Lead score case study

Anurag Choudhary
Ananya
DS-60 batch

PROBLEM STATEMENT

An education company X Education , selling online courses to industry professionals The company markets its courses through several websites and search engines like Google. When people browse through these websites or fill up forms providing their email address or phone number, they are classified as a lead. The company is facing a problem of low lead conversion rate of 30% which means out of 100 leads contacted in a day, only 30 gets converted.

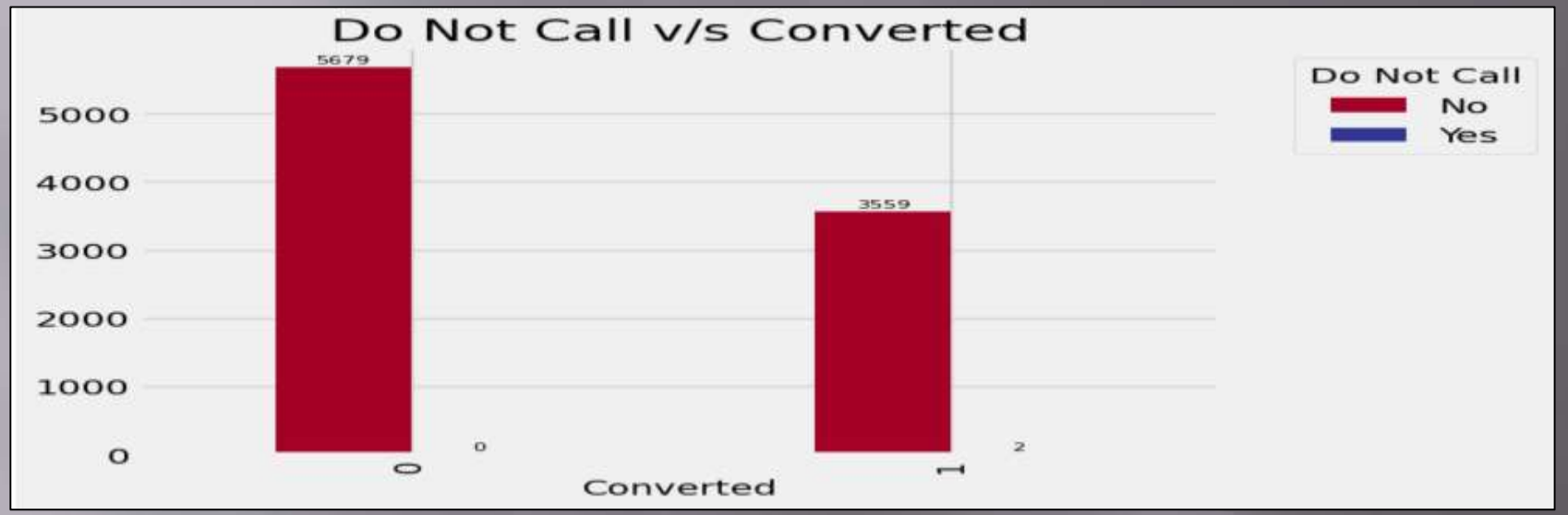
BUSINESS OBJECTIVES

The objective of case study is to help the company in making its lead identification process more efficient by identifying leads which are most likely to convert into paying customers or in other words hot leads and thus improve its lead conversion rate to 80%. To achieve this, a logistic regression model is to be built to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher lead score will mean higher chance of lead converting to paying customers and a lower lead score will mean cold lead and low chances of lead conversion.

