LEAD SCORE PREDICTION

SUBMITTED BY:

ASHISH SRIVASTAVA
ADITYA CHOPRA
AVNISH SHARMA
SAHIL BANSAL

BUSINESS OBJECTIVES

X Education sells online courses to industry professionals. It markets it's courses on several websites and search engines.

The company receives lots of leads, but it's conversion rate, at around 30%, is pretty low

Objectives:

- Identify most promising leads (Hot leads)
- Help sales/marketing teams to strategize and focus on fewer more significant areas/leads
- Create a model to achieve the above 2 points with a precision of at least 80%
- Prepare a strategy high and low seasons

METHODOLOGY

Data Preparation & Analysis

- Gather & review metadata
- Inspect for quality and consistency
- Fix data quality issues, impute missing values etc.
- Create various plots to visually depict the relationship between different dimensions and measures

Pre-Processing

- Split into Training and Test set
- Standardize Data
- Identify and eliminate insignificant and multi-collinear features

Model Creation and Evaluation

- Build model on training data
- Perform Predictions on test data
- Derive various metrics and plots to compare models with different hyper-parameters
- Choose a model and it's related parameters

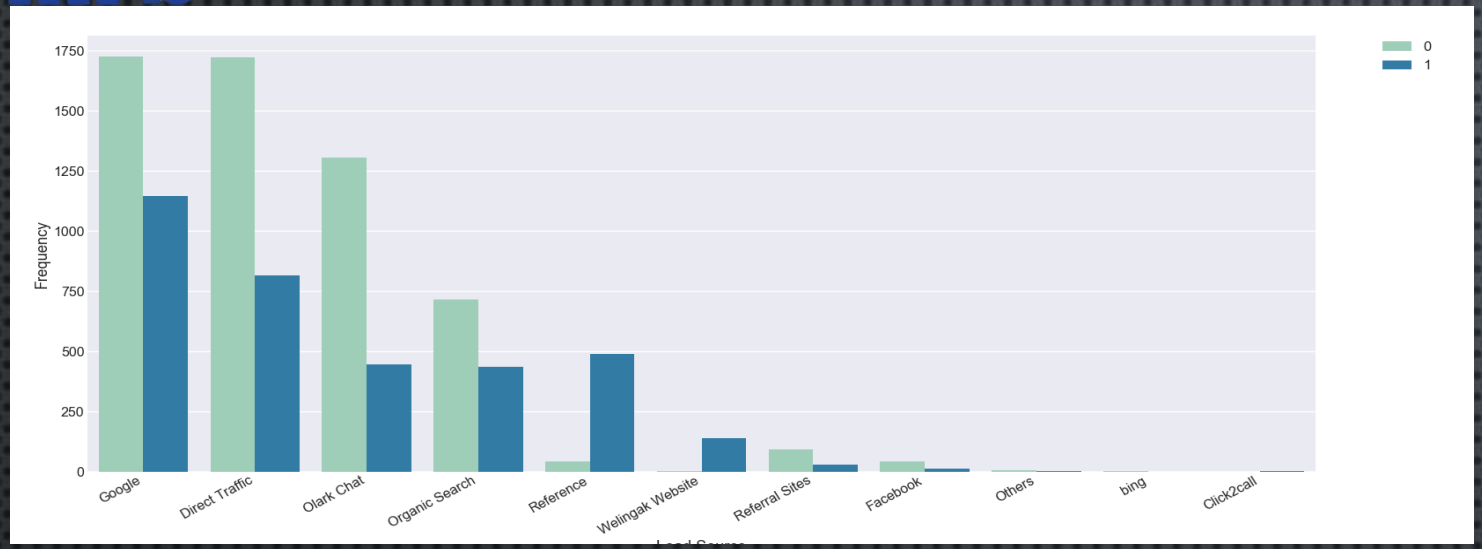
DATA EXPLORATION

Data Quality

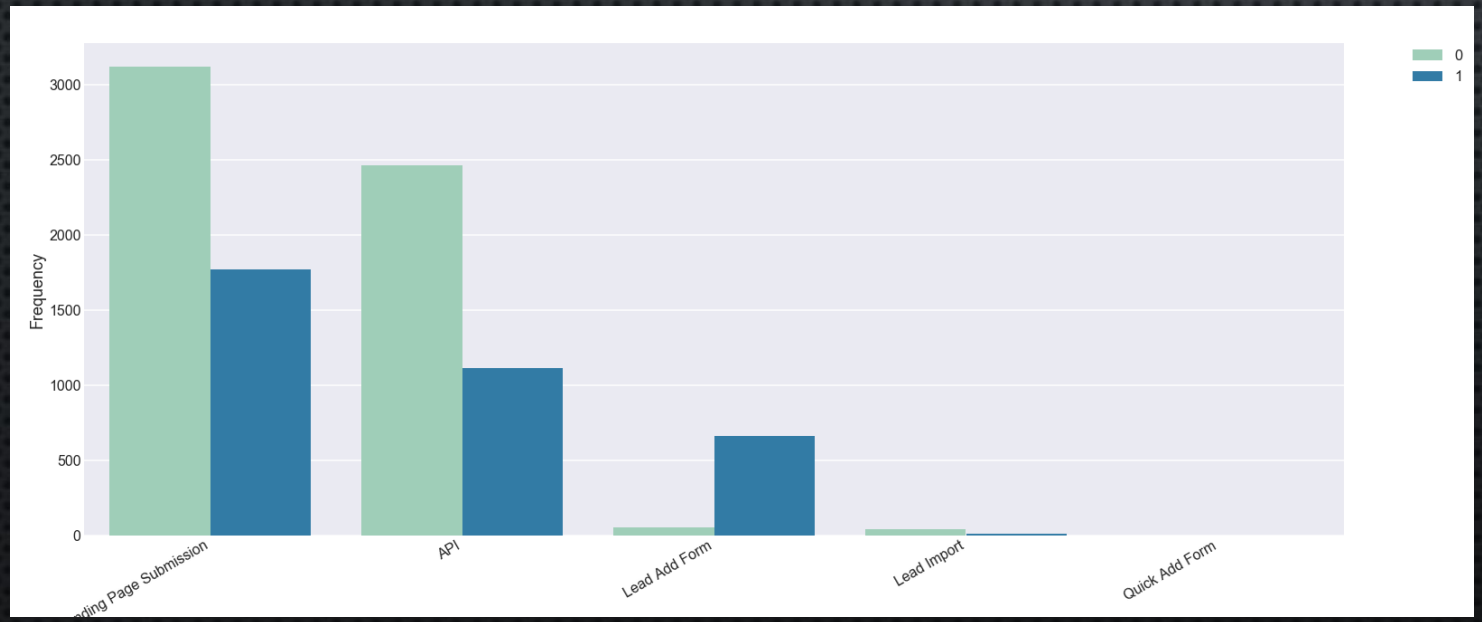
- Original Dataset – 9240 rows, 37 features

