iiit-b

UpGrad

Lead Scoring Case Study

By
Vyom Bhatt
Nisha Kambli

The Problem Statement

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Now, although X Education gets a lot of leads, its lead conversion rate is very poor. To make this process more efficient, the company wishes to identify the most potential leads, also known as **Hot Leads**.

The company requires a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Methodology – Data Preparation

1. Initially performing **missing values treatment** on the different variables provided in the dataset
   - Rows with more than 30% null values removed, and columns with more than 30% null values removed
   - Imputation done on the ‘Lead Source’ column based on the ‘Lead Origin’ column data
   - Dropping categorical columns that have massive imbalance in the levels
2. Exploratory Data Analysis
   - Visualizing numerical variables using scatter plots to identify trends between them
   - Univariate analysis, bivariate analysis and segmented univariate analysis on the categorical variables to understand what is the distribution of the categories, and the levels within each category contributing to highest conversions
3. Outlier Analysis – boxplot visualizations for every numerical variable to treat outlier values
4. Creating dummy variables and encoding of categorical variables, and dropping dummy variables showing massive imbalance
5. Splitting the data into test-train, then performing feature scaling on the numerical variables (MinMaxScaler())
6. Identifying multicollinearity using the correlation heatmap

Methodology – Model Building

7. Perform Logistic regression on the entire train set initially to get an idea of what the accuracy looks like
8. Performing Logistic Regression using RFE to limit the model to 25 most significant variables
9. Fine-tune the model by repeating the model building after dropping 1 variable at a time based on p-value (indicating variable significance) and VIF (variable inflation factor)
10. Finally, we're left with 19 variables giving an accuracy of 0.79 as the final model
11. Run the prediction on the same train set to identify the confusion matric, and evaluate the model based on accuracy, precision and recall. Also evaluating based on ROC curve which returned an area of 0.86
12. Identified optimal cut-off probability as 0.4 based on accuracy, sensitivity and specificity for different probabilities
13. Make predictions on the test data set using 0.4 as cut- off probability - accuracy on test set was 0.8
14. Use our model to answer the question raised by X Education i.e. assigning lead score from 0-100 for every lead in the data and identifying the number of leads that will lead to 80% of conversion rate

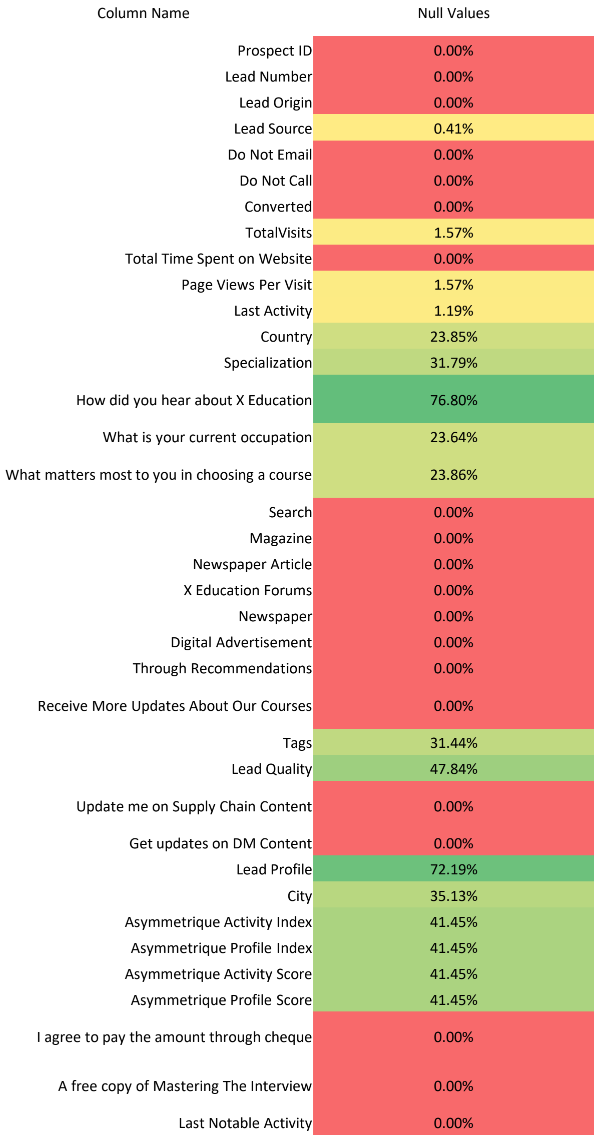
Missing Value Treatment

Within the dataset provided, looking at each of the column individually, we can see that there are missing values that need to be treated.

