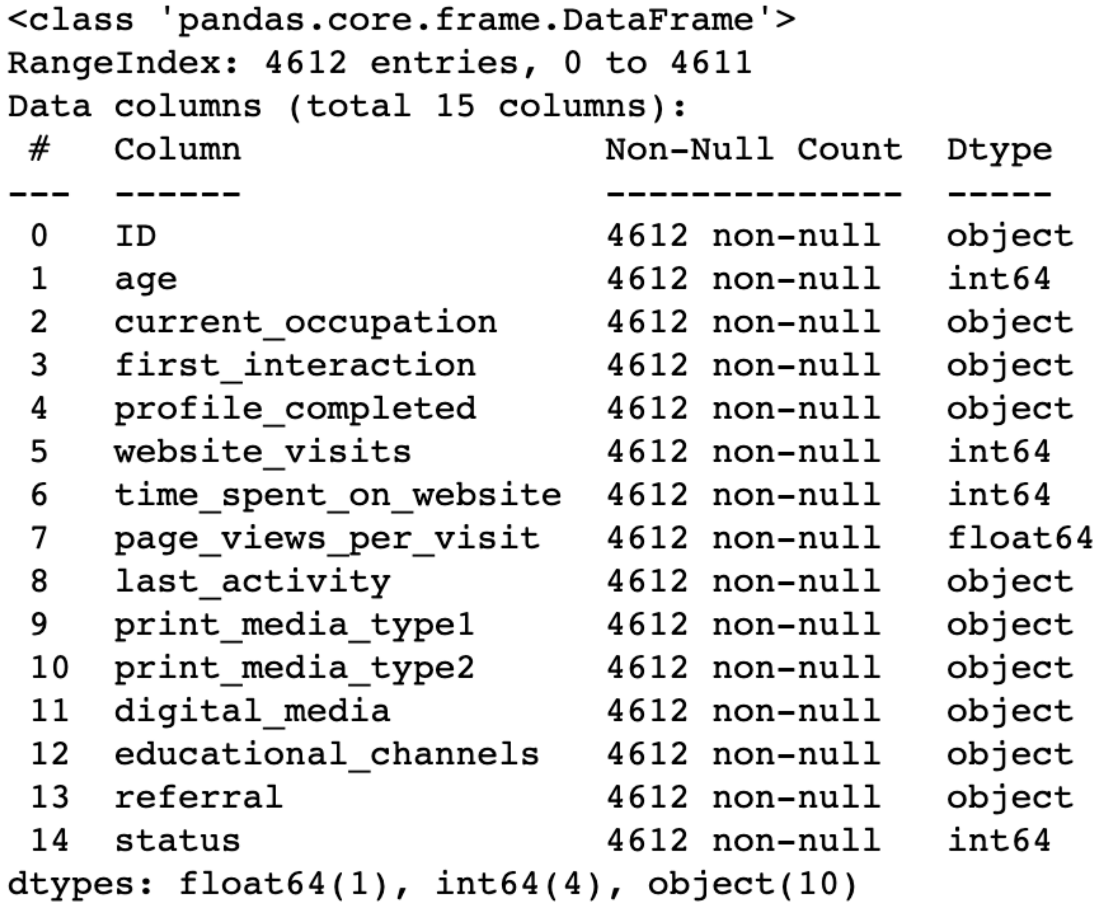
ExtraaLearn Potential Customers Prediction

Classification, Ensemble Methods

‘Femi Bolarinwa

Data Snapshot



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4612 entries, 0 to 4611
Data columns (total 15 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   ID                     4612 non-null   object 
 1   age                    4612 non-null   int64  
 2   current_occupation     4612 non-null   object 
 3   first_interaction      4612 non-null   object 
 4   profile_completed      4612 non-null   object 
