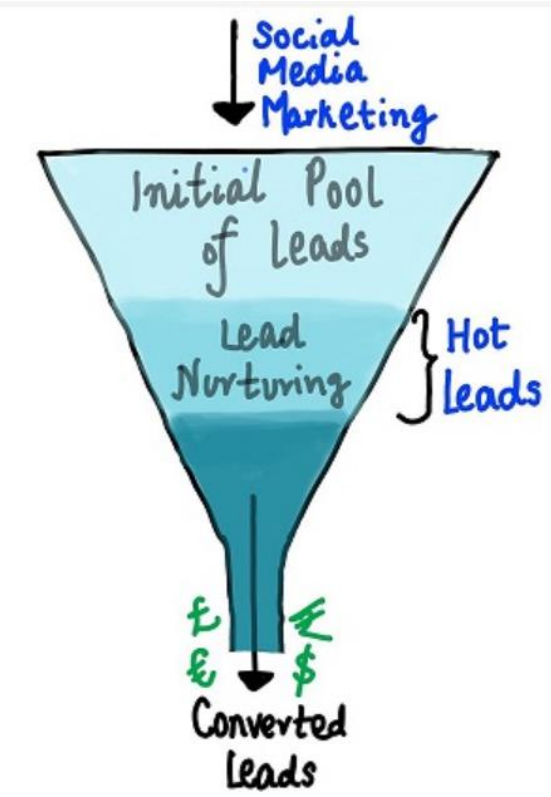
Lead Scoring Case Study

IDENTIFYING ‘HOT LEADS’

By: Rishik Patel & Vandit Sadaphale

Problem Statement

- ‘X Education’ sells online courses to industry professionals and markets it through several websites and search engines.
- A customer filling a form for a course is termed as a ‘Lead’, to which the sales team further try to convert.
- As explained in diagram, out of a large initial pool of leads, only a few become Hot Leads that are further nurtured to get Converted Leads.
- Existing conversion rate at X Education is around 30%
- In order to improve the efficiency, we are required to identify the ‘Hot Leads’ i.e., the potential leads having higher probability of conversion, so that business can straight away invest and focus only on these customers.
- The CEO has given a ballpark of the target lead conversion rate to be around 80%.