Data Quality Issues	Features	Resolution
Very Low/No variance	Prospect ID, Magazine, 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, Country	These features had very low variance (only 1 value predominantly) and were dropped
Too Many NULL's	Lead Profile, Lead Quality, Asymmetrique Profile Score Asymmetrique Activity Score, Asymmetrique Profile Index, Asymmetrique Activity Index, City	>30% values missing with no logical way of imputation. Features were dropped
Data Imputation	Specialization, TotalVisits, Page Views per Visit	Continuous variables were imputed with mean value Categorical variables were imputed with 'Unknown'

LEAD SOURCE AND LEAD ORIGIN



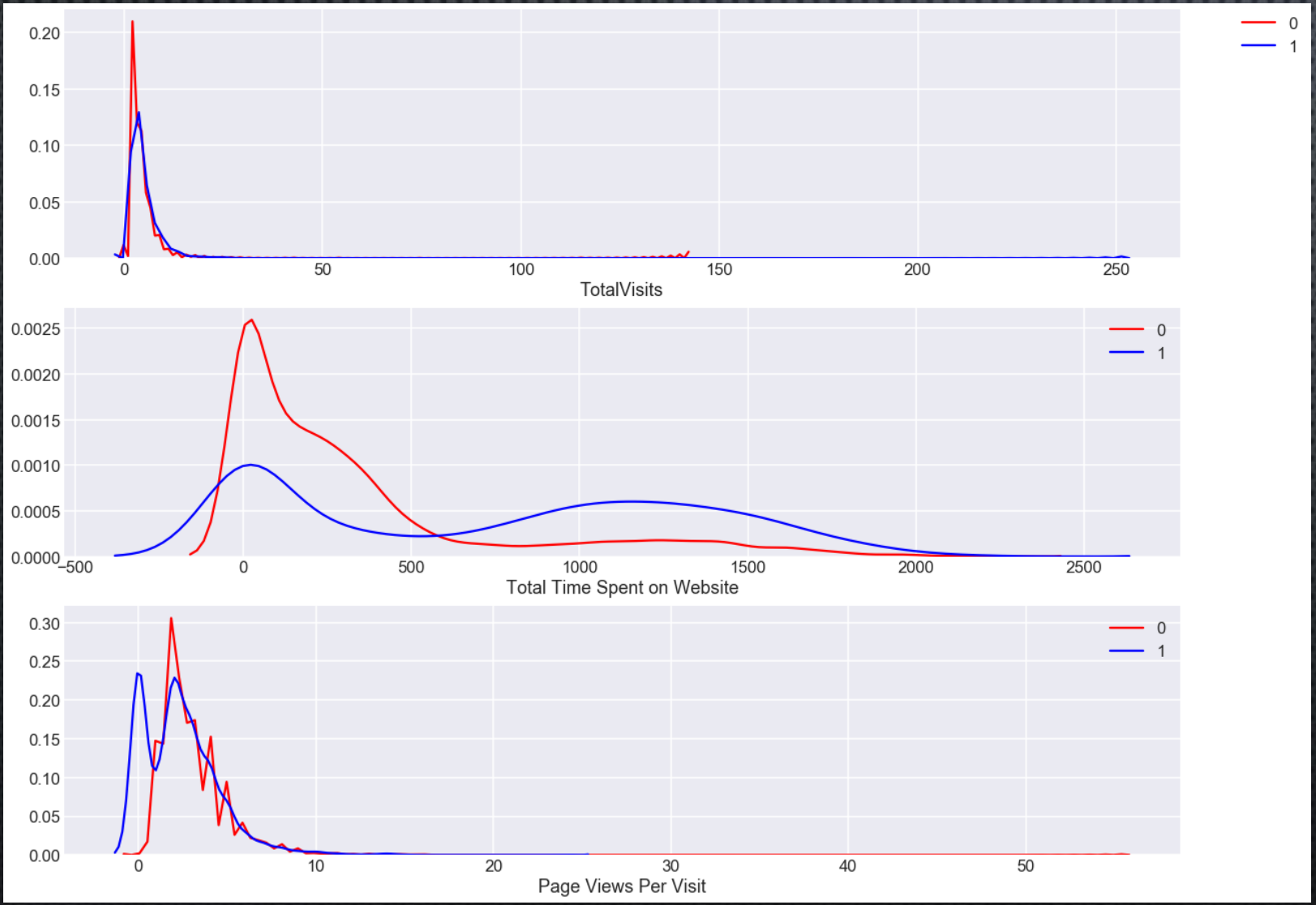
Lead Source



Lead Origin

- The 2 plots depict the distribution of Lead Source (Top) and Lead Origin (bottom)
- From the plots, we see that Google, Direct Traffic, Olark Chat and Organic Search generate the maximum leads, but References has the highest conversion
- Most number of Leads originate from the Landing Page followed by API (search engines), but the most conversions happen through Lead Add form

