CREDIT ASSIGNMENT
LEAD SCORING CASE STUDY

Group Members:

1. Dipanwita Kanchan Kumar Biswas
2. Saumalya Ghosh

Problem Statement

- X Education is an education company that sells online courses to industry professionals.
- The Company gets a lot of leads, but 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.

Business Objective

- The Company want to know the Promising Leads
- We need to create a model which identifies the hot leads
- The model should work with future data as well

Approach for the Solution - EDA

- Read and understand the data
- Check for the missing values
  - *Deleting the data which has more than 40% missing values*
  - *If the NA values are most frequently used, then keep the value as is so that it does not hamper the calculation*
  - *Impute if*
    - Continuous data
      - *Has outlier- median*
      - *Does not have outlier- mean*
    - Categorical Data- Mode
- Delete the Unwanted / non relevant data
- Check for outliers
- Check if there is any data imbalance
- Perform Univariate/Bivariate analysis
- Get the data ready

Approach for solution- Model Building

- Create Dummies of categorical variable, divide the data to train test dataset
- Perform Feature scaling to bring all of the data to the similar scale
- Perform the Feature selection- RFE or manual method of elimination
- Create model using the train data
  - *Check the summary and identify the columns which have pvalue >.05*
  - *Remove 1 column*
  - *Repeat the steps until all p values <0.05*
  - *Check the multicollinearity of the data using the Variance Inflation factor(VIF) VIF<5*
- Check the accuracy of the final model
  - *Take a threshold, calculate the Accuracy, Confusion Matrix, Precision, Recall, ROC_AUC score, and the ROC curve*
  - *We can perform this on multiple threshold values*
  - *Finalize the threshold which has the minimum difference between the precision and the recall (precision, recall tradeoff)*
- Predict the model in the test data , use the threshold decided to calculate the final observation
- Perform Accuracy, Confusion Matrix, Precision, Recall, ROC_AUC score, and the ROC curve for the test data as well

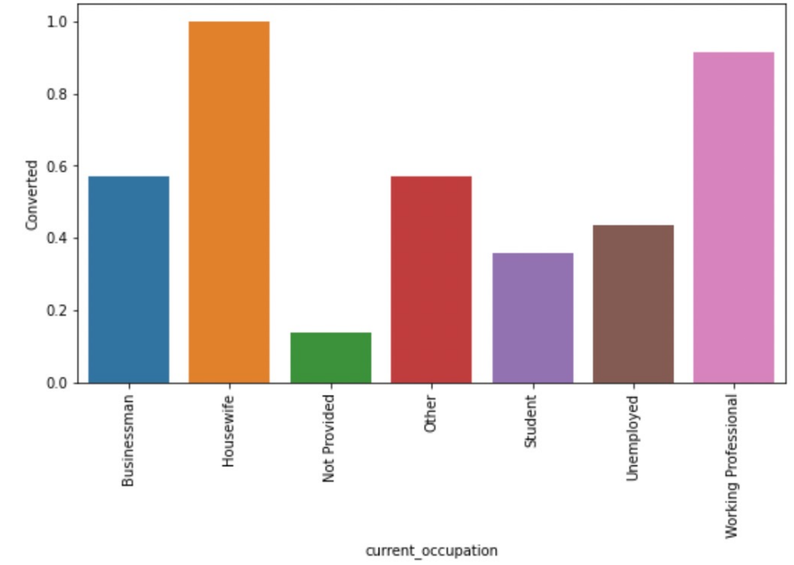
EDA

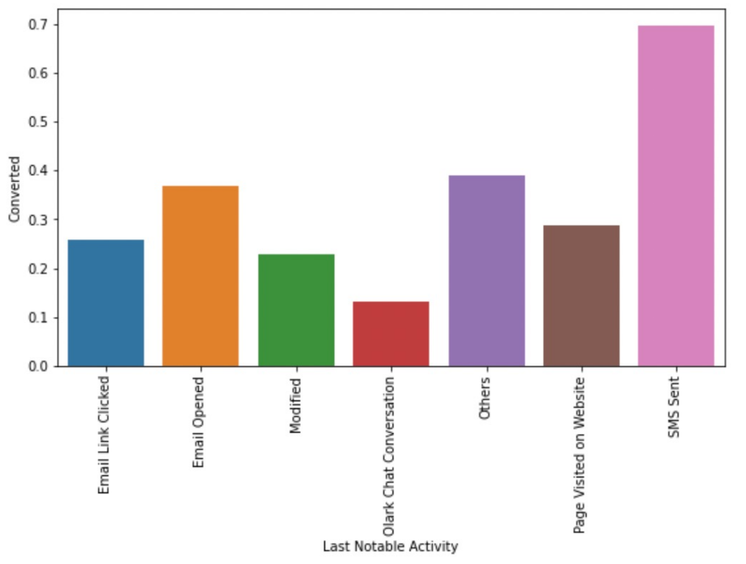
- Based on the check on the data a few columns were highly skewed ie 1 category had more than 95% of the data value
  - *Do Not Call*
  - *Country*
  - *What matters most to you in choosing a course*
  - *Search*
  - *Magazine*
  - *Newspaper Article*
  - *X Education Forums*
  - *Newspaper*
  - *Digital Advertisement*
  - *Through Recommendations*
  - *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*

