upGrad

Lead Scoring Case Study

Ajeevansh Gautam
Ashish Pandey

Problem Statement

upGrad

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. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Objective:

Build a logistic regression model 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 score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Cleaning & Treating Missing Values

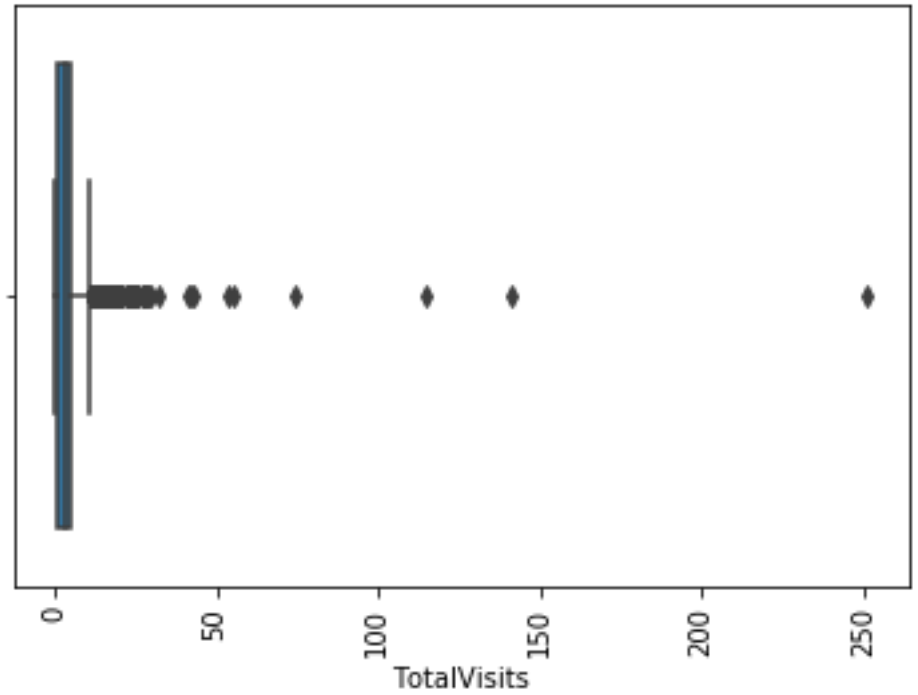
upGrad

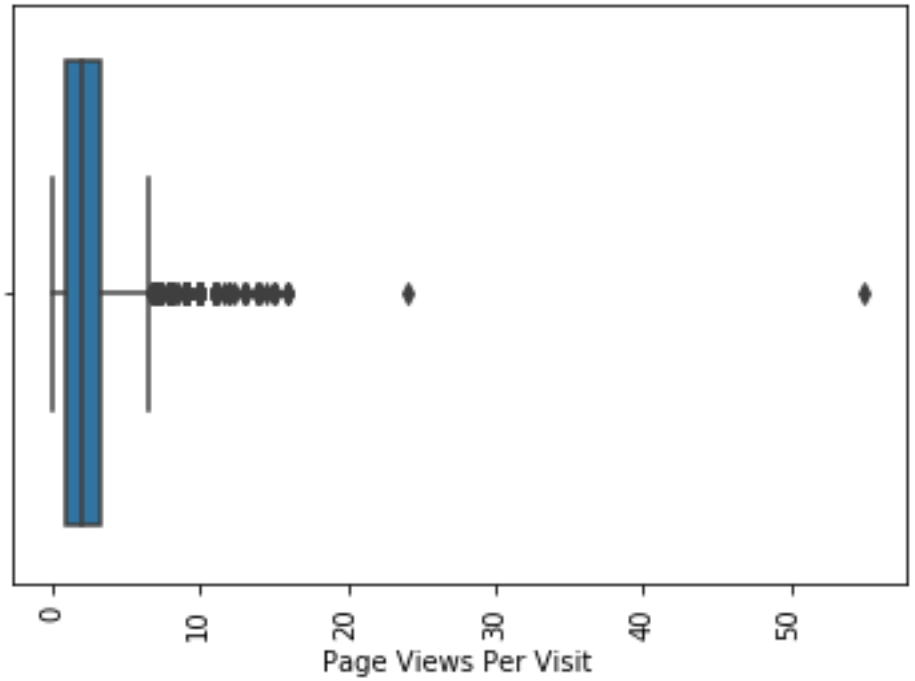
- The data contained 9240 rows and 36 columns out which only 7 columns were numerical.
- Many of the columns had more than 20% of null values but some of these columns were important from the business point of view.
- Lead Quality even though had 50% of Null Values but seemed like an important column as per business point of view.
- Columns which had more than 90% of the same values like: Country, Do Not Call, Search, Magazine, Newspaper Article, X Education Forum, etc could be dropped straight away because these columns won't add meaning to the analysis.
- We can club some non performing values of some features like: Tags, Last Activity and Lead Source to "Others", so that we can reduce some of the features, while creating dummies.

