# PROJECT 2

## PRESENTATION

# AGENDA

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

# BUSINESS

# PROBLEM

Objective : to Identify which leads are more likely to convert into paying customers to optimize marketing efforts and increase the efficiency

Business Context : ExtraaLearn is a fast growing EdTech startup offering programs to students and professionals all. With a high volume of incoming leads, they want to prioritize the most promising leads and allocate resources effectively

# DATA OVERVIEW

**Data Description**:

- Data shape: (4612, 15)
- Target Variable: status (converted or not)
- 15 independent features:
  - Categorical: (first_interaction, current_occupation, profile_completed, etc.)
  - Numerical: (time_spent_on_website, website_visits, etc.)

**Key Variables**:

- Behavioral data (website interaction, last activity)
- Source (referral, media exposure)

# EDA - LEAD BEHAVIOR INSIGHTS

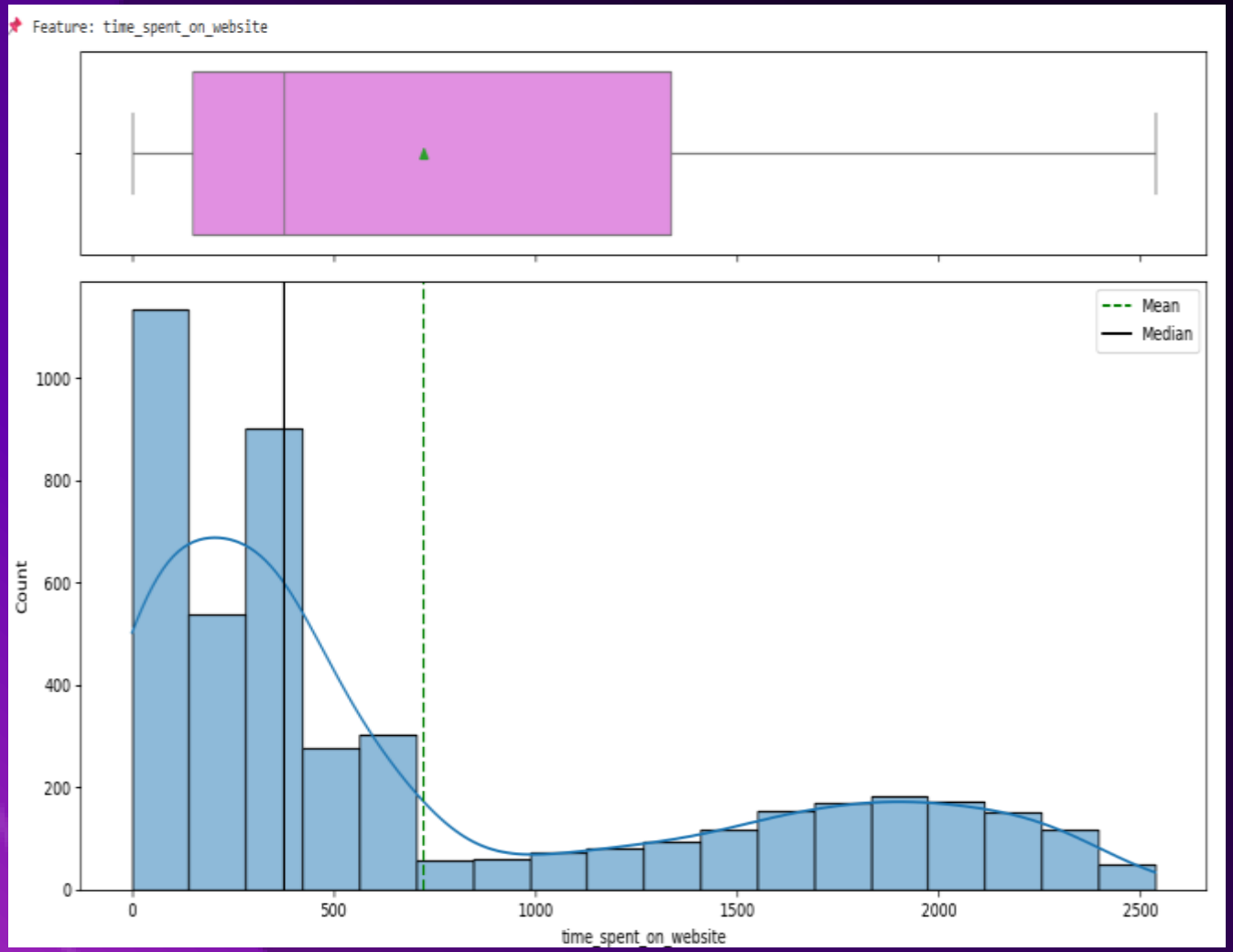
**VISUALS INCLUDED**:

HISTOGRAM-BOXPLOTS FOR TIME-BASED VARIABLES

BARPLOTS FOR CATEGORICAL FEATURES

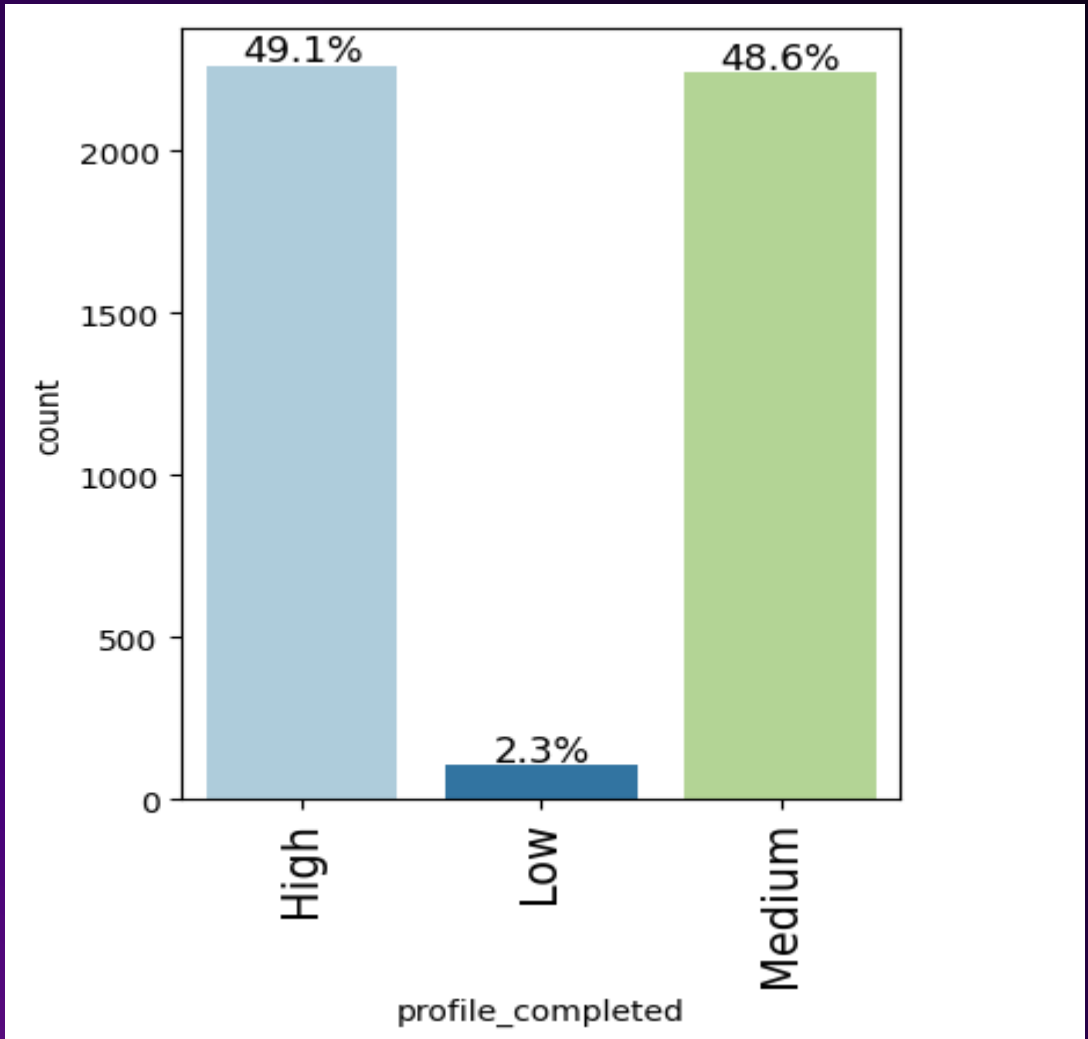
STACKED BAR CHARTS WITH TARGET VARIABLE

# EDA - LEAD BEHAVIOR INSIGHTS




Leads with **higher website visits and time spent** are more likely to convert

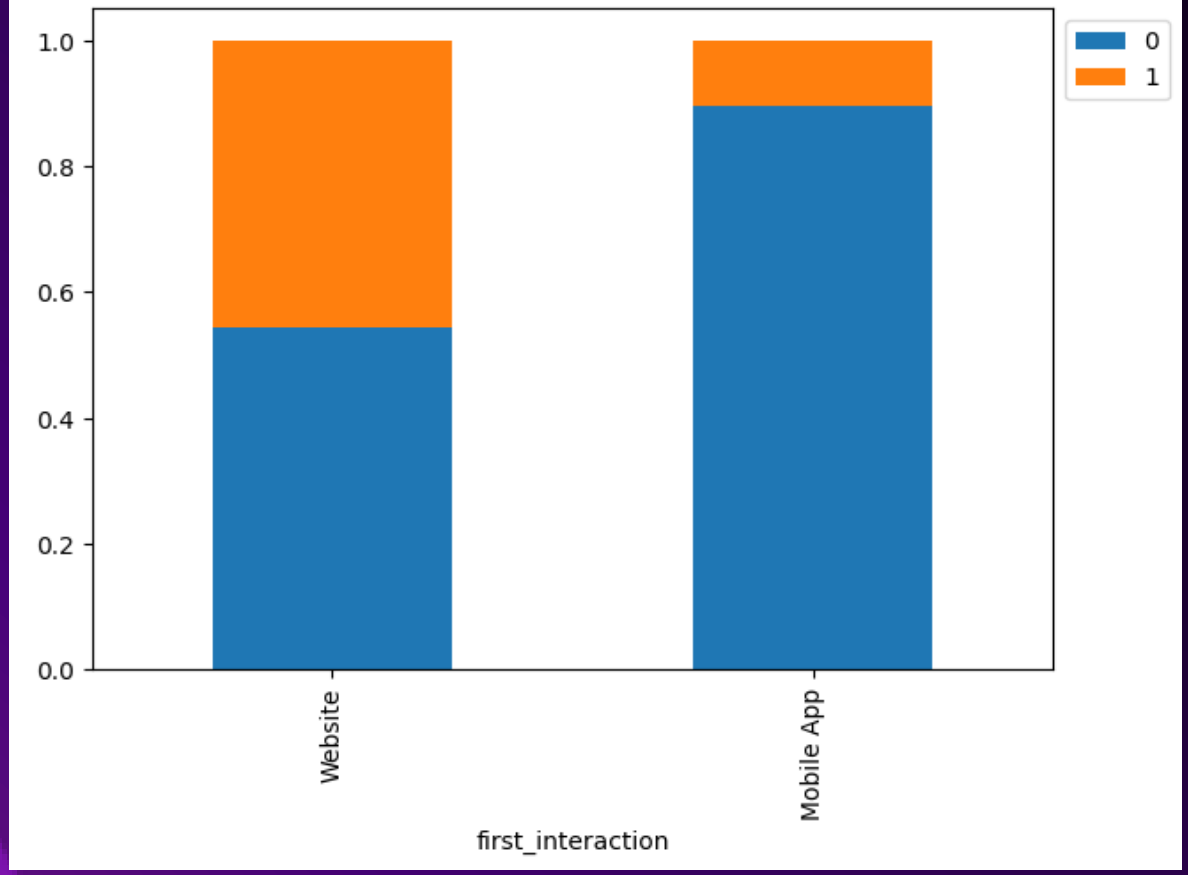
