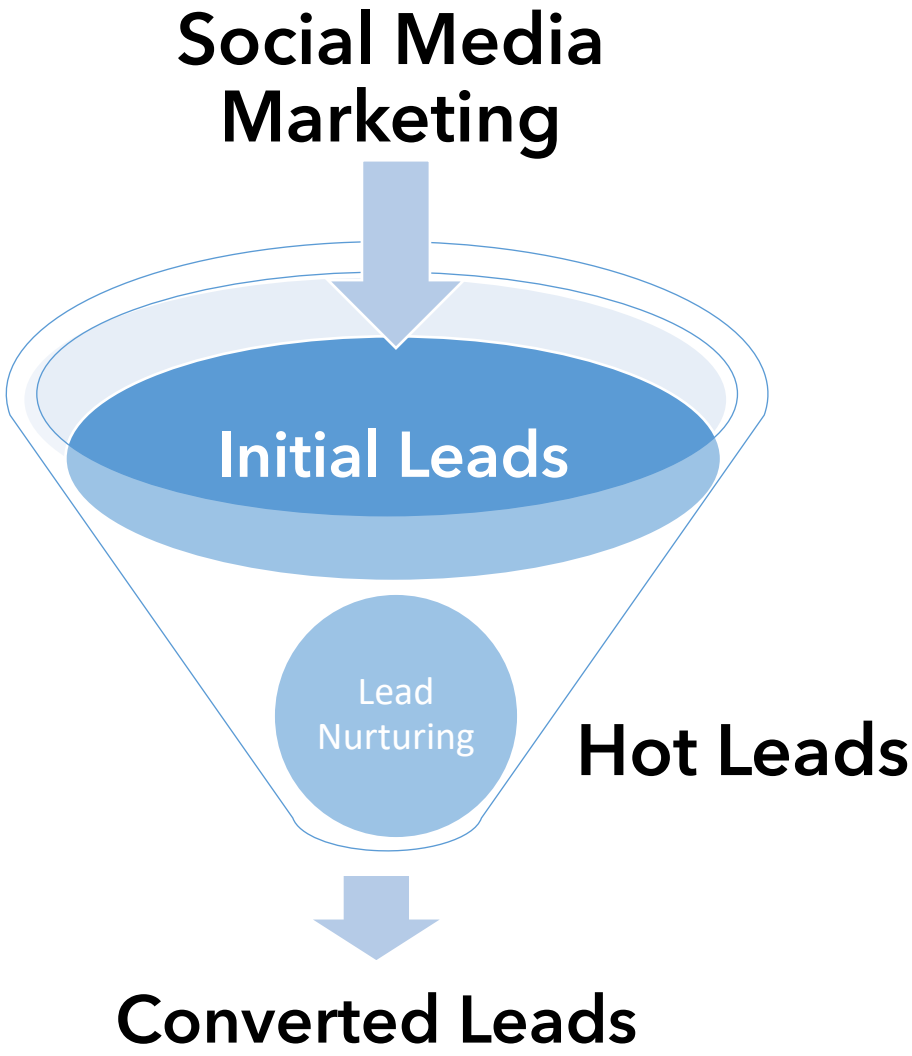
X EDUCATION
LEAD SCORING CASE STUDY

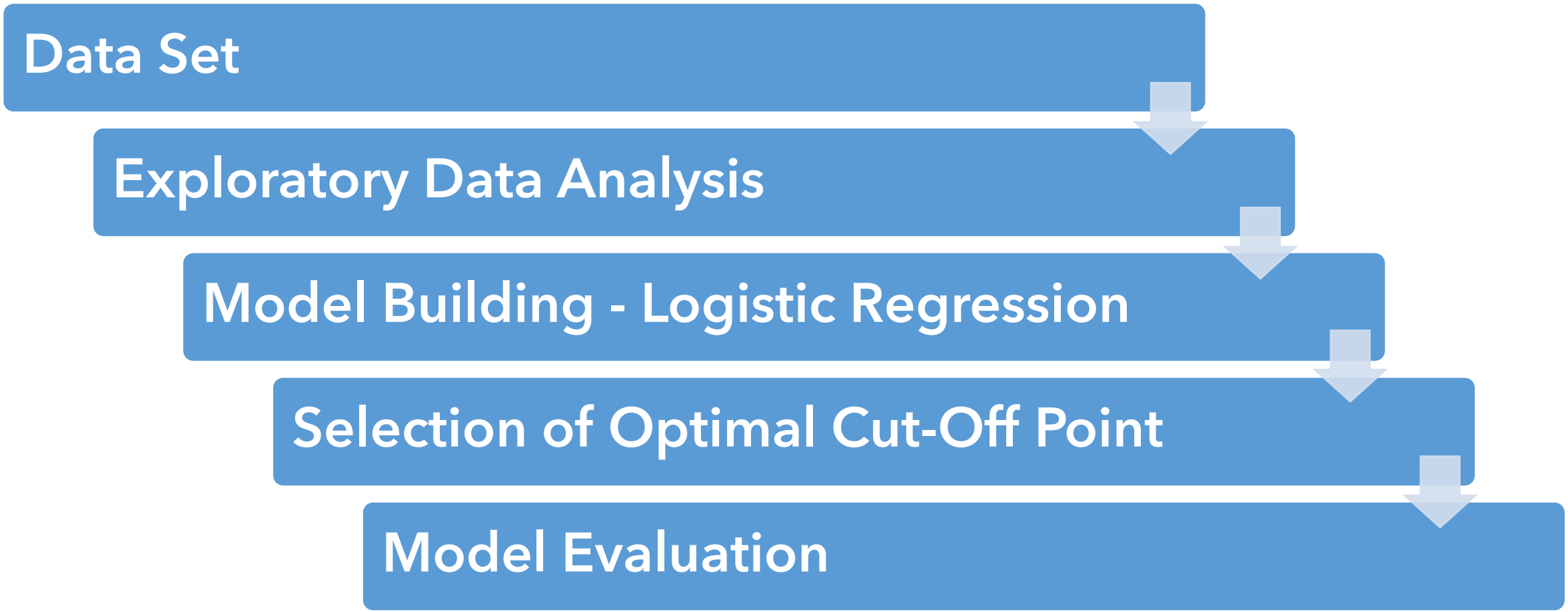
Submission Team:
Amit Raheja
Rajesh Kumar Mishra
Sachita Chauhan
Shailendra Jha

Business Objectives

- Increasing the Lead Conversion rate from around 30% to around 80%
  - Current Lead conversion is around 30%
  - Building the right model to identify and classify the most potential leads tagged as "Hot Leads"
  - The conversion rate from the "Hot Leads" should be around 80%
- The model should be adjustable to include company's requirement changes



Solution Approach



Data Understating and Manipulation

- Lead Data Set provided for analysis: 9240 rows and 37 columns.
- Converting “Select” to “Null” (“Select” implies that user has not selected any value)
- Drop columns which have only one Unique value, OR the columns which have very less variation.
- The outlier treatment is not done, as it can impact Lead Conversion
- Imputing the “NULL” across different variables, with respective values
- Converting categorical variables to dummy features
- Test - train split of the data
- Data normalization, and dropping variables in scenarios showing highly co-related variables

Logistic Regression Model

- After EDA, Logistic Regression Model is built in python using **GLM()** function, under statsmodel library.
- The model contained all the variables, some of which had insignificant coefficients
- Such variables are removed using
  - Automated Approach: RFE (Recursive feature elimination) with number of features = 25.
  - Manual approach based on VIFs and p-values.
- The final tally of variables with their respective values
  - Significant p-values near to zero
  - VIFs < 3