Removing Outliers

upGrad

- A few of the columns had outliers (Page Views Per Visit, TotalVisits), so we removed them using 0.05 and 0.95 percentile bound values.
- The quantile values here are taken as 0.05 and 0.95





Analysis Approach

upGrad

- The first step is to perform Univariate Analysis on the data. From the initial analysis a lot of insights can be generated.
- After performing EDA, we can easily figure out what all columns have an impact on the business and what columns can safely be dropped.
- Outlier removal only happened for two variables.
- After performing the initial steps of analysis, we can move on with the Test-Train Split and Feature Scaling of the variables using a Standard Scalar.
- Model Building: Initially we will use all the columns to generate a regression model and then use RFE Technique to shred down the number of features. (In this case RFE using 15 features from 85)
- After removing "Tags_invalid number" we can see that the p-value and VIF of the rest of the features are in accordance with the model parameters (p-value<0.05 & VIF<5)
- For the train dataset we will now generate the predicted values and use a 0.5 cutoff for mapping.

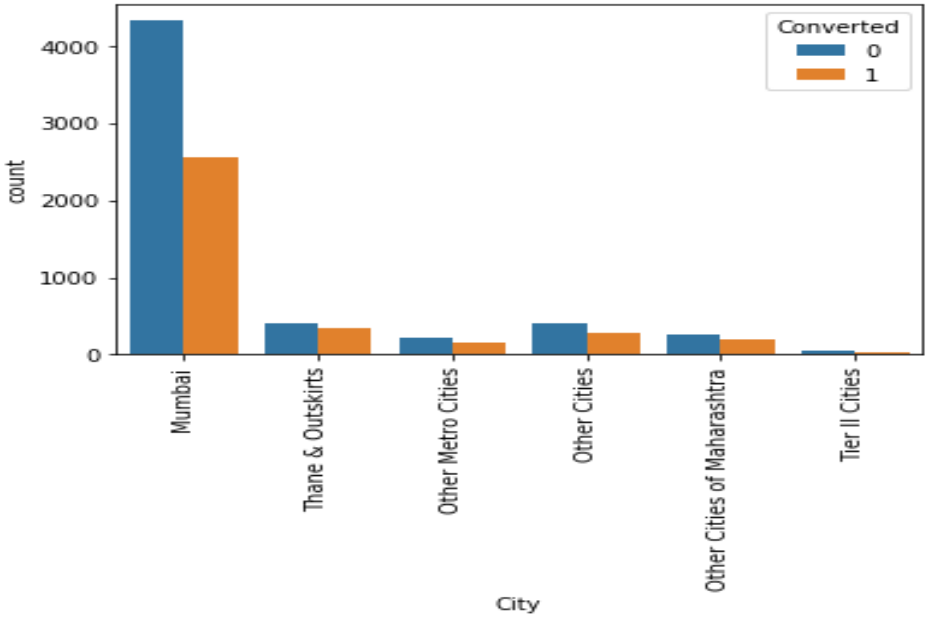
Analysis Approach

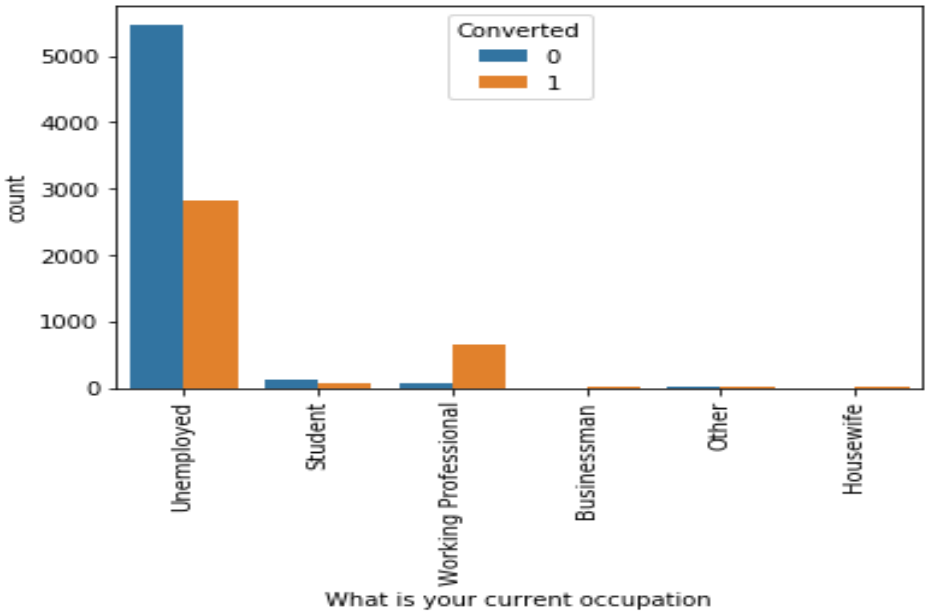
upGrad

- Checking the confusion matrix and overall accuracy (0.919) in this case.
- Checking Sensitivity, Specificity, False Positive Rate, Positive Predicted Value.
- ROC Curve will be plotted, and we will find the Optimal Cutoff Probability (0.2 in this case) by plotting Probability with Accuracy, Specificity, Sensitivity.
- 0.2 Cutoff Probability means that any value that has a probability of more than 0.2 will considered to be converted.
- We can then assign a Lead Score based on the Predicted Probability.
- After checking the relevant True Positive, True Negatives, False Positives, and False Negatives we can then plot a Precision Recall Curve.
- After plotting the precision recall curve, we can finally move forwarding with predicting on the test dataset and using metrics to evaluate the same.

