Summary Report

The steps involved in solving the lead scores case study is listed below:

Data Understanding and Data Cleaning:

- The data is loaded into a data frame and is inspected for null values, shape and summary statistics.
- It's observed that some of the columns have '**Select**' as an input. This has been considered as a null value and null percentages are calculated across all columns.
- The data frame is treated to remove null values. Columns having no importance have been dropped.
- Outlier treatment is performed on continuous variables.

Data Preparation:

- Created dummy variables for categorical variables.
- Binary mapping has been carried out for columns with Yes/No values.

Test Train Split:

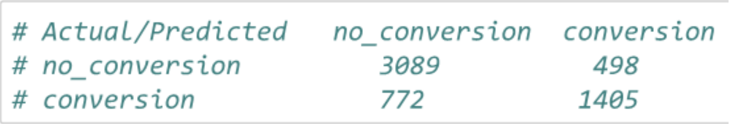
- Splitting the data frame into testing set and training set.
- Creating data frames for the target and response variables.

Feature Scaling:

- Standard scaler is applied on continuous variables to bring all the variables onto a common scale.

Building a Logistic Regression Model:

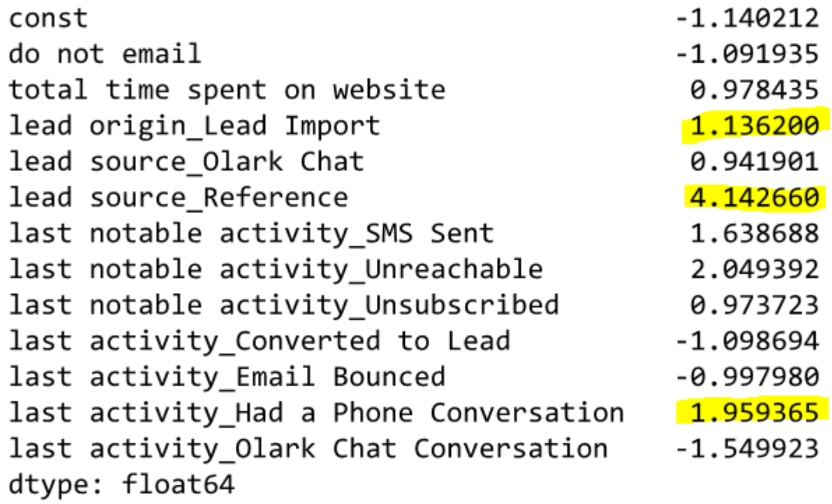
- A logistic regression model is built using all the variables.
- RFE is applied for the process of feature elimination to obtain top 15 features.
- Manual elimination is performed to identify the significant variables with least correlations. The final model consists of 12 features and the model accuracy is computed by manually selecting 0.5 as the cut-off probability.
- Confusion Matrix is obtained as follows



```
# Actual/Predicted   no_conversion   conversion
# no_conversion          3089            498
# conversion             772            1405
```

Identifying Top 3 features:

- The top 3 features which would increase the probability of lead conversion have been reported based on the coefficient values.

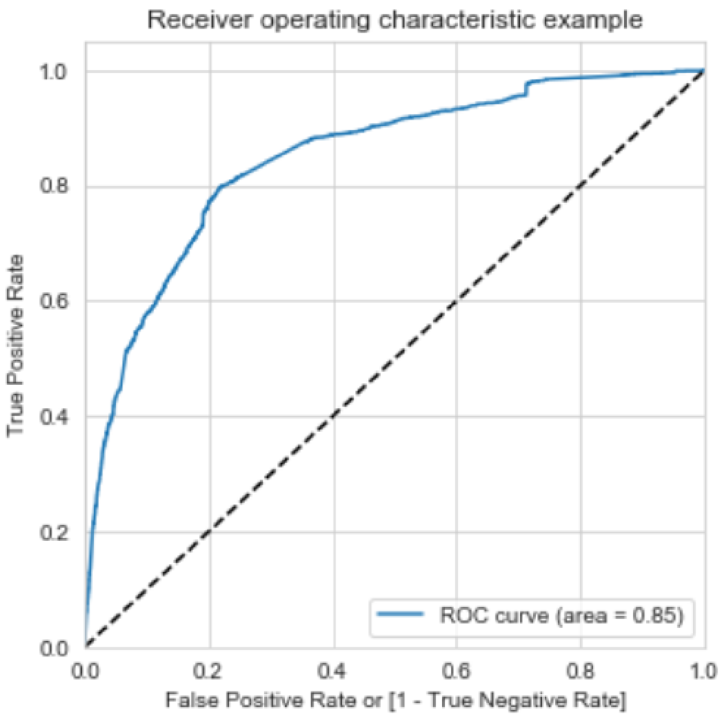


```
const                                   -1.140212
do not email                            -1.091935
total time spent on website              0.978435
lead origin_Lead Import                  1.136200
lead source_Olark Chat                   0.941901
lead source_Reference                    4.142660
last notable activity_SMS Sent           1.638688
last notable activity_Unreachable        2.049392
last notable activity_Unsubscribed       0.973723
last activity_Converted to Lead         -1.098694
last activity_Email Bounced             -0.997980
last activity_Had a Phone Conversation   1.959365
last activity_Olark Chat Conversation   -1.549923
dtype: float64
```