All the columns that have over 30% of missing values i.e. Specialization, How did you hear about X Education, Tags, Lead Quality, Lead Profile, City, Asymmetrique Activity Index, Asymmetrique Profile Index, Asymmetrique Activity Score, Asymmetrique Profile Score have been eliminated as imputing these many values would create a lot of bias/assumptive data

The missing values in Lead Source column have been imputed based on the maximum Lead Source values within the different Lead Origin values.

The missing values in the numerical columns i.e. Total Visits and Page Views per Visit are minimal, and we've proceeded with mean value imputation instead of removing the column so that we're not losing any datapoints

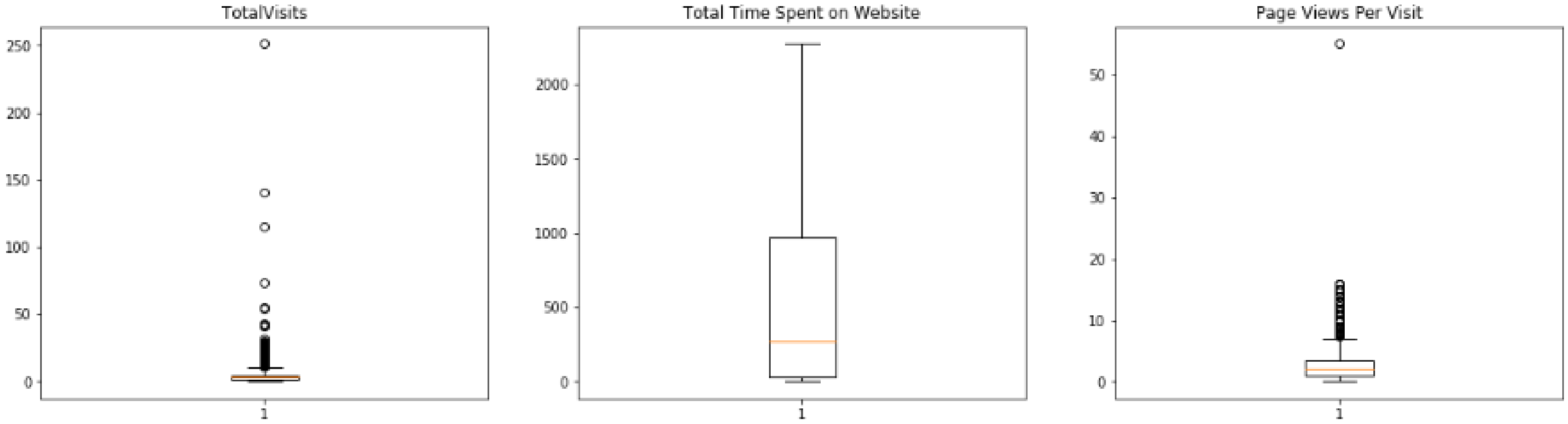


Column Name	Null Values
Prospect ID	0.00%
Lead Number	0.00%
Lead Origin	0.00%
Lead Source	0.41%
Do Not Email	0.00%
Do Not Call	0.00%
Converted	0.00%
TotalVisits	1.57%
Total Time Spent on Website	0.00%
Page Views Per Visit	1.57%
Last Activity	1.19%
Country	23.85%
Specialization	31.79%
How did you hear about X Education	76.80%
What is your current occupation	23.64%
What matters most to you in choosing a course	23.86%
Search	0.00%
Magazine	0.00%
Newspaper Article	0.00%
X Education Forums	0.00%
Newspaper	0.00%