# EDA - LEAD BEHAVIOR INSIGHTS




**HIGH PROFILE COMPLETION (75%-100%)** HAS THE HIGHEST CONVERSION.

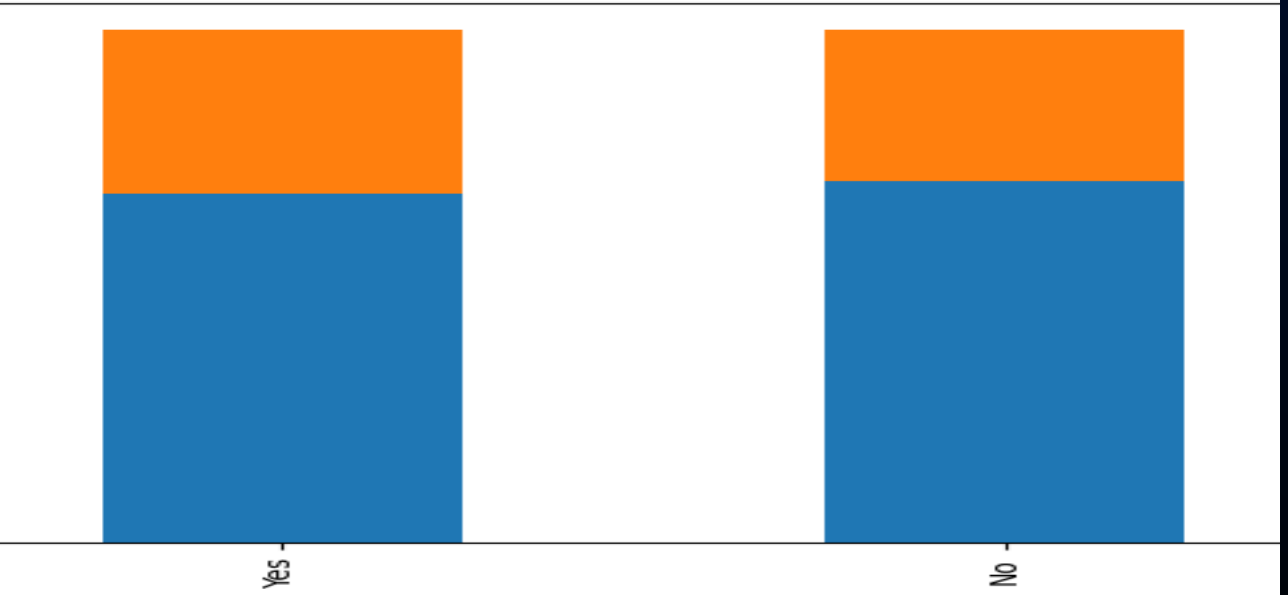
# EDA - LEAD BEHAVIOR INSIGHTS

| status | 0 | 1 | All |
|---|---|---|---|
| first_interaction | | | |
| All | 3235 | 1377 | 4612 |
| Website | 1383 | 1159 | 2542 |
| Mobile App | 1852 | 218 | 2070 |

