The first step was to inspect the provided data. As a preliminary step, we replaced all the 'Select' values in categorical columns with NaN. As part of Exploratory Data Analysis, we did the basic data quality check (checking for duplicates, null value percentage, outlier treatment etc.), univariate and bivariate analysis. Initially, columns with more than 70% values are null were dropped. A few other columns were dropped based on their inferred significance. For ex., Columns where all records have the same value don't contribute any information to the model. Following are the major learnings from the EDA:

- Overall conversion rate - ~38%
- Lead Origin - We need to focus more on improving lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form.
- Lead Source - Focus should be on improving lead conversion of Olark chat, organic search, direct traffic, and google leads and on generating more leads from reference and welingak website.
- Leads spending more time on the website are more likely to be converted. Website should be made more engaging to make leads spend more time.
- Most of the lead have their Email opened as their last activity. Conversion rate for leads with last activity as SMS Sent is almost 60%.
- Working Professionals going for the course have high chances of joining it. Unemployed leads are the most in numbers but has around 30-35% conversion rate.
- Leads tagged as 'Will revert after reading the email' have a very high volume of conversion, and should be pursued aggressively.
- Most leads are from Mumbai with around 30% conversion rate.

Post EDA, dummy variables were created for all the categorical columns. The dataset was then split into train and test data in the ratio 7:3.

Feature scaling was then done using the Standard scaler. We began with RFE to find out the variables which contribute the most to the model. Once we got the columns after carrying out RFE, we built a model using StatsModel. We then iteratively dropped the columns with high P-Values and VIFs, while checking P-Values and VIFs at each step.

We then used the model for doing prediction on the train set. We plotted the Sensitivity-Specificity and Precision-Recall trade-off plots to find the optimal cut-off probability. We settled on a cut-off probability of 0.55 (given by Precision-Recall trade-off). After this, we made predictions on the test set. We also calculated the metrics like Accuracy, False Positive Rate, Positive/Negative predictive value etc.

As a final step, we assigned a lead score by multiplying the predicted probability by 100 and merged the results back to the original dataset.

Following are the metrics from the predictions done on the test data:

Overall Accuracy: 92.71%
Sensitivity: 86.98%
Specificity: 96.12%
False Positive Rate: 3.88%
Positive Predictive Value: 93.04%
Negative Predictive Value: 92.53%
Precision: 93.04%
Recall: 86.98%