Lead Scoring Case Study

UpGrad – DS C58

15 Jan 2024

Study Group: Anjana , Krishna Arjun, Sundaravelmurugan

Analysis Journey

- Business Objective
- Datasets
- Null values, Outliers
- Data Univariate analysis
- Bivariate analysis
- Model Building
- Model Evaluation
- Insights

Business Objective

- Apply Logistic Regression model on the assigned lead score on the Leads
- Identify potential leads who could be converted into a student for X Education
- Identify driving factors behind lead conversion

Datasets

- Datasets provided
  - *'Leads.csv'* - Leads dataset
- 9240 observations and 37 features provided
- 'Converted' is the Target variable
- Variables Prospect ID, and Lead Number are similar to index variables
- Variables 'Receive More Updates About Our Courses', 'Update me on Supply Chain Content', 'Get updates on DM Content', 'I agree to pay the amount through cheque', 'Magazine' has only one Class and don't add much value
- Analysis and Model building would be done with the remaining variables

EDA Approach

- Identify Missing/Null values and approach to handle the deal the issue
- Identify Outliers and approaches to handle outliers
- Data Imbalance analysis
- Univariate and Bivariate analysis and plots
- Correlation analysis
- Model building using Logistic regression
- Model Evaluation using various metrics
- Derive business insights

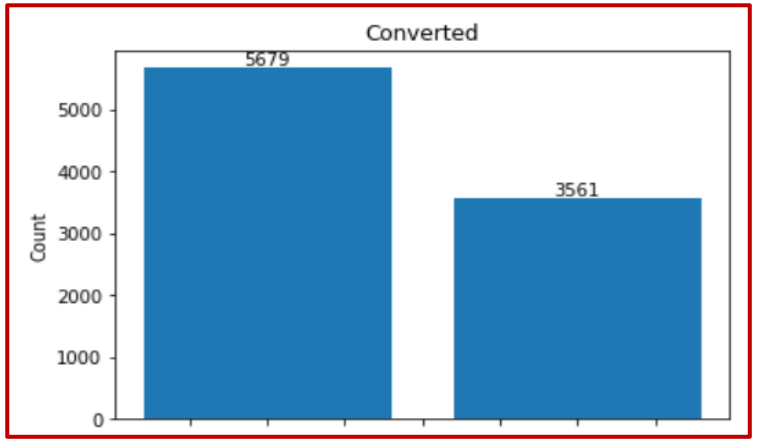
Missing Values

- A total of 7 columns have missing values > 40%. Those features are dropped before further analysis steps.
- There are 4 columns with missing values ranging from as low as ~140 rows (< 1%). These Features are all categorical and imputed with Mode values.
- Features ‘Country’, ‘Specialization’, ‘What is your current occupation’, ‘What matters most to you in choosing a course’, ‘Tags’, ‘City’ has missing values in the 26 to 39% range. For these features, the missing values would be created as a separate category.
- Variables Country is binned into three categories India, Unknown and Other Countries

Outliers

- Outliers are values that lie far away from other values. By definition, any value outside of the 1.5 * InterQuartile range in an Outlier.
- Outlier treatment is necessary in the Data Preprocessing step before proceeding to model building. Outliers could be treated in different ways:
- There are three numerical features in the dataset: ‘TotalVisits’, ‘Total Time Spent on Website’, and ‘Page Views Per Visit’.
- Histogram of the three variables shows that the variables are skewed to the right.
- Remove all outliers outside of 1 percentile and 99 percentile.

Univariate Analysis – Leads Dataset



- The target variable looks fairly balanced
- Minority class percentage is ~ 40%
- Lead Origin: Landing Page Submission and API dominated the
- Feature
- Lead Source: Google, Direct Traffic, and Olarkchat dominate the feature with a list of other values contributing to the remaining values
- Do Not Email: Most people have answered No. Do Not Call: Most People have opted for No.
- Last Activity: Email Opened and SMS sent seem the most occurring Last Activity Country: Most people are from India Specialization: The values are contributed across multiple specializations Current Occupation: Unemployed is the mostly occurring Feature class

Bivariate Analysis – Leads Dataset

Features that Influence Target Variable

- Added form is more effective way to convert people
- Olark chat and referral sites perform lowest in the conversion of people
- Reference helps most in converting people.
- SMS sending has very good responses from people.
- Indian people are showing a positive response in conversion count compared to out of India.
- Management professions like Finance, HR, Marketing, and Operations have a very good count of conversion compared to other specializations. Working professionals show an excellent count of conversion whereas unemployed people have a higher count for being converted.
- People asking for Better Career Prospects show the highly positive response in conversion and People who didn't search about X Education courses have good chances for conversion
- People who haven't seen any ads through Recommendation as well as didn't demand a free copy of "Mastering the Interview" have a good count of conversion
- SMS sent followed by Email Opened have the highest conversion count compared to other activities

Model Building

Dummy variables are created for Categorical variables

Train Test dataset created with a train test ratio of 70-30.

RFE technique applied to select the top 15 significant variables

VIF applied on each step of model building to check for multi collinearity

At each iteration of model building, least significant variables was identified using p-value and removed from the model.