Digital Advertisement	0.00%
Through Recommendations	0.00%
Receive More Updates About Our Courses	0.00%
Tags	31.44%
Lead Quality	47.84%
Update me on Supply Chain Content	0.00%
Get updates on DM Content	0.00%
Lead Profile	72.19%
City	35.13%
Asymmetrique Activity Index	41.45%
Asymmetrique Profile Index	41.45%
Asymmetrique Activity Score	41.45%
Asymmetrique Profile Score	41.45%
I agree to pay the amount through cheque	0.00%
A free copy of Mastering The Interview	0.00%
Last Notable Activity	0.00%

Outlier Analysis

Outlier Analysis

Looking at the distribution of the different features provided:



We observe that there exists a number of datapoints that are above the upper whisker of the box plot which can be treated.

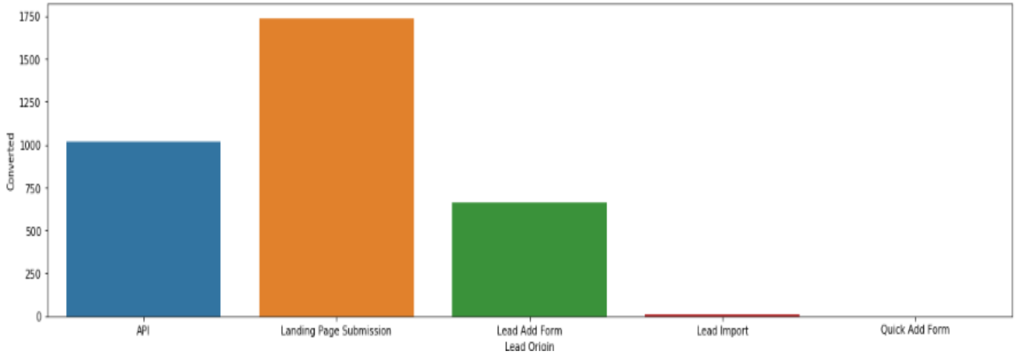
Also, on observing these datapoints, removing them doesn’t affect our converted user share, hence we can go ahead and drop these datapoints.

Outlier treatment was done for both the ‘Total Visits’ and the ‘Page Views per Visit’ variables.

EDA Insights

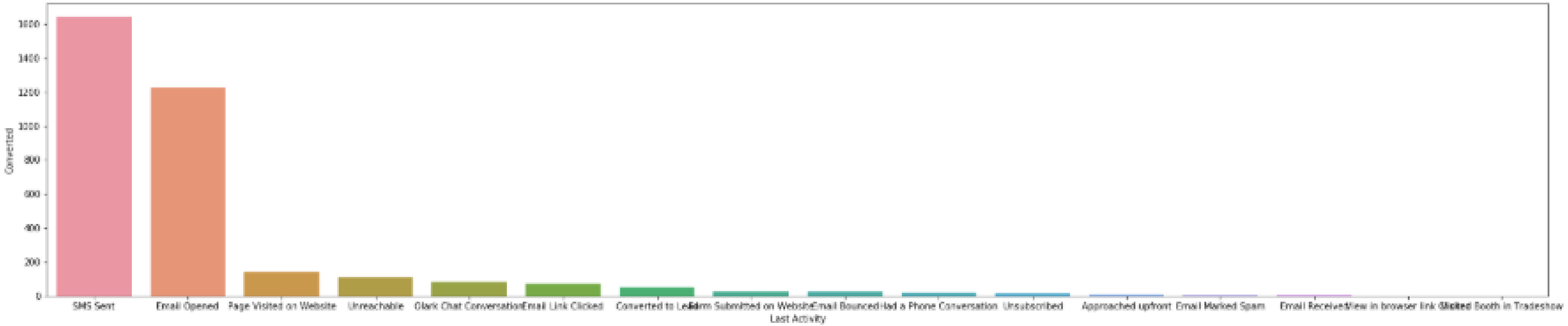
On performing exploratory data analysis on the datasets, we can derive some really interesting insights that give us an indication of what our converts look like, which factors are important, if there exists class imbalance etc.