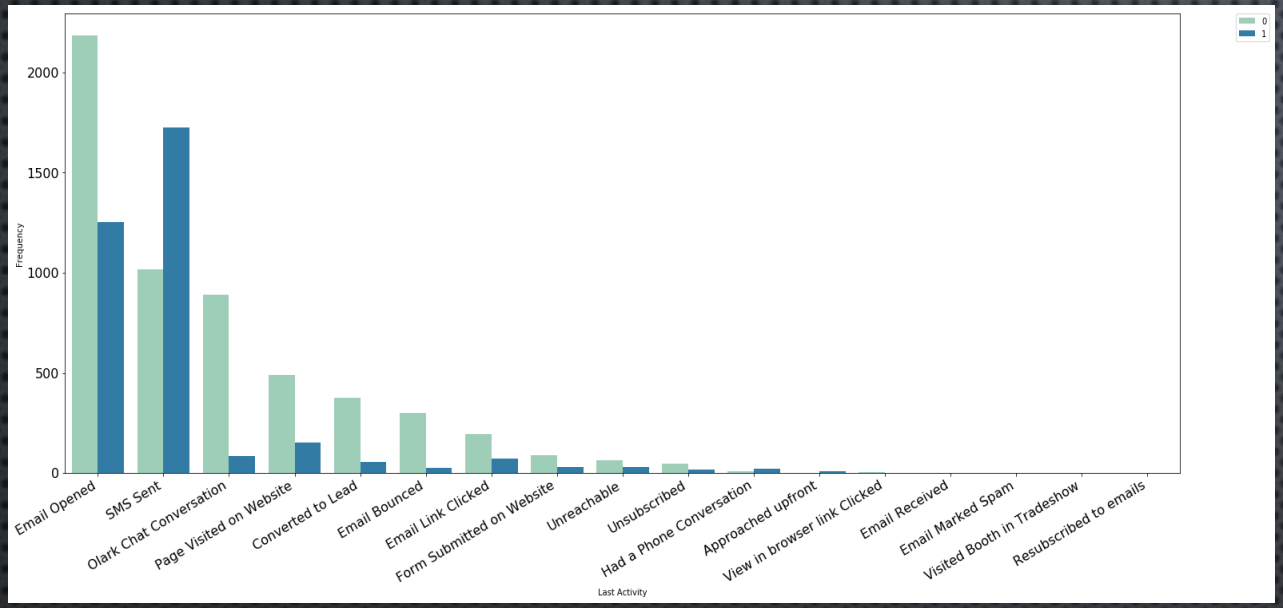
PAGE VIEW METRICS



Distribution of various page metrics

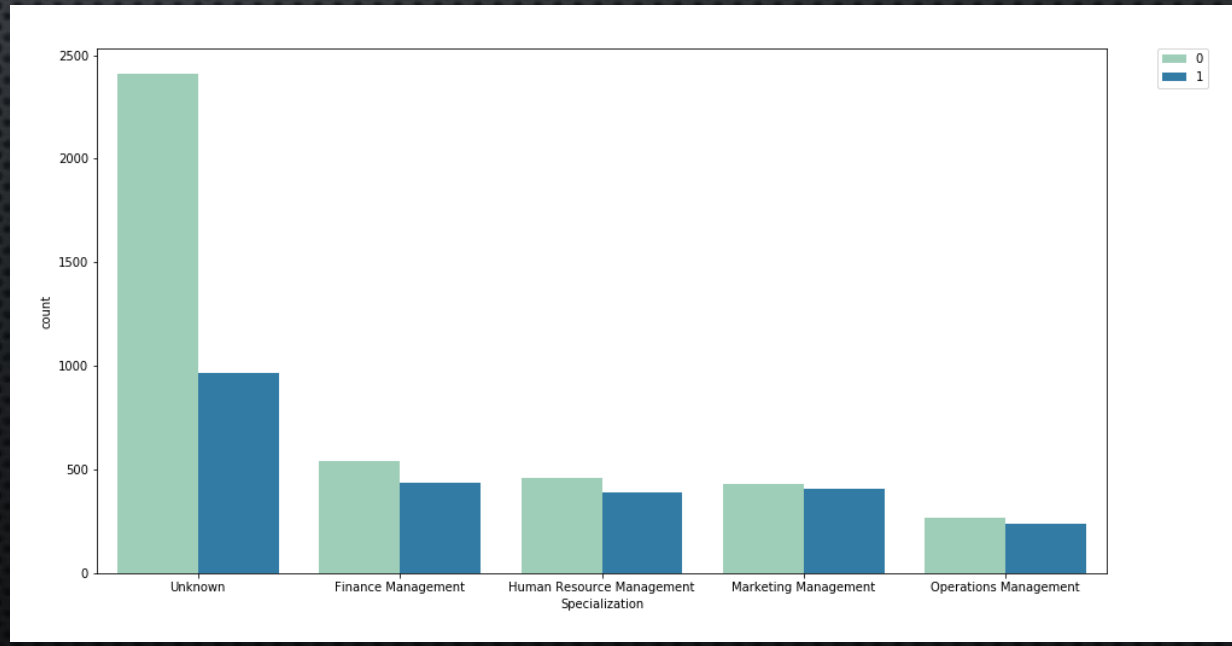
- This plot depicts the distribution of various page related metrics, like Total Visits, Time Spent on Website and Number of pages seen per visit
- Total time spent on Website – For converted leads, this is spread out wide, meaning people possibly spent more time before taking a positive decision
- They also visited fewer pages compared to unconverted leads, probably because their search was more directed and they were not just browsing