S. No.	Feature	coef	std err	z	P>|Z|	VIF
1	const	-1.9025	0.099	-19.181	0	5.43
2	What is your current occupation_UNEMPLOYED	1.4457	0.111	13.035	0	2.07
3	Tags_WILL REVERT AFTER READING THE EMAIL	3.8817	0.182	21.293	0	1.99
4	What is your current occupation_WORKING PROFESSIONAL	1.8109	0.332	5.459	0	1.75
5	Tags_RINGING	-4.1159	0.236	-17.43	0	1.6
6	Tags_CLOSED BY HORIZZON	6.3264	1.009	6.267	0	1.23
7	Last Notable Activity_SMS SENT	2.4448	0.126	19.433	0	1.23
8	Last Activity_OLARK CHAT CONVERSATION	-1.489	0.214	-6.965	0	1.22
9	Tags_INTERESTED IN OTHER COURSES	-2.5135	0.332	-7.572	0	1.22
10	Lead Quality_WORST	-3.1879	0.541	-5.89	0	1.21
11	Lead Origin_LANDING PAGE SUBMISSION	-0.5348	0.105	-5.093	0	1.17
12	Tags_NOT DOING FURTHER EDUCATION	-3.083	1.038	-2.969	0.003	1.15
13	Tags_SWITCHED OFF	-4.5531	0.525	-8.666	0	1.14
14	Lead Source_WELINGAK WEBSITE	3.2842	0.736	4.46	0	1.07
15	Last Activity_EMAIL BOUNCED	-1.2627	0.356	-3.543	0	1.06
16	Tags_INTERESTED IN FULL TIME MBA	-2.3305	0.729	-3.196	0.001	1.06
17	Tags_INVALID NUMBER	-4.3231	1.033	-4.186	0	1.05
18	Last Activity_CONVERTED TO LEAD	-1.5017	0.327	-4.594	0	1.05
19	Tags_LOST TO EINS	6.7975	0.834	8.153	0	1.04
20	Tags_OPP HANGUP	-2.553	0.812	-3.144	0.002	1.02

Model Evaluation and Optimization

ROC Curve

- ROC Curve demonstrates
  - Tradeoff between sensitivity and specificity
  - Closer the curve follows the left-hand border and then the top border of ROC space, the more accurate the test
  - Closer the curve comes to 45° diagonal of the ROC space, the less accurate the test
- For our model, ROC curve is towards the upper left corner, and area under the curve is more as displayed in Fig 1. Thus, our model is an optimal choice to move forward with the analysis