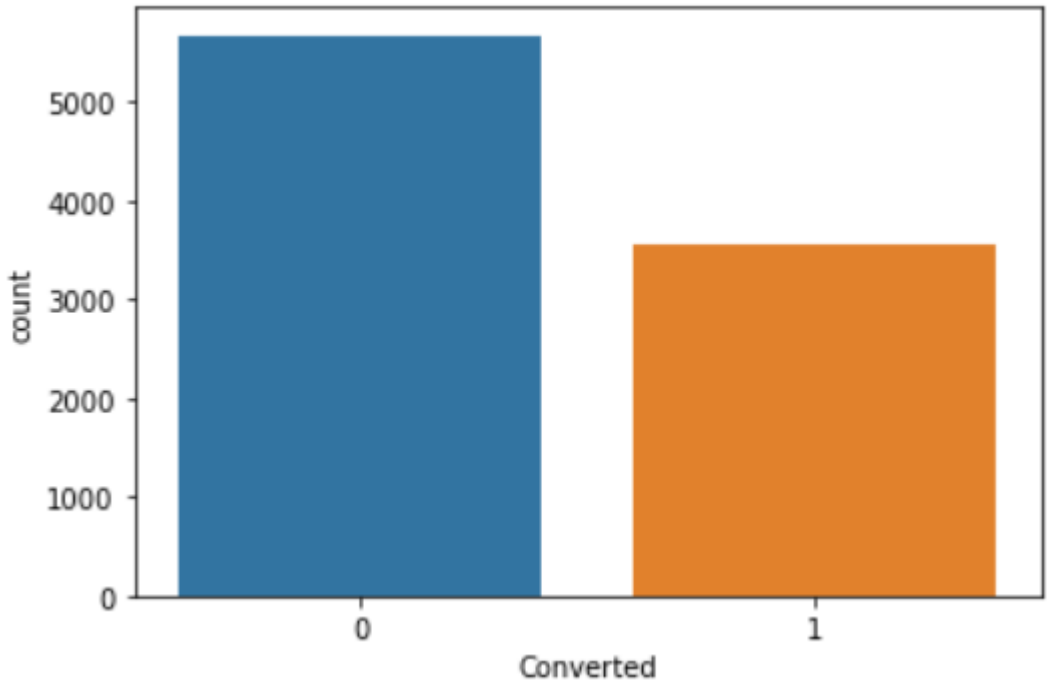
Analysis Approach

- We started with performing thorough EDA to verify nature and relationship of the available data, fixed outliers & derived business & statistical inferences.
- Structured the data well by treating the missing values.
- Prepared data for modelling with binary encoding (for Yes/No variables) & OHE (One Hot Encoding for categorical variables)
- We chose mixed approach i.e., RFE (Recursive Feature elimination) and Manual to select efficient features based on significant p-value, VIF's and optimal cut off point to build suitable Logistic Regression Model.
- We evaluated the achieved model on various metrics like Accuracy, Sensitivity, Specificity etc

EDA

Conversion Rate

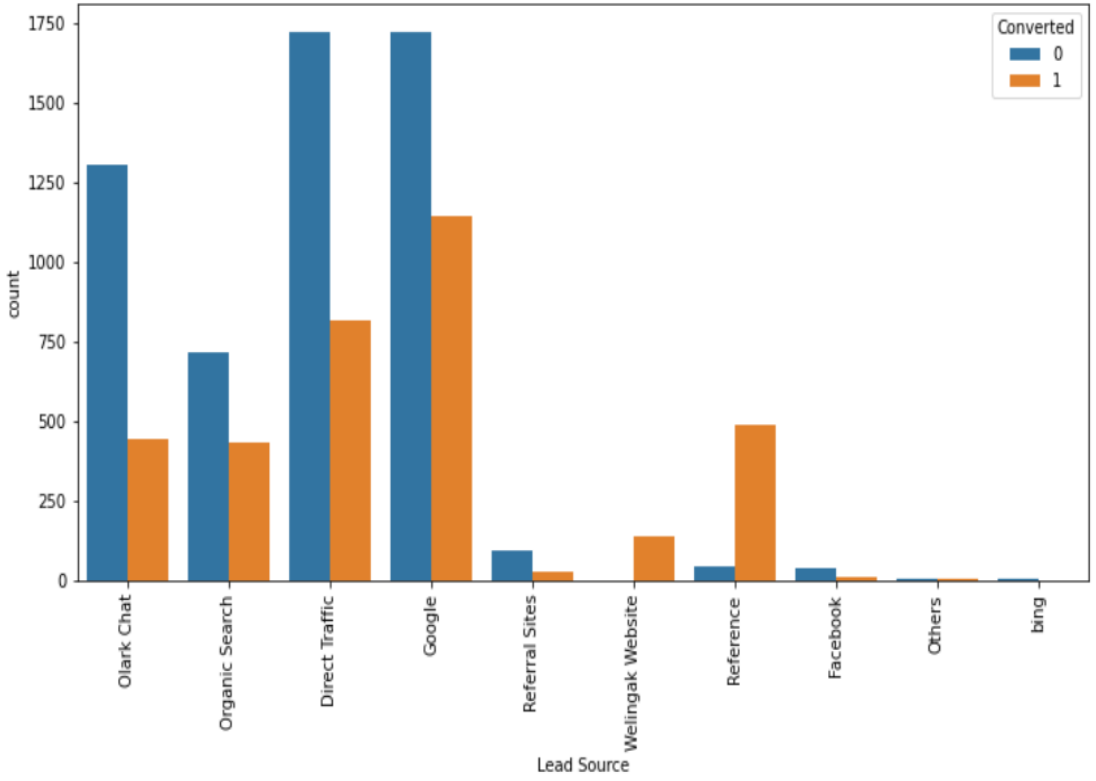
- One third conversion observed on the available data.
- Out of 9240 customers, 3561 (38.5%) customers finally converts while the rest churned.
- Highlighting the stated conversion rate of 30% and needed scope of improvement.