Univariate analysis

- People who have SMS Sent, opened email have high conversion rate

- Occupation wise, Housewife, Working Professional, Businessmen have high conversion rate

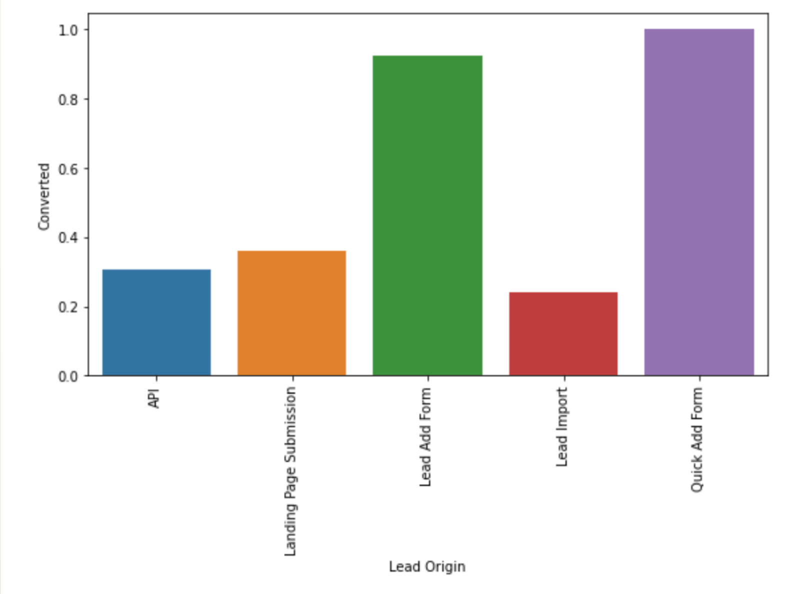


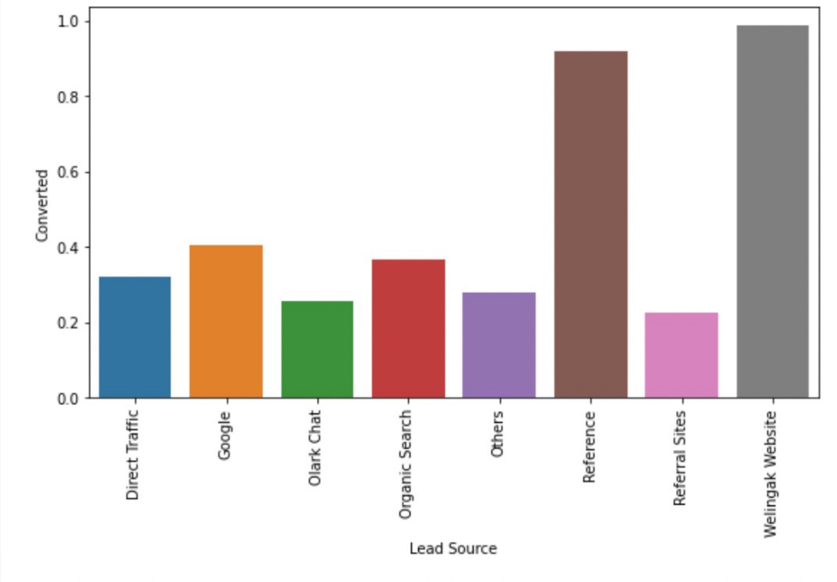


Univariate analysis

- People who filled the Lead add form, Quick add form have a very high conversion