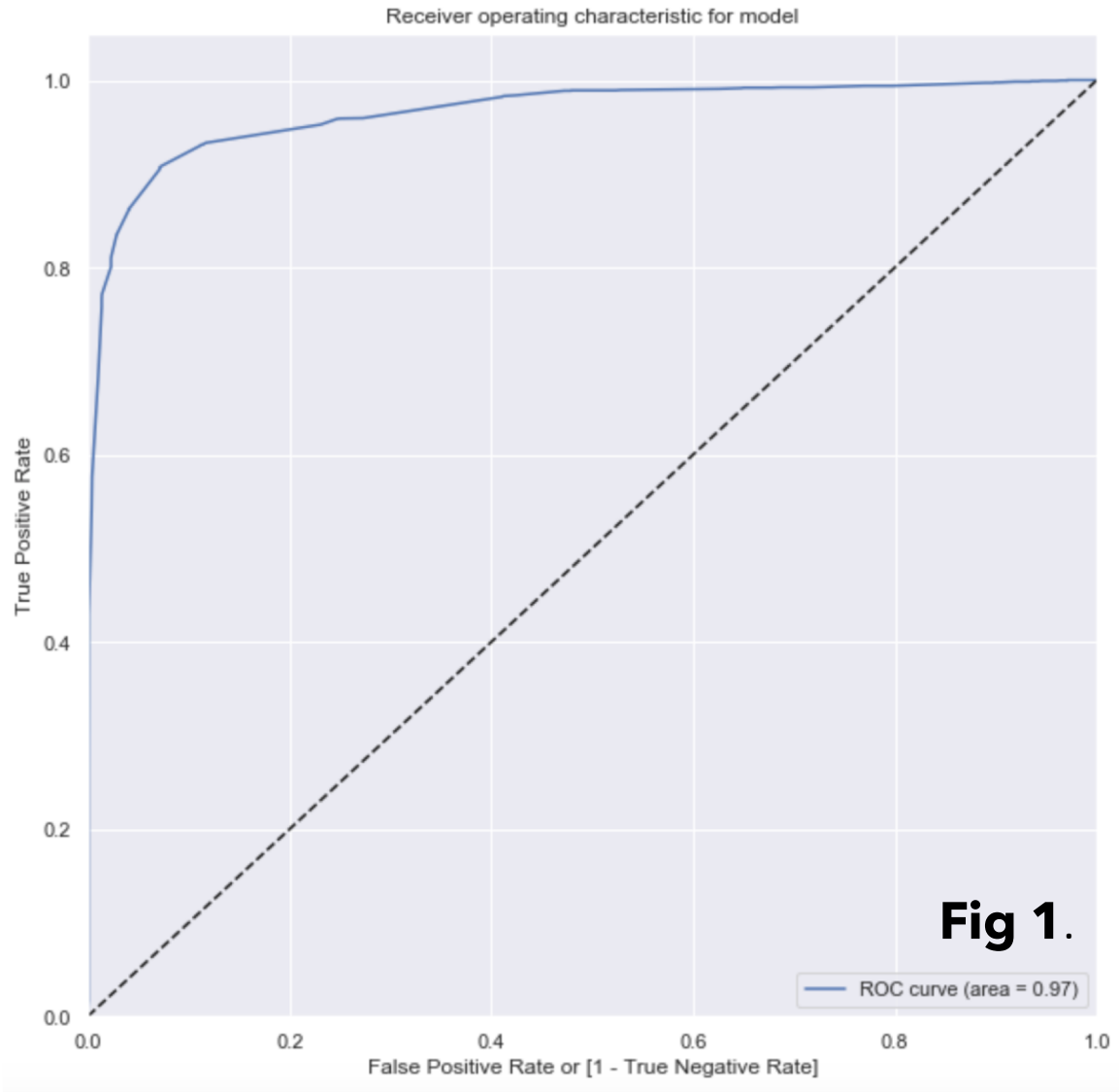


Fig 1.

Finding optimal cut off point

- Plotting accuracy, sensitivity and specificity for various probabilities. Fig 2.
- Cut-Off point is ~0.28, where all three coincide, resulting:
  - Specifity: ~92%
  - Sensitivity: ~90%
  - Accuracy: ~91%
  - False Positive rate: ~7%
  - Positive Predicted Value: ~88%
  - Negative Predicted Value: ~94%
- Precision: ~88%
- Recall: ~90%

