Lead Scoring Case Study

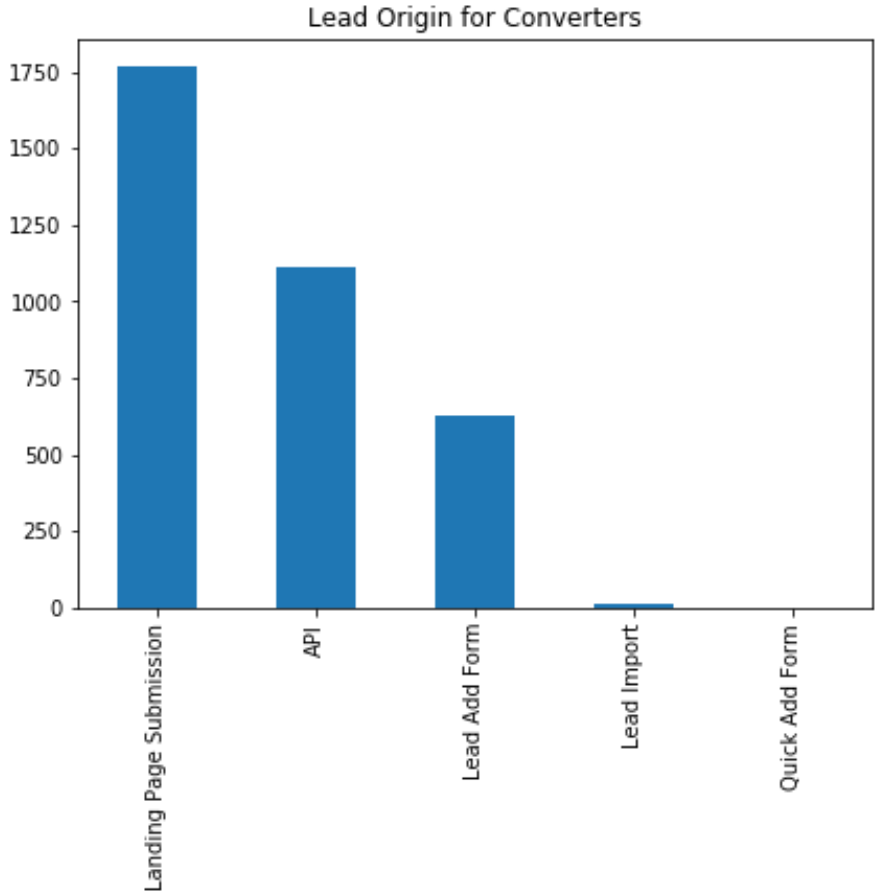
Problem Statement:

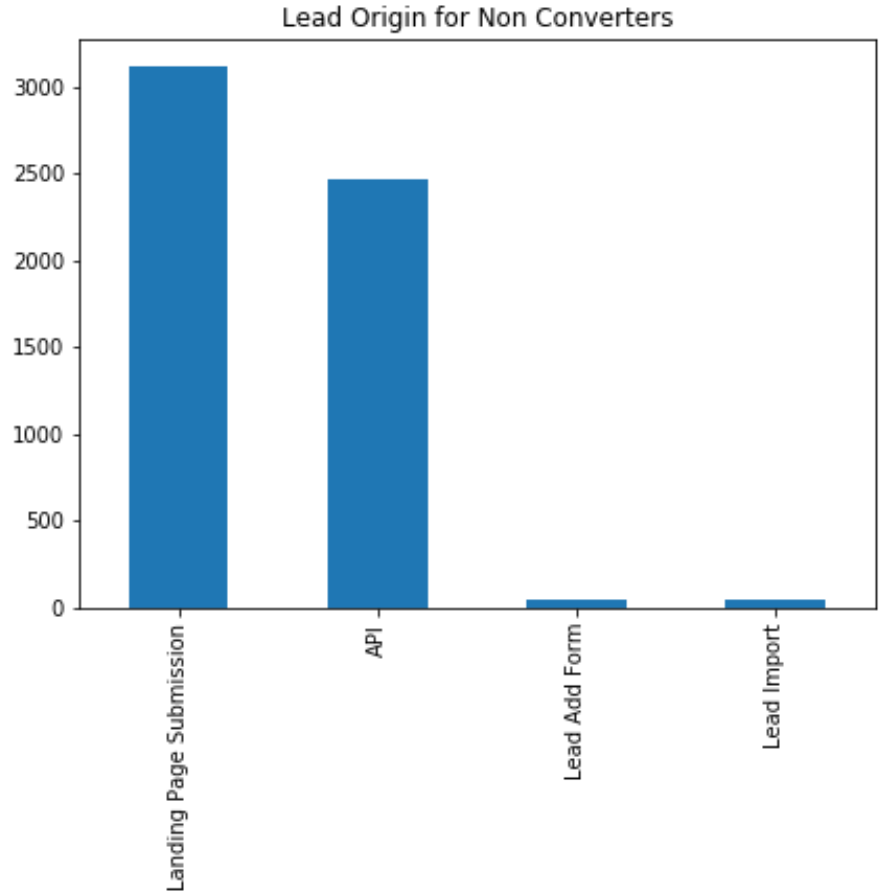
- X Education is an online education company that sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google.
- Once people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When 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. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- company requires a model which will assign a lead score to each of the leads. Customers with higher lead score have a higher conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Analysis Approach:

- Load dataset
  - Import the data from .csv file to a Data frame.
- Data Understanding
  - Check various parameters of Data frame.
- Data Cleaning
  - Treat “Select” and “null” values.
  - Drop unnecessary columns
  - Handle Outliers.
- Exploratory Data Analysis
  - Drop Prospect ID and Lead Number.
  - Perform Univariate, Bivariate analysis.