EDA

Lead Source

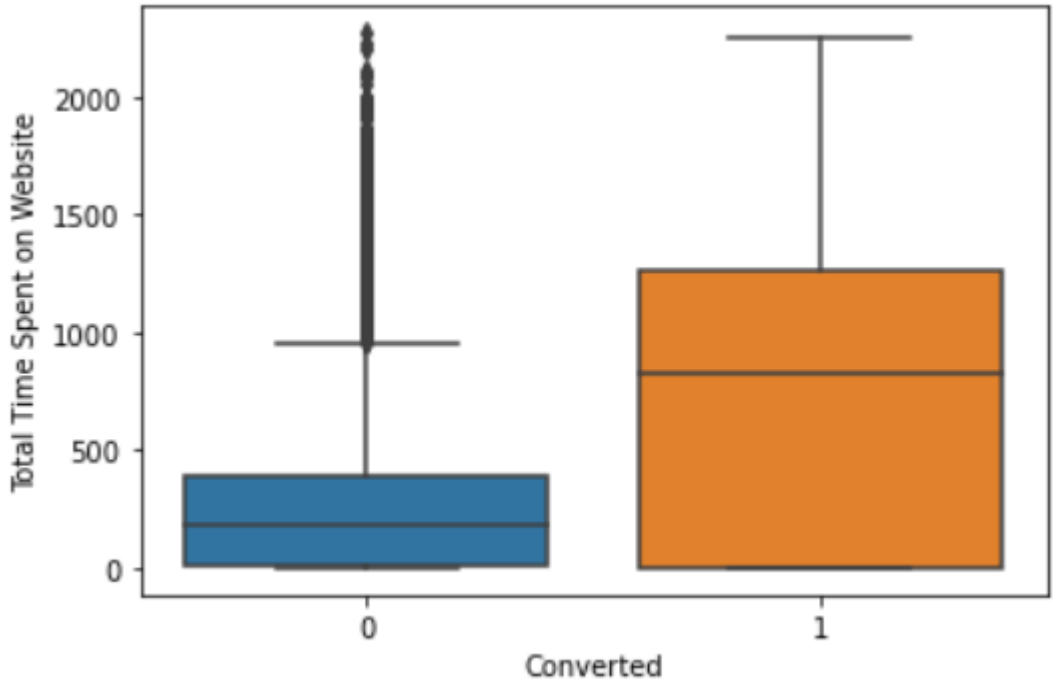
- Leads from 'Reference' & 'Welingak Website' are highly convertible.
- Leads from 'Organic Search', 'Google' also show around one third conversion rate.
- Leads from 'Olark Chat', 'Direct Traffic' are OK and should be focused more on for enhancement.
- Leads from other fields don't show much response and might be dropped off.



EDA

Total Time Spent on Website

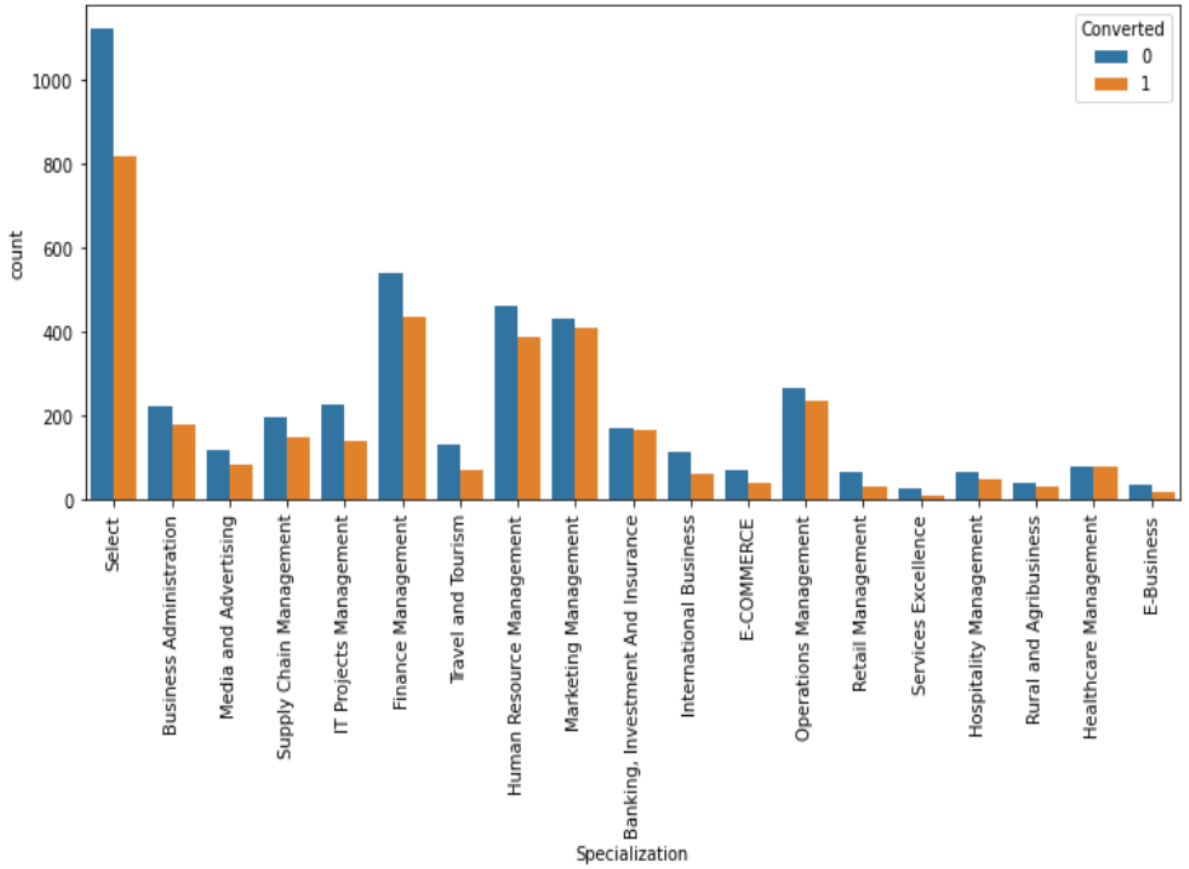
- Converted customers spend more time (median ~900) on Website compared to non converting ones (median ~250).
- More useful and perhaps customized content should be catered to gain catch customers attention which may lead to increased conversions.