 5   website_visits         4612 non-null   int64  
 6   time_spent_on_website  4612 non-null   int64  
 7   page_views_per_visit   4612 non-null   float64
 8   last_activity          4612 non-null   object 
 9   print_media_type1      4612 non-null   object 
 10  print_media_type2      4612 non-null   object 
 11  digital_media          4612 non-null   object 
 12  educational_channels   4612 non-null   object 
 13  referral               4612 non-null   object 
 14  status                 4612 non-null   int64  
dtypes: float64(1), int64(4), object(10)
```

Model Evaluation Criterion

Precision V Recall

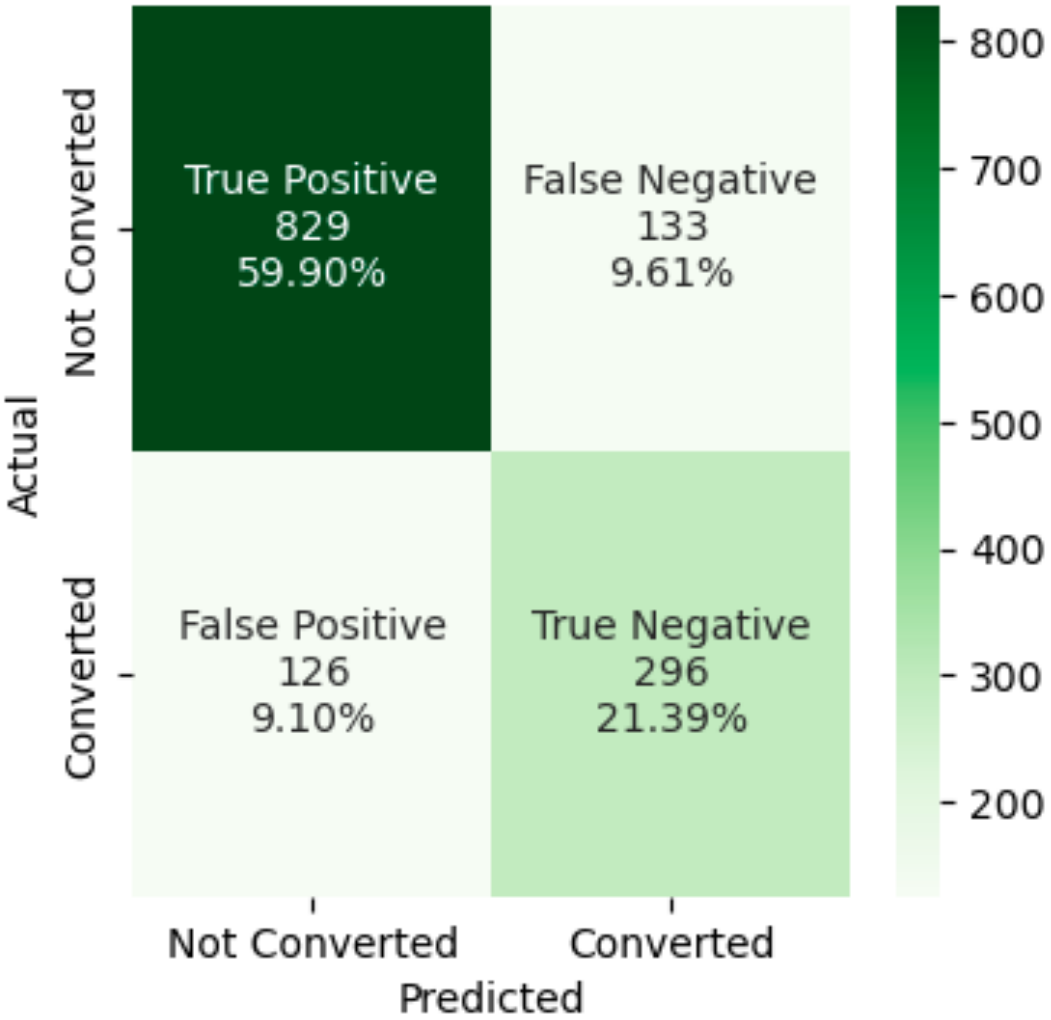
- Positive conversion prediction that turns out to be false means a waste of resources by the company.
- Negative conversion prediction that turns out could have been positive means the company loses a potential customer.
- Second scenario above is a greater loss. Therefore False Negative needs to be minimized
- Recall is the main evaluation metric.
- Original data randomly split into training and test datasets at 70:30 ratio.