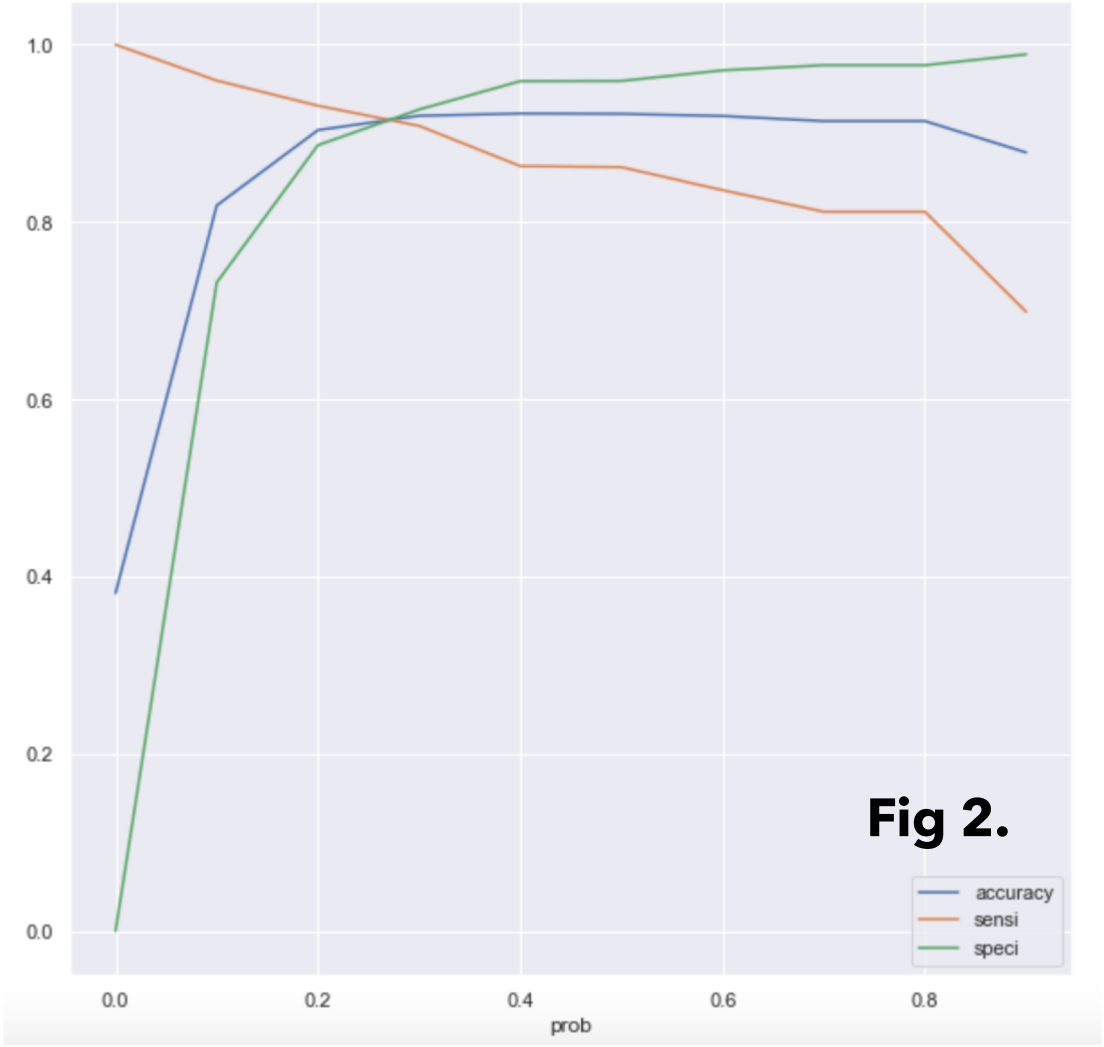


Fig 2.

Finding optimal cut off point

- As the requirement is to select the most promising leads, i.e. Leads that are most likely to convert in paying customer. Thus,
  - We want to increase the accuracy and precision
  - Plotting accuracy, positive predicted value and negative predicted value for various probabilities. Fig 3.
- Cut-Off point is ~0.39, where all three coincide, resulting:
  - Specifity: ~95%
  - Sensitivity: ~86%
  - Accuracy: ~92%
  - False Positive rate: ~4%
  - Positive Predicted Value: ~92%
  - Negative Predicted Value: ~91%
- Precision: ~92%
- Recall: ~86%