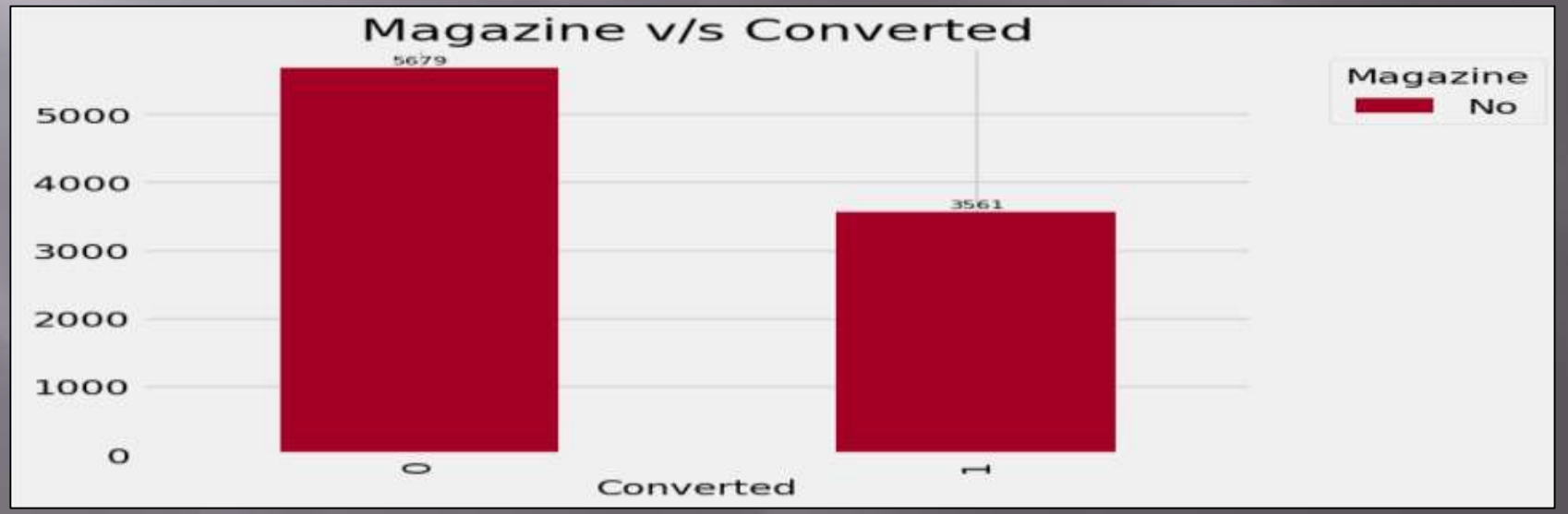
KEY STEPS UNDERTAKEN FOR ANALYSIS

1. Understanding the Problem statement, Business objectives and gaining necessary domain knowledge.
2. Observing the datasets , checking its structure.
3. Data cleaning of application dataset: missing values and outlier handling, checking datatypes.
4. Carried out Exploratory data analysis (EDA) for better understanding of data and gaining useful insights.
5. Data preparation: create dummy features (one-hot encoded), Train-test split, Feature scaling and looking for co-relations between variables.
6. Feature selection using RFE (Recursive feature elimination) algorithm.
7. Building logistic regression model.
8. Determining the optimum cut of value using ROC Curve, sensitivity-specificity curve and precision-recall curve.
9. Making prediction on train dataset and evaluating the model using metrics: accuracy, sensitivity, specificity, precision, recall.
10. Making prediction on test dataset and evaluating the model using metrics mentioned above.
11. Generating the lead score.

KEY FINDINGS OF EDA (EXPLORATORY DATA ANALYSIS)