Looking at the different types of **Lead Origin's** contributing to conversions:



The Landing Page Submission type of **Lead Origin** was the most effective as it garnered the most number of conversions

Looking at the **Last Activity**, we observe that ‘SMS Sent’ and ‘Email Opened’ were the ones post which we saw most of conversions

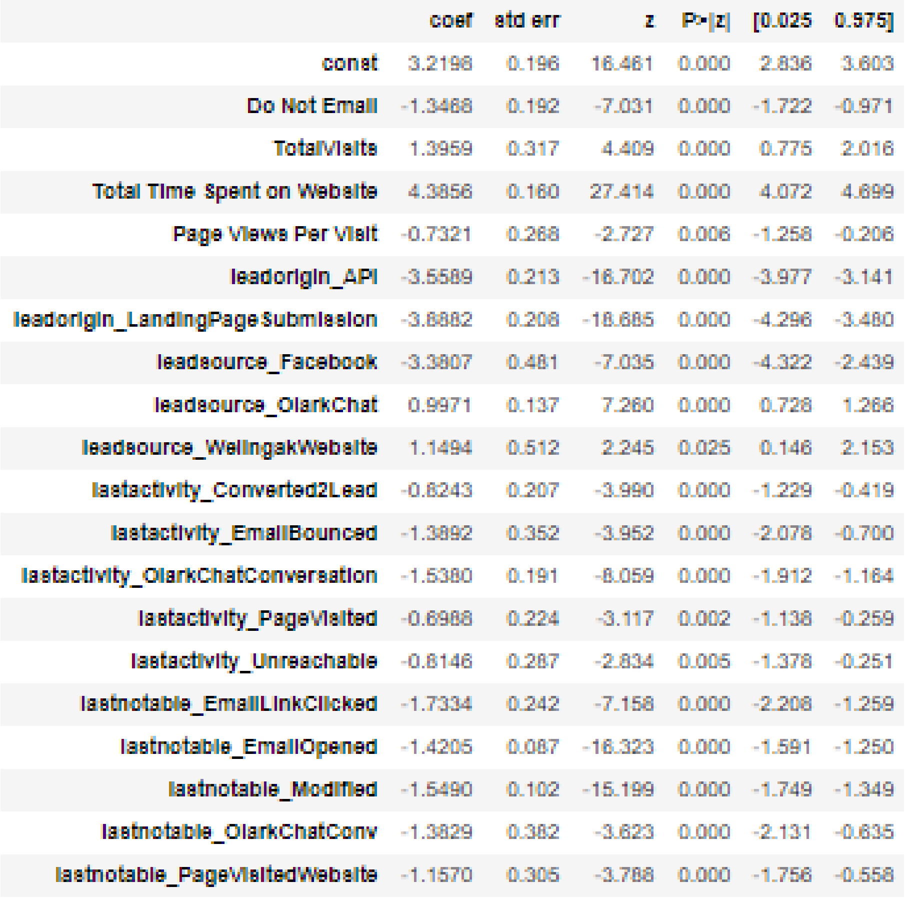


Model Building

Post Exploratory Data Analysis, we performed encoding of the categorical variables into dummy variables, and then went ahead with a test-train split (30-70). Min-Max scaling was performed on the numerical variables to bring them to the same scale.

We performed **Logistic Regression using RFE** to bring the features down to 25 first, and then fine-tuning was done based on p-values and VIF values to eventually drill down to 19 variables on which the model finally showed impressive performance.