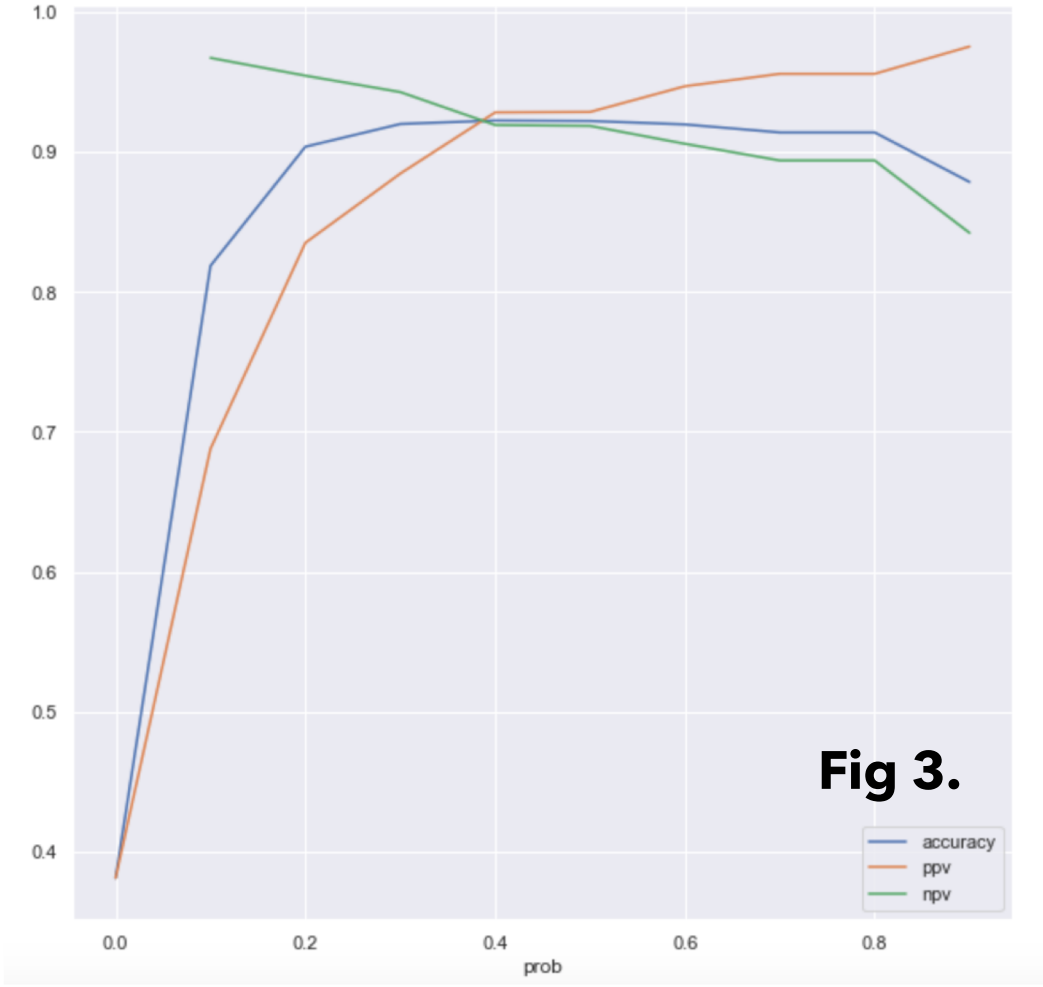
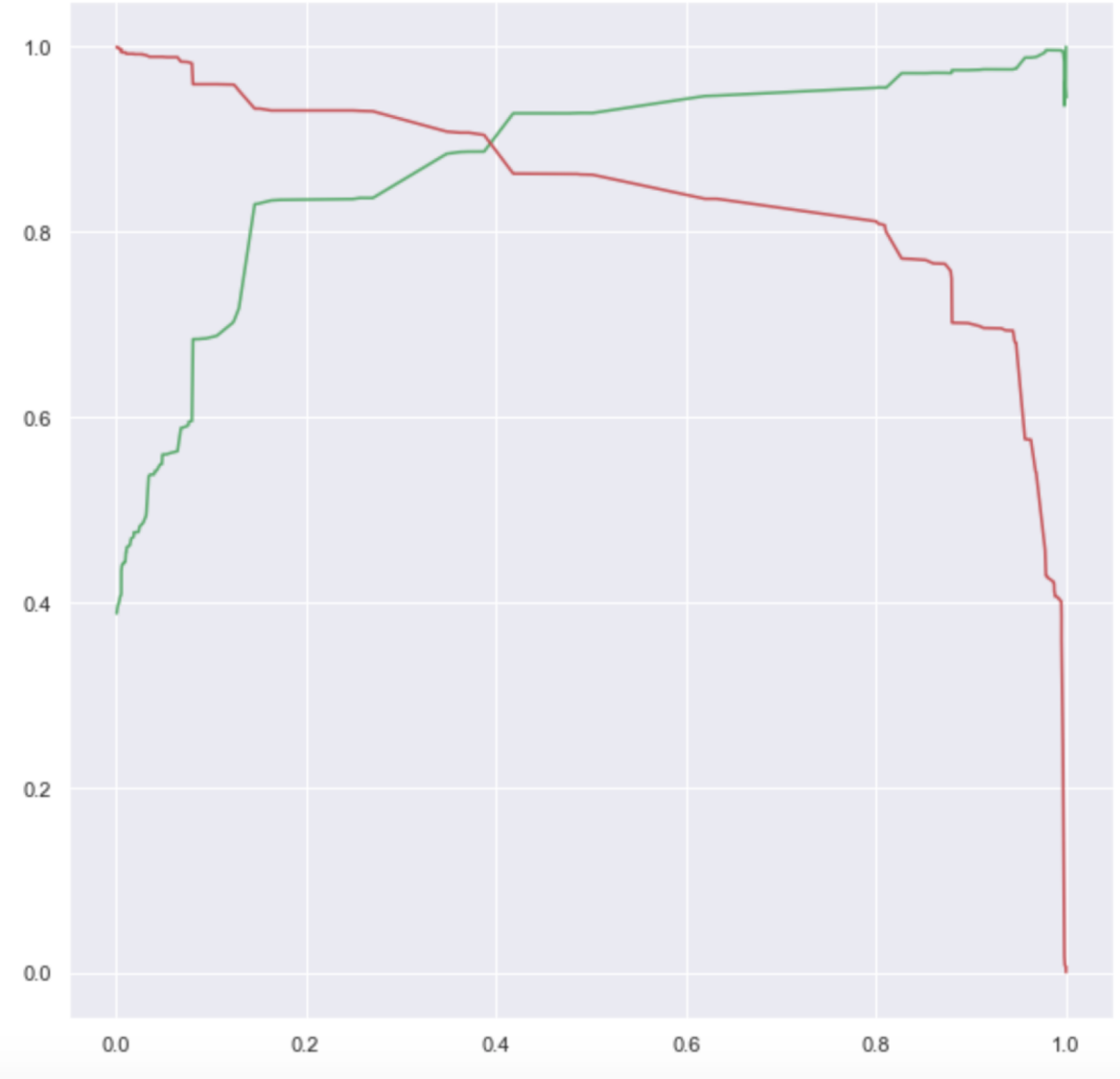