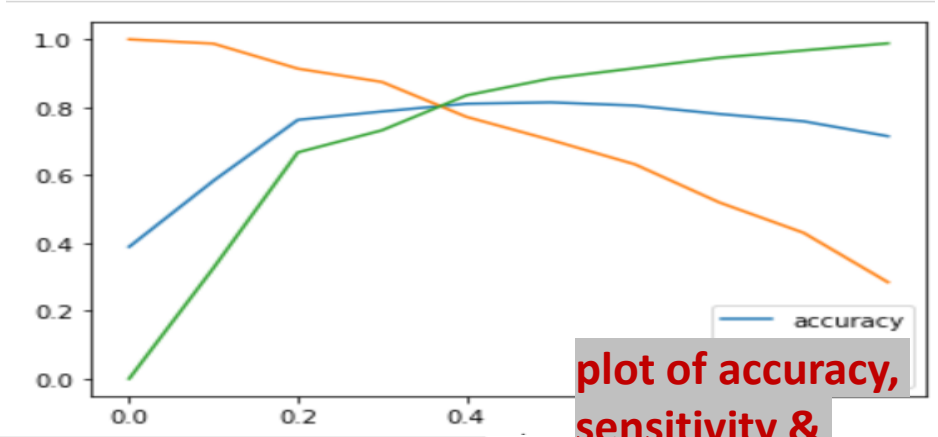
Final model has 12 dependent variables

Model Evaluation

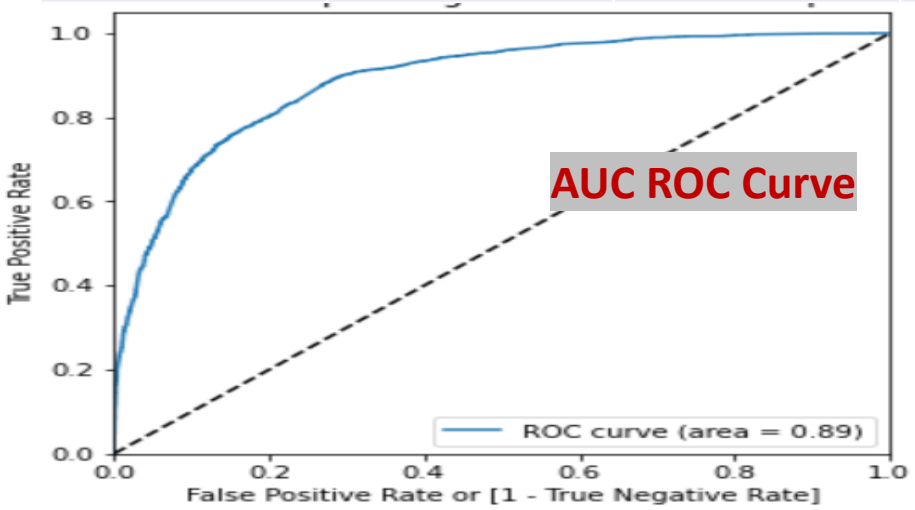
Model was evaluated using various metrics.

Accuracy using confusion matrix, Sensitivity, Specificity, Precision, Recall, AUC-ROC Curve

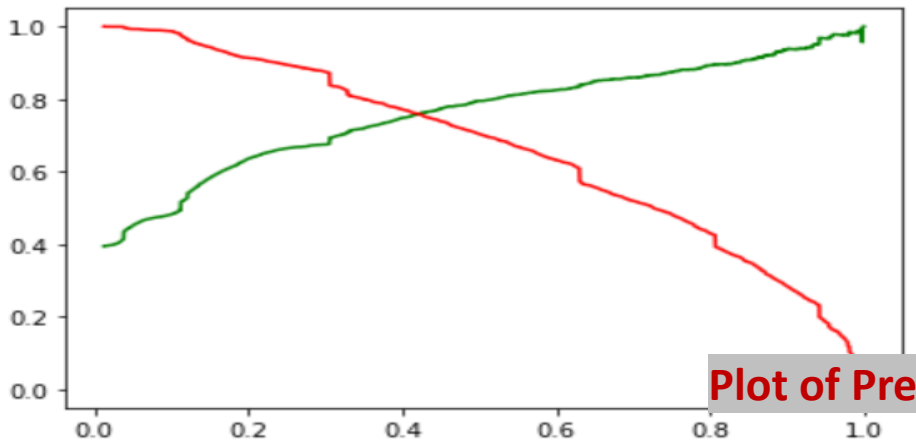
Data	Accuracy	Sensitivity	Specificity	AUC
Train data	81.03	77.17	83.49	.89
Test data	81.1	77.08	83.54	.88



plot of accuracy, sensitivity & specificity for different cutoff probabilities



AUC ROC Curve



Plot of Precision vs Recall

Business Insights

Variables that are important in Converting Potential Leads:

total time spent on the Website

Total number of visits

Lead source: Olark Chat

Last activity: SMS, Olark chat conversation

So, if a Potential lead has features with more time spent on the website, More number of visits, and Source is Olark Chat, then there is a high change that these leads could be converted into new Students.

Thank You