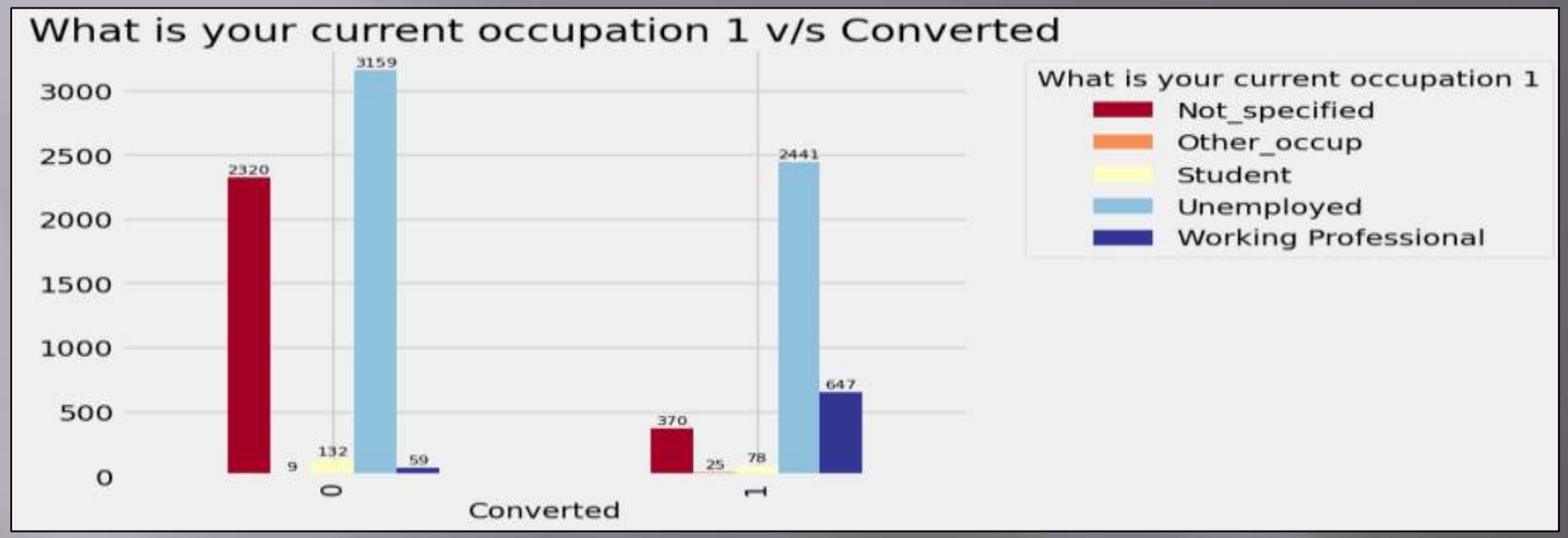


Most leads prefer not to be called over phone. Maximum percentage of unique value is “No”.

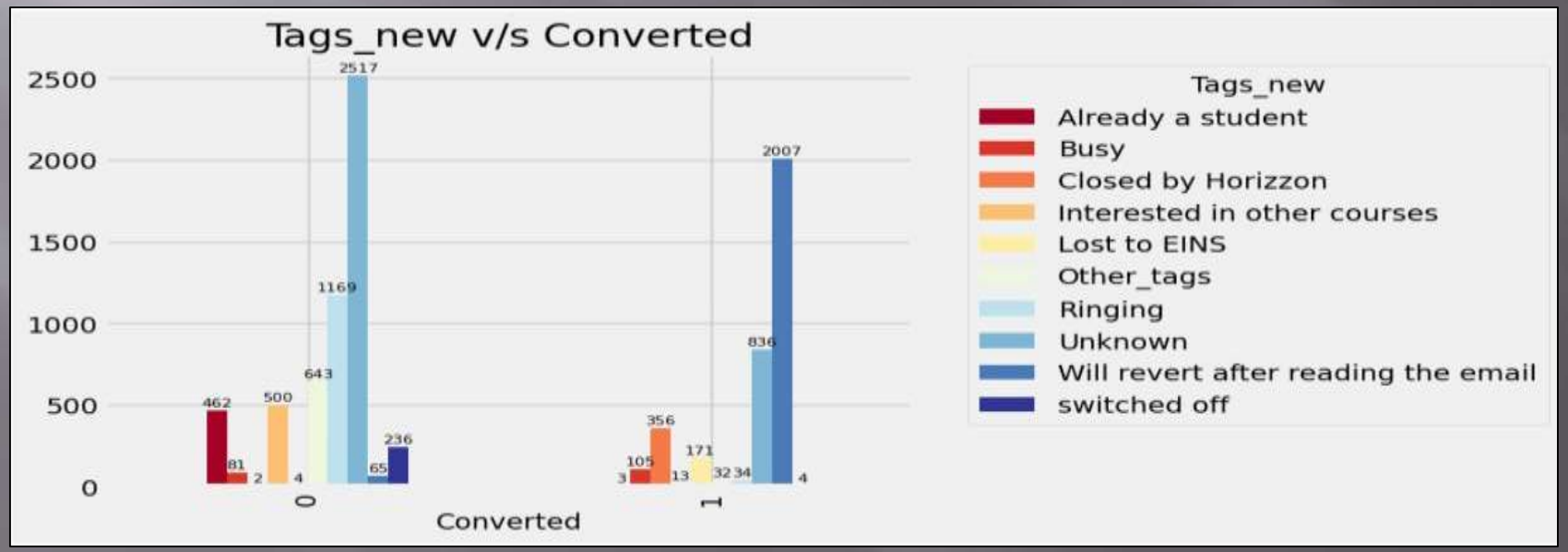


None of the leads have seen company’s ads through magazine

KEY FINDINGS OF EDA (EXPLORATORY DATA ANALYSIS)