Fig 3.

Model Evaluation: Precision and Recall

- As per business requirement, we have chosen 0.39 as a Cut-Off value, which gives better results for both accuracy and precision
- Accuracy: ~92%
- Precision: ~92%
- Recall: ~86%
- The graph shows a trade-off between Precision and Recall



Prediction on Test Data

- With a chosen cut-off value of 0.39; we get below results:
  - Specifity: ~95%
  - Sensitivity: ~88%
  - Accuracy: ~92%
  - False Positive rate: ~4%
  - Positive Predicted Value: ~92%
  - Negative Predicted Value: ~92%

Actual/Predicted	Not Converted	Converted
Not Converted	1597	80
Converted	131	964

- As Precision (positive predicted value) > 80%, we can use the same model for achieving our objective of increasing the conversion rate

Top 3

- **Top 3 most contributing variables towards the probability of a lead getting converted**
  - Tags
    - CLOSED BY HORIZZON
    - LOST TO EINS
    - WILL REVERT AFTER READING THE EMAIL
  - Lead Source
    - WELINGAK WEBSITE
  - Last Notable Activity
    - SMS SENT

- **Top 3 categorical/dummy variables in the model which should be focused the most to increase the probability of lead conversion**
  - Tags_LOST TO EINS
  - Tags_CLOSED BY HORIZZON
  - Tags_ WILL REVERT AFTER READING THE EMAIL

Additional Use Case-1

- Interns are hired for a period of 2 months every year. The sales team, in particular, has around 10 interns allotted to them. So during this phase, target is to make lead conversion more aggressive. So, almost all of the potential leads (i.e. the customers who have been predicted as 1 by the model) are targeted to be converted. Hence, phone calls are to be made to as much of such people as possible. For this, following is applicable:
  - For this problem we can choose a low Cut-Off point of 0.28, which gives a good result for accuracy, sensitivity and specificity.
  - Though, the precision is less; but as this cut-off includes more people that can be called for conversion and since we have a large team we can take follow up for maximum people.
  - Rest all results for this cut-off value is shown in Slide 8 (Finding Optimal Cut-Off Point)

Additional Use Case-2

- At times, the company reaches its target for a quarter before the deadline. During this time, the company wants the sales team to focus on some new work as well. So, during this time, the company's aim is to not make phone calls unless it's extremely necessary, i.e. they want to minimize the rate of useless phone calls. For this, following is applicable:
  - For this problem we can choose people who has probability between 0.39 and 0.6 because these are the people who has significant chances of conversion but need follow up. So these people need more follow up to make it convert as compared to people who has high probability. Since we need to make limited calls in order to support the deals, we can call only those people whose probability is wavering close to cut-off.

Conclusion

- The model is prepared for prediction of the conversion of the leads. The probability values are generated by the model. The cut-off decided for the model is 0.39. All leads whose probability is generated above this threshold value can be classified as Hot Lead.