LAST ACTIVITY & SPECIALIZATION



Plot of Last Activity

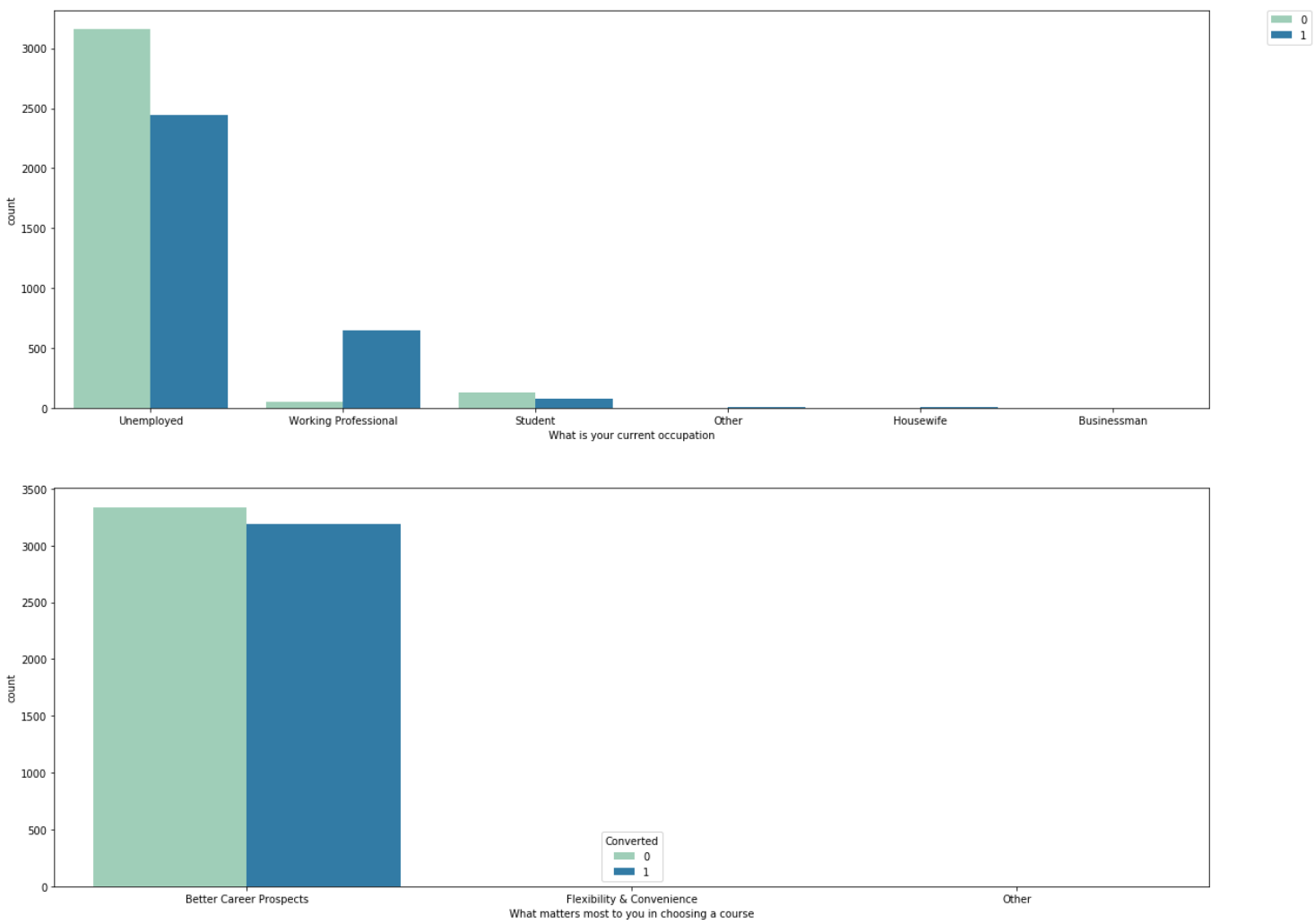
- This plot (on left), displays the last activities performed on a lead. We see that most of the lead conversion occurs via email and SMS communication.
- We also notice that the conversion rate for Olark chat channel is low, something that can be improved



- This plot (on right), displays the specialization/course offering that people have shown interest in.
- It is interesting to note, that for all the segments, the conversion ratio is approximately equal.
- Maybe the choice of specialization does not have a bearing on lead conversion