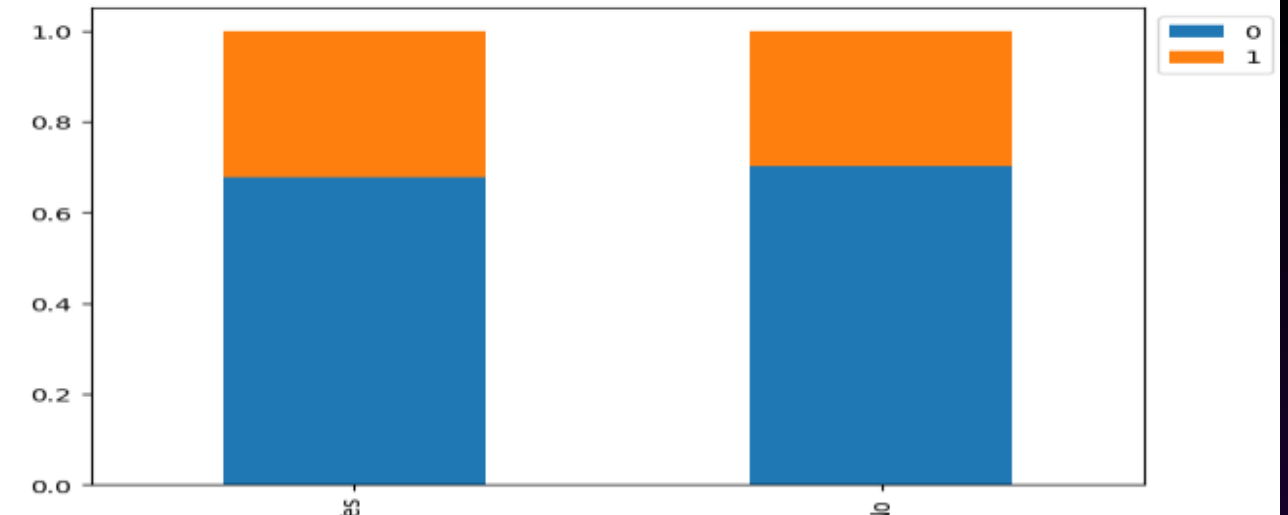



FIRST INTERACTION VIA **WEBSITE** PERFORMS BETTER THAN MOBILE APP

| ia_type1 | 0 | 1 | All |
|---|---|---|---|
| | 3235 | 1377 | 4612 |
| | 2897 | 1218 | 4115 |
| | 338 | 159 | 497 |




| status | 0 | 1 | All |
|---|---|---|---|