- GridSearchCV used to tune parameters to optimize classifiers

Decision Tree Model

Performance on unseen test data

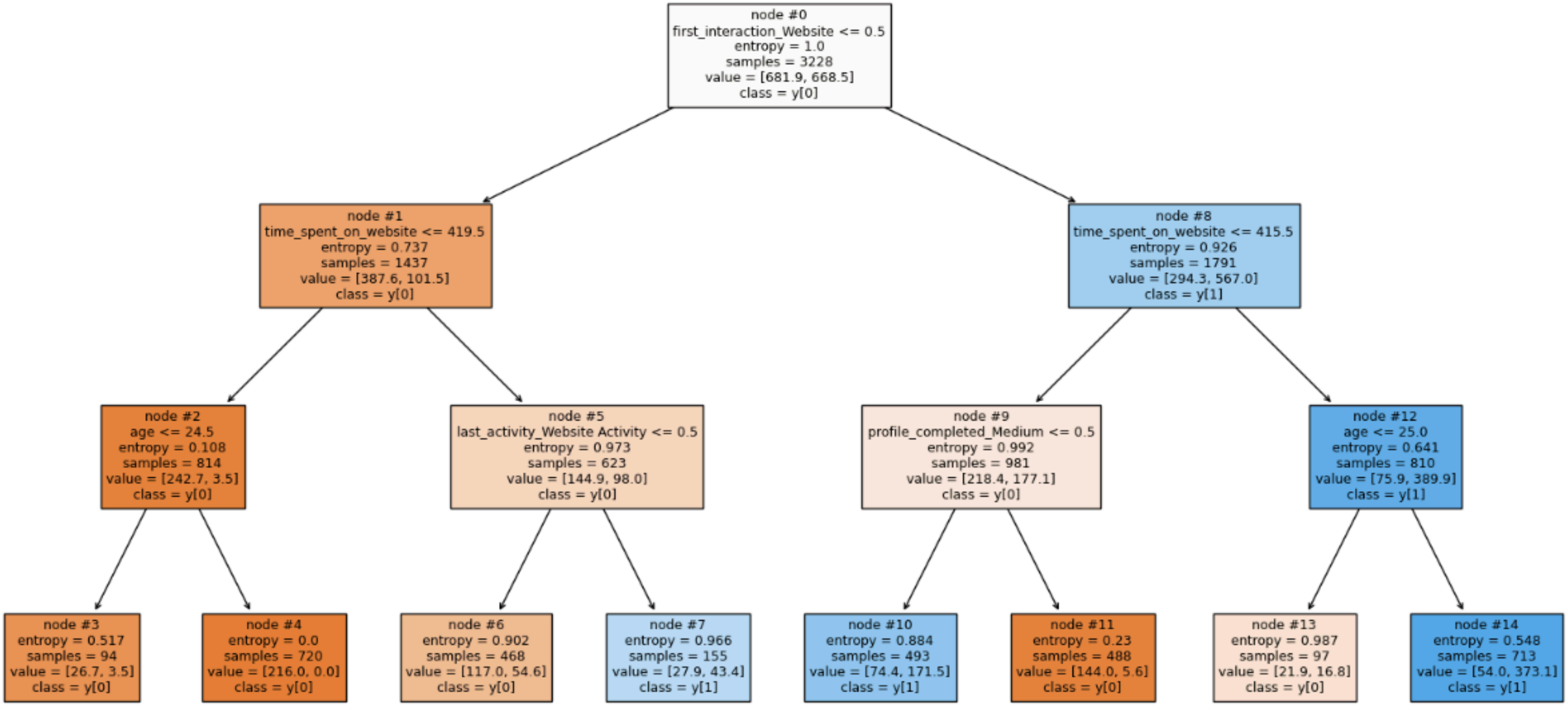
- Recall: 82%
- Precision: 77%
- f1-score: 78%
- Accuracy: 80%

Confusion Matrix



Actual \ Predicted	Not Converted	Converted
Not Converted	True Positive 829 59.90%	False Negative 133 9.61%
Converted	False Positive 126 9.10%	True Negative 296 21.39%

Decision Tree Visualization

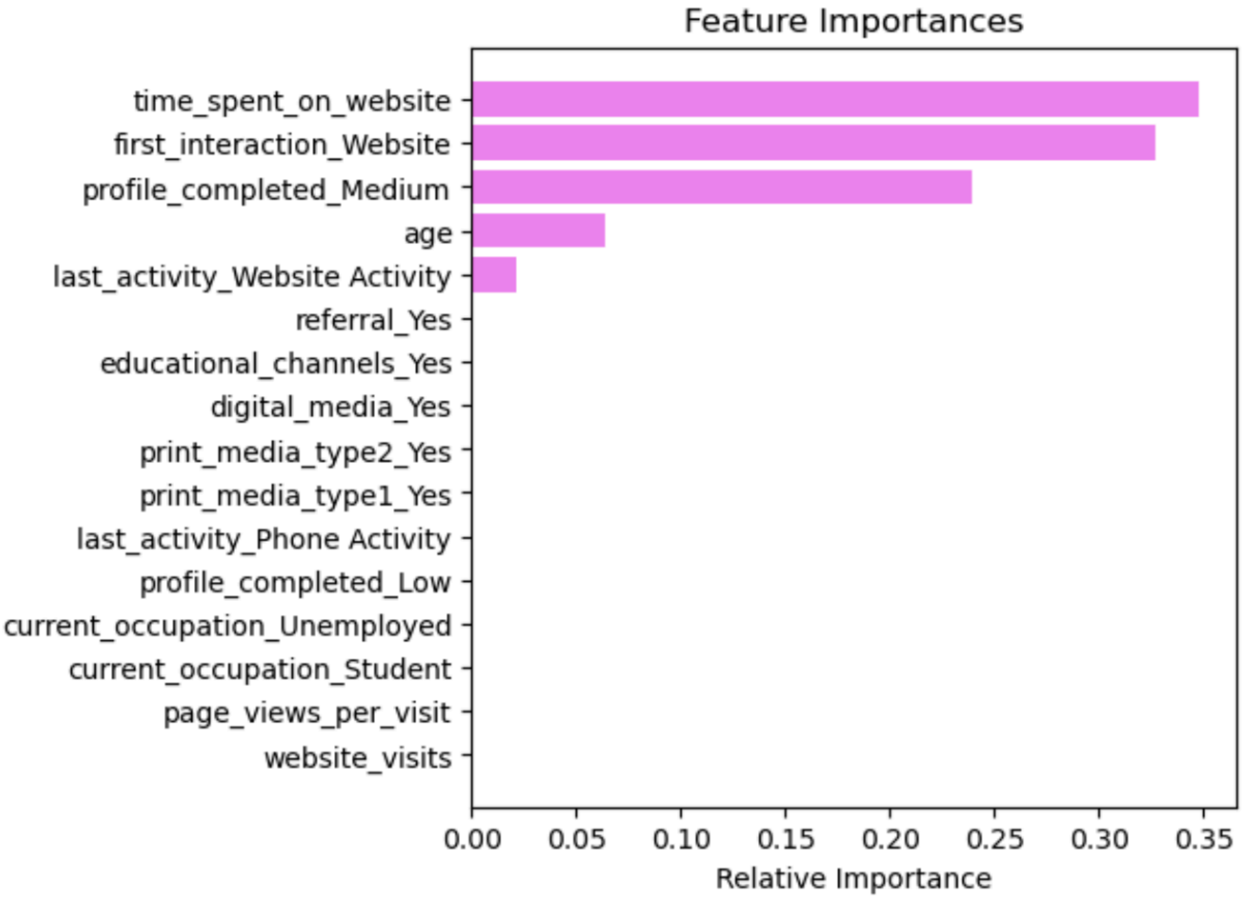


Decision Tree Model

Feature Importance

- Certain features are significantly more important in determining leads converted to paying customers

