Presentation Title

Project and Course Name

Date

Contents / Agenda

- 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

Problem Statement: Identifying Potential Conversion of Leads in ExtraaLearn

The EdTech startup, ExtraaLearn, is experiencing rapid growth, generating a substantial number of leads. However, the challenge lies in identifying and focusing on leads more likely to convert to paid customers. To address this, the data science team aims to build a predictive model that can accurately predict lead conversions. The objective is to optimize resource allocation by targeting leads with a higher likelihood of conversion, thus enhancing the efficiency of marketing efforts and improving overall customer acquisition.

- **Key Objectives:**

1. **Lead Conversion Prediction:** Develop a machine learning model that predicts the likelihood of a lead converting into a paid customer.
2. **Resource Optimization:** Identify and prioritize leads with a higher probability of conversion to maximize resource allocation, including marketing efforts, follow-ups, and personalized engagement strategies.
3. **Insight Generation:** Extract insights into the factors driving lead conversions. Identify key features or attributes that significantly influence the conversion process to aid in strategic decision-making and resource allocation.

- **Approach:**
- Conducted comprehensive exploratory data analysis (EDA) to understand lead characteristics and their relationship with conversions.
- Prepared and cleaned the data for modeling by handling missing values, encoding categorical variables, and scaling numerical features.
- Built several classification models, including Decision Tree and Random Forest, to predict lead conversions.
- Evaluated model performance using appropriate metrics and techniques, aiming for high accuracy and robust predictions.
- Extracted feature importance to understand the key factors contributing to lead conversions.

- **Expected Outcome:**
- A robust predictive model that accurately identifies leads with a higher likelihood of converting into paying customers.
- Insights into the factors driving lead conversions, enabling targeted marketing strategies and resource allocation.

Data Overview

The dataset comprises of 4612 observations and 15 features.

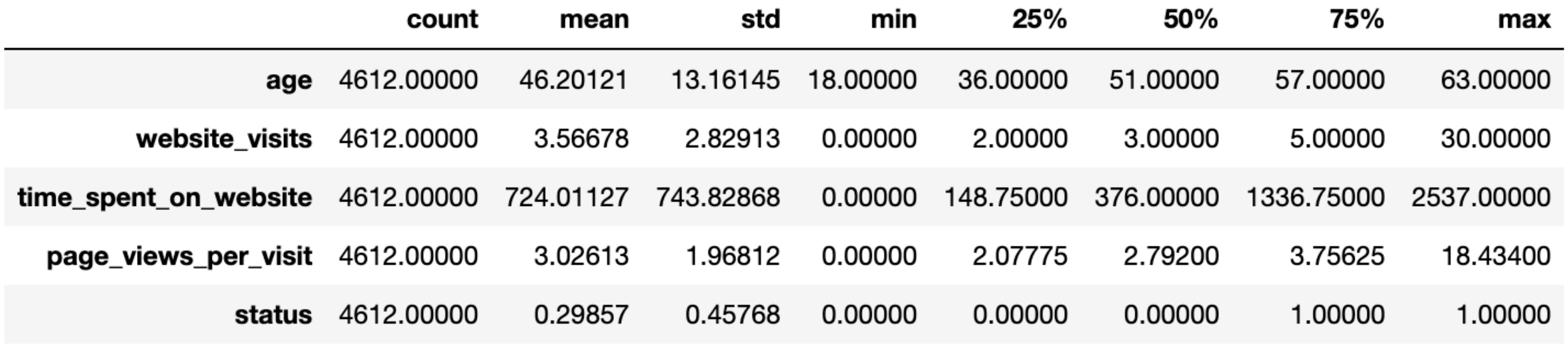
The data set contains 10 categorical features and 5 numerical:

- **Categorical Features:** ID, current occupation, first interaction, profile completed, last activity, print media type 1, print media type 2, digital media, educational channels and referral.
- **Numerical Features:** Age, website visits, time spent on the website, pages views per visit, and status .

The target variable in this context is ''Status, representing whether a lead converted into a paying customer or not.

There are no missing values.