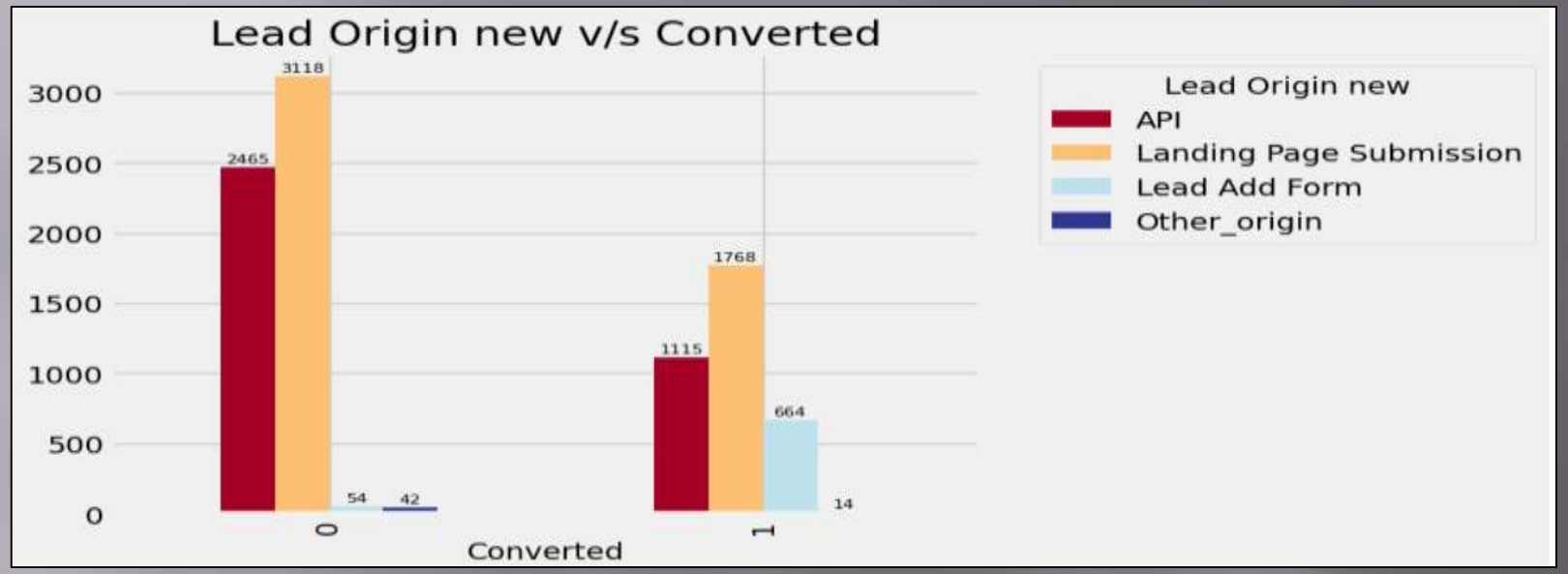


Working professionals have a good conversion rate

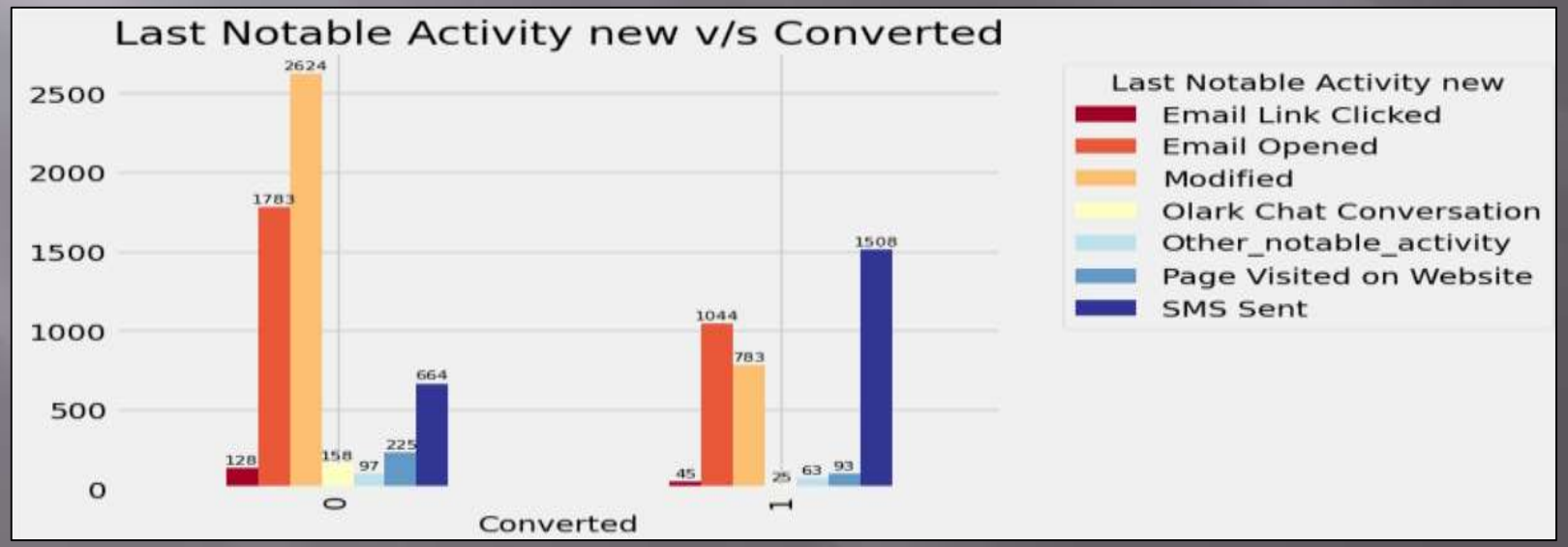


Tags: Closed by horizon, lost to EINS, will revert after