- Some features have to influence on leads conversion to paid customers

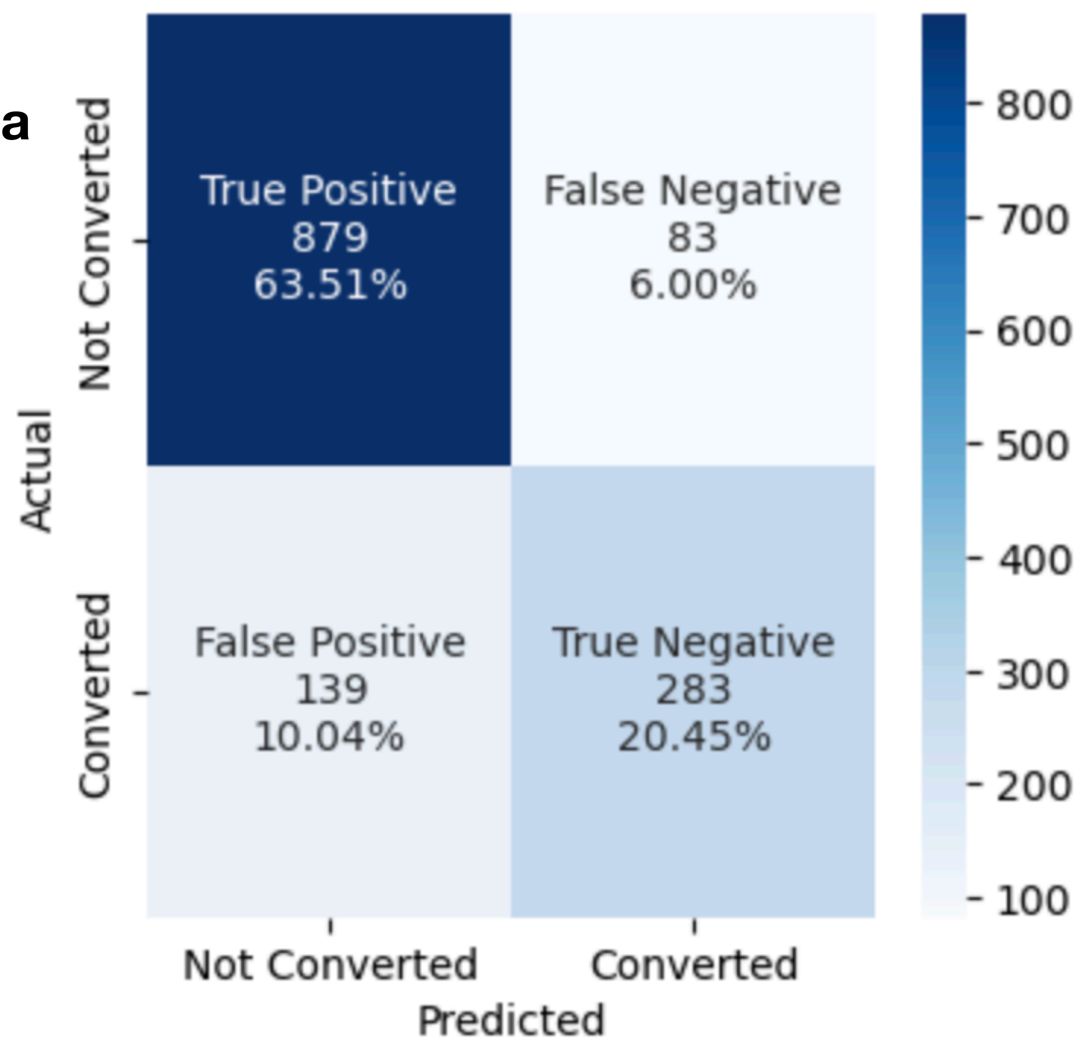


Bagging Model

Performance on unseen test data

- Recall: 79%
- Precision: 82%
- f1-score: 80%
- Accuracy: 84%

Confusion Matrix



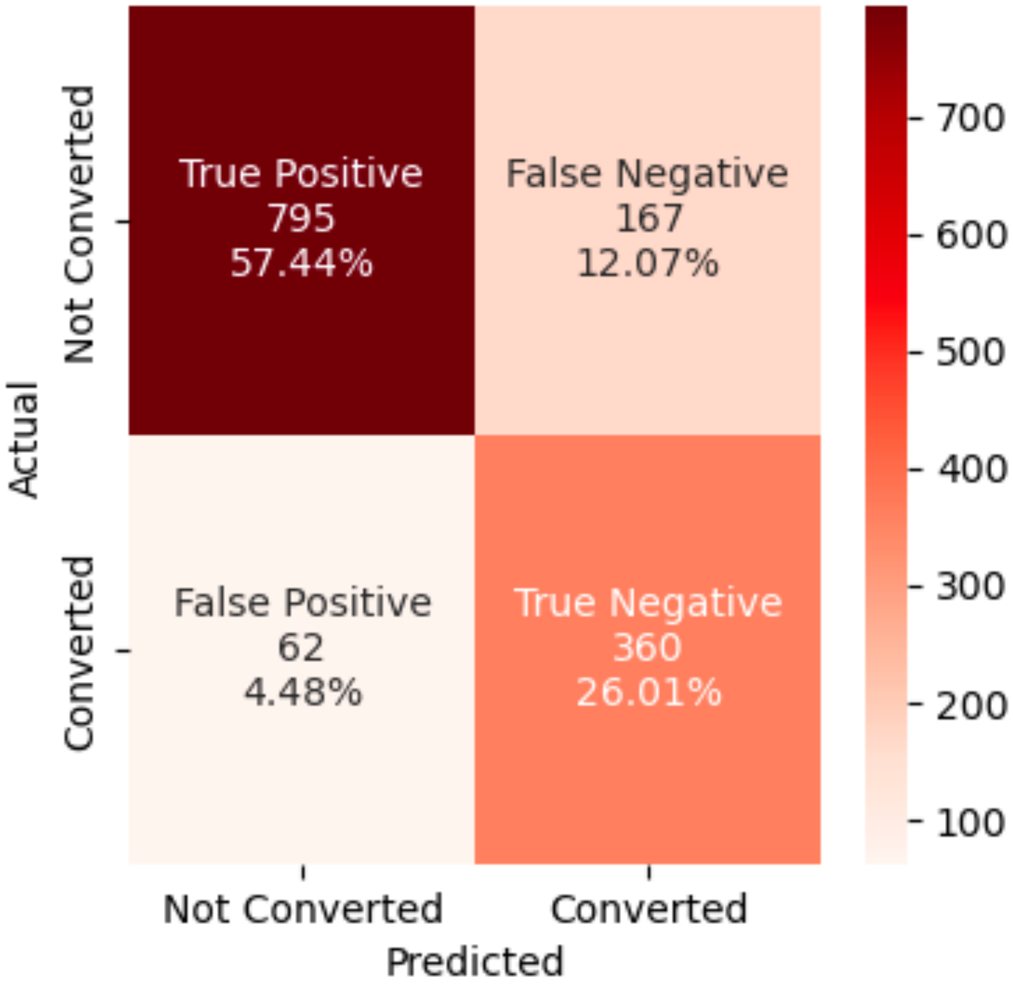
Actual \ Predicted	Not Converted	Converted
Not Converted	True Positive 879 63.51%	False Negative 83 6.00%
Converted	False Positive 139 10.04%	True Negative 283 20.45%

Random Forest Model

Performance on unseen test data