The final list of variables that were selected are as shown in the model summary on the right

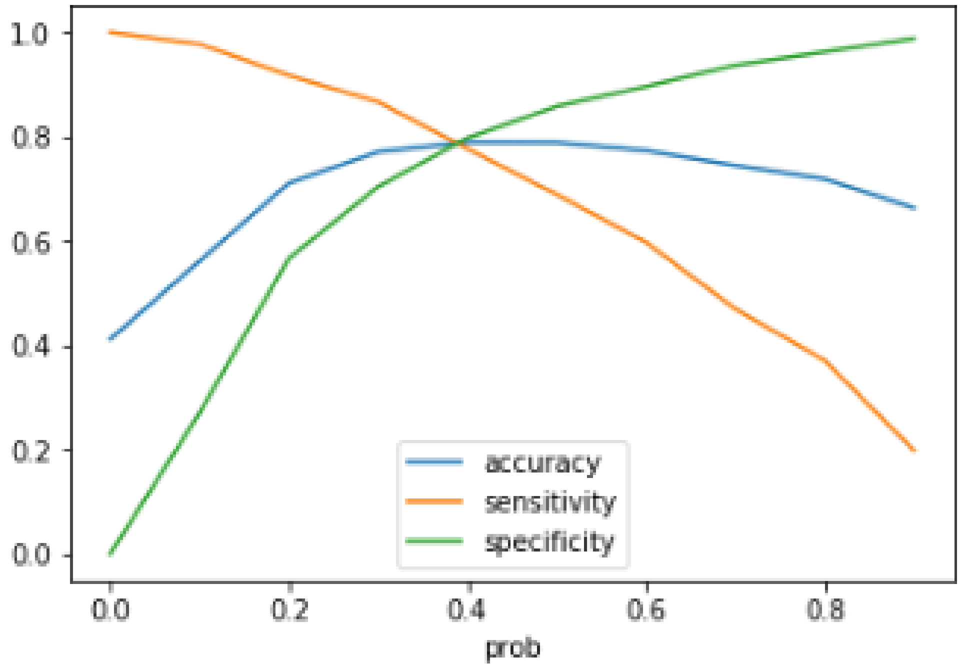


	coef	std err	z	P>|z|	[0.025	0.975]
const	3.2198	0.196	16.461	0.000	2.836	3.603
Do Not Email	-1.3468	0.192	-7.031	0.000	-1.722	-0.971
TotalVisits	1.3959	0.317	4.409	0.000	0.775	2.016
Total Time Spent on Website	4.3856	0.160	27.414	0.000	4.072	4.699
Page Views Per Visit	-0.7321	0.268	-2.727	0.006	-1.258	-0.206
leadorigin_API	-3.5589	0.213	-16.702	0.000	-3.977	-3.141
leadorigin_LandingPageSubmission	-3.8882	0.208	-18.685	0.000	-4.296	-3.480
leadsource_Facebook	-3.3807	0.481	-7.035	0.000	-4.322	-2.439
leadsource_OlarkChat	0.9971	0.137	7.260	0.000	0.728	1.266
leadsource_WelingakWebsite	1.1494	0.512	2.245	0.025	0.146	2.153
lastactivity_Converted2Lead	-0.8243	0.207	-3.990	0.000	-1.229	-0.419
lastactivity_EmailBounced	-1.3892	0.352	-3.952	0.000	-2.078	-0.700
lastactivity_OlarkChatConversation	-1.5380	0.191	-8.059	0.000	-1.912	-1.164
lastactivity_PageVisited	-0.6988	0.224	-3.117	0.002	-1.138	-0.259
lastactivity_Unreachable	-0.8146	0.287	-2.834	0.005	-1.378	-0.251
lastnotable_EmailLinkClicked	-1.7334	0.242	-7.158	0.000	-2.208	-1.259
lastnotable_EmailOpened	-1.4205	0.087	-16.323	0.000	-1.591	-1.250
lastnotable_Modified	-1.5490	0.102	-15.199	0.000	-1.749	-1.349
lastnotable_OlarkChatConv	-1.3829	0.382	-3.623	0.000	-2.131	-0.635
lastnotable_PageVisitedWebsite	-1.1570	0.305	-3.788	0.000	-1.756	-0.558

Prediction

Finding the optimal cut-off point

The output of our model was a list of probabilities that convey how likely is a lead to convert, with a higher probability indicating a greater chance of the lead converting. In order to do this, we plotted the accuracy, sensitivity and specificity values at different probability cut-off values from 0 to 1