| print_media_type2 | | | |
| All | 3235 | 1377 | 4612 |
| No | 3077 | 1302 | 4379 |
| Yes | 158 | 75 | 233 |




# EDA - SOURCE AND CHANNEL PERFORMANCE

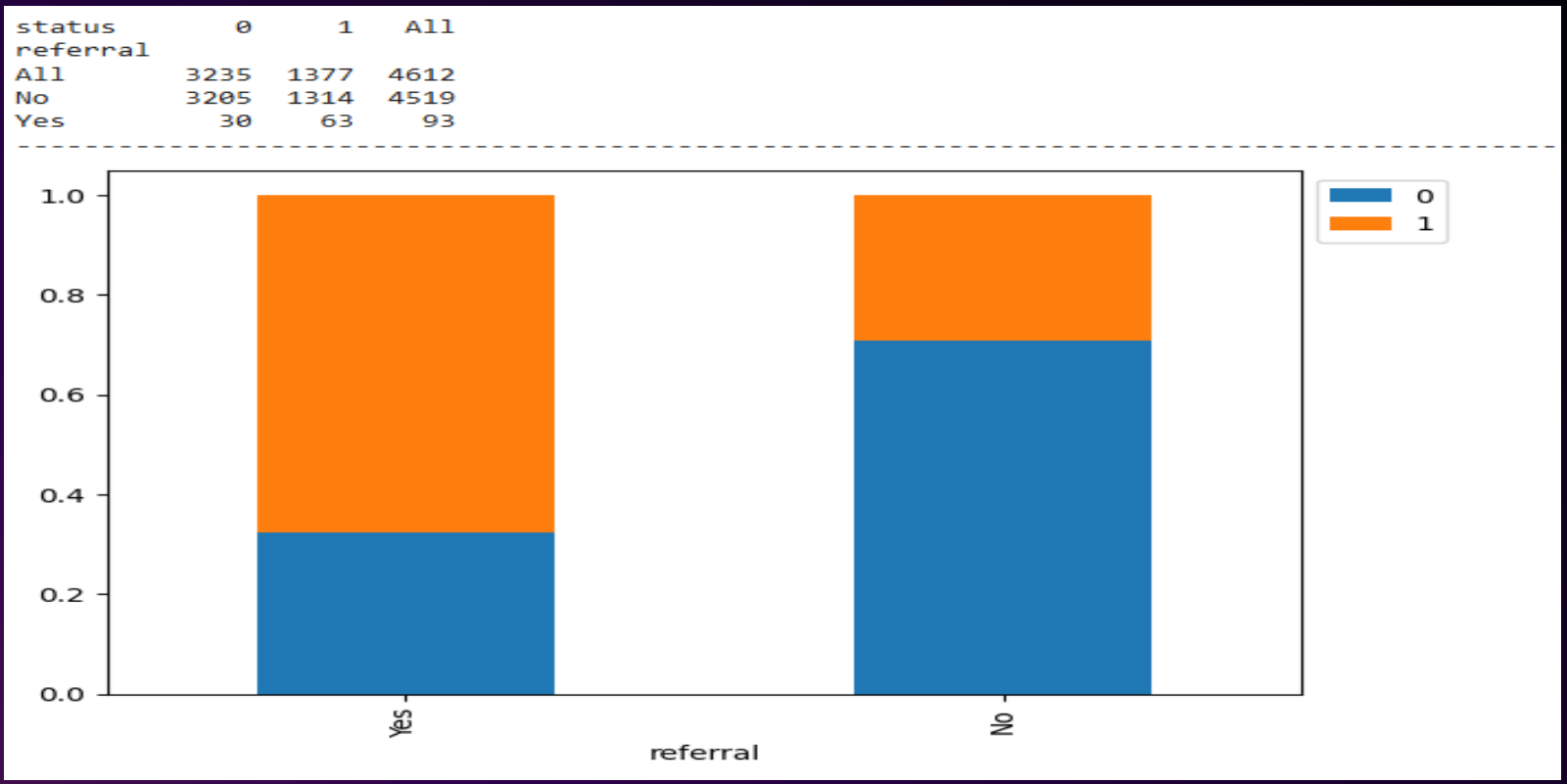
**Referral** and **educational channels** drive the highest conversions.

