LEAD SCORING CASE STUDY

Identify the ‘hot’ leads which are more likely to result in a lead conversion

PROBLEM STATEMENT

- An education company X Education, sells online courses to industry professionals.
- The company markets its courses in various ways so that various people come to its platform and fill up a form showing their interest in the courses.
- When people fill up forms by providing email id or phone number, they are classified to be a lead. The company also gets leads from past referrals.
- Currently, the lead conversion rate of X Education is quite low, which means very few people end up signing up for their courses.
- The company wants to identify those ‘hot’ leads, which would lead to conversion and improve the lead conversion rate.
- In technical terms, the company requires a model, which will assign a lead score to each lead. A high lead score would mean that, that lead is a ‘hot’ lead and a low score means that, that lead is a ‘cold’ lead.

ANALYSIS OVERVIEW

- The input data had around 9240 leads with 37 features.
- There were a lot of missing values in the data. The following techniques were used for missing value treatment.
  - Features with high missing value percentage were dropped.
  - The categorical columns were imputed based on mode values
  - The numerical columns were imputed based on median values.
  - A few of the columns were having ‘Select’ as one of the values. Such values were replaced with NaN and were imputed accordingly.
  - Some features had 1% missing values. Such leads were removed.
  - A few of the categorical features were imputed with logical values.

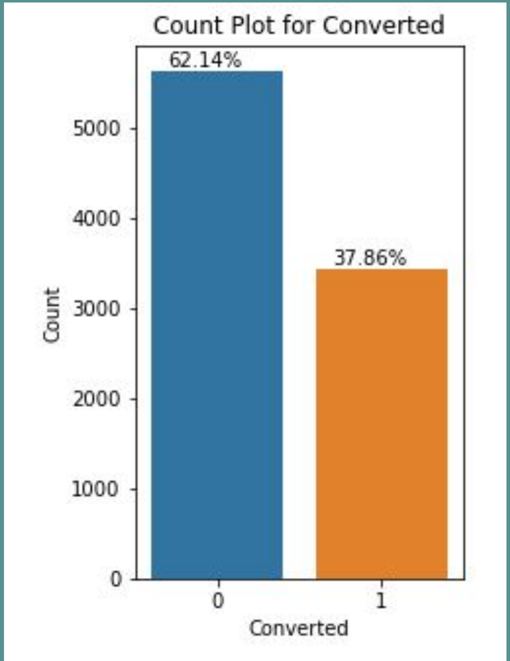
ANALYSIS OVERVIEW

- During EDA, it was observed that the dataset also had some features for which more than 98% of the leads belonged to a single category. Such features were also dropped.
- The lead conversion rate was around 37%.
- The categorical features containing True/False values were converted to 1/0 and those features containing multiple levels were converted to dummy columns using one-hot encoding.
- Heatmap was used to visualize the correlation and highly correlated columns were removed. The final dataset had 38 unique features.
- The numerical columns were scaled using MinMaxScaler after the train test split was done.
- RFE was used to select top 20 features and then manual fine tuning was done to arrive at the final model.

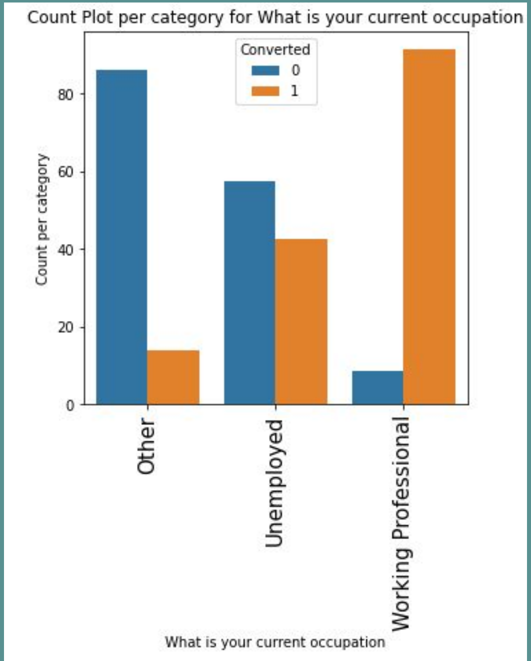
ANALYSIS OVERVIEW

- ➔ The final model was evaluated on the train data using precision/recall and sensitivity/specificity metrics, which had reasonable values.
- ➔ The final probability threshold was decided by plotting the sensitivity, specificity and accuracy for different probability values. The value came out to be 0.332.
- ➔ There was also no presence of multicollinearity among the final features selected which was confirmed by calculating the variance inflation factor.
- ➔ After finalizing the model on training data, it was used to predict the values on the test dataset.
- ➔ The model also had reasonable values of sensitivity/specificity and precision/recall on the test dataset.
- ➔ Therefore it was concluded that the final model would be able to identify the hot leads accurately.