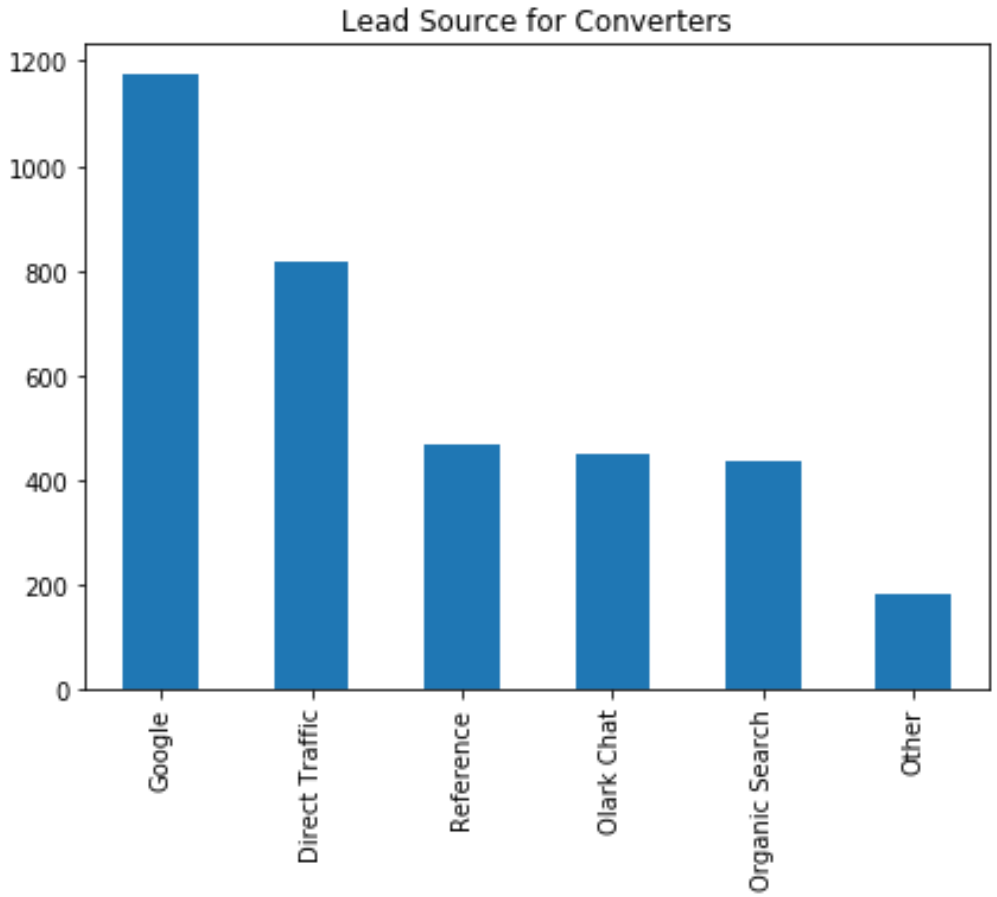
Bar plot of “Lead Origin”:

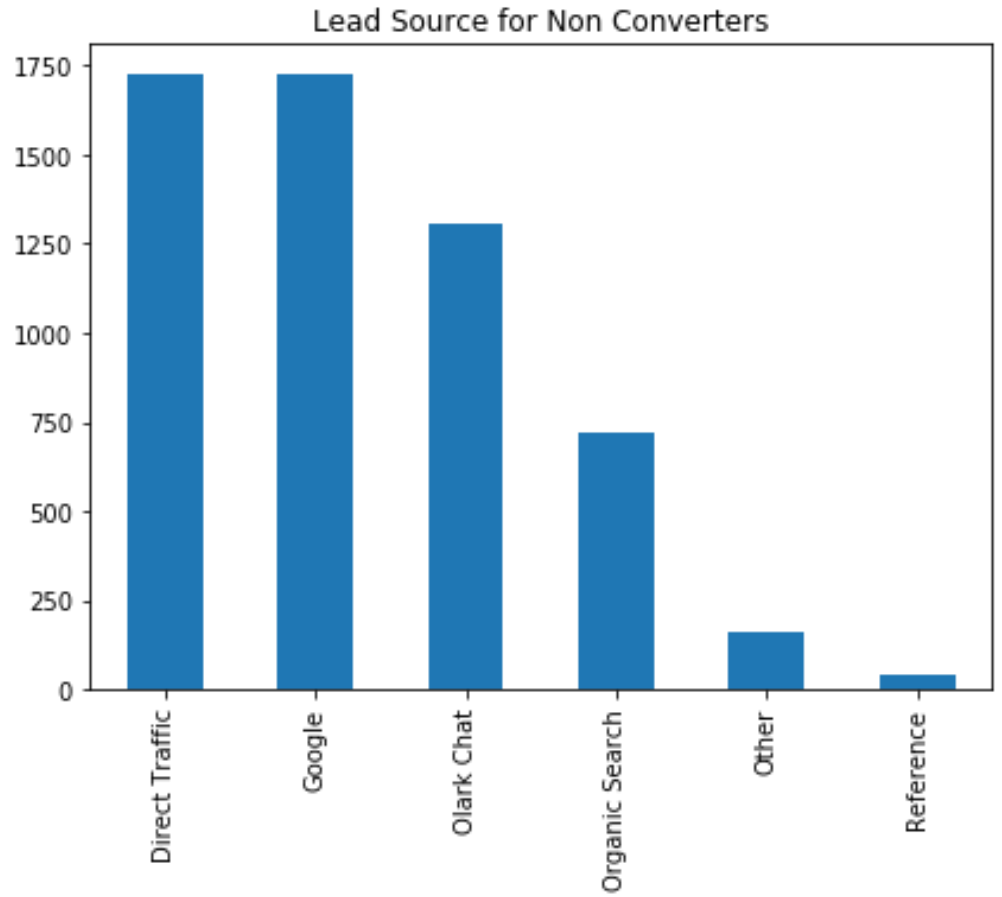




From above plot, origin is "Landing Page Submission" for most of the leads who have converted. The other take away is almost all the leads with origin "Lead Add Form" have converted and registered for online course. In addition to this, there is 100% coversion rate among the leads whose origin is "Quick Add Form".

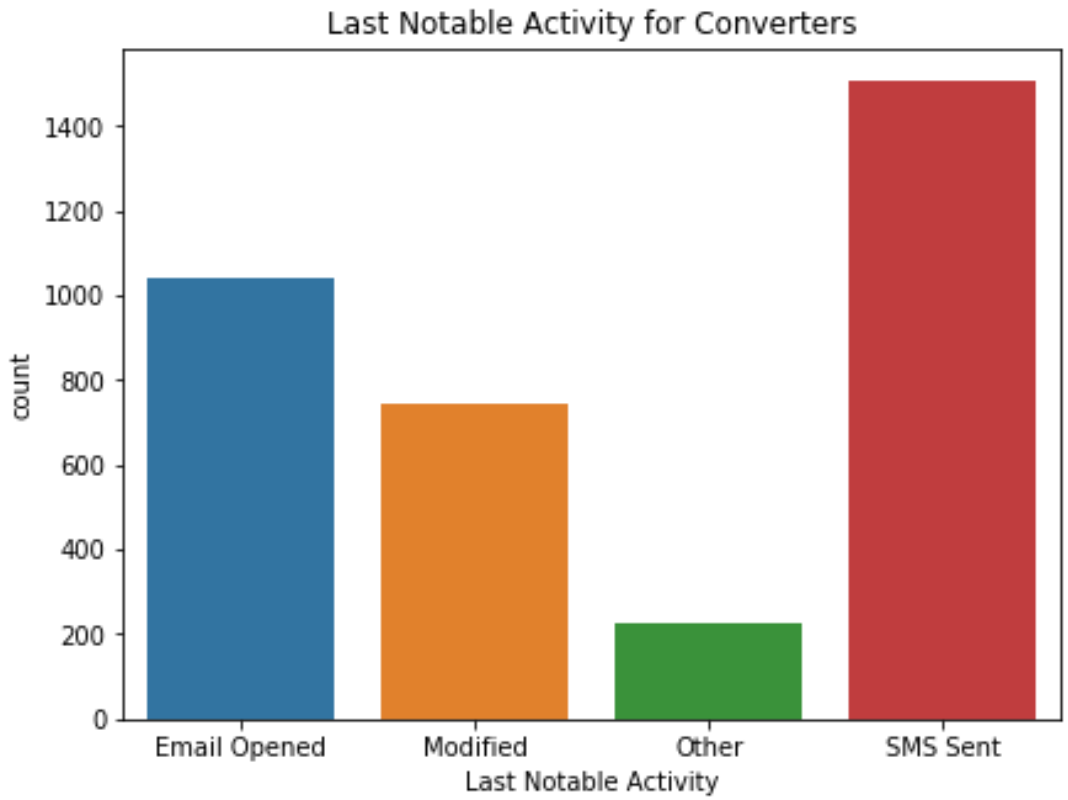
Bar plot of “Lead Source”:

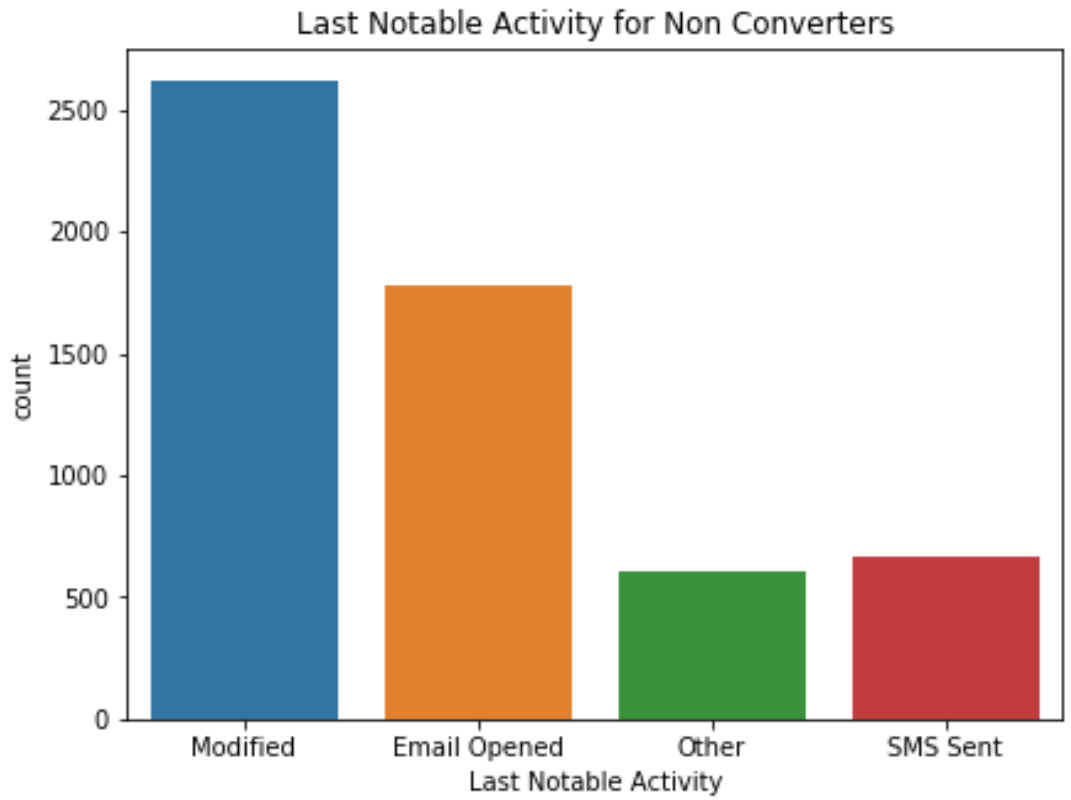




From above plot, Source is "Google" for most of the leads who have converted. The other take away is, almost all the leads with Source "Reference" have converted and registered for online course.

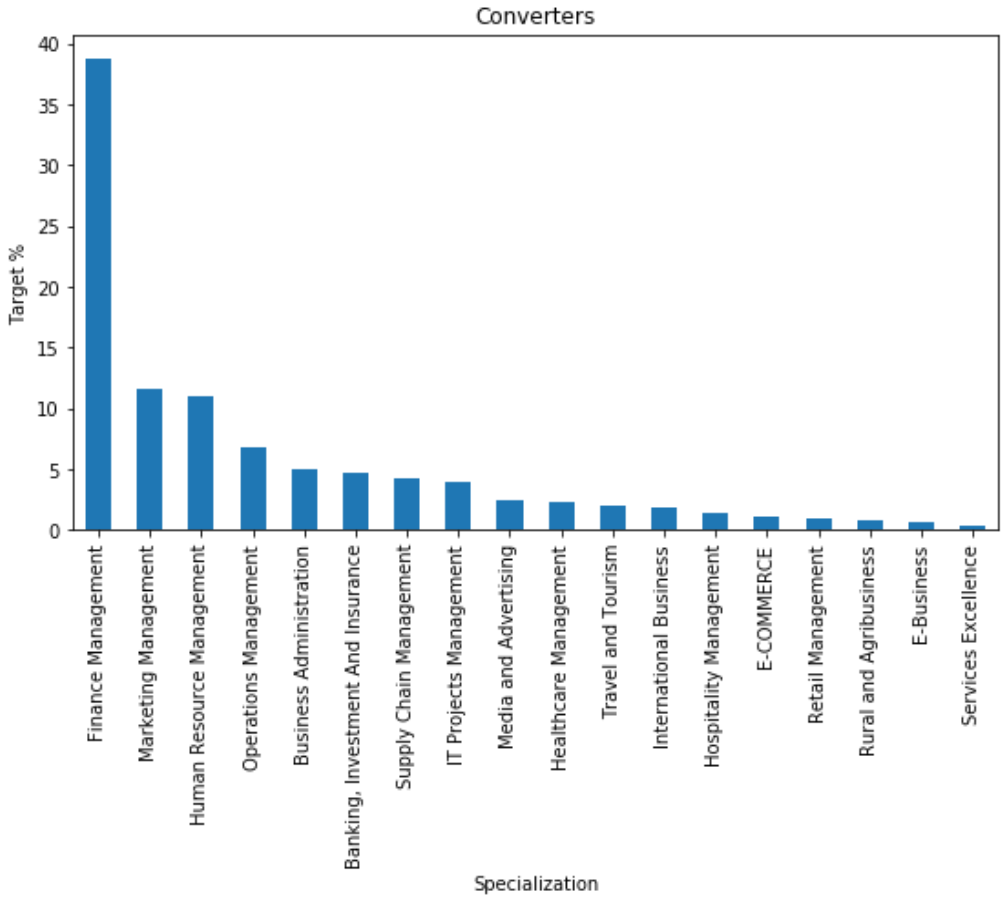
Count plot of “Last Notable Activity”:

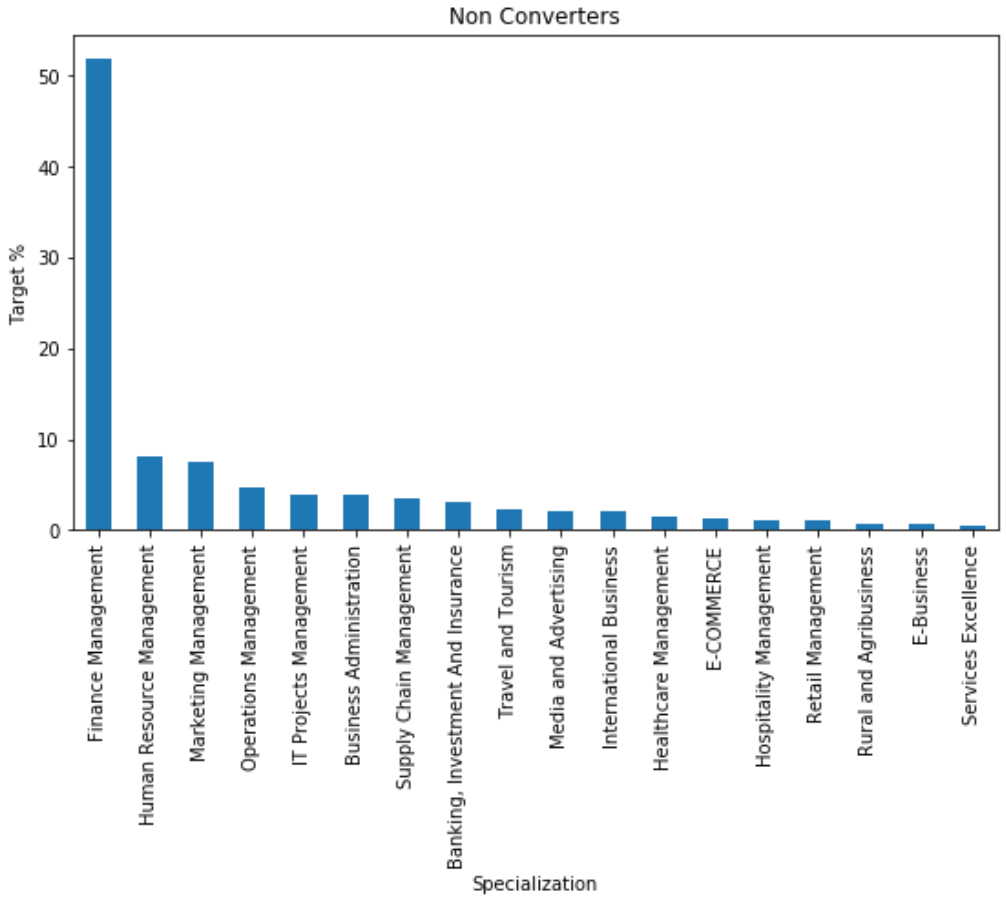




From above plot, it is clear that "SMS Sent" is the most significant last notable activity for leads who have converted and the same "SMS Sent" is least significant activity among the leads who have not converted.