Plot of Specialization

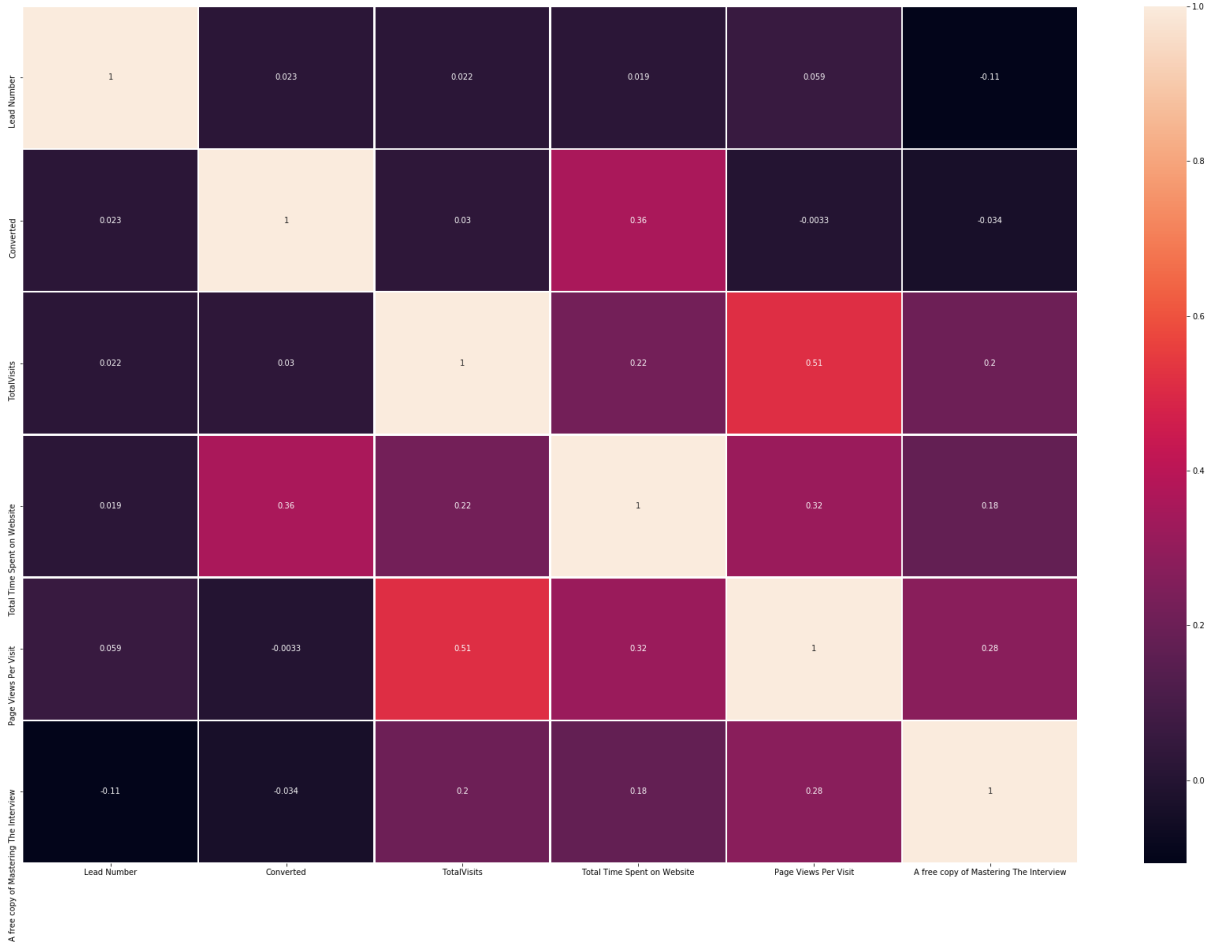
CURRENT OCCUPATION & MOTIVATION



Plot of Current Occupation and Motivation

- This plot depicts the employment status of the enquirers and what their motivation/aspiration is
- Occupation – We see most of the inquiries coming from Unemployed people, which is quite logical. This segment has a very high rate of conversion as well.
- Working professionals have less volume but extremely high conversion, probably because they are motivated to accomplish something.
- Student volume is less, but has a decent conversion rate. There should be some directed focus on this segment to increase the volume
- Motivation – 'Better Career Prospect' seems to be the only motivation behind pursuing education

CORRELATION BETWEEN FEATURES



Heat map of different features

- The plot depicts correlation between various numerical features of the dataset
- Some features are positively correlated, like TotalVisits, Total Time Spent on Website and Page views per visit.
- There seems to be a high correlation with ‘A free copy of Mastering The Interview’. Maybe freebies and career advancement advices can bring in more volume

ELIMINATING AND RETAINING FEATURES

- So far, we inspected, cleansed, eliminated and visualized the data.
- We also Standardized the continuous variables, one-hot encoded categorical variables and divided the dataset into training and test set
- However, there are still large number of variables, all of which may not be significant, or may have a high multi-collinearity. We will eliminate them systematically
- Using Recursive Feature Elimination and Variance Inflation Factor analysis, we identify the insignificant and multi-collinear variables and remove them.
- The resulting dataset thus consists of features that are significant for the regression modelling

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6395
Model:	GLM	Df Residuals:	6377
Model Family:	Binomial	Df Model:	17
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2592.8
Date:	Sun, 03 Mar 2019	Deviance:	5185.5
Time:	00:44:32	Pearson chi2:	6.73e+03
No. Iterations:	6	Covariance Type:	nonrobust

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.5360	0.158	-9.748	0.000	-1.845	-1.227
Total Time Spent on Website	1.1015	0.040	27.204	0.000	1.022	1.181
Lead Origin_Landing Page Submission	-0.9230	0.135	-6.828	0.000	-1.188	-0.658
Lead Origin_Lead Add Form	3.5867	0.218	16.422	0.000	3.159	4.015
Lead Source_Direct Traffic	-0.3228	0.089	-3.632	0.000	-0.497	-0.149
Lead Source_Olark Chat	1.1789	0.123	9.551	0.000	0.937	1.421
Last Activity_Email Bounced	-2.0357	0.414	-4.921	0.000	-2.846	-1.225
Last Activity_Email Opened	0.5346	0.092	5.794	0.000	0.354	0.715
Last Activity_Olark Chat Conversation	-0.9172	0.175	-5.256	0.000	-1.259	-0.575
Specialization_Hospitality Management	-0.7816	0.340	-2.299	0.021	-1.448	-0.115
Specialization_Unknown	-0.9700	0.126	-7.683	0.000	-1.217	-0.723
Last Notable Activity_Email Bounced	1.6392	0.605	2.710	0.007	0.454	2.825
Last Notable Activity_Had a Phone Conversation	3.4330	1.163	2.951	0.003	1.153	5.713
Last Notable Activity_SMS Sent	1.7753	0.101	17.569	0.000	1.577	1.973
Last Notable Activity_Unreachable	2.0800	0.538	3.868	0.000	1.026	3.134
What is your current occupation_Student	0.9241	0.238	3.890	0.000	0.458	1.390
What is your current occupation_Unemployed	1.0106	0.087	11.550	0.000	0.839	1.182
What is your current occupation_Working Professional	3.4090	0.205	16.623	0.000	3.007	3.811