EDA

Specialization

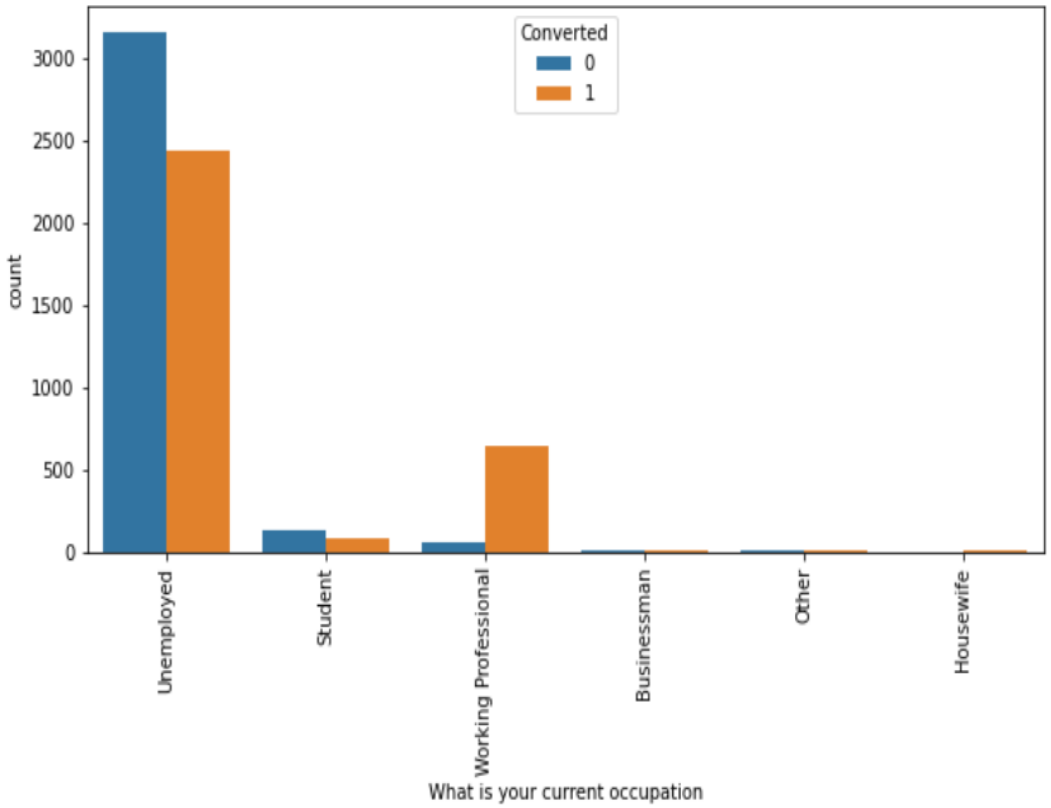
- 'Select' column is equivalent to null value. Thus, this is treated with appropriately during data preparation.
- Most conversions are from Unidentified sectors (null value) and recognized sectors like Healthcare Mgmt., Banking, Insurance & Investment, Marketing Mgmt.
- Almost all sectors have shown reasonable conversion rate and should thus be equally promoted well.