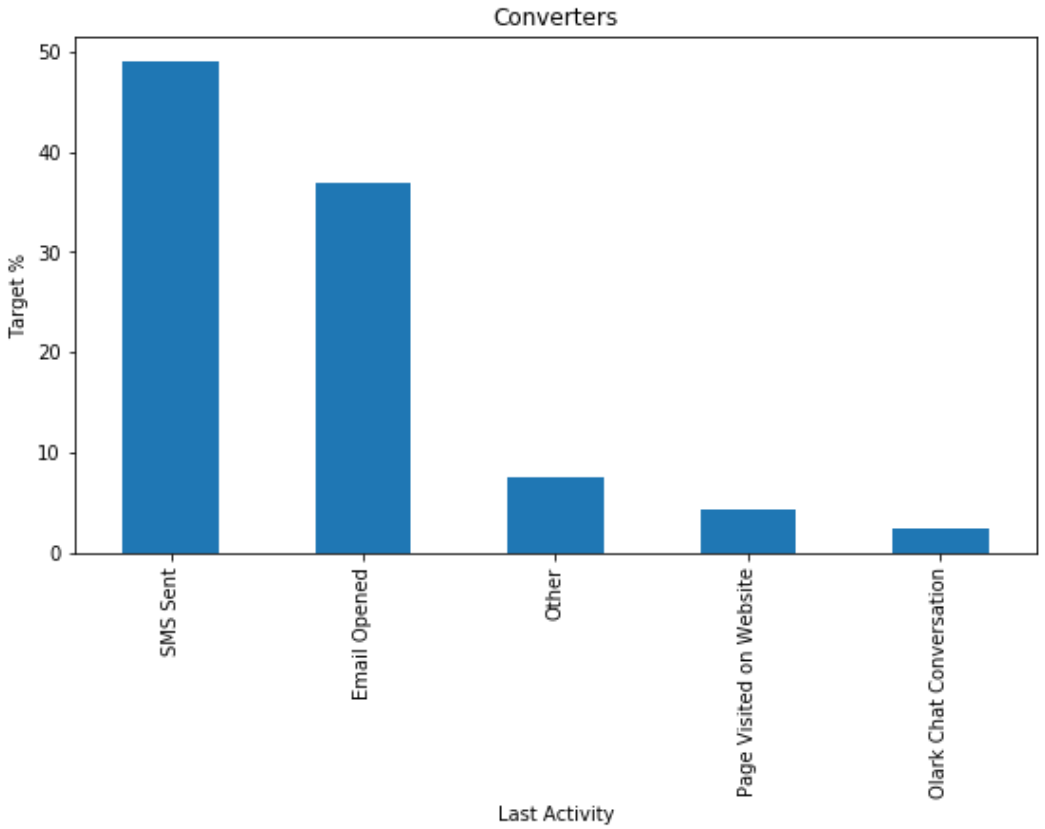
Bar Plot: Specialization vs Target %

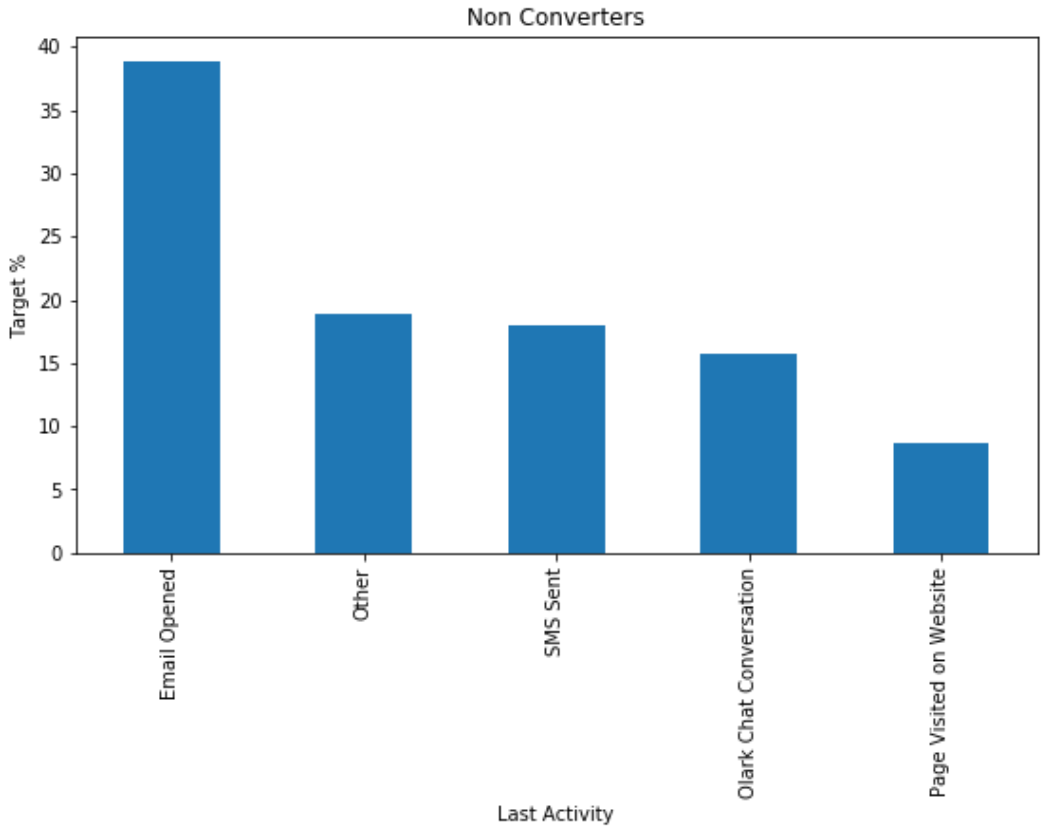




From above plot, there are high percentage of leads with "Finance Management" specialization which is almost triple than second best who have converted and registered for course.

Bar Plot: Last Activity vs Target %





From above plot, "SMS Sent" and "Email Opened" are the 2 most significant last activies among the converted leads where "SMS Sent" among Converted is almost triple then non converters.

Analysis Approach (cont):

- Data Preparation
  - Dummy variable creation, train-test split , Scaling variables.
- Building a Model
  - Use RFE to select 15 variables.