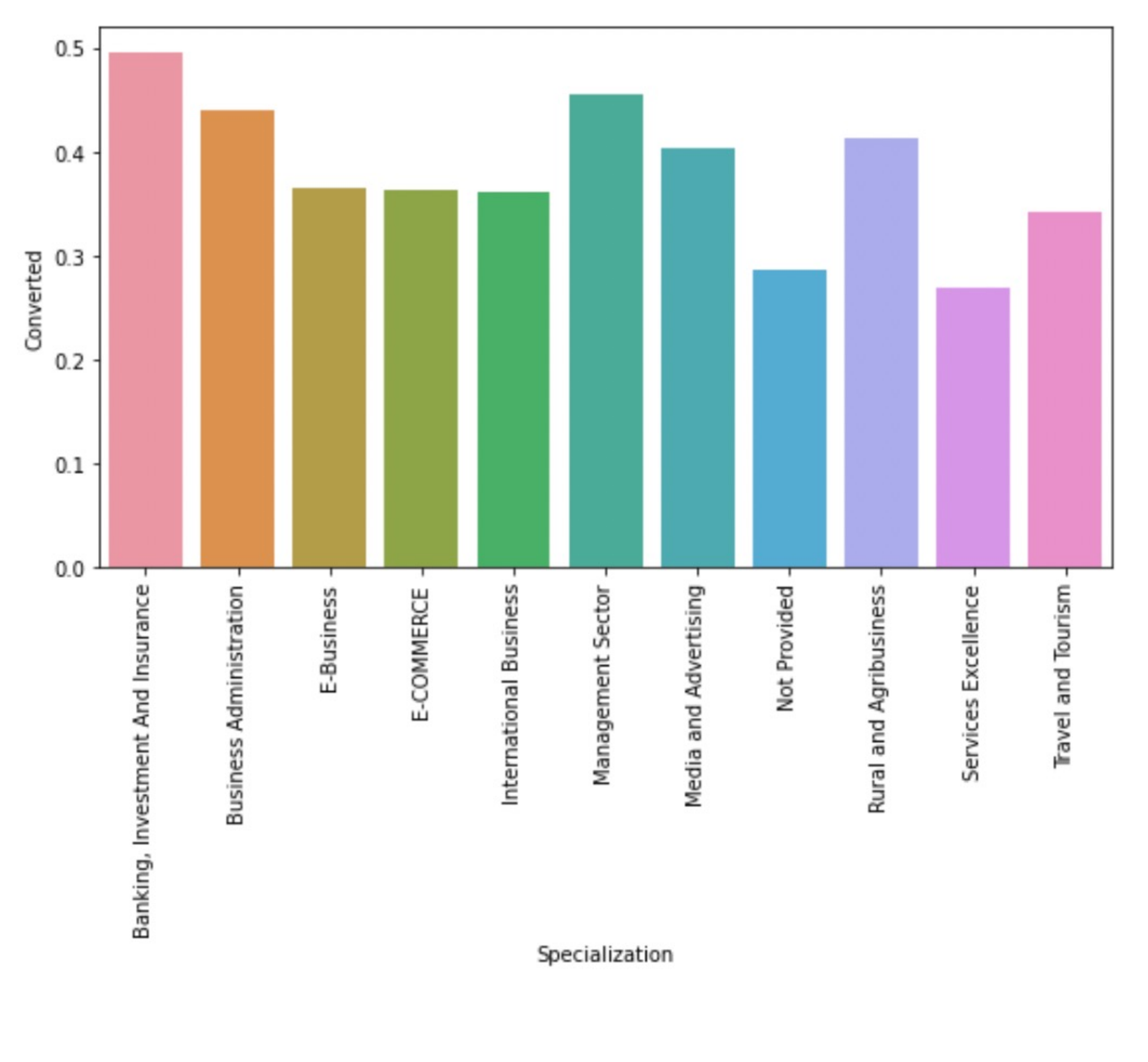
- Leads from reference, Welingak, google, organic Search have a high conversion rate





Univariate Analysis

- People under Banking sector or management specialization are more inclined to get converted