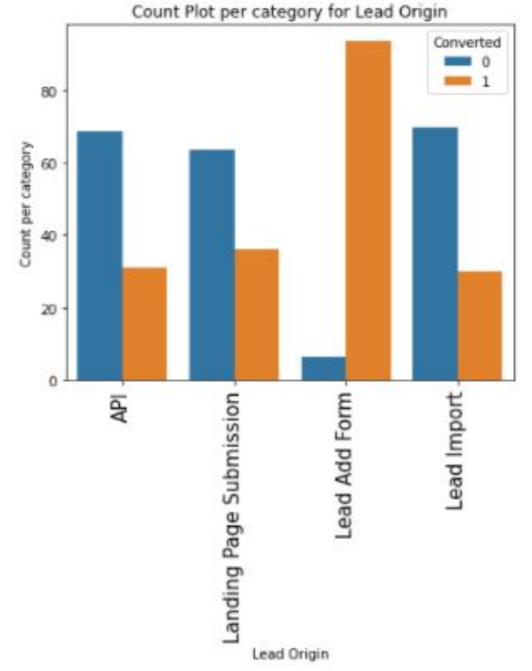
EXPLORATORY DATA ANALYSIS



There is a 38% conversion rate in the dataset.

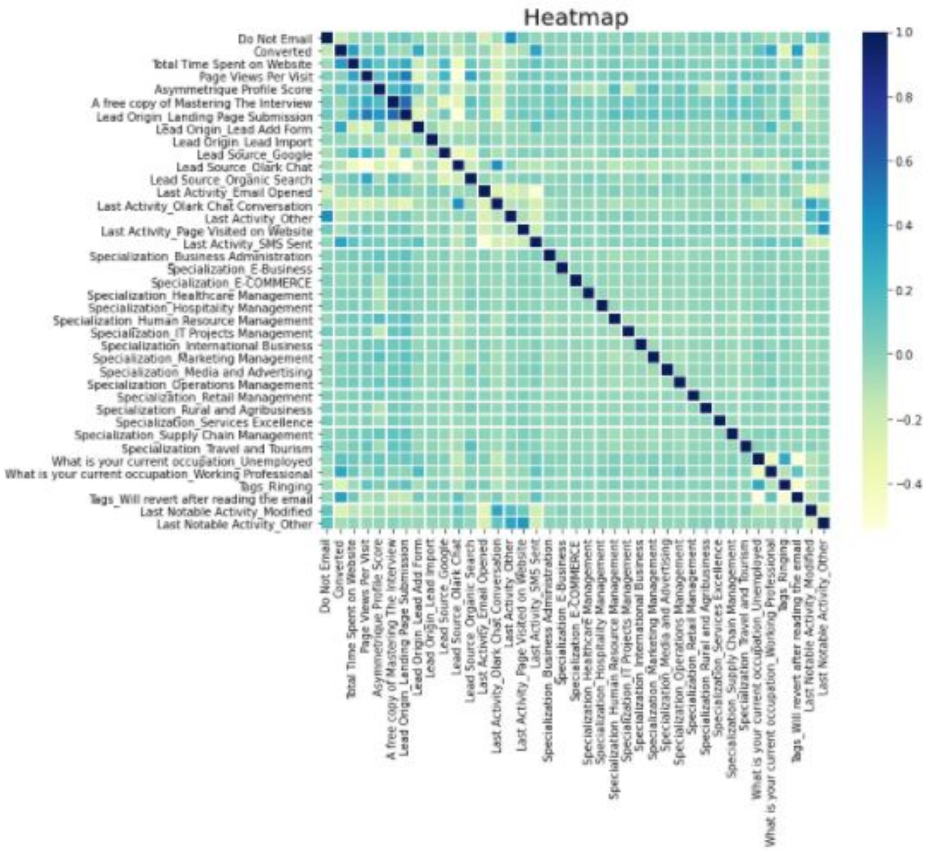


Working Professionals and Unemployed have a high conversion rate.



Leads originating from Lead Add Form and Landing Page Submission have a high conversion rate.