    ['Total Time Spent', 'Lead Origin_API', 'Lead Origin_Landing Page Submission',
    'Lead Origin_Lead Add Form','Lead Origin_Lead Import', 'Lead Source_Olark Chat',
    'Lead Source_Reference', 'Last Activity_Email Opened', 'Last Activity_Olark Chat Conversation',
    'Last Activity_SMS Sent', 'Specialization_Finance Management',
    'Specialization_Hospitality Management', 'Specialization_Retail Management',
    'Last Notable Activity_Modified', 'Last Notable Activity_SMS Sent']

    are the variables used in building the model.
  - Drop insignificant variables one at a time using p-value and VIF.

    [Lead Origin_API, Lead Origin_Lead Import, Lead Source_Reference,
    Specialization_Retail Management, Last Notable Activity_SMS Sent]

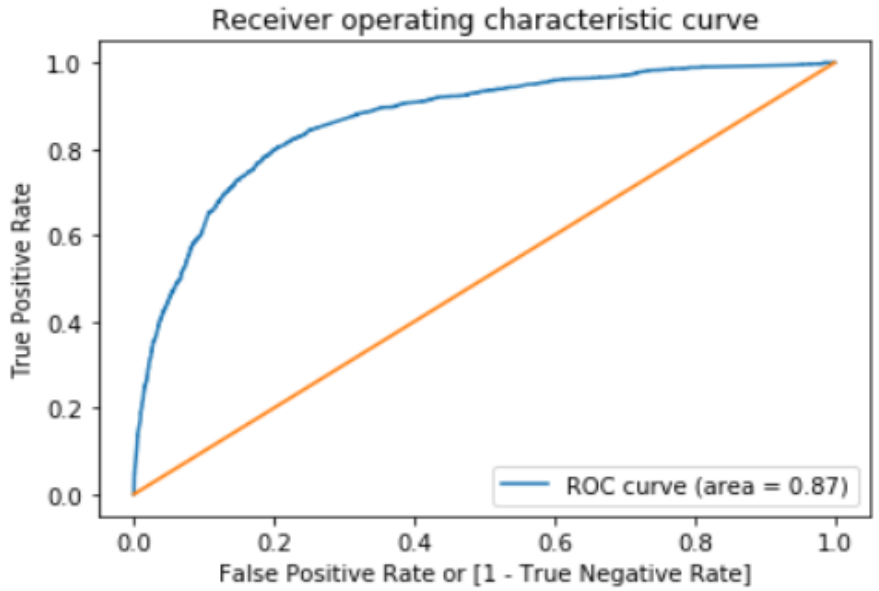
    are the variables which were dropped due to their insignificance.

Analysis Approach (cont):

- Model Evaluation
  - Evaluate model on various parameters like Accuracy, Sensitivity, Specificity etc.
  - Below are performance parameters calculated with an arbitrary probability cut off point 0.5.
    - Accuracy - 0.802
    - Sensitivity - 0.67
    - Specificity - 0.87
    - Positive Predictive Rate - 0.77
  - Even though accuracy is 0.802, model identified only 67 percent of converted data correctly.

Analysis Approach (cont):

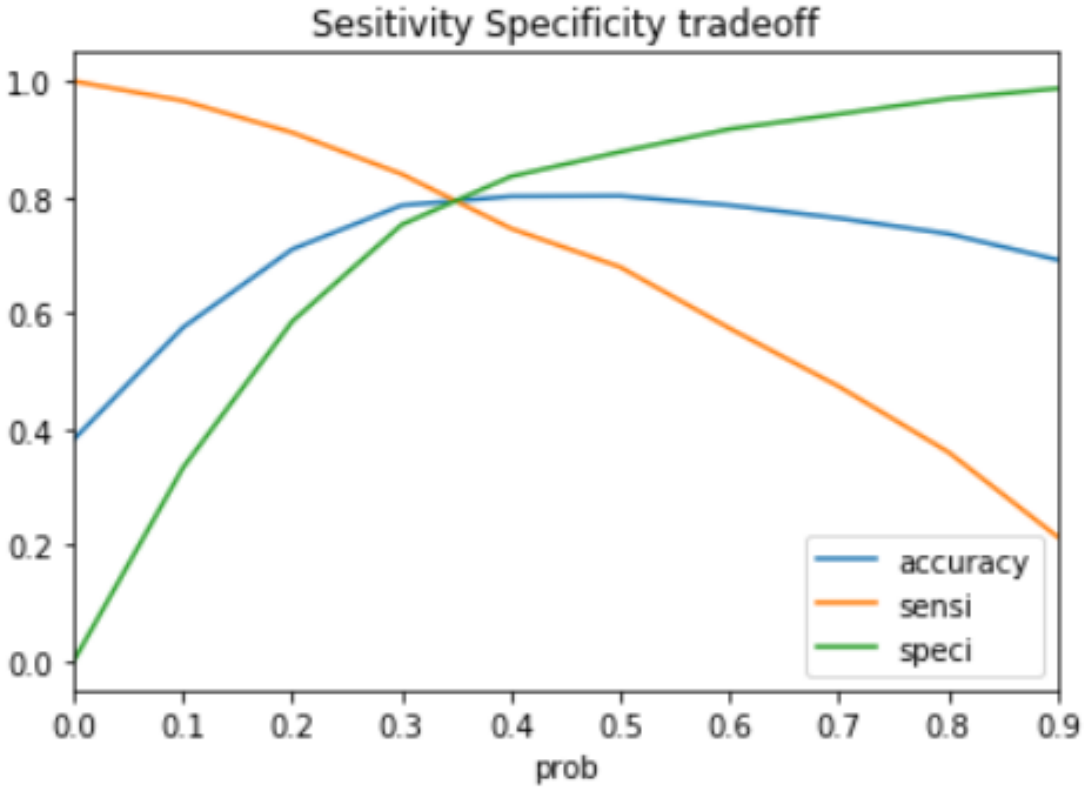
- ROC Curve
  - Plot ROC curve to verify weather our final model is a random model or not.