reading email have good conversion rates

KEY FINDINGS OF EDA (EXPLORATORY DATA ANALYSIS)

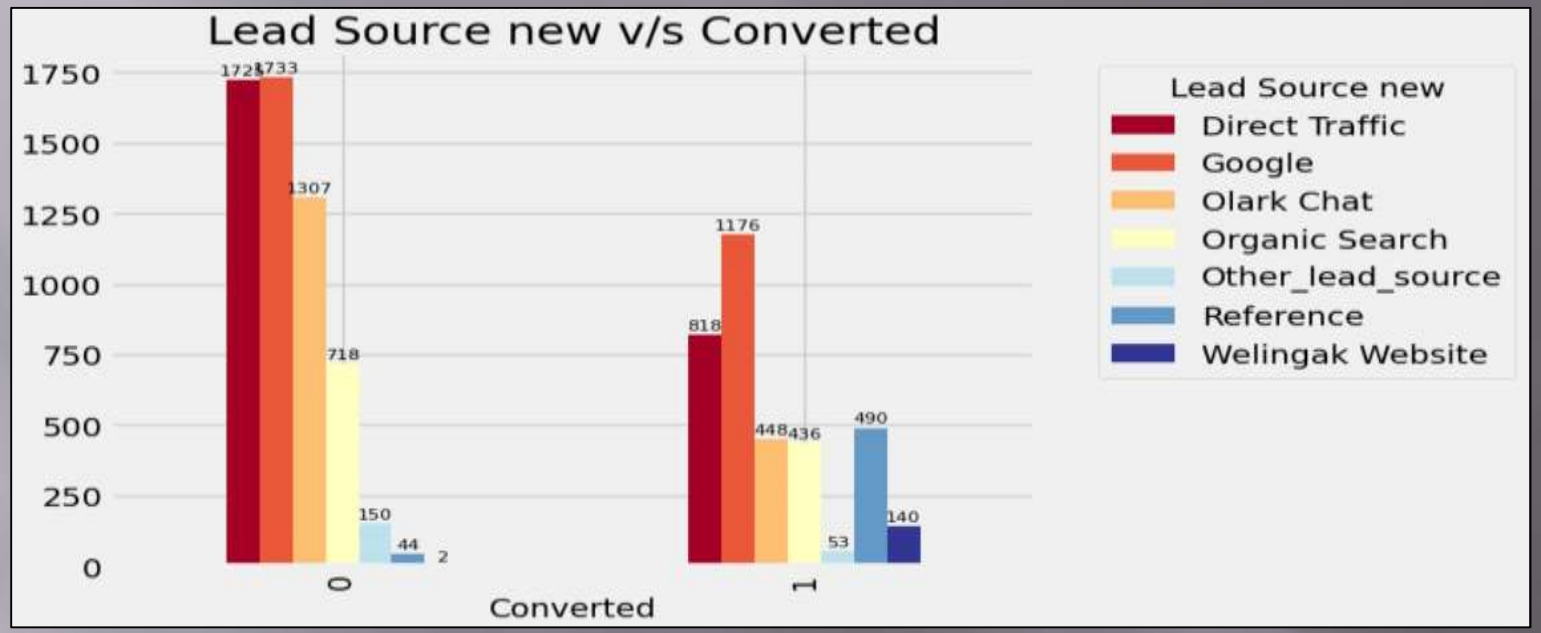


Leads originating from lead add form have good conversion rates.