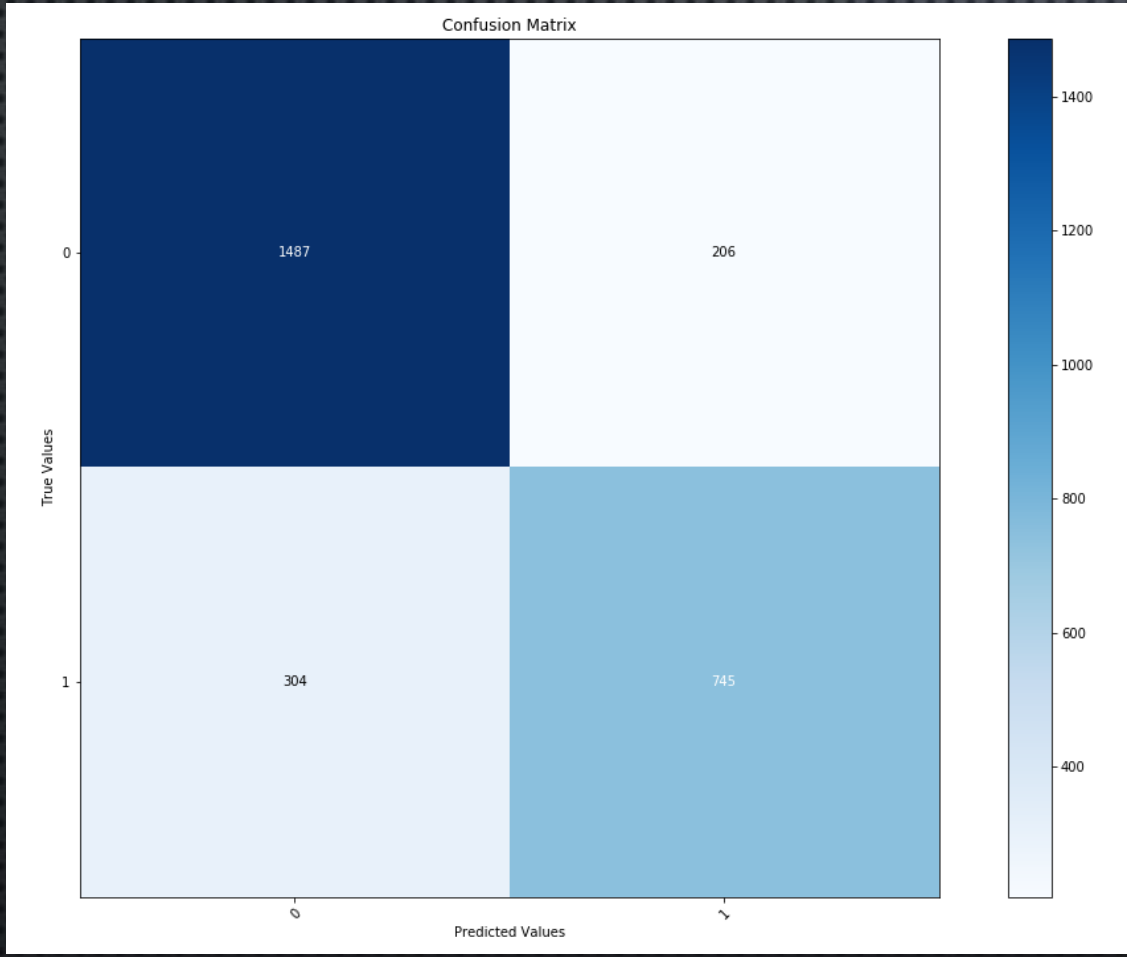
MODEL BUILDING & EVALUATION

- Using RFE, we reduced our feature list from 103 to 17. These features can now be used to train the logistic regression model and make predictions on unseen data

- Building a model on Training dataset yields an accuracy of **~82.30%**, and validating it on Test dataset yields a model accuracy of **~81.4%**

- **Model Evaluation**:
  - **Confusion Matrix**: This is used to derive various other metrics such as Precision-Recal, Sensitivity-Specificity
  - **Receiver Operating Characteristic (ROC) Curve**: This is a plot of True positive rate against False positive rate, or signal v/s noise. The higher the area under curve (AUC), the better the model.
  - **Precision-Recall/ Sensitivity-Specificity plot**: These plots depict the ability of the model to provide predictions with minimal error, and facilitate business to choose the right metric that fits their need

CONFUSION MATRIX



Confusion Matrix

- Confusion Matrix describes the performance of a classification model and helps in deriving meaningful insights into the effectiveness of the model.
- It is made up of the following elements, which help in deriving further metrics:

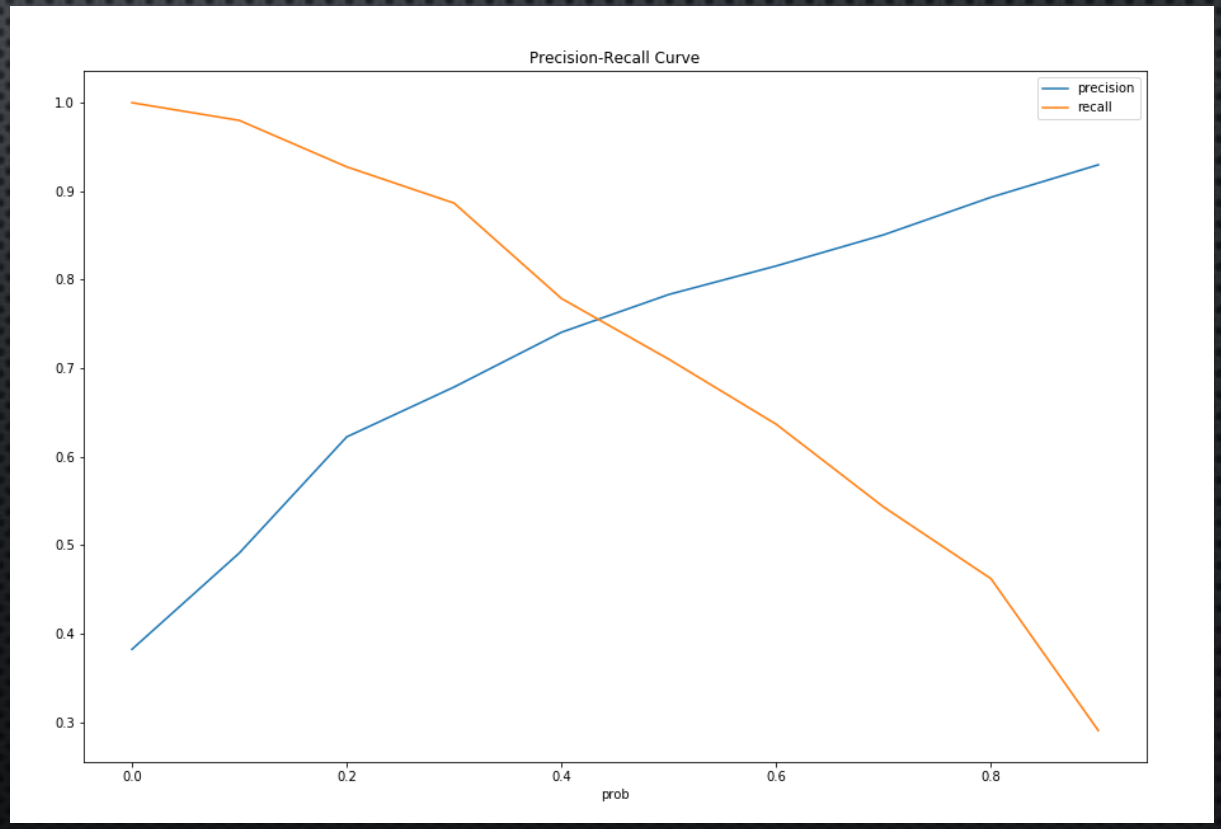
True Positive (TP)	Predicted and Actual values are 'Yes'
True Negative (TN)	Predicted and Actual Values are 'No'
False Positive (FP)	Predicted as 'Yes', but actually 'No' (Type I error)
False Negative (FN)	Predicted as 'No', but was actually 'Yes' (Type II error)