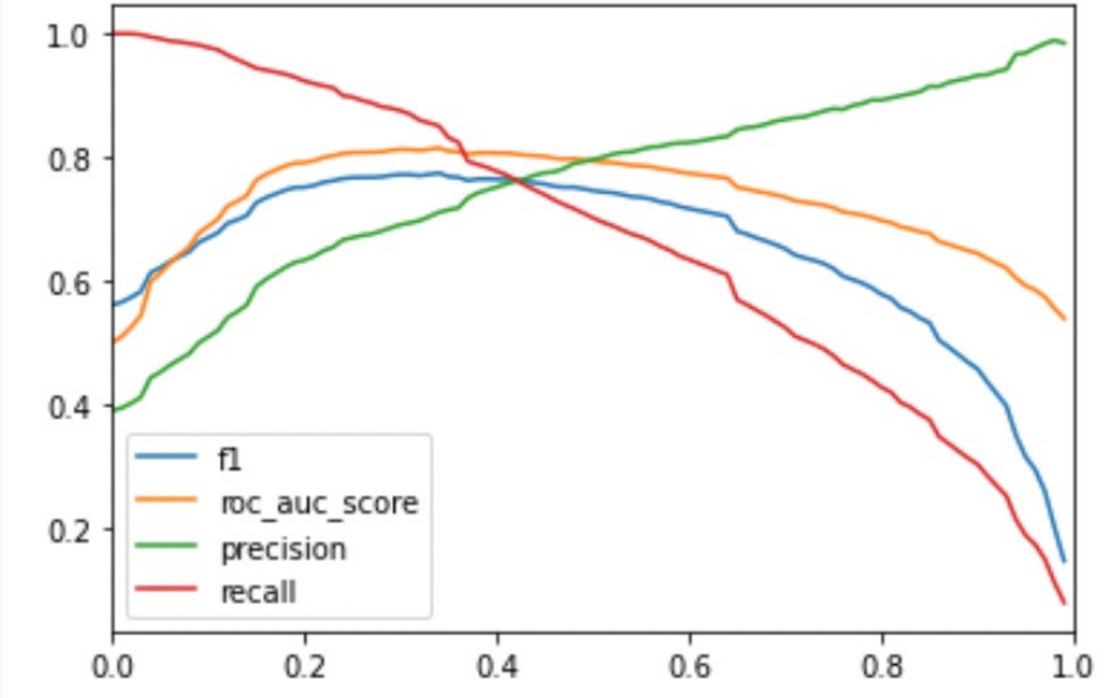


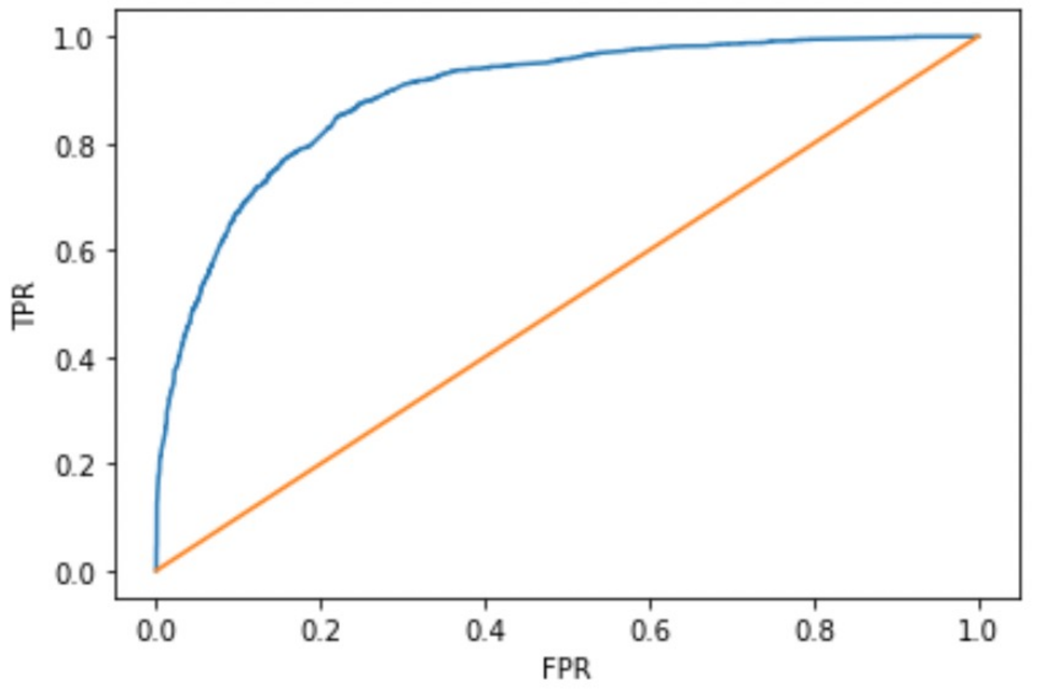
FINAL LIST IDENTIFIED BY THE MODEL

Current Occupation	Unemployeed
	Working Professional
	Student
	Others
Last Activity	SMS Sent
	Olark Chat Conversation
Last Notable Activity	Email Opened
	Others
Lead Origin	Lead Add Form
Lead Source	Olark Chat
	Welingak Website
Specialization	Management Sector
Total Time Spent	
Do Not Email	
Freebee Received	