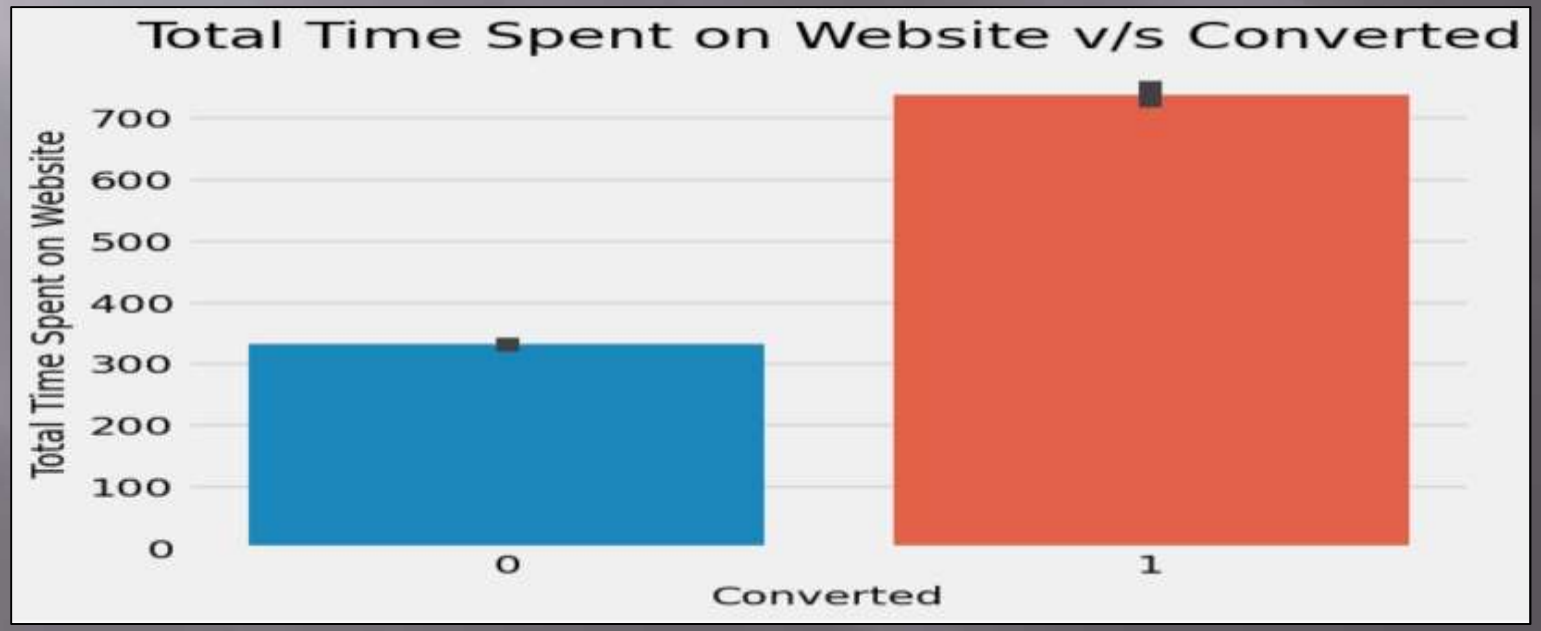


Leads having SMS sent as last notable activity have good conversion rates and leads having last notable activity as modified and olark chat conversation have less conversion rates

KEY FINDINGS OF EDA (EXPLORATORY DATA ANALYSIS)