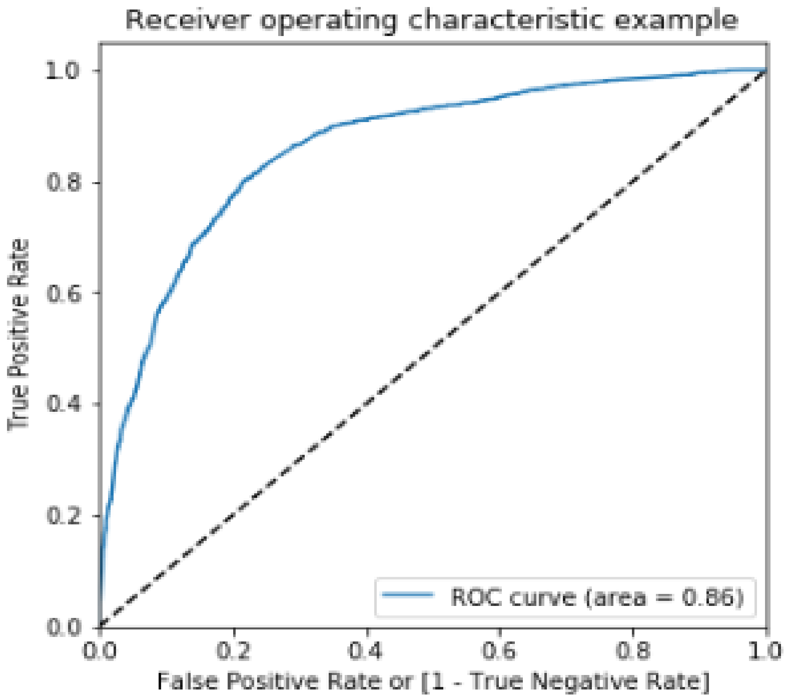
Based on the plot, we can see that the optimal cut-off probability (the point where all the three metrics intersect) is approximately 0.4.

We used this cut-off value to determine whether the predicted value for the lead would be 0 or 1

Model Evaluation

Area under ROC curve

On plotting the ROC curve, which shows the **tradeoff between the True Positive Rate (TPR) and the False Positive Rate (FPR),** we see that the area under the curve is **0.86** which is very close to 1, hence shows that the model is good



Other Evaluation Parameters

Accuracy = 0.79

Sensitivity = 0.79

Specificity = 0.80

Positive Rate = 0.20

Positive predictive value = 0.72

Negative predictive value = 0.86

Precision Score = 0.72

Recall Score = 0.79

Key Variables for Lead Conversion

Based on the model, the key variables that impact the conversion of a lead include:

- Total Time Spent on Website
- Total Visits
- Lead Source (Wellingak Website, Olark Chat)

The above variables have been selected based on the magnitude of the coefficient and the sign of the coefficient. As a result, if the above variables are high, that would lead to a much higher probability for the lead to convert

Recommendations

X Education CEO has asked us to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

The model we have built assigns a lead score to each of the leads given in the dataset. In order to achieve the ballpark figure, we can use the predicted conversions data to identify the total number of potential converts that are most likely to convert.

In our case, we managed to predict that a total of 1087 users would be converting. In order to ensure we're meeting at least 80% of users that convert, it would be ideal if the company could reach out to about 1400 of the customers with the highest probability to convert to ensure 80% of those would be the **Hot Leads**

Recommendations

After identifying the **Hot Leads** using our model, the company can implement several strategies to try and persuade them to convert.

This could include:

- Referral bonuses
- Incentives on reaching different stages of the enrollment process – for example, they could initiate 100% refund for the first 7 days of the course to the user post enrollment incase they don't like the course
- Grouping similar courses into 1 and providing the course as an entire package with a discount
- Scholarship tests to get the students interested in applying for the course based on the scholarship amount
- Better marketing of the opportunities that await the consumers if they end up enrolling for the course
- Sending them regular testaments of students who have previously been part of the course