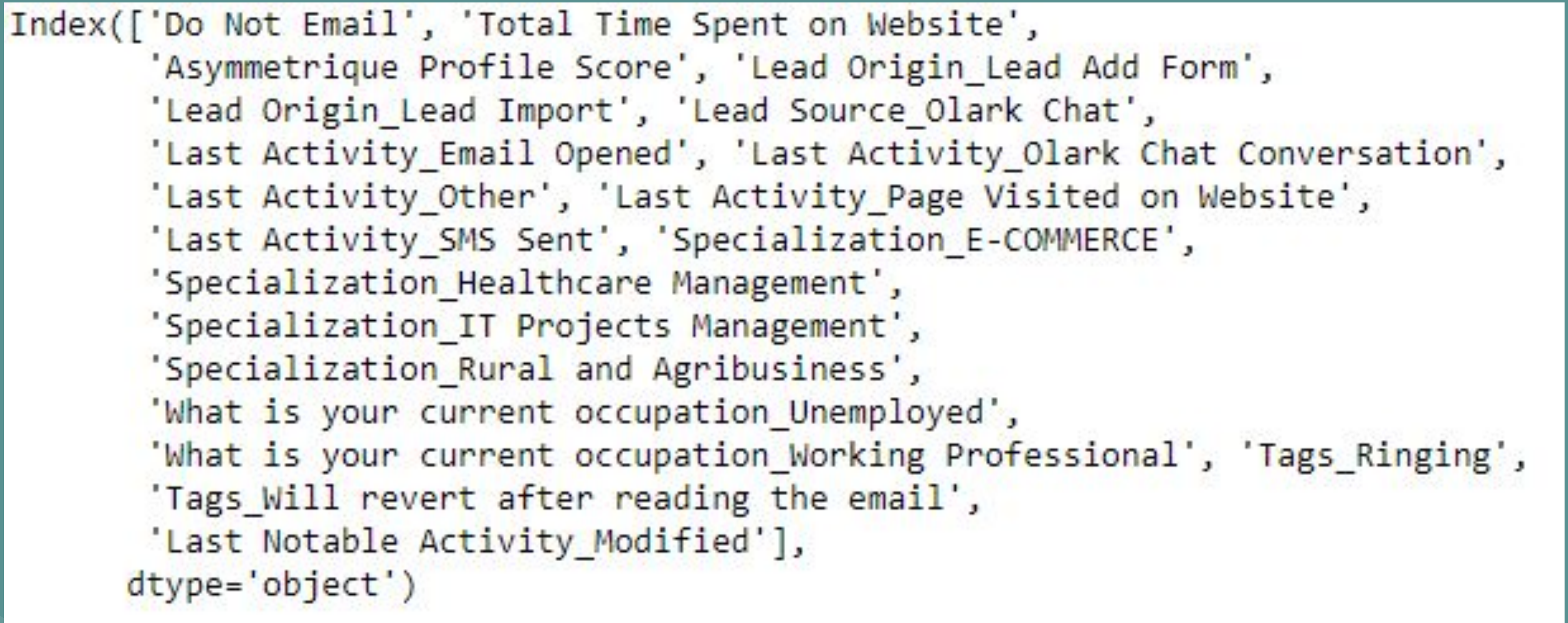
EXPLORATORY DATA ANALYSIS



There was a high correlation among a few dummy variables, which were removed so as to avoid multicollinearity.