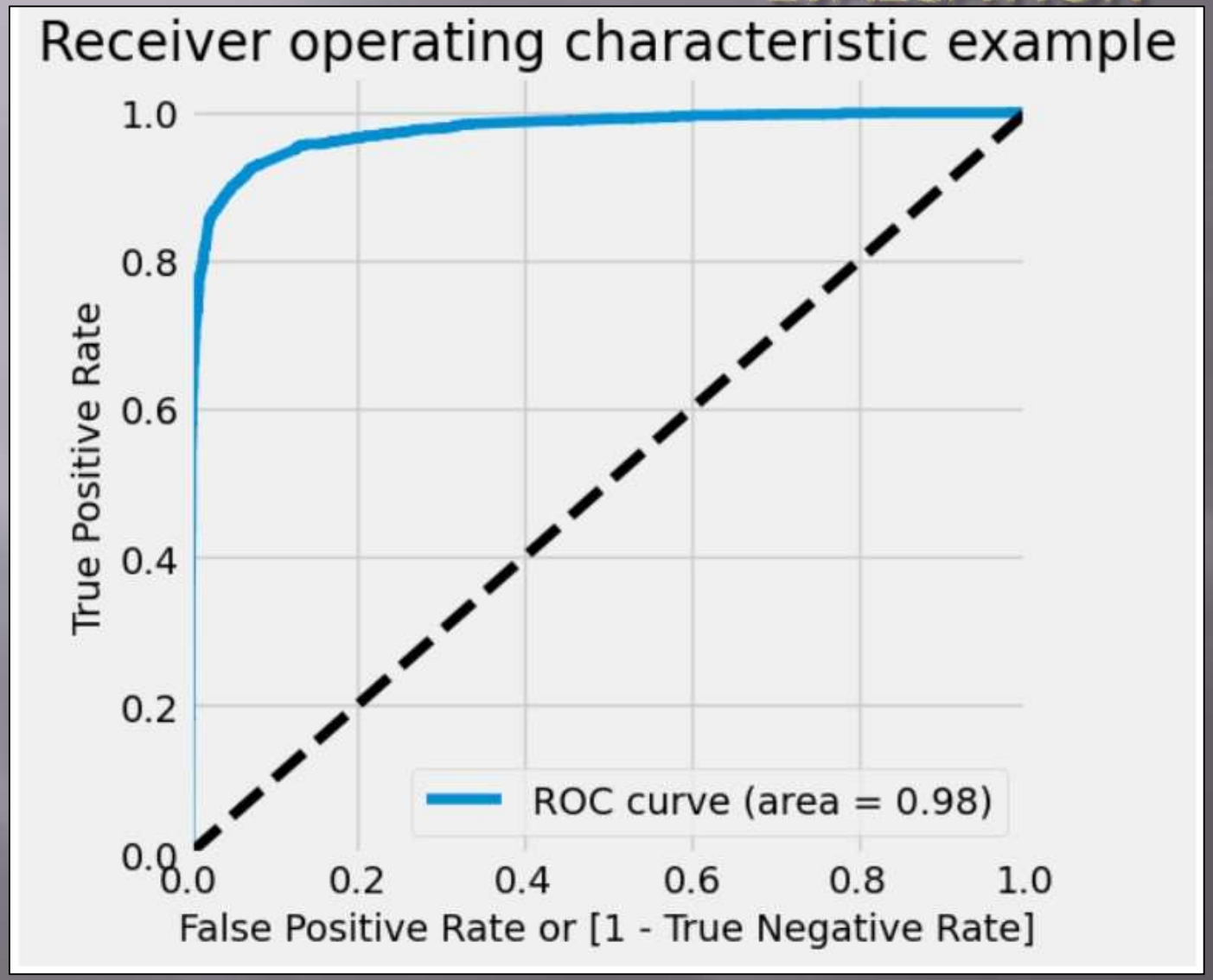


Leads sourced from Reference, Welingak Website have good conversion rates.