Table 1.- Distribution Summary



	count	mean	std	min	25%	50%	75%	max
age	4612.00000	46.20121	13.16145	18.00000	36.00000	51.00000	57.00000	63.00000
website_visits	4612.00000	3.56678	2.82913	0.00000	2.00000	3.00000	5.00000	30.00000
time_spent_on_website	4612.00000	724.01127	743.82868	0.00000	148.75000	376.00000	1336.75000	2537.00000
page_views_per_visit	4612.00000	3.02613	1.96812	0.00000	2.07775	2.79200	3.75625	18.43400
status	4612.00000	0.29857	0.45768	0.00000	0.00000	0.00000	1.00000	1.00000

EDA Results

After the bivariate analysis, we got that from the three current occupations the professionals are the group with the biggest percentage of converted, being this group the biggest too, with 56.7%.

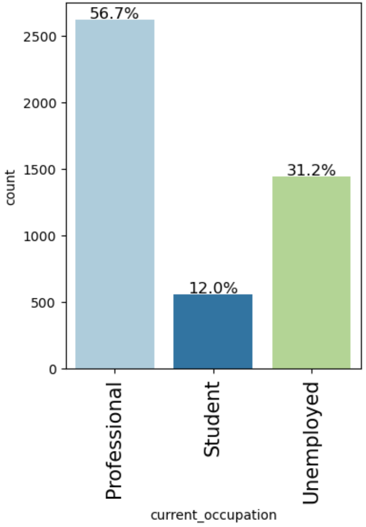


Figure 1.- Barplot on current_occupation

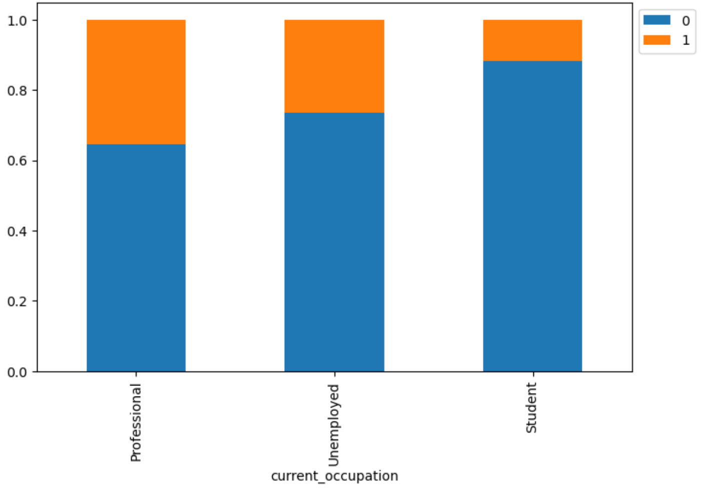


Figure 2.- Stacked barplot on current_occupation and status

The leads that contacted ExtraaLearn by their website converted more tan the ones that meet ExtraaLearn by their app. The difference between this two is very notorious.

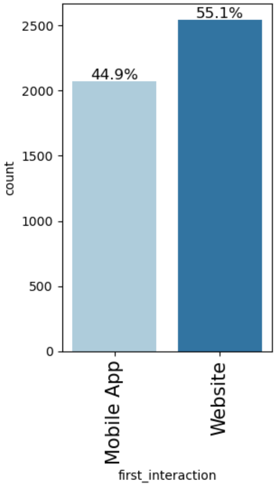


Figure 3.- Barplot on first_interaction

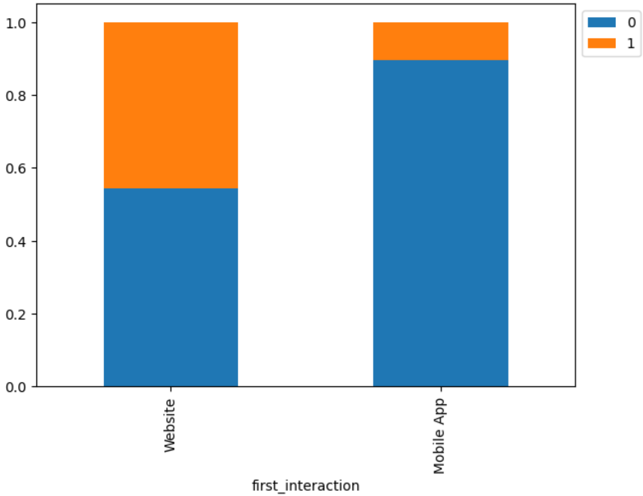


Figure 4.- Stacked barplot on first_interaction and status

The higher the profile is completed is directly related to the status of the lead. It’s worth mentioning that the high and medium have a very small differnce.