Model Accuracy

Based on the roc_auc_score threshold will consider the threshold to be at 0.34



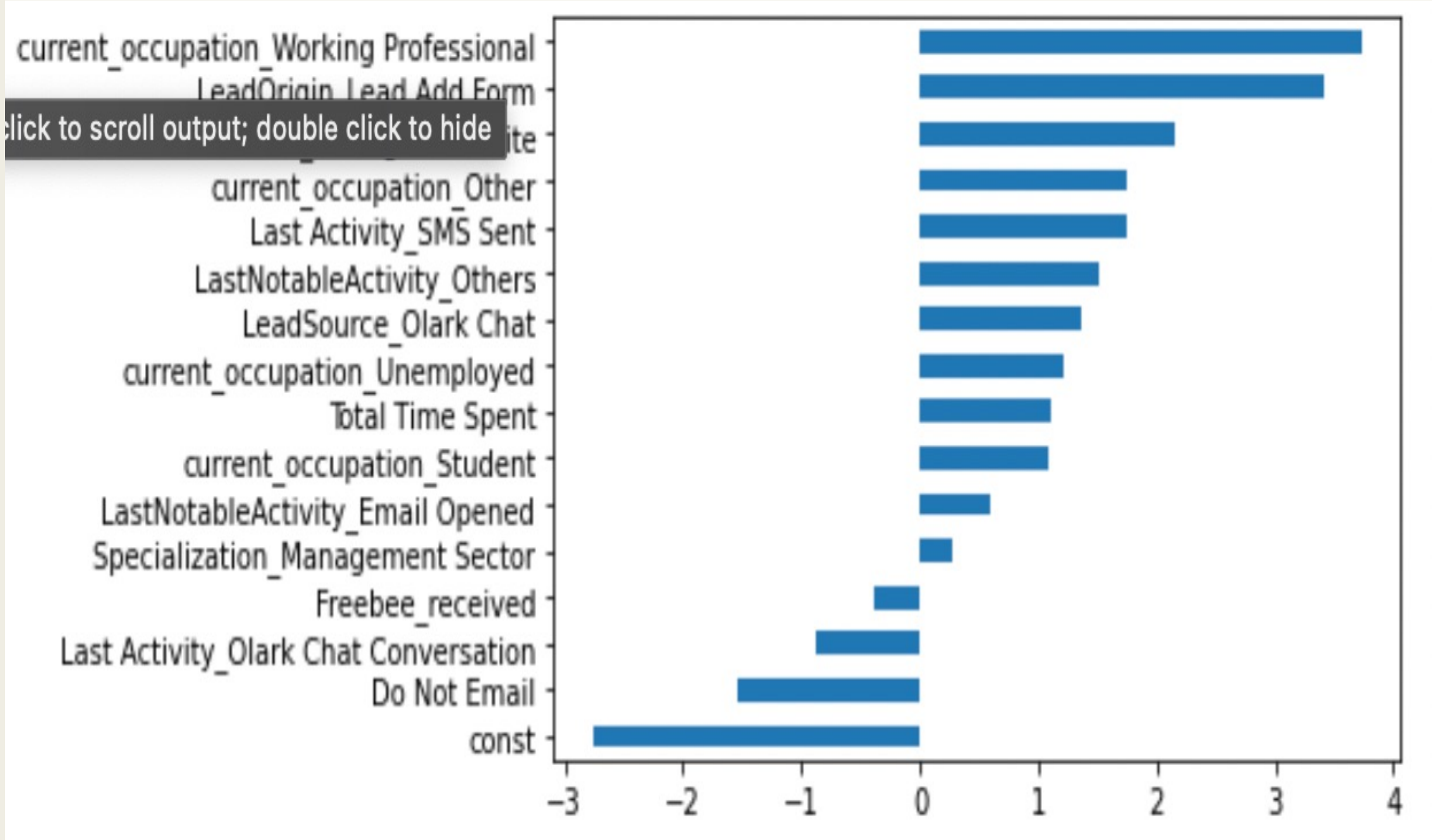


Final Accuracy for test and train data

- Train Accuracy
  - *Accuracy- 80.69%*
  - *F1 score= 77.44%*
  - *Precision- 71.07%*
  - *Recall- 85.05%*
  - *ROC_AUC_Score- 81.48%*

- Test Accuracy
  - *Accuracy- 80.39%*
  - *F1 score= 76.26%*
  - *Precision- 69.80%*
  - *Recall- 84.03%*
  - *ROC_AUC_Score- 81.28%*

- There is very less difference between the accuracy scores between the Train data and the test data.
- ROC Scores are also about 80% so we can consider this as a good model



Conclusion

So the mentioned columns matter the most while identifying the potential business

- Current Occupation
- Last Activity
- Last Notable Activity
- Lead Origin
- Lead Source
- Specialization
- Total Time Spent
- Do Not Email
- Freebee Received