Print media (type1 and type2) show minimal impact.

Last activities involving **emails and phone calls** correlate strongly with positive conversion.

# EDA - SOURCE AND CHANNEL PERFORMANCE

| status | 0 | 1 | All |
|---|---|---|---|
| referral | | | |
| All | 3235 | 1377 | 4612 |
| No | 3205 | 1314 | 4519 |
| Yes | 30 | 63 | 93 |




**Referral** and **educational channels** drive the highest conversions.

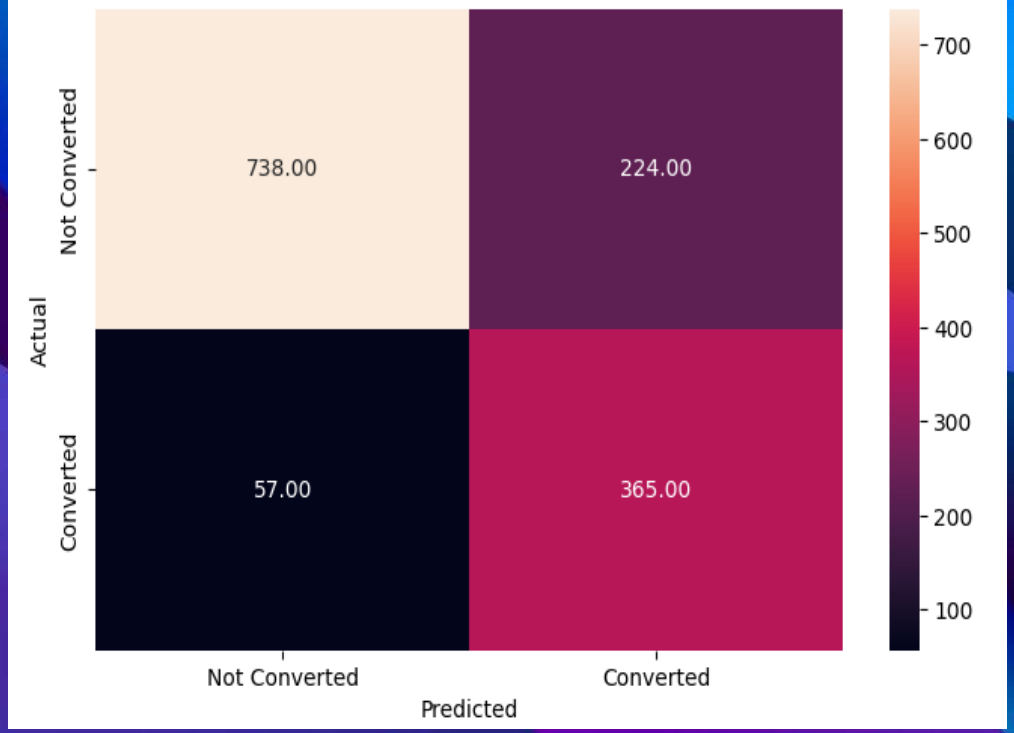
Print media (type1 and type2) show minimal impact.

Last activities involving **emails and phone calls** correlate strongly with positive conversion.