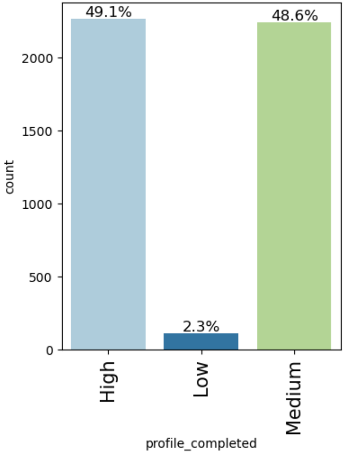


Figure 5.- Barplot on profile_completed

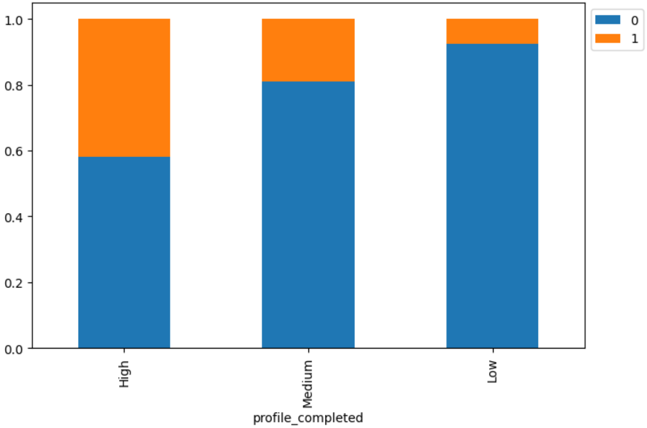


Figure 6.- Stacked barplot first_interaction and status

From the three chanels of interaction, the website interactions seems to be the one which works the best even thought is the least common.

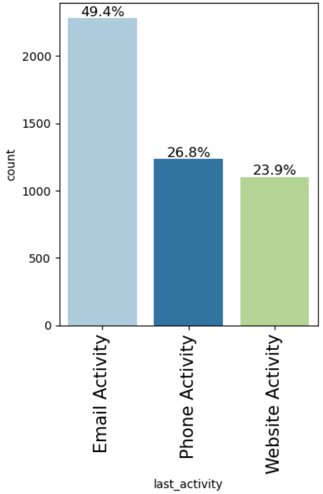


Figure 7.- Barplot on last_activity

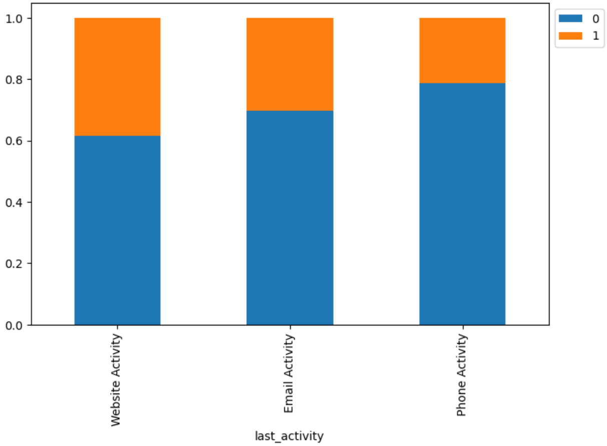


Figure 8.- Stacked barplot on last_activity and status

Out of the five flags, the referral is the one that has the most converted leads with only two percent of the observations.

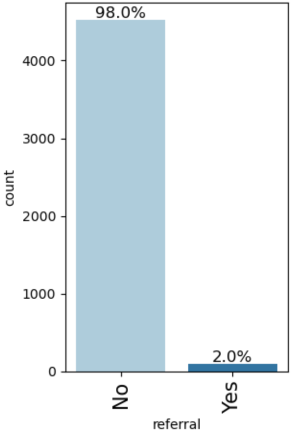


Figure 9.- Barplot on referral

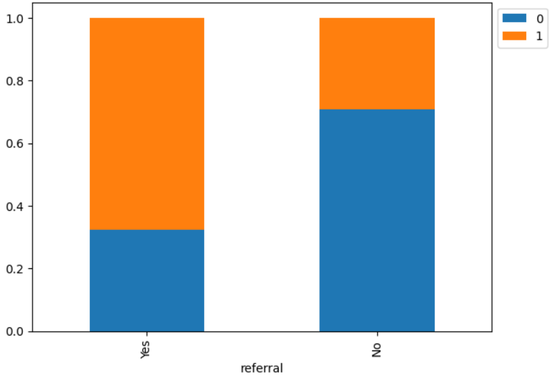


Figure 10.- Stacked barplot on referral and status

The figure 11 shows us that the website visits and the page views per visit are the numerical atributes with outliers while age and time spent on website don’t have any.