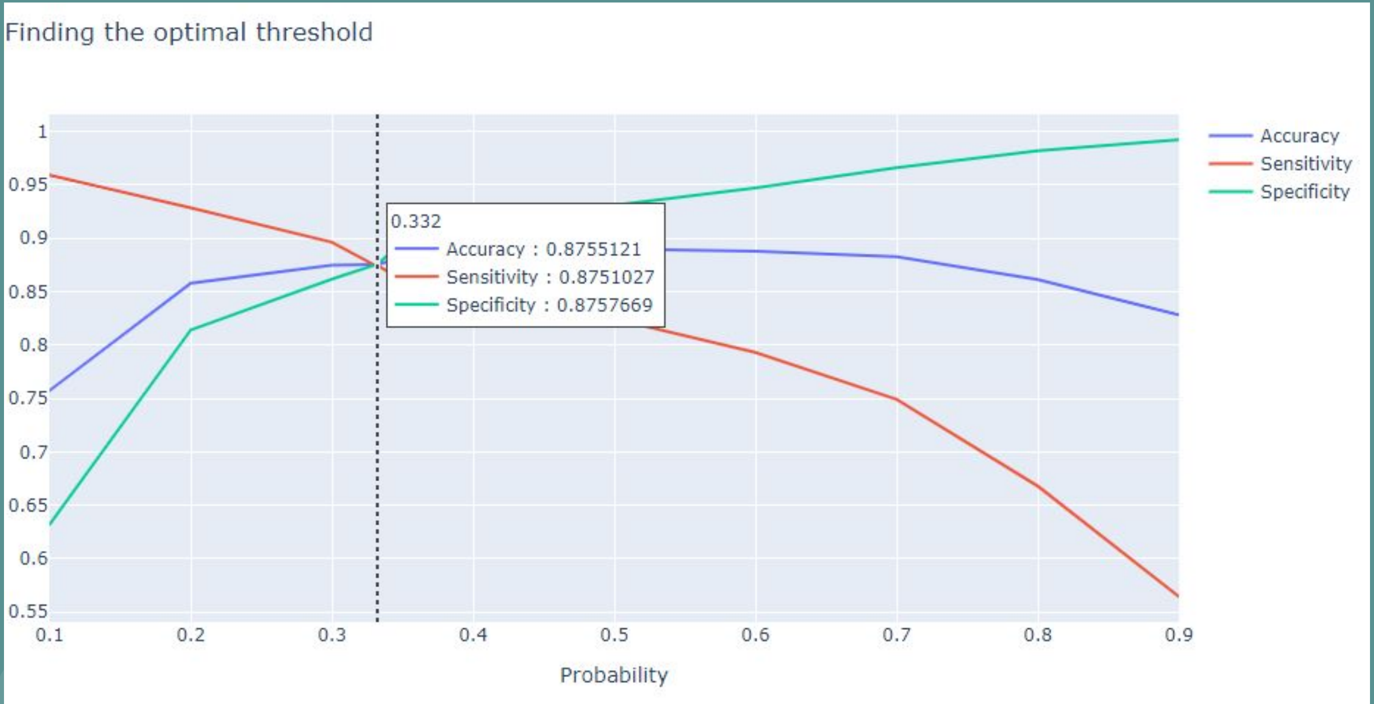
FEATURE SELECTION USING RFE

RFE was used to select top 20 features for model building.



```
Index(['Do Not Email', 'Total Time Spent on Website',
       'Asymmetrique Profile Score', 'Lead Origin_Lead Add Form',
       'Lead Origin_Lead Import', 'Lead Source_Olark Chat',
       'Last Activity_Email Opened', 'Last Activity_Olark Chat Conversation',
       'Last Activity_Other', 'Last Activity_Page Visited on Website',
       'Last Activity_SMS Sent', 'Specialization_E-COMMERCE',
       'Specialization_Healthcare Management',
       'Specialization_IT Projects Management',
       'Specialization_Rural and Agribusiness',
       'What is your current occupation_Unemployed',
       'What is your current occupation_Working Professional', 'Tags_Ringing',
       'Tags_Will revert after reading the email',
       'Last Notable Activity_Modified'],
      dtype='object')
```