Metric	Implication	Formula	Model Value
Accuracy	How often is the classifier correct	(TP+TN)/Total	81.4%
Sensitivity/ Recall	Probability that an actual 'Yes' is predicted 'Yes'	TP/(TP+FN)	71.02%
Specificity	Proportion of 'No' correctly predicted	TN(/TN+FN)	87.83%
Precision	Probability that a predicted 'Yes' is actually 'Yes'	TP/(TP+FP)	78.34%

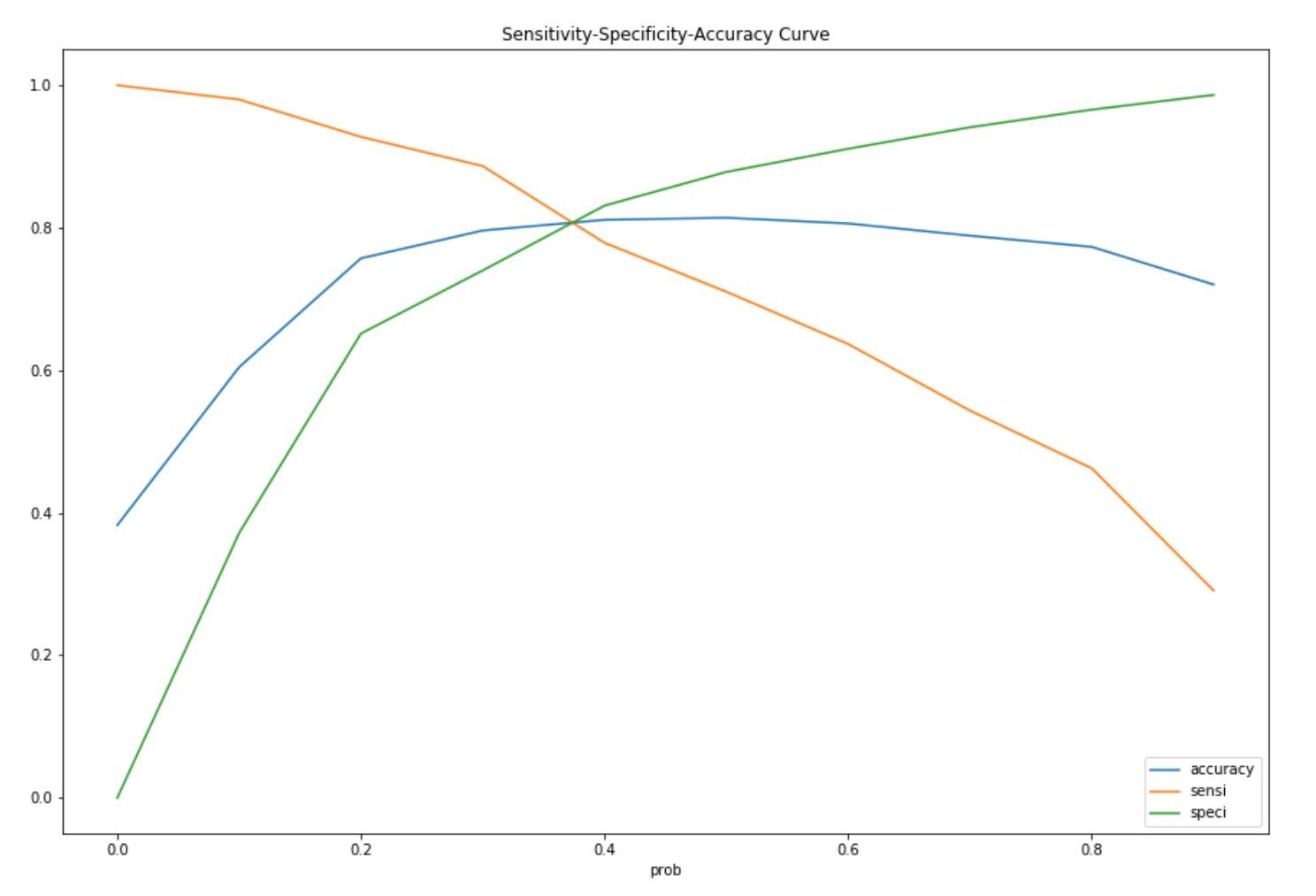
COMPARING MODEL CUT-OFFS

- Default model cut-off is at 0.5, however, this is not ideal, and needs to be tested for a range of probabilities.
- The probability cut-off at which out business objective of achieving 80% precision is met, should be chosen.
- The below table and chart illustrate the effect of using different probability cut-offs on the model. From this, we see that a cut-off of **0.6** is optimum

	prob	accuracy	sensi	speci	precision	recall
0.00	0.00	0.38	1.00	0.00	0.38	1.00
0.10	0.10	0.60	0.98	0.37	0.49	0.98
0.20	0.20	0.76	0.93	0.65	0.62	0.93
0.30	0.30	0.80	0.89	0.74	0.68	0.89
0.40	0.40	0.81	0.78	0.83	0.74	0.78
0.50	0.50	0.81	0.71	0.88	0.78	0.71
0.60	0.60	0.81	0.64	0.91	0.82	0.64
0.70	0.70	0.79	0.54	0.94	0.85	0.54
0.80	0.80	0.77	0.46	0.97	0.89	0.46
0.90	0.90	0.72	0.29	0.99	0.93	0.29