- Recall: 84%
- Precision: 81%
- f1-score: 82%
- Accuracy: 83%

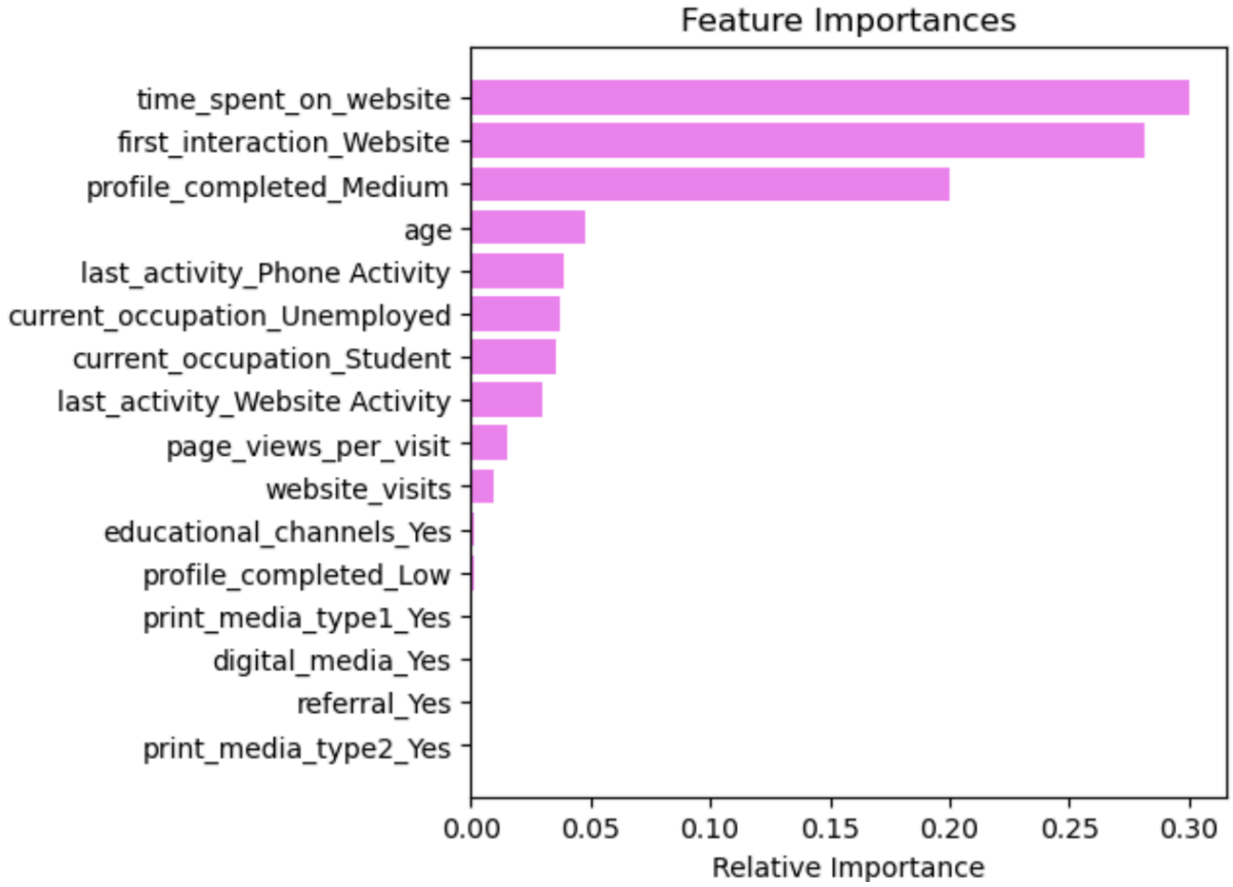
Confusion Matrix



Random Forest Model

Feature Importance

- Feature importance similar to Decision Tree

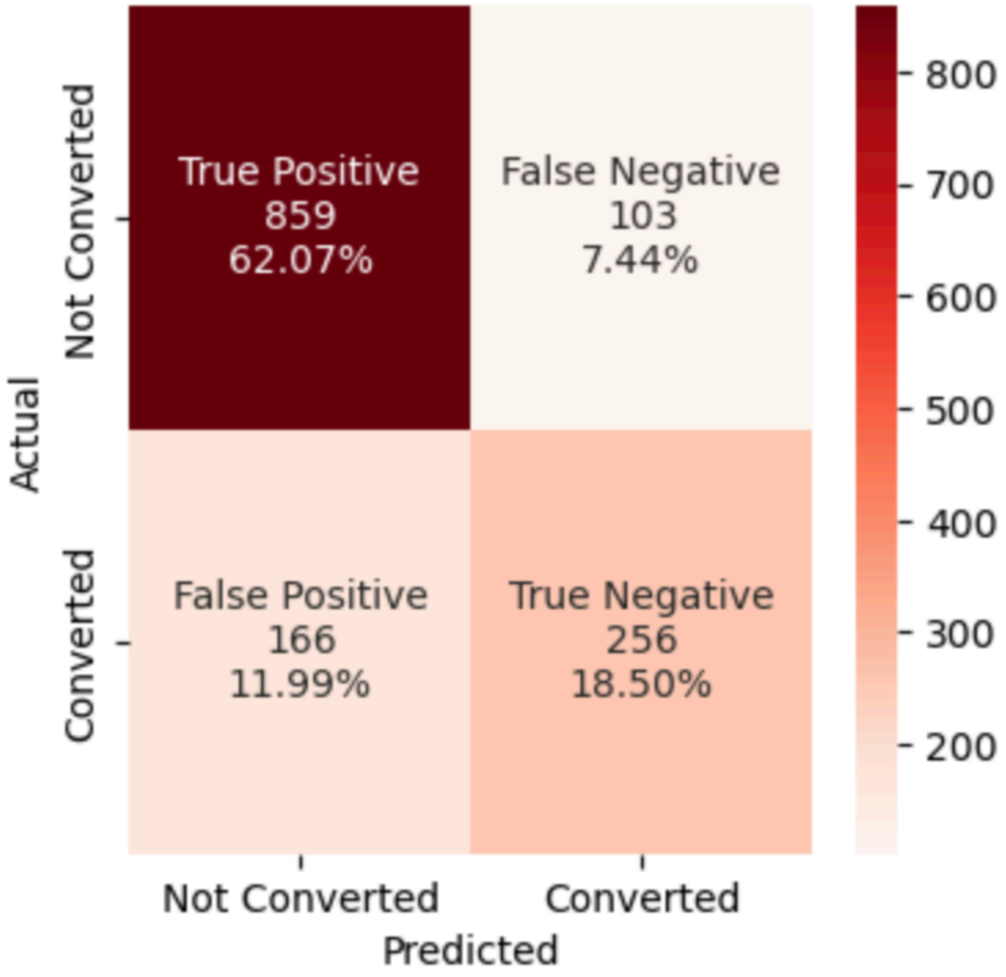


Logistic Regression Model

Performance on unseen test data

- Recall: 75%
- Precision: 78%
- f1-score: 76%
- Accuracy: 81%

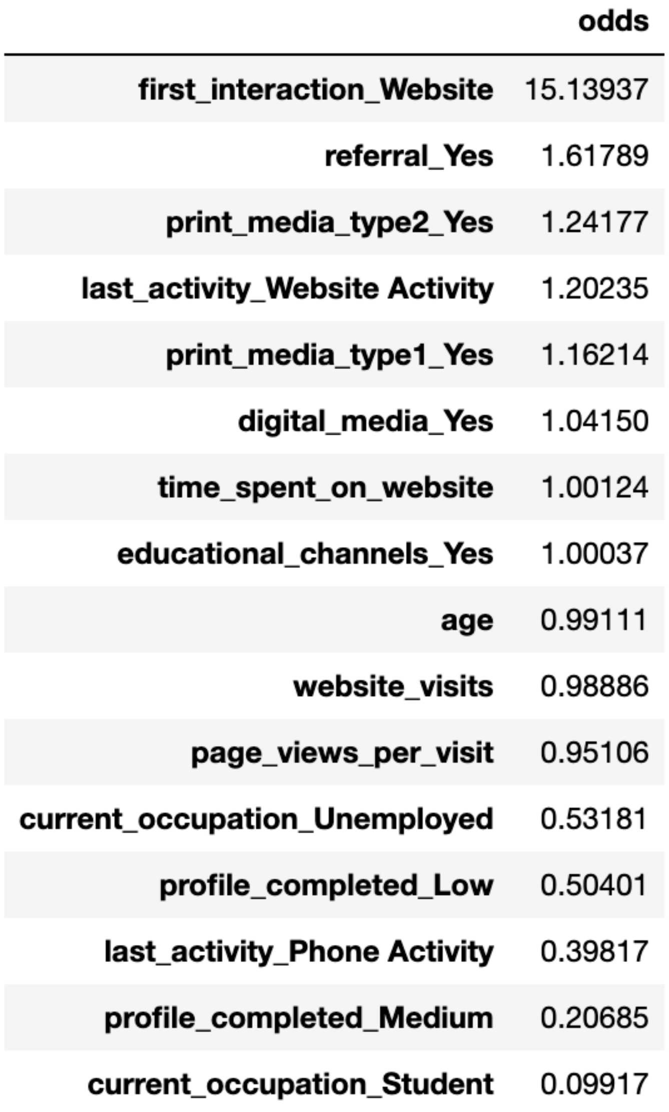
Confusion Matrix



Logistic Regression Model

Odds

- Odds of each feature causing a positive conversion of potential customer shown.