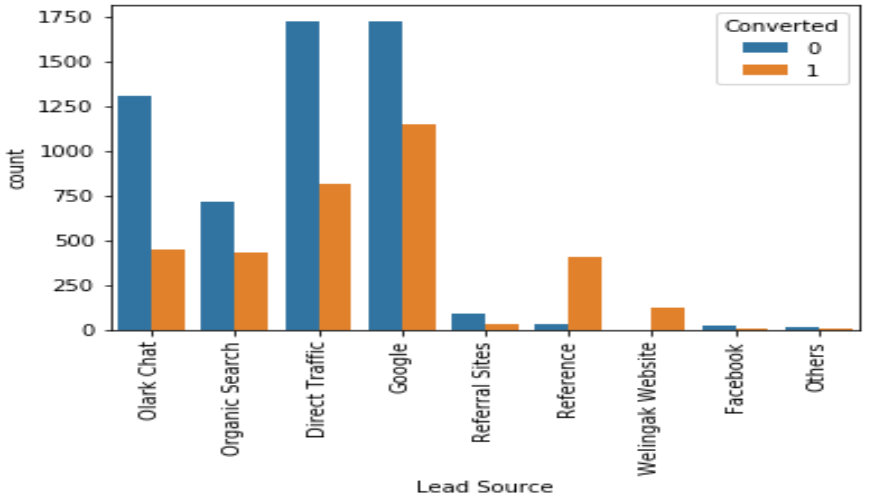
Analysis Results – EDA

upGrad





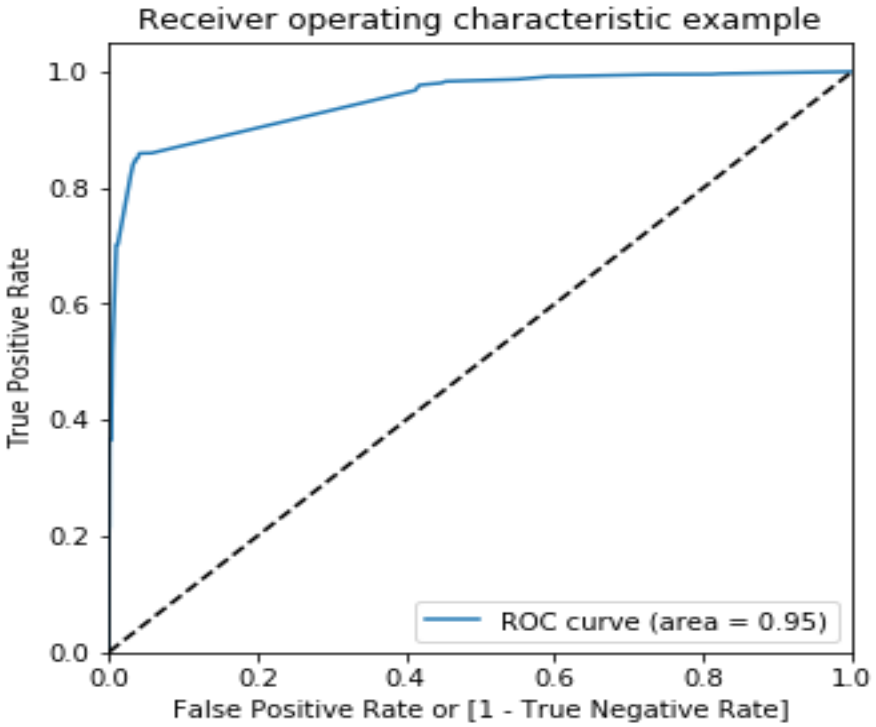
1. The most conversions have happened from Mumbai.
2. The most conversions have been when the occupation is unemployed, but the conversion ratio is higher for Working Professionals.
3. The maximum conversions have happened when Lead Source was Google.