EDA

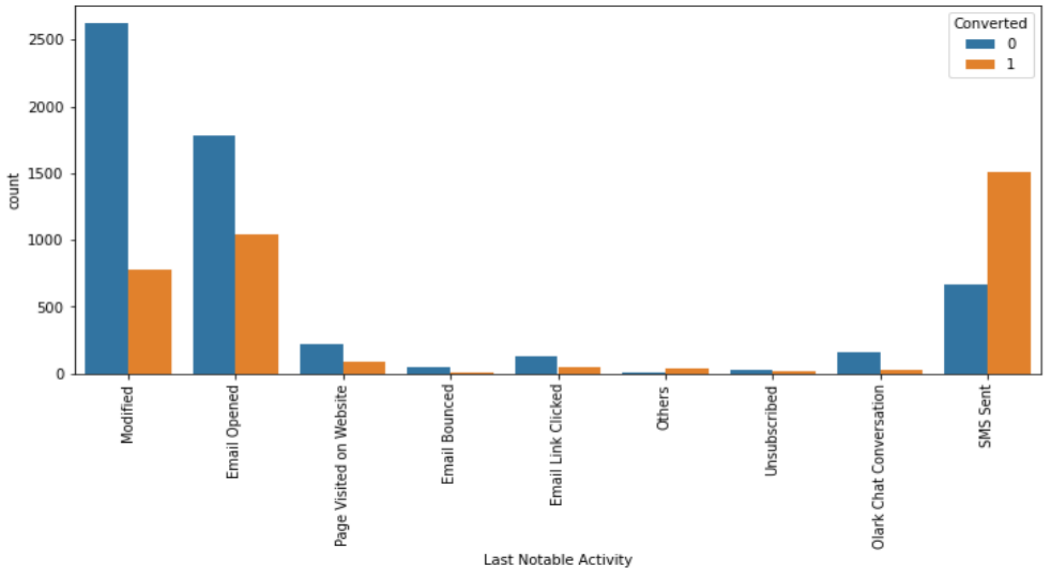
What is your current Occupation

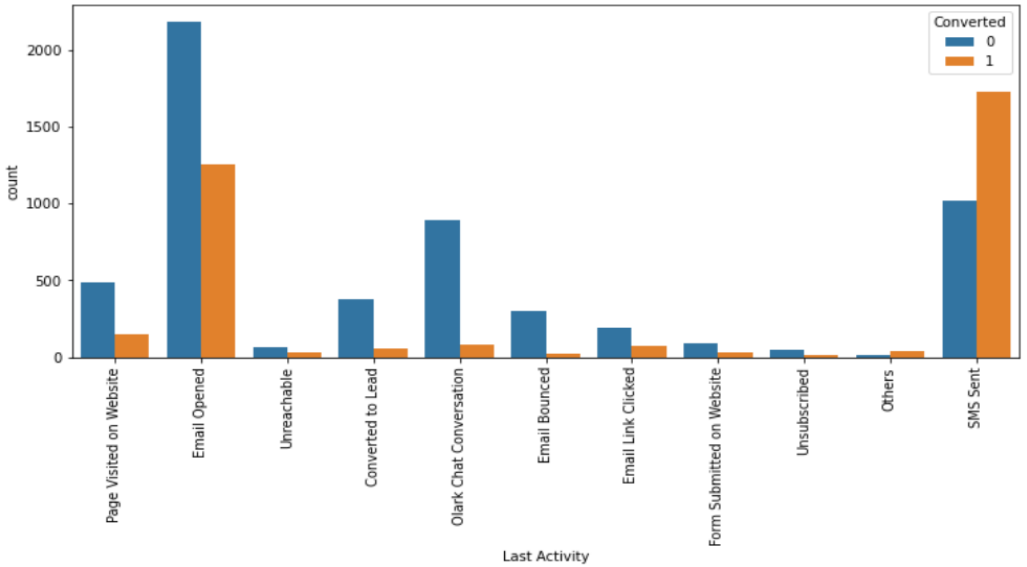
- Working Professionals are highly likely to convert as they seek to sharpen their skill set with evolving industry.
- Unemployed personals are also highly prone to conversion as they desperately look for job opportunity.
- Students also show fair conversion ratio how ever frequency is less, perhaps because unassociated content. Relevant content can be created.