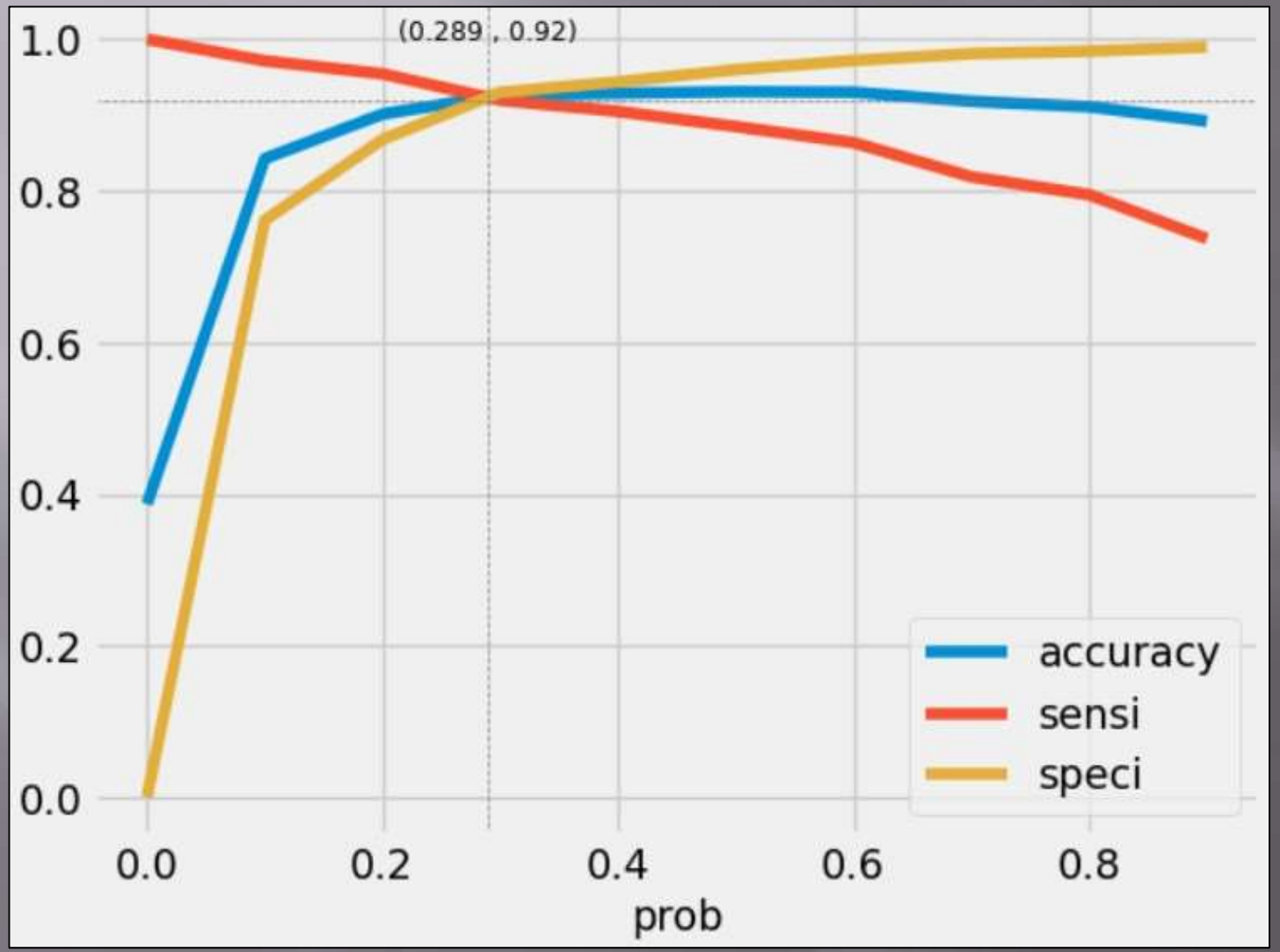
Leads spending more time on website have good conversion rates.

KEY RESULTS OF MODEL EVALUATION



The ROC curve is hugging the top-left corner of the graph, indicating high sensitivity (True Positive Rate) and low false positive rate simultaneously. Also AUC is 0.98 which is near to ideal value of 1.

KEY RESULTS OF MODEL EVALUATION



With optimum cut off obtained from accuracy-sensitivity-specificity curve following metrics were obtained for train and test data:

```
Accuracy of train data: 0.927
Sensitivity of train data: 0.9229
Specificity of train data: 0.9297
False positive rate of Train Data: 0.0703
False negative rate of Train Data: 0.0771
Precision of Train Data: 0.8922
Recall of Train Data: 0.9229
```

```
Accuracy of test data: 0.921
Sensitivity of test data: 0.9122
Specificity of test data: 0.9264
False positive rate of Test Data: 0.0736
False negative rate of Train Data: 0.0878
Precision of Test Data: 0.8846
Recall of Test Data: 0.9122
```

CONCLUSIONS AND RECOMMENDATIONS

- **Highly recommended Groups which should be targeted to increase conversion rates are:**

1. Leads having tags Closed by Horizzon, Tags_Lost to EINS, Will revert after reading the email.
2. Leads having last activity as SMS sent should be targeted.
3. Leads spending more time on website.
4. Leads originating from Lead Add Form.
5. Leads sourced from Olark chat, Welingak Website.
6. Working professional leads.

- **The following groups should not be focused on as their chances of conversion are low:**

1. Leads having tags switched off , already a student, ringing, Interested in other course and other tags
2. Leads with last notable activity as modified, Olark Chat Conversation
3. Leads with do not email.