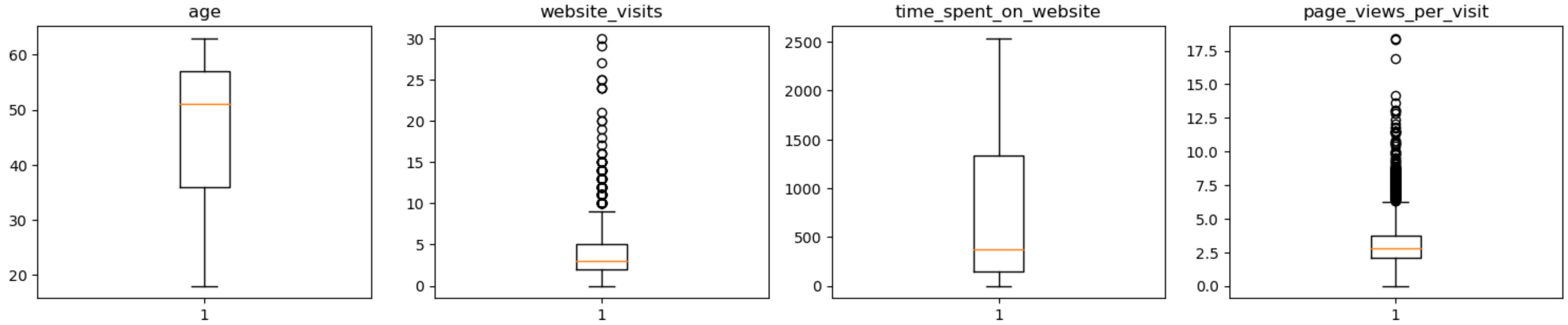
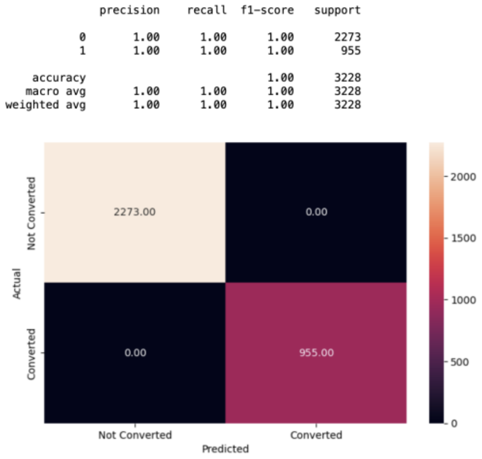


Figure 11.- Outliers detection with barplots

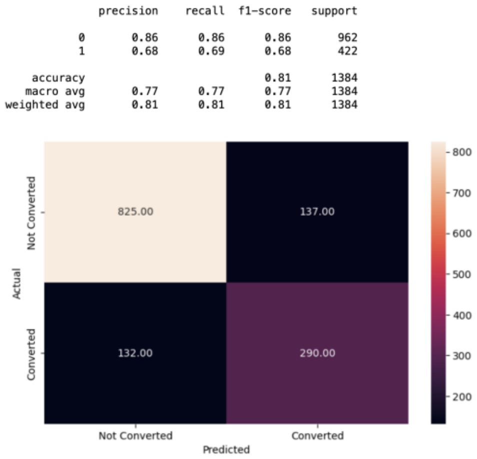
Model Building – Decision Tree

The decision tree was overfitted at the beginning.



```
              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
```

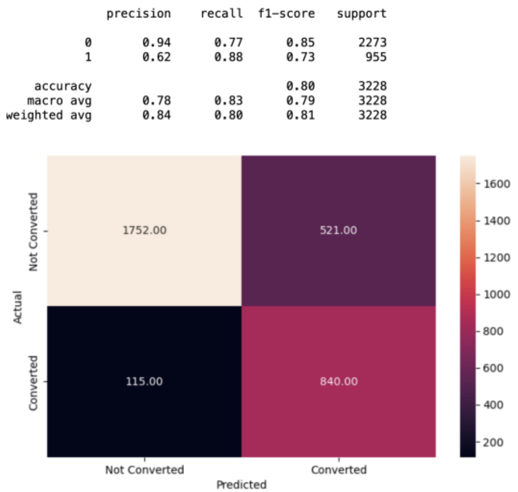
Figure 12.- Performance on the train data



```
              precision    recall  f1-score   support

           0       0.86      0.86      0.86       962
           1       0.68      0.69      0.68       422

    accuracy                           0.81      1384
   macro avg       0.77      0.77      0.77      1384
weighted avg       0.81      0.81      0.81      1384
```