- The odds appear to be consistent with relative feature importances from the decision tree and random forest models



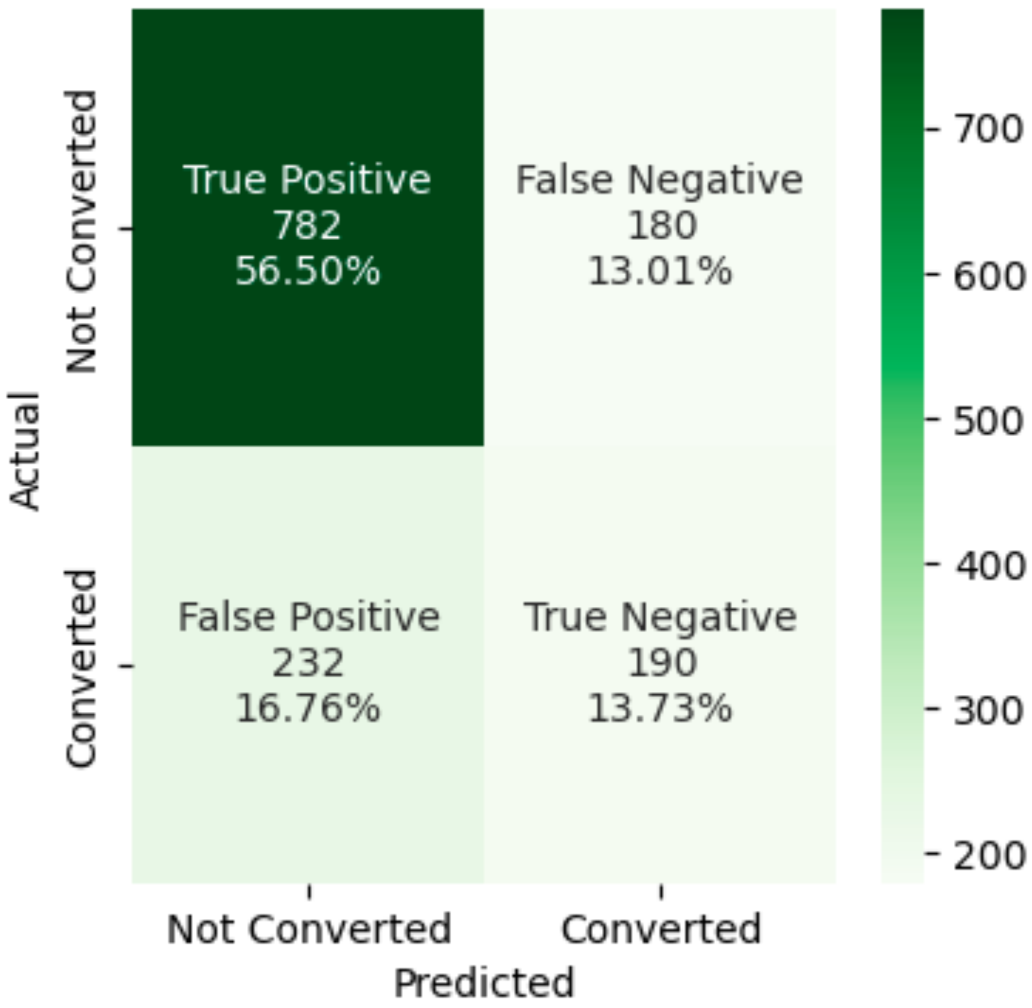
	odds
first_interaction_Website	15.13937
referral_Yes	1.61789
print_media_type2_Yes	1.24177
last_activity_Website Activity	1.20235
print_media_type1_Yes	1.16214
digital_media_Yes	1.04150
time_spent_on_website	1.00124
educational_channels_Yes	1.00037
age	0.99111
website_visits	0.98886
page_views_per_visit	0.95106
current_occupation_Unemployed	0.53181
profile_completed_Low	0.50401
last_activity_Phone Activity	0.39817
profile_completed_Medium	0.20685
current_occupation_Student	0.09917

KNN Model

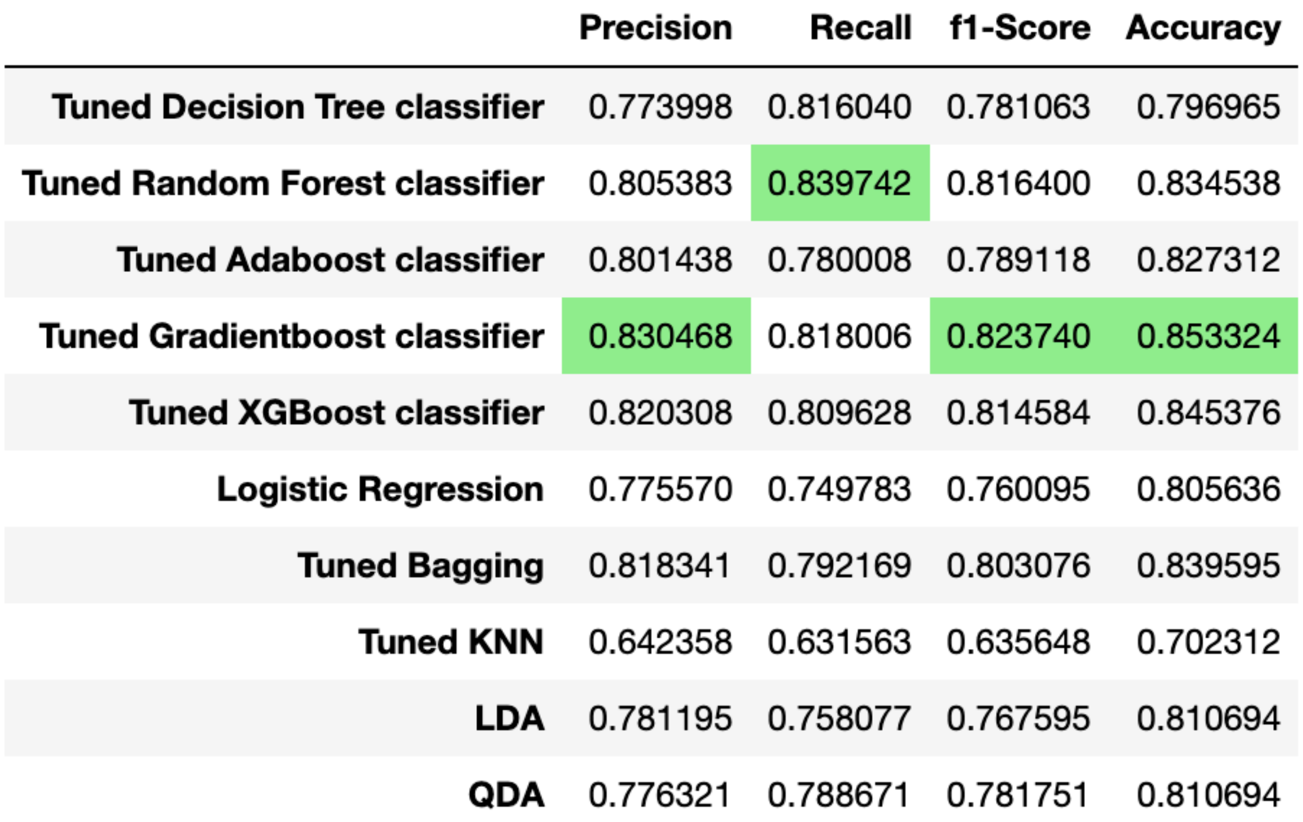
Performance on unseen test data

- Recall: 63%
- Precision: 64%
- f1-score: 64%
- Accuracy: 70%

Confusion Matrix



Performance Summary of Various Classifier Models



	Precision	Recall	f1-Score	Accuracy
Tuned Decision Tree classifier	0.773998	0.816040	0.781063	0.796965
Tuned Random Forest classifier	0.805383	0.839742	0.816400	0.834538
Tuned Adaboost classifier	0.801438	0.780008	0.789118	0.827312
Tuned Gradientboost classifier	0.830468	0.818006	0.823740	0.853324
Tuned XGBoost classifier	0.820308	0.809628	0.814584	0.845376
Logistic Regression	0.775570	0.749783	0.760095	0.805636
Tuned Bagging	0.818341	0.792169	0.803076	0.839595
Tuned KNN	0.642358	0.631563	0.635648	0.702312
LDA	0.781195	0.758077	0.767595	0.810694
QDA	0.776321	0.788671	0.781751	0.810694

Business Insight and Recommendation

- Random Forest Model with tuned parameters give the best performance. 84% of actually converted leads were predicted correctly for the unseen test data. Performance across all metrics are well balanced for the tuned random forest.

- Relative feature importances of the Decision tree and Random Forest models suggest time spent on the website and first_interaction_website are the most important factors in causing a lead conversion followed by profile_completed, age, and last_activity. ExtraaLearn need to ensure and improve customer experience on the website to increase chance of lead conversion.

- Logistic regression odds confirms the above mentioned factors.