# DATA PREPROCESSING

1. Removed duplicates and handled missing values
2. Encoded categorical variables using one-hot encoding
3. Train-Test split (70:30)
4. Addressed imbalance in status with class_weigh

| | 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 |

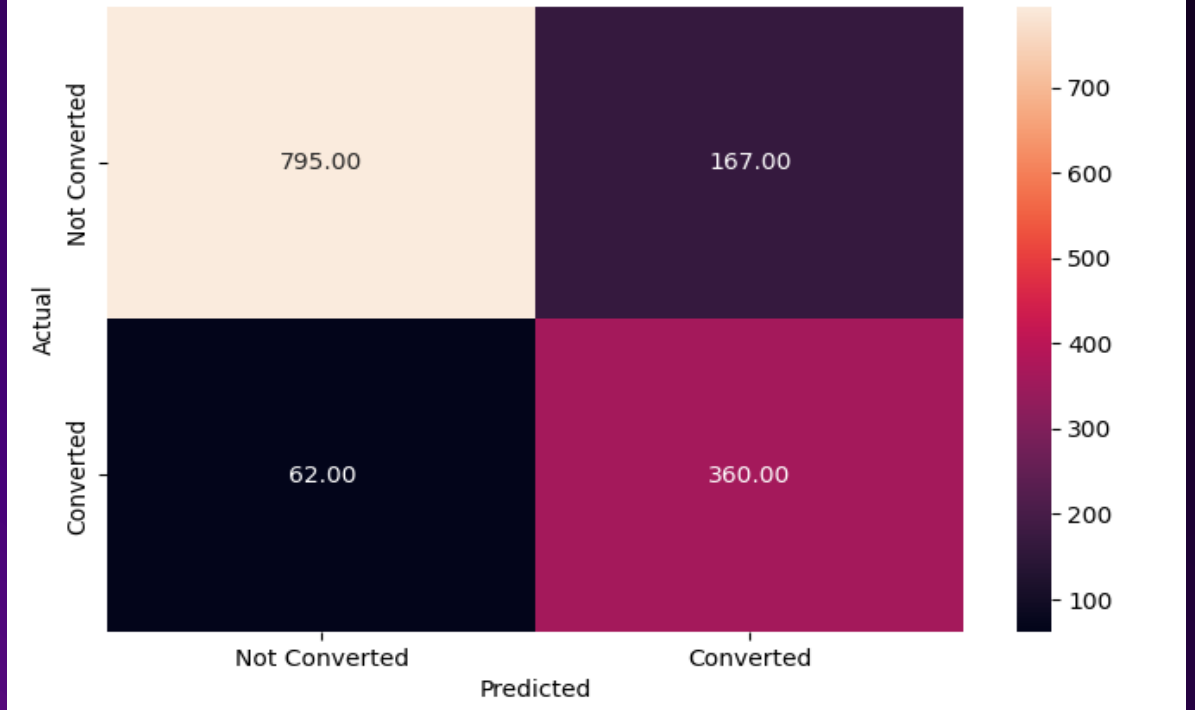



# MODEL BUILDING - DECISION TREE

- **Base model** showed overfitting (high train recall, low test recall)
- **Tuned version** with GridSearchCV improved generalization
- Parameters tuned: max_depth, min_samples_leaf, criterion

# MODEL BUILDING - RANDOM FOREST

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.93 | 0.83 | 0.87 | 962 |
| 1 | 0.68 | 0.85 | 0.76 | 422 |
| accuracy | | | 0.83 | 1384 |
| macro avg | 0.81 | 0.84 | 0.82 | 1384 |
| weighted avg | 0.85 | 0.83 | 0.84 | 1384 |