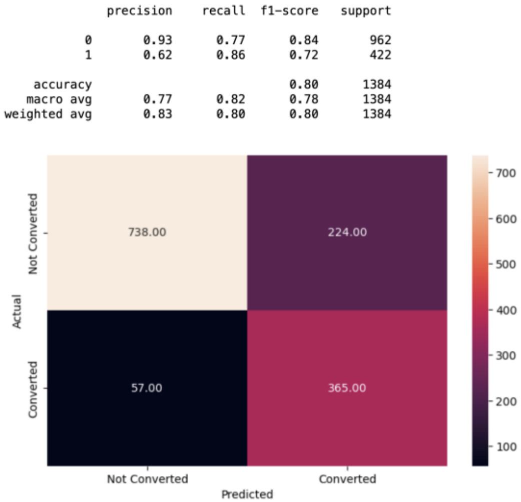
Figure 13.- Perforance on the testing data

We will use the class_weight hyperparameter with the value equal to {0: 0.3, 1: 0.7} which is approximately the opposite of the imbalance in the original data.



```
              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
```

Figure 14.- Perforance on the testing data after GridSearchCV



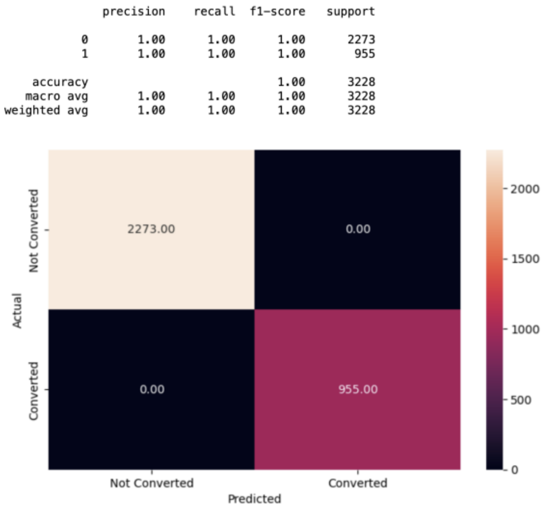
```
              precision    recall  f1-score   support

           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
```

Figure 15.- Perforance on the testing data after GridSearchCV

After tuning the hyperparameter, the model isn't overfitting anymore, for the testing the data the recall decreased, as well as the f1-score. The precision and accuracy increased.

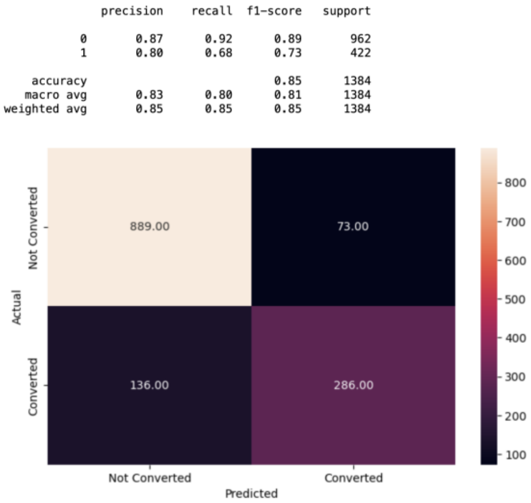
Model Building – Random Forest Classifier

The training data was overfitted at the beginning and the accuracy of the test prediction was around 85%.



```
              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
```

