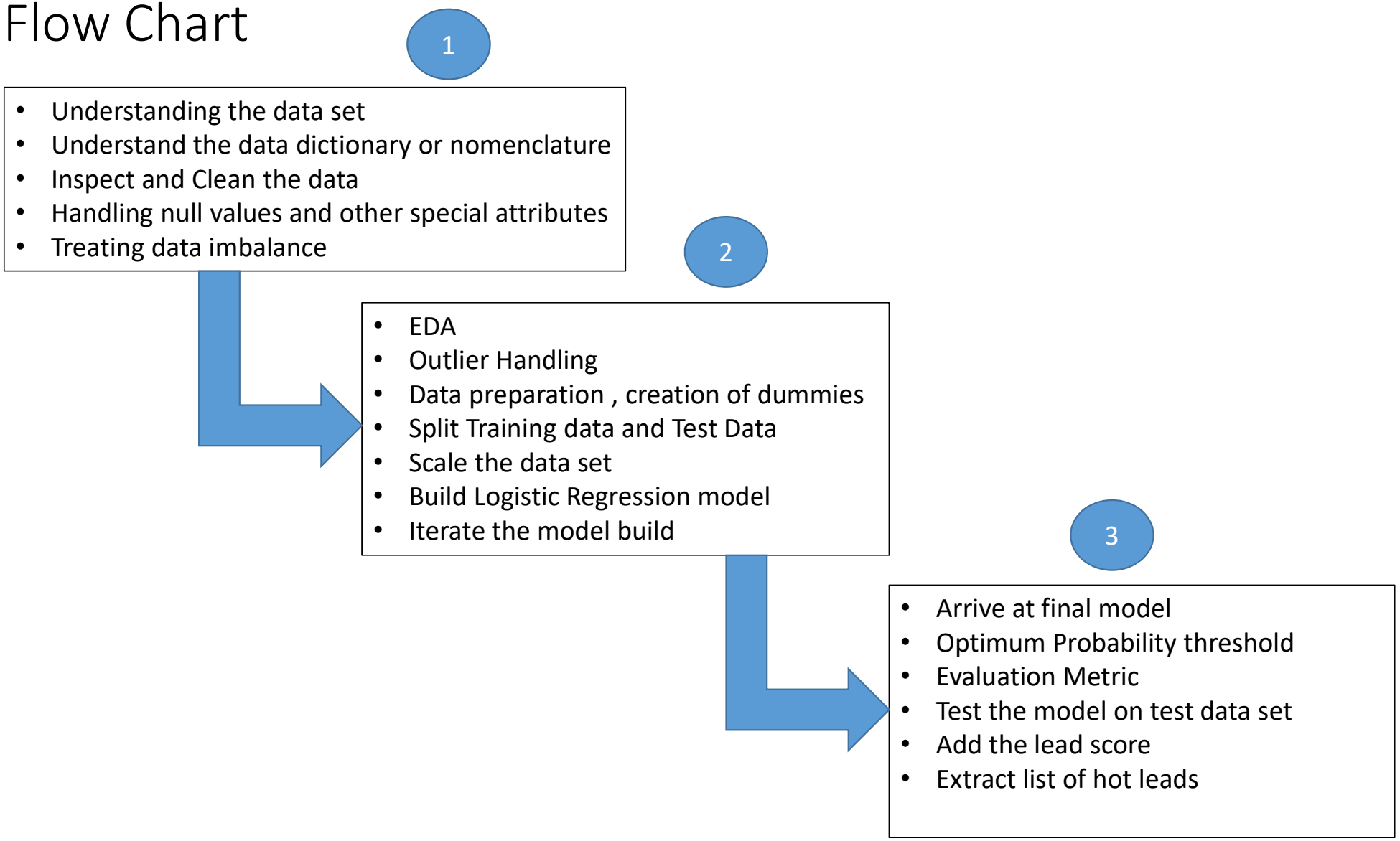
Summary Report

-Vinay Patil

-Dheeraj Annapragada

-Shiny Garikimukku

Flow Chart


1
Understanding the data set
Understand the data dictionary or nomenclature
