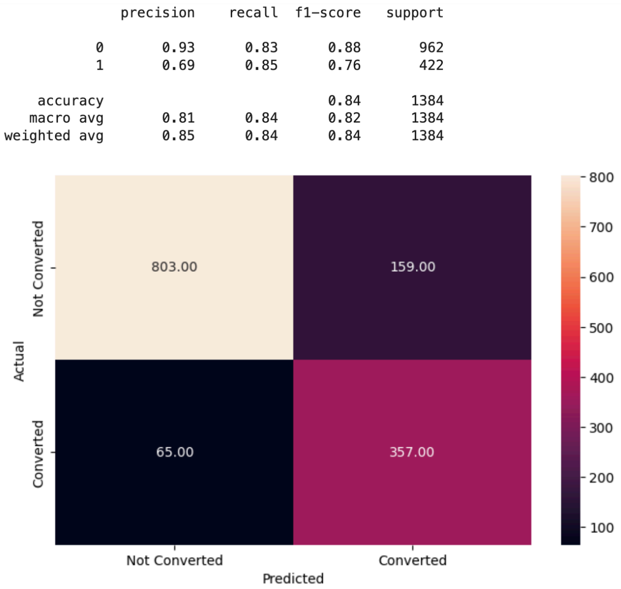
- The Random Forest model achieves an accuracy of 85% on the test data, with strong precision (80%) for the Converted class but relatively lower recall (69%), indicating it misses some actual converted leads.
- The performance drop from the training data (100% accuracy) to the test data (85% accuracy) confirms overfitting, though the model still generalizes reasonably well.

Random Forest - Hyperparameter Tuning (Using GridSearchCV)



```
              precision    recall  f1-score   support

           0       0.94      0.83      0.88      2273
           1       0.69      0.88      0.77       955

    accuracy                           0.85      3228
   macro avg       0.81      0.85      0.83      3228
weighted avg       0.87      0.85      0.85      3228
```

Training data



```
              precision    recall  f1-score   support

           0       0.93      0.83      0.88       962
           1       0.69      0.85      0.76       422

    accuracy                           0.84      1384
   macro avg       0.81      0.84      0.82      1384
weighted avg       0.85      0.84      0.84      1384
```

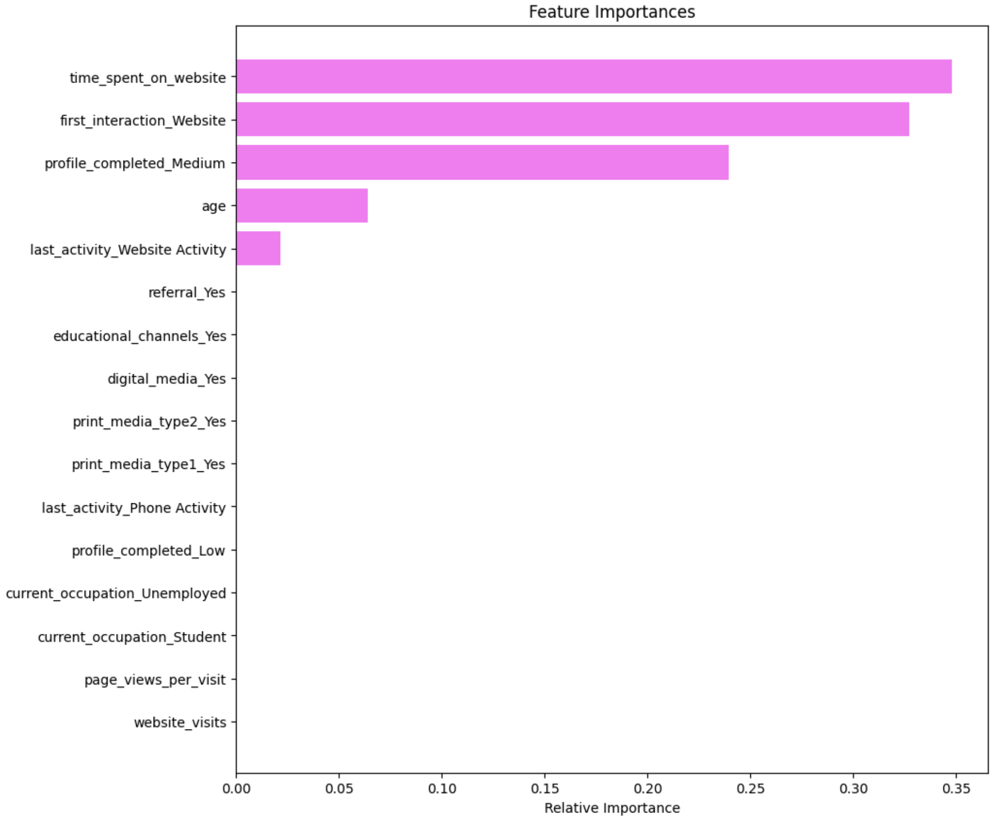
Test data

- The tuned Random Forest model achieves consistent accuracy of 85% on the training data and 84% on the test data, indicating good generalization without significant overfitting.
- It performs well in minimizing false negatives for the Converted class (recall of 88% on training and 85% on test), which aligns with the business goal of reducing lost potential customers.
- However, the model shows relatively lower precision for the Converted class (69% on both training and test data), leading to more false positives and potential resource inefficiencies.