SENSITIVITY, SPECIFICITY & ACCURACY

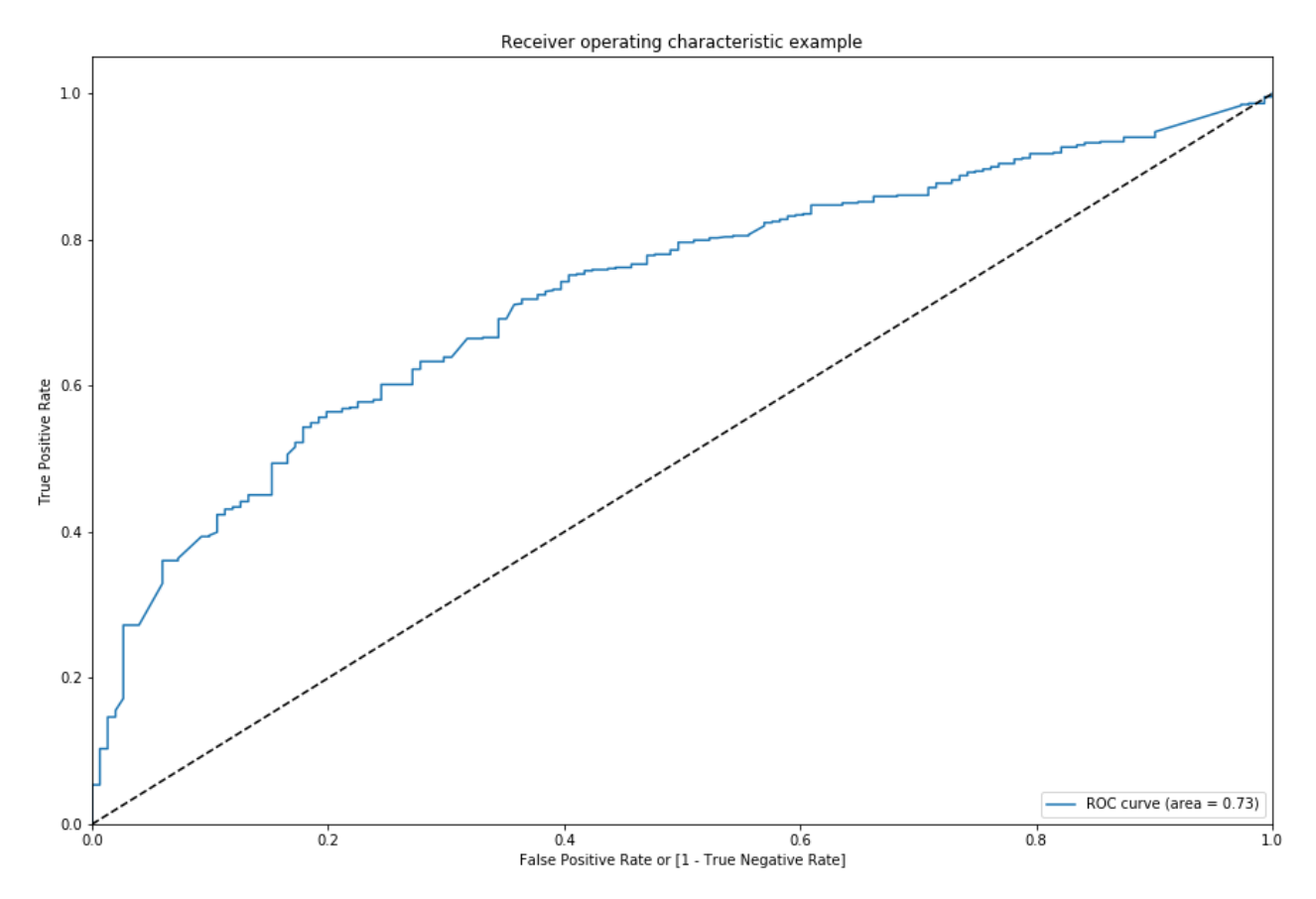


Sensitivity, Specificity and Accuracy for different cut-off

- This plot depicts some more model metrics over a range of probability cut-offs.
- Sensitivity and Specificity are inverse of each other
- Sensitivity measure the goodness of model to classify a lead conversion as 'Yes' when it actually is 'Yes'
- Similarly Specificity measure the goodness of the model to predict a lead conversion as 'No' when it is actually 'No'
- These 2 metrics measures the capability of the model to identify the potential leads as 'hot' or 'cold' with a high degree of confidence.

RECEIVER OPERATING CHARACTERISTIC CURVE

- ROC Curve tells us how much a model is capable of distinguishing between classes.
- Higher the Area Under Curve (AUC), the better it is in identifying 'No' as 'No' and 'Yes' as 'Yes'



ROC Curve for cut-off of 0.6

- The ROC curve has a Area Under Curve of 0.73
- This is a acceptable number and demonstrates that ~73% of the time, the model is able to correctly distinguish between 'Yes' and 'No'

BUSINESS OBJECTIVES IN TERMS OF MODEL METRICS

- At a probability cut-off of 0.6, we have a Precision of 0.82, Sensitivity of 0.64 and Specificity of 0.91
  - 82% of the times when model predicted a Lead would 'convert', it did
  - 91% of the time when model predicted a Lead would 'not convert', and it did not
- If probability cut-off is reduced:
  - Precision Decreases – Model would predict a lead as 'converted', but it will not convert
  - Specificity Decreases – Probability of predicting a lead as 'not converting' reduces, which leads to more follow ups with reduced positive outcome
- If Probability cut-off is increased:
  - Sensitivity reduces: Our model is predicting less 'Yes' and has become more restrictive.
  - Precision Increases: Outcome of Model prediction is good, and leads marked as 'Yes' do indeed get converted.
  - Specificity Increases: Model increasingly marks non-promising leads as No, so less effort in chasing them

SUMMARY AND RECOMMENDATION

- From the analysis and plots in previous slides, we identified that the following features are most significant in predicting the conversion rate of a lead:
  - Total Time Spent on Website
  - Lead Origin
  - Current Occupation of aspirant
  - Last Notable Activity
- We also analyzed the effect of different cut-offs on various metrics and how they can be adjusted to meet different business objectives:
  - Higher conversion rate can be achieved by aiming for higher Precision and raising the cut-off
  - Higher number of leads can be pursued by reducing the cut-off thereby increasing the specificity
- Some other aspects which stood out during the analysis:
  - Digital advertising channels can be improved (negligible leads from search, newspaper digital adv.)
  - Students should be targeted and strategies planned around them
  - Improve engagement through chat platform

Thank You