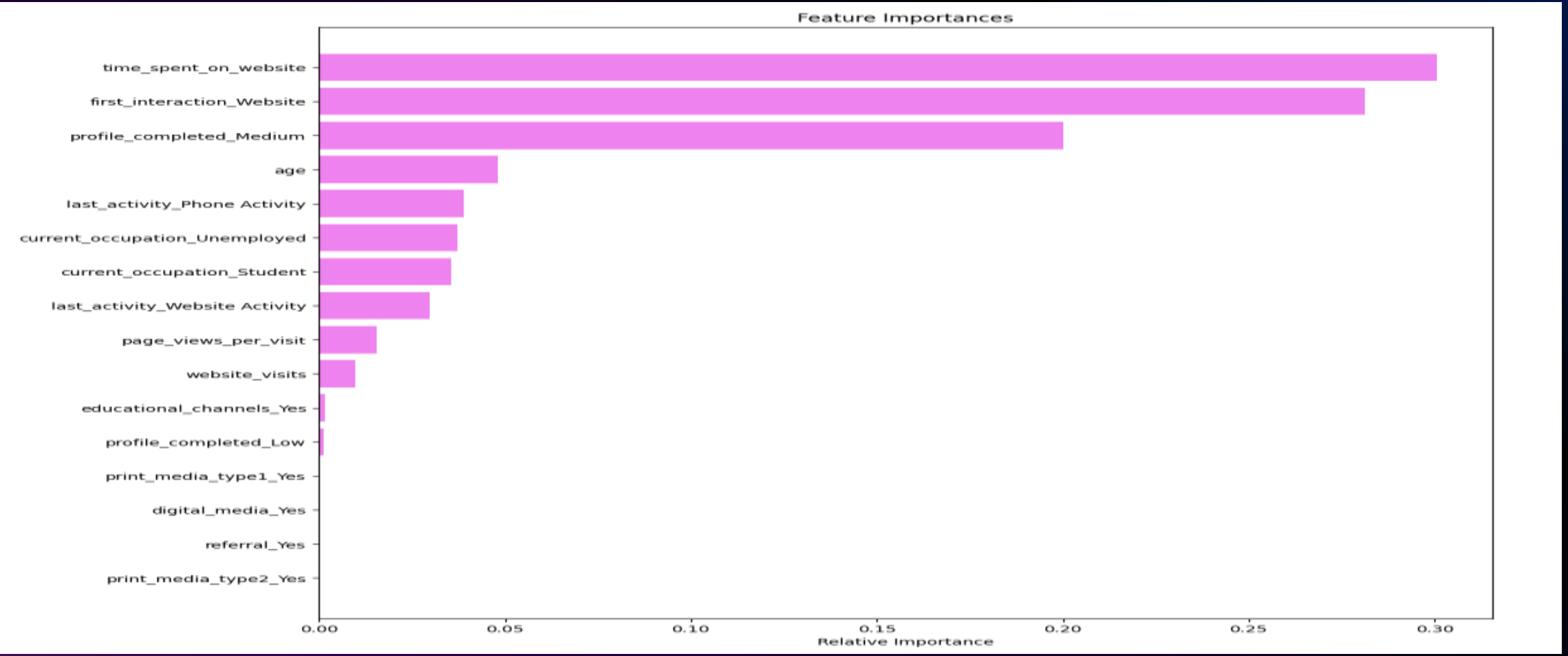

- Stronger baseline performance than decision tree
- Hyperparameter tuning included:
  - n_estimators, max_depth, max_samples, max_features
- Final model: RandomForestClassifier with class_weight = {0: 0.3, 1: 0.7}

# MODEL EVALUATION SUMMARY

The tuned Random Forest model was chosen for its high recall on test data (0.83), the model effectively minimize false negatives, which is crucial for identifying potential customers. It balanced performance well across training and test sets, handled class imbalance, and provided valuable feature insights, making it the most reliable and business aligned model.

| Model | Train Recall | Test Recall | Notes |
|---|---|---|---|
| Decision Tree | 0.91 | 0.68 | Overfit |
| Tuned DT | 0.79 | 0.72 | Improved |
| Random Forest | 0.94 | 0.80 | Strong performance |
| Tuned RF (Final) | 0.90 | **0.83** | Best performer |

# FEATURE IMPORTANCE



# FEATURE IMPORTANCE

- **Top Predictive Features:**
  - time_spent_on_website
  - first_interaction_website
  - profile_completed
  - Age
  - last_activity

**Low impact**:

- print media
- digital media flags

# KEY OBSERVATIONS

- Leads with more engagement convert better
- Print media is not an effective channel
- Referrals and education-related sources are highly effective
- First interaction and profile completion should be prioritized in lead nurturing strategy

# RECOMMENDATIONS

1. **Conclusion :**

- Developed a predictive model with **83% recall**
- Key drivers of conversion identified
- Model can be used for **real-time lead scoring** and targeted nurturing
- Future work: A/B testing on marketing strategies based on model output

**Optimize onboarding**:

Encourage complete profile creation

Offer tailored content to website-first users

**Resource allocation**:

Focus sales efforts on referral-based and website-based leads

**Channel investment**:

Reduce dependency on print media

Increase visibility in educational forums

**Deploy model in CRM**:

Integrate lead scoring to prioritize follow-ups

# THANK YOU

Nawaf Alzahrani

Nawaf4zh5@gmail.com