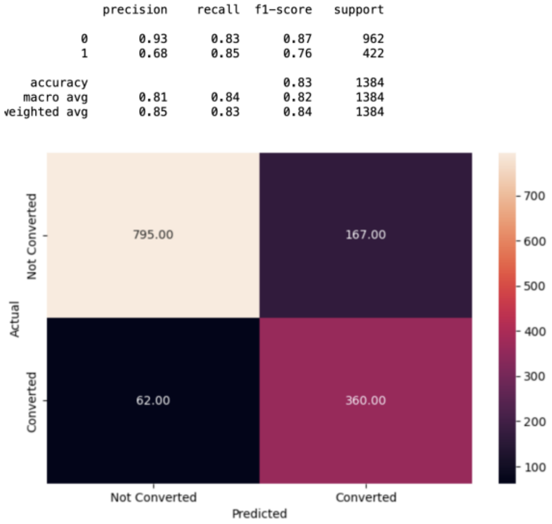
Figure 16.- Performance on the train data



```
              precision    recall  f1-score   support

           0       0.87      0.92      0.89       962
           1       0.80      0.68      0.73       422

    accuracy                           0.85      1384
   macro avg       0.83      0.80      0.81      1384
weighted avg       0.85      0.85      0.85      1384
```

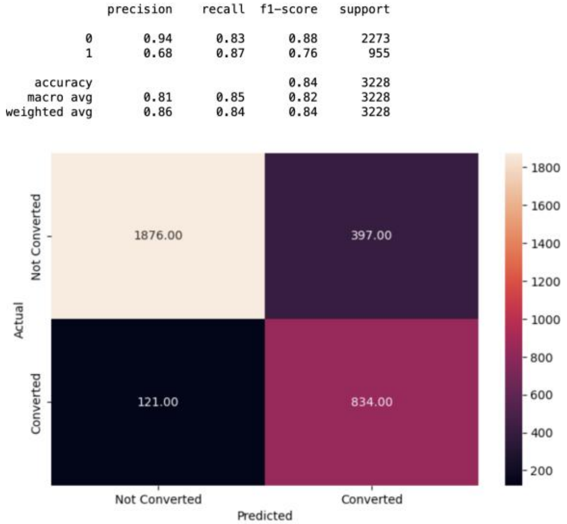
Figure 17.- Performance on the train data

After the tunning the accuracy is 84% and the models weren't overfitting anymore and the recall for the two were 85% and 87%.



```
              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
```

Figure 18.- Perforance on the testing data after GridSearchCV



```
              precision    recall  f1-score   support

           0       0.94      0.83      0.88      2273
           1       0.68      0.87      0.76       955

    accuracy                           0.84      3228
   macro avg       0.81      0.85      0.82      3228
weighted avg       0.86      0.84      0.84      3228
```

Figure 19.- Perforance on the testing data after GridSearchCV

- Based on the preforming of the models and the objective of the prediction, the performance metric that we are going to give the more value to, is accuracy. This is because the similarities between the values of all the other metrics don’t show a significant difference between the two models, so we are going to focus on showing the results that can be more accurate.

Feature Importance – Decision Tree

The most important feature is the time the users spend on the website, followed by their first interaction and how completed their profile is.

The age reflects the current situation of the users, showing what the different groups use the platform for.