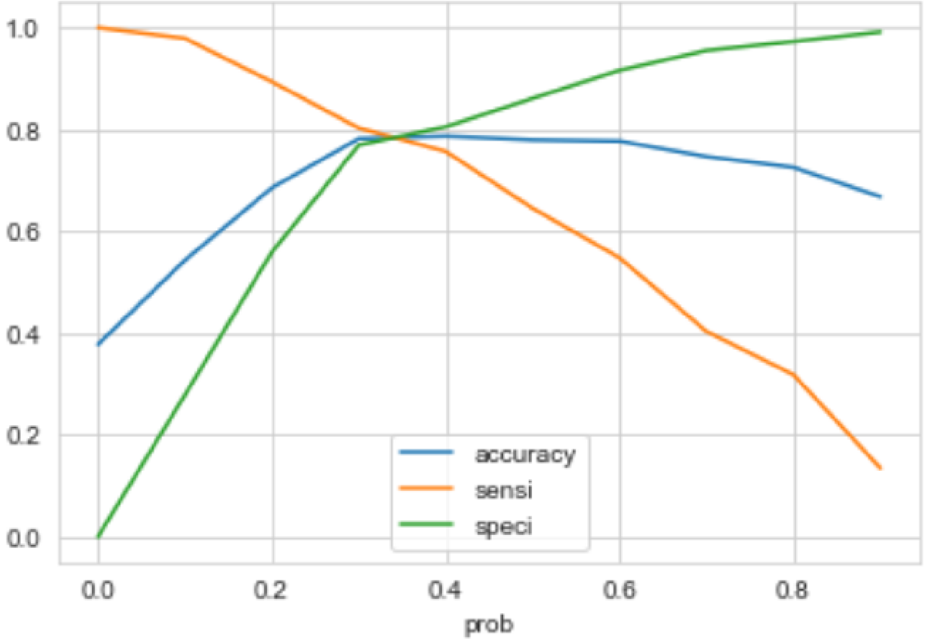
- Top 3 categorial features are also reported. These are same as the top 3 features.

Metrics beyond Accuracy:

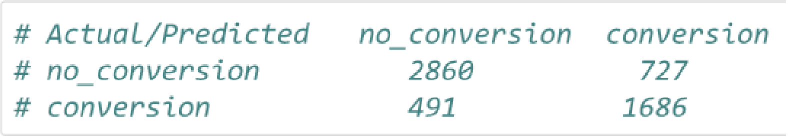
- Sensitivity, specificity & false positive rates are computed for conversion probability >= 0.5.
- ROC curve is plotted. Area under the curve is 0.85.



- Identified optimal cut-off point to be 0.35 by plotting sensitivity, specificity and accuracy for various probabilities.



- Recomputed confusion matrix, accuracy, specificity and false positive rate for the optimal cut-off.



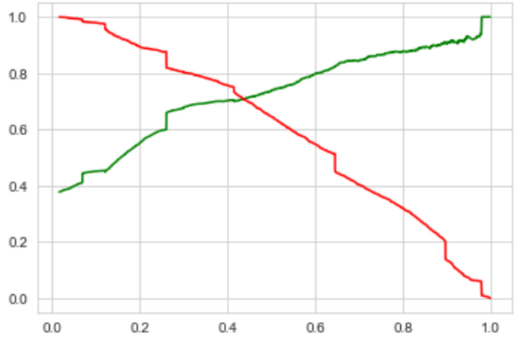
```
# Actual/Predicted     no_conversion    conversion
# no_conversion            2860            727
# conversion                491           1686
```

Metric results

Metric	Result
Accuracy	78%
Sensitivity	78%
Specificity	80%
False positive rate	21%

Precision and Recall:

- Precision and Recall trade-off curve are plotted and the optimal cut-off probability is 0.44.
- Precision and recall values are computed for the optimal cut-off of 0.44.



Metric results at cut-off point 0.44

Metric	Result
Precision	71%
Recall	70%

Solution to Business Objective:

- Objective is to have 80% conversion i.e. Of all the hot leads that the model generates 80% of them should convert i.e. the precision of the model should be 80%.
- Lead Score is calculated as (conversion probability) *100.
- For a lead score >=61 or conversion probability >=0.61, we have a precision of 80%.

Train Dataset Result

Cut-off Point	Precision	Recall
0.44	71%	70%
0.5	74%	65%
0.6	79%	55%
0.61	80%	54%

Making Predictions on Test Set:

- For a lead score >=61 or conversion probability >=0.61, the model precision on the test set is 77.5% which is pretty close to the conversion rate we are aiming for.

Test Dataset Result

Metric results at cut-off point 0.61	Result
Precision	78%
Recall	52%

Addressing Problems presented by X Education:

- **Strategy for aggressive lead conversion:** Here, the goal is to increase the sensitivity of the model so that all the potential leads can be identified as hot leads. Let's assume that we need to identify at least 90% of the actual leads i.e. the sensitivity is 90% so that conversion is maximum.

  We would recommend X Education sales team to contact customers with lead score >=19 as this would enable them in reaching out to 90% of the potential leads.

- **Strategy to minimize rate of useless phone calls:** Here, our goal is to increase the conversion rate i.e. the precision of the model. From the original business objective, an 80% conversion is achieved if we target customers with lead score >=61 or conversion probability >=0.61.

  We would recommend X Education sales team to only contact customers with lead score >=95. This would result in 92.8% of chance that the person contacted would result in conversion. This way, the sales team could focus on other work as there wouldn't be many customers above a lead score of 95.