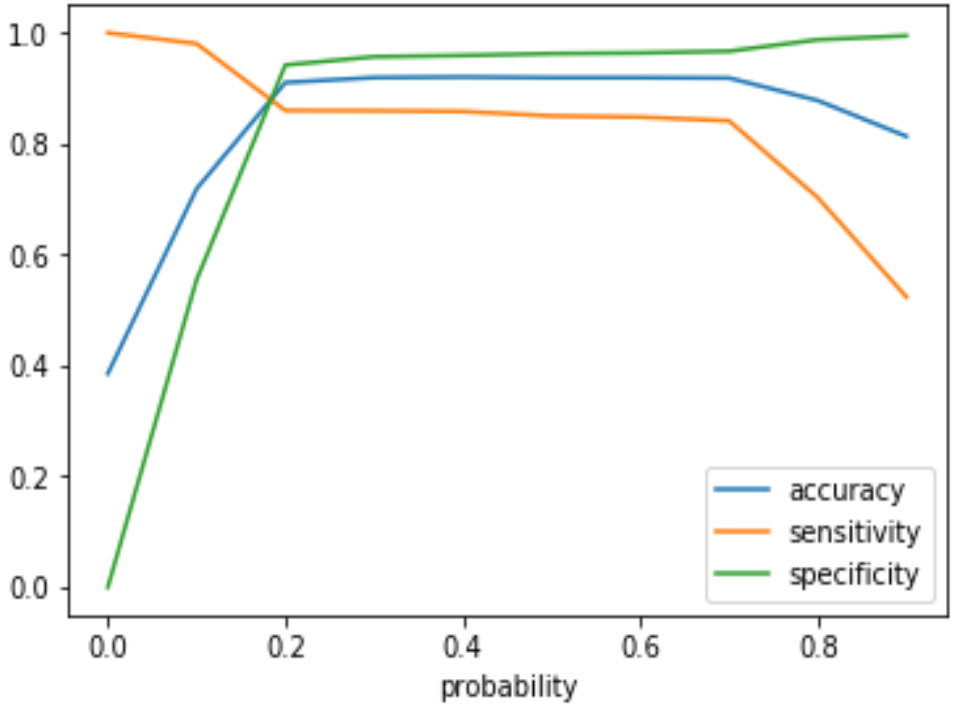


Analysis Results – Metrics Accuracy

upGrad

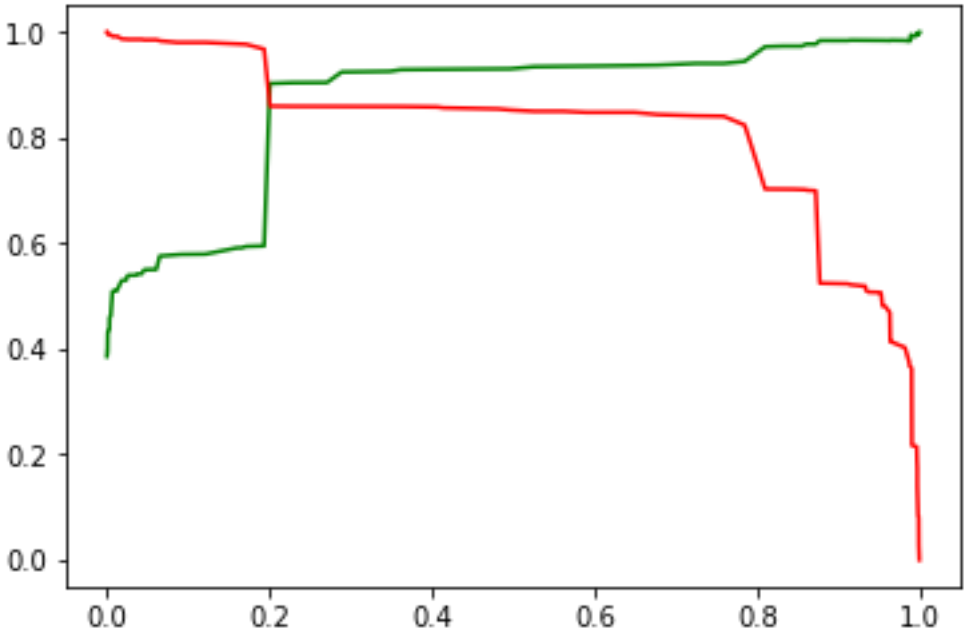
- Sensitivity: 85.9%
- Specificity: 94.1%
- False Positive Rate: 0.058%
- Positive Predictive Value: 90.2%
- Negative Predictive Value: 91.4%





Analysis Results – Overall Accuracy with Test Data

upGrad



Precision Recall Curve

- Overall Accuracy: 90.7%
- Sensitivity: 85.9%
- Specificity: 94.1%

Analysis Results – Final Factors Impacting Business

upGrad

- The factors that affect the conversion rate are:
  - *Do Not Email*
  - *Lead Origin_Lead Add Form*
  - *Lead Source_Welingak Website*
  - *What is your current occupation_Working Professional*
  - *Tags_Busy*
  - *Tags_Closed by Horizzon*
  - *Tags_Lost to EINS*
  - *Tags_Ringing*
  - *Tags_Will revert after reading the email*
  - *Tags_switched off*
  - *Lead Quality_Not Sure*
  - *Lead Quality_Worst*
  - *Asymmetrique Activity Index_03.Low*
  - *Last Notable Activity_SMS Sent*