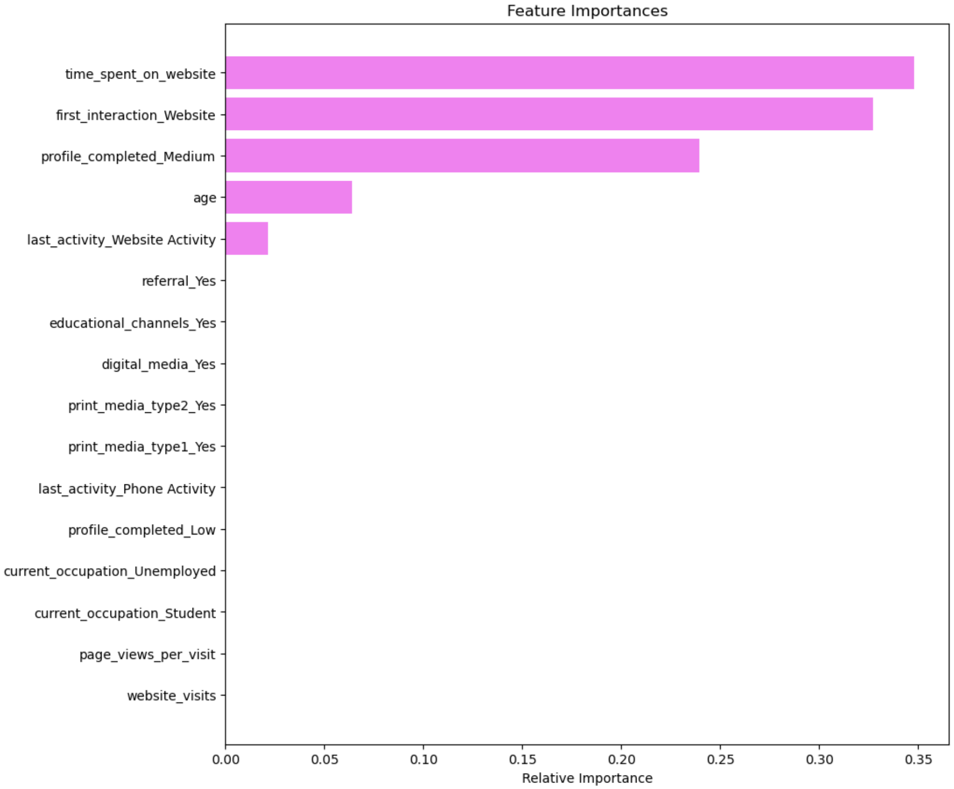


Figure 20.- Feature importance according to the decision tree

Feature Importance – Random Forest

The number of features is bigger, and it takes into consideration a more than half of them.

The time spent on the website, remains as the most important feature, having the next three features with the same importance as the decision tree.

It's interesting how the unemployed status appears right after the last activity on the phone, this is telling us what group is the most interested on the site.

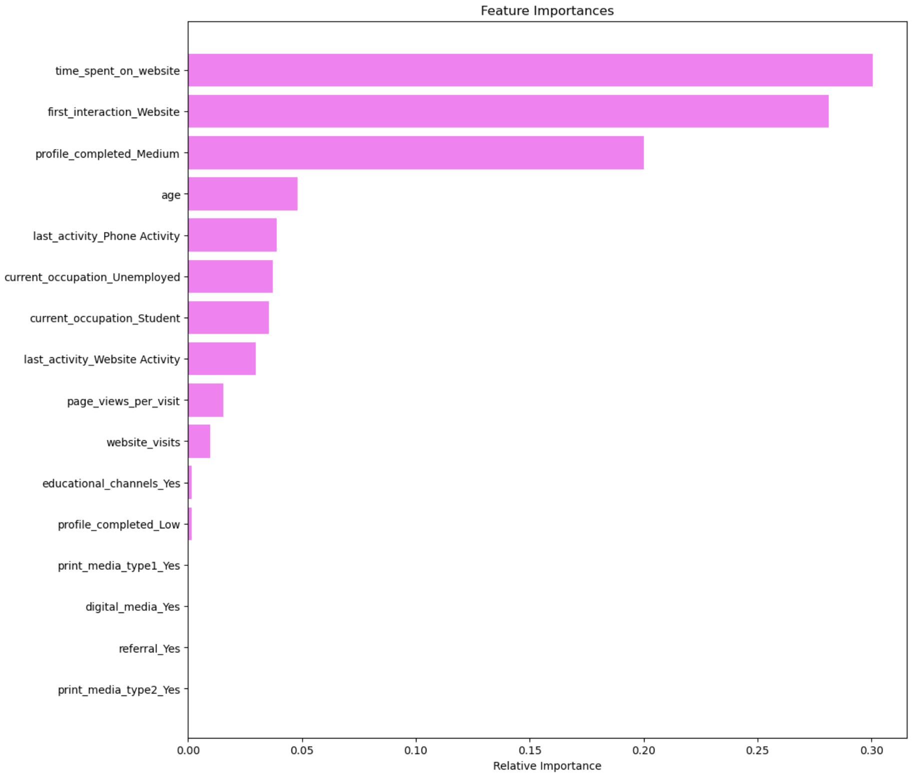


Figure 21.- Feature importance according to the random forest

Model Performance Summary

Table 2.- Models performance summary

	Accuracy	Precision	Recall	F1 Score
Decision Tree	0.80	0.620	0.860	0.720
Random Forest	0.83	0.680	0.850	0.760

Following the table 2, we can see the similarities on the two models, being the random forest one more accurate this is the one that is choosen to be the main model, it can proides us more insight of all the important features so we can make a better predicion.

Recomendations

- It is suggested that the company focused on increasing the time spent by the users on the website, maybe some incentives could help achieve this, like virtual rewards.

- The first interaction has been found to be key for the purposes of converting the users, a welcome video or an interactive guide for the website might be a friendly way to welcome new users.

- The competition of the profile is the third most important feature. Following the same ideas on the first bullet, an incentive for this could be a good way to make the users convert.

Thank you for your attention

Great Learning

Happy Learning !