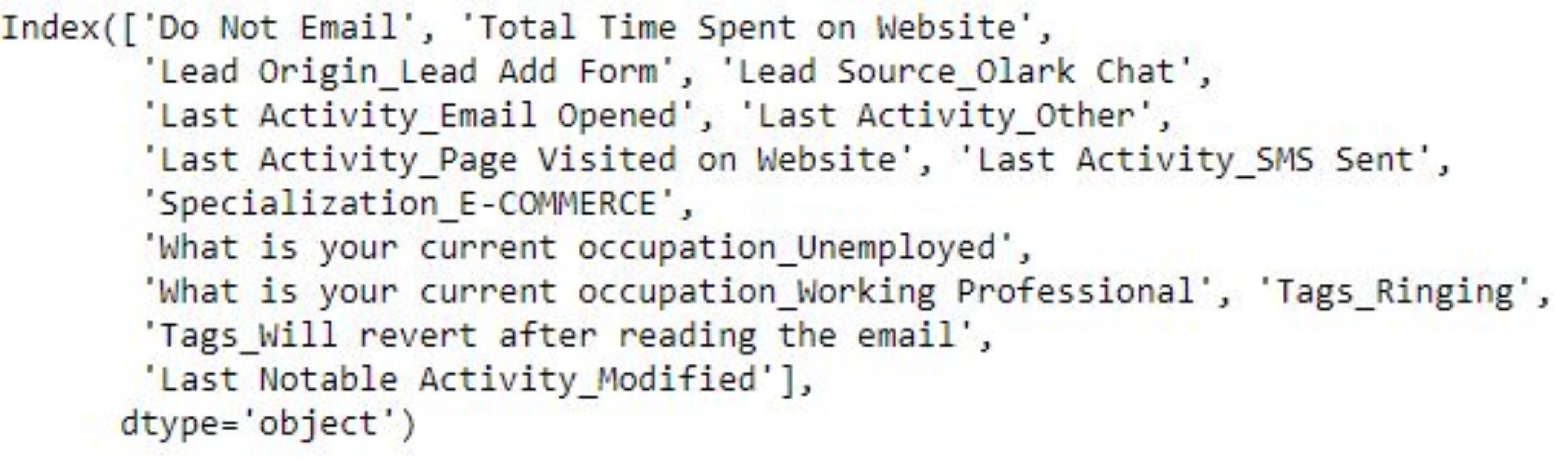
OPTIMAL THRESHOLD



Plotting the sensitivity, specificity and accuracy for various probability values revealed that the optimal threshold value should be 0.332

FINAL FEATURES LIST

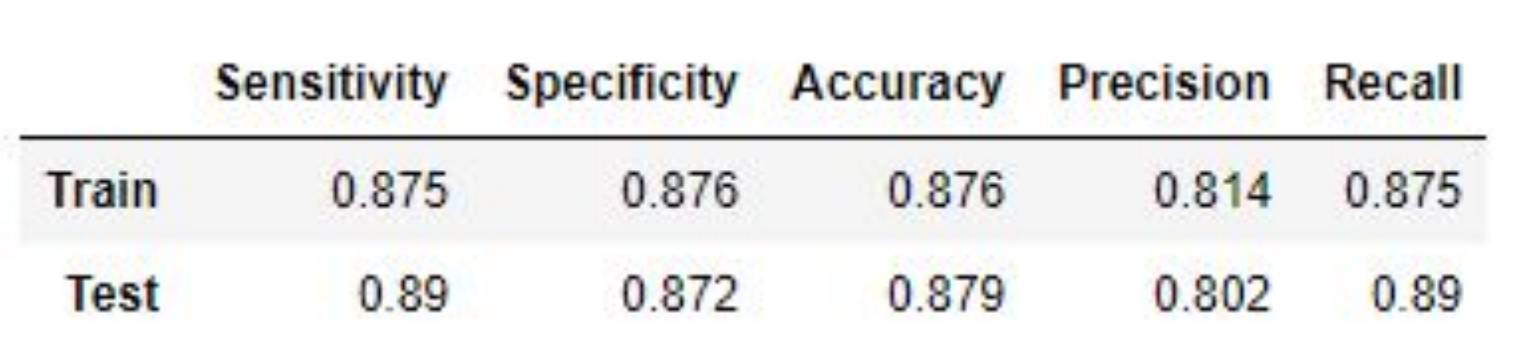
The final model had 14 features.



```
Index(['Do Not Email', 'Total Time Spent on Website',
       'Lead Origin_Lead Add Form', 'Lead Source_Olark Chat',
       'Last Activity_Email Opened', 'Last Activity_Other',
       'Last Activity_Page Visited on Website', 'Last Activity_SMS Sent',
       'Specialization_E-COMMERCE',
       'What is your current occupation_Unemployed',
       'What is your current occupation_Working Professional', 'Tags_Ringing',
       'Tags_Will revert after reading the email',
       'Last Notable Activity_Modified'],
      dtype='object')
```

MODEL EVALUATION

The below table reveals the values of various metrics for train and test data. The sensitivity that was observed for the test data was 0.89 which confirmed that our model was correctly able to identify the hot leads from a given set of leads.



	Sensitivity	Specificity	Accuracy	Precision	Recall
Train	0.875	0.876	0.876	0.814	0.875
Test	0.89	0.872	0.879	0.802	0.89

SUMMARY

- ➔ The dataset had many null values which were handled appropriately.
- ➔ Outliers present in the numerical columns were also removed.
- ➔ **EDA revealed that working professionals and unemployed had the highest lead conversion rate.**
- ➔ Dummy variables were created for appropriate categorical variables.
- ➔ RFE was used to pick out top 20 features.
- ➔ The final model had **14 features**.
- ➔ Since the CEO wanted to accurately identify the converted leads, we focussed more on the sensitivity and recall metrics.
- ➔ The **sensitivity** and **recall** metrics of the model on the **train dataset was around 87.5%** and on the **test dataset was 89%**.
- ➔ The **accuracy** of the model on the **train dataset was 88%** and on the **test dataset was 87%.**
- ➔ **This model was finalized as it was correctly able to predict the hot leads**.