Model Building - model performance metric

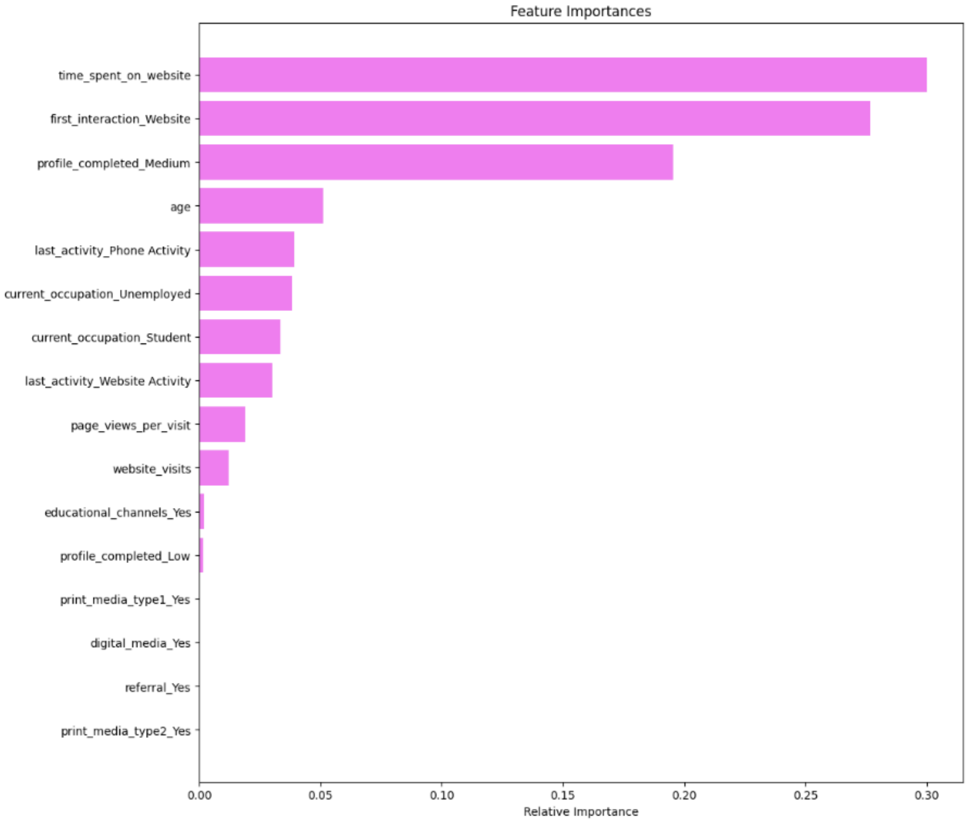
- Recall is chosen as the primary performance metric because the business goal is to minimize false negatives (missed potential customers).
- Higher recall ensures that most potential customers are identified, aligning with resource allocation priorities and reducing the opportunity cost of losing leads.

Feature Importance: Tuned Decision Tree



- Time spent on the website and first interaction via the website are the most significant features, followed by profile completion, age, and last activity.
- The remaining variables have no influence on the model's decision regarding lead conversion.

Feature Importance: Tuned Random Forest



- Like the decision tree model, time spent on the website, first interaction via the website, profile completion, and age are the top four features distinguishing converted and non-converted leads.
- However, unlike the decision tree, the random forest also considers other variables, such as occupation and page views per visit, indicating that it accounts for a broader range of factors.

Model Performance Summary

Model	Accuracy	Precision (Class 1)	Recall (Class 1)	F1-Score (Class 1)	Observations
Decision Tree (Initial)	81%	69%	70%	70%	Overfits training data; recall needs improvement.
Decision Tree (Tuned)	80%	62%	86%	72%	Minimizes false negatives but increases false positives
Random Forest (Initial)	85%	80%	69%	74%	Slight overfits training data; better test performance; recall needs improvement
Random Forest (Tuned)	84%	69%	85%	76%	Strong recall and reduced overfitting with balanced generalization

Model Performance Summary

- Summary:
  - The initial decision tree model overfits the training data and lacks generalization, making it less reliable for unseen data. Tuning the decision tree improved recall significantly for class 1 (converted leads), but at the cost of precision due to higher false positives.
  - The initial random forest model provided stronger generalization than the decision tree but had lower recall for class 1. Tuning the random forest balanced performance, maintaining a high recall (85%) while slightly reducing precision.
- Best Model Recommendation:
  - The tuned random forest model is the best choice because it balances recall and precision effectively, aligns with the business objective of minimizing false negatives (missed potential customers) and generalizes well to unseen data.

Great Learning

Happy Learning !