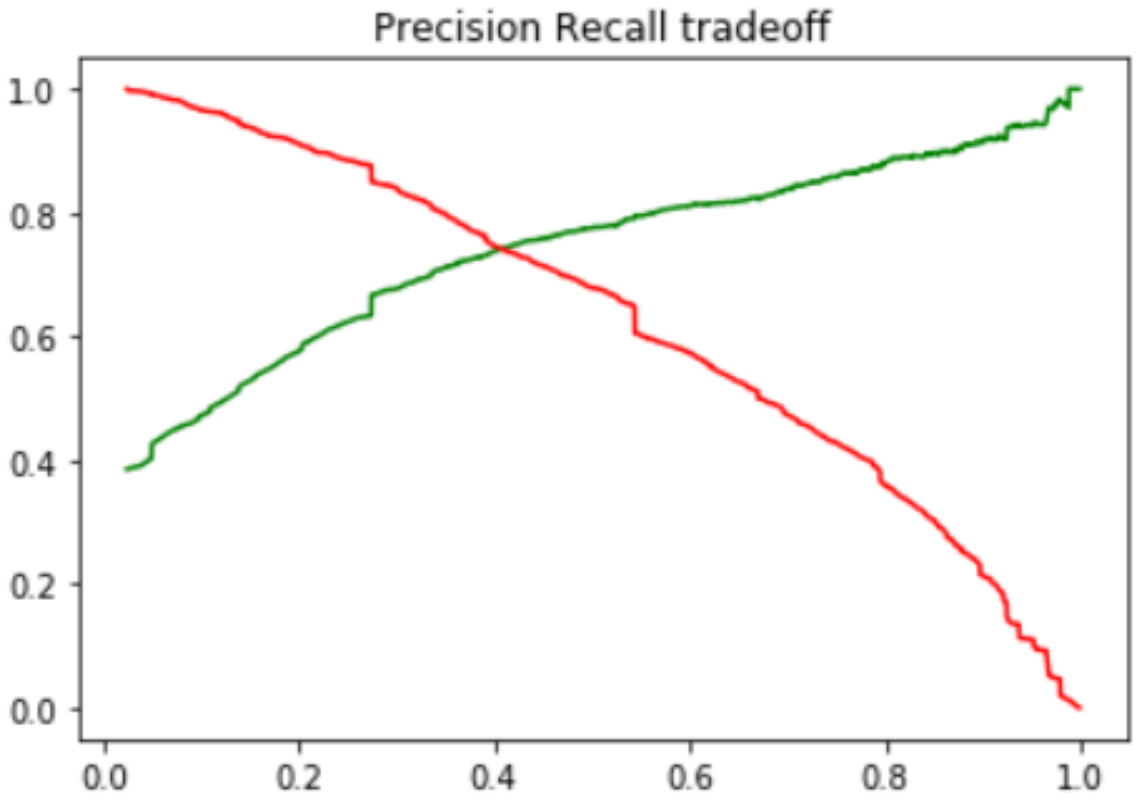
  - We have good amount of area under the curve. So, our model is not a random model.

Analysis Approach (cont):

- Finding Optimal cut off
  - Plot Accuracy, Sensitivity, Specificity for various probabilities to find optimal cut off.
  - From “Sesitivity Specificity tradeoff”, we can say that optimal cutoff prob is at around 0.35.



- From “Precision Recall tradeoff”, we can say that optimal cut off prob is at around 0.40.



Analysis Approach (cont):

- Model Re-evaluation
  - Predicted the final predication value with cut-off 0.35 probability at which accuracy, sensitivity and Specificity curves are intersecting.
  - Below are performance parameters calculated with an optimal probability cut off of 0.35.
    - Accuracy - 0.8
    - Sensitivity - 0.8
    - Specificity - 0.8
    - Precision – 0.71

Analysis Approach (cont):

- Validate the model
  - Verify the model on test dataset.
  - With the final model and with the cut-off value 0.35, predicted converted hot leads in the test data with below performance parameters
    - Accuracy - 79%
    - Sensitivity - 79%
    - Specificity - 79%.
- Target Leads
  - Lead Numbers whose probability is >= 0.35 are the set of leads the institution should target in order to get a conversion rate of 80%