Problem:

Increase conversion rate by 80% by only focusing on hot leads (Leads most likely to convert).To build a model to find Hot leads(Leads with very high chance of conversion) using logistic regression and to find important factors to be considered while determining a lead as 'Hot lead'.

Initial data with 9240 records in leads.csv file has 37 columns which include 30 categorical and 7 numerical columns are available
In data cleaning: As ‘Select’ is not a valid class, instead is actually a null value. Several columns in this dataframe which contain 'Select' value have been replaced with a null value to perform the analysis correctly. Dropped the columns having more than 25% missing value because the data size was small.
Visualizing Data and EDA: Helped us identify outliers, and redundant columns

Performing Variance Thresholding:Variance thresholding helps find features that hardly change in data as those features are not useful and they could be removed.

Next is Model Building: Our final model has 13 variables. Our top 3 variables were: ‘Lead Source_Welingak Website’, ‘Total Time Spent on Website’, ‘Lead Source_Reference’. Thus making Lead source our most important categorical variable.

Prediction & Model Evaluation: Accuracy, sensitivity/Recall, Specificity, Confusion matrix and ROC curve were the most important metrics used. We got low recall and high specificity because only for variables with probability greater than 80% does the model predict as positive because we have taken conversion probability cutoff as 80% as the question specifies a requirement for ‘Hot Leads’.
Boat Test and Train Data have similar metrics values.
Possible Solutions based on Model to increase conversion rate:

- The biggest influencing factor in the list of 13 factors we have considered in our model is ‘Lead Source_Welingak Website’. Which means that people who reach the company through this website have a higher chance of converting, so the company could invest in this avenue by increasing advertisements to this website.
- The next important contributing factor is ‘Total Time Spent on Websites’ which is understandable. The more time a customer spends on the website , the more the customer is interested in the program. So the sales team could target customers who spend more time on the website.
- The next important factor is ‘Lead Source_Reference’. If a person is referred to the program by someone they know and trust they seem to join the program
- Factors like ‘Do Not Email’,’ Last Activity_Olark Chat Conversation’, have negative coefficients thus these factors contribute to the customer not getting converted to lead. By doing root cause analysis we could find possible causes and eliminate the same.

Learning Outcome:
In this Case study we learned:

- About framing problem statements, analyzing problems & Solution and Implementing solutions.
- We learned how to build a logistic regression model after cleaning data and data visualization.
- We learned how to handle null equivalent values (Select).
- How to analyze categorical variables.
- We learned Performing Variance Thresholding.
- We learned how to evaluate logistic regression model
- And we learned how to prioritize solutions and factors.