EDA

Last Notable Activity & Last Activity





- Having sent an SMS at last moment (as last resort) had often turned out customer being converted (SMS' are being readily accessible and it shows their eagerness)
- Mostly customers where dropped an email which fairly received good conversion rate.
- In 'Others', specifically 'Had a Phone Conversation' shows fair converted customers, however it's not attempted much by sales team. It should be focused more on.

EDA, Data Cleaning & Preparation

Overall Summary

- EDA Conclusion:
  - Dropped 14/37 zero variance features.
  - Clustered insignificant labels under a single label ('Others') within some features.
  - Derived some proportional and relevant inference to be emphasized upon

- Data Cleaning: (retained ~70% of data at the end having 13 features)
  - There were 15 columns with missing values ranging from 0.389% to 78.46%.
  - Dropped 3 features having more than 50% missing values.
  - Dropped 6 features due to skewness, business insignificance & too much diversity in data.
  - Dropped 29% rows against missing values for 'What is your Current Occupation' to avoid too much loss of features.
  - Dropped 150+ more rows for features having less than 2% missing values.
  - Imputed missing values in ‘Specialization with ‘Others’.

- Data Preparation:
  - Binary Transformation: Converted 2 Yes/No columns to 1/0
  - OHE: One Hot Encoding to create dummy variables for 6 categorical variables with multiple levels. Dropped actual features as well as one of the dummy feature for all to check Multicollinearity.

Data Modelling

Overall Summary (1/3)

- Train-Test Split:
  - Divided data set into 2 set → X: Feature variables ; y: Target variable
  - Split the two into Train & Test data set each in ratio of 70:30.
  - *Train* data set shape: (4473,58) || *Test* data set shape: (1918,58)

- Scaling (Train Set) : Scaled 3 numerical columns using StandardScaler (fit_transform)

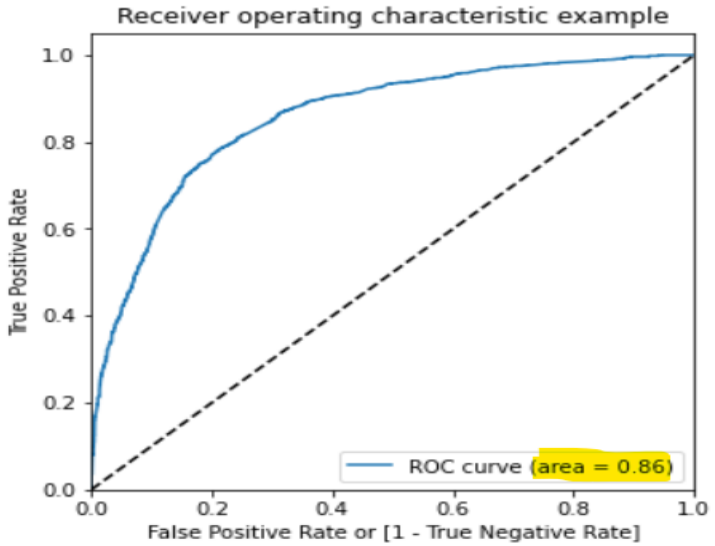
- Model Building (Train Set) :
  - Created 1st Logistic Regression model using Statsmodels GLM and derived coefficients and p-value for all features
  - Performed feature selection via RFE to get top 20 significant features out of 58 and stored them in a list (prf).
  - Again built model with above selected top 20 features using Statsmodel and got coefficients and p-value for all.
  - Now we dropped features with p-value>0.05 or a VIF>5 and rebuild the model iteratively.
  - After dropping overall 7 features, we achieved model with good p-value and VIF for all retained features.
  - Did predictions on y (y_train) based on achieved model on X (X_train) & created Conversion prediction column 'Converted_Pred' with a probability threshold of >0.5.

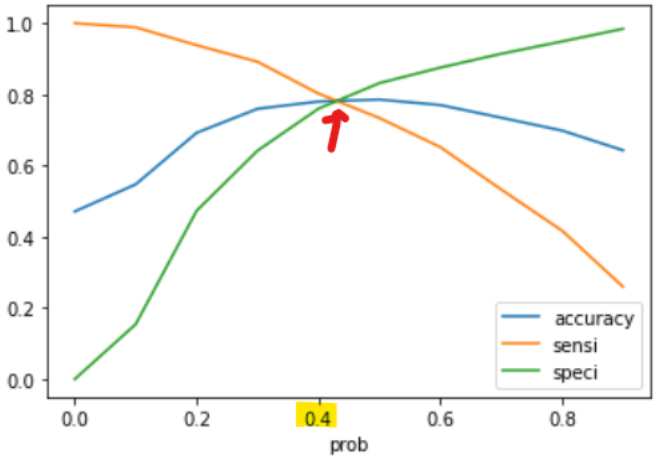
- Evaluation: (Achieved metrics on **Train data set**)
  - **Accuracy**:78.5% | **Precision**: 79.5% | **Sensitivity/Recall**: 73.4% | **Specificity**: 83.2%

Data Modelling

Overall Summary (2/3)

- ROC Curve:
  - Shows the tradeoff between Sensitivity & Specificity. The closer the curve is to the top left corner, greater area the occupied by it, better the model.
  - A good ROC curve area of 0.86 is achieved.





- Optimal Cut off Point:
  - Is the probability at which we achieve balanced sensitivity and specificity
  - From the curve beside, 0.4 seemed to be the optimum (balance between sensitivity & specificity) point to take as a cutoff probability.

Data Modelling

Overall Summary (3/3)

- Model Building (Train Set): (with optimal threshold)
  - Created another Conversion prediction column 'Converted_Pred_Final' with a updated probability threshold of >**0.4**.

- Evaluation: (Final Achieved metrics on **TRAIN data set**)
  - **Accuracy**:78% | **Precision**: 74.9% | **Sensitivity/Recall**: 80.3% | **Specificity**: 76%

- **TEST SET PREDICTIONS**:

  - Scaling : Scaled same 3 numerical columns (as in train set) using StandardScaler (transform).
  - Selecting finalized set of features (from X_train) on test set (X_test).
  - Predicted y_test probability ('Converted_Prob') based on created & verified model on train set run over X (X_test) & created Conversion prediction column 'Conv_Pred_Final' with same probability threshold of >0.4

- Evaluation: (Achieved metrics on **TEST data set**)
  - **Accuracy**:77.1% | **Precision**: 74.4% | **Sensitivity/Recall**: 82.5% | **Specificity**: 71.7%

Recommendation

LEAD SCORE

	Prospect ID	Converted	Converted_Prob	Conv_Pred_Final	Lead Score
0	594	0	0.144426	0	14
1	101	1	0.237537	0	23
2	8123	1	0.943012	1	94
3	2661	1	0.981634	1	98
4	5666	1	0.364378	0	36
5	5271	0	0.702404	1	70
6	958	1	0.083022	0	8
7	4232	1	0.801392	1	80
8	4301	0	0.482256	1	48
9	100	0	0.708344	1	70
10	9219	1	0.743200	1	74
11	3837	1	0.989755	1	98
12	4515	0	0.504590	1	50
13	8699	0	0.183906	0	18
14	7167	1	0.966442	1	96

- Lead score is a score assigned to each of the leads (denoted by ‘Prospect ID’) 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 score ranges from 0-99.
- As per optimal model criteria, any score above 40 is considered a ‘HOT LEAD’ and should be targeted.
- If sales team have more bandwidth and want to make conversions more aggressively, we can reduce the optimal cut off point (say 0.2), so that they can even take a hunch on derived potential leads with a score as low as 20.
- On other side, if they want to minimize the effort spent in calling so may leads, but still maintaining a high conversion ratio, we could raise the optimal cut off point (say 0.7 or 0.8) i.e., featuring the potential leads with a conversion probability of more than 70 or 80%.

Thanks!

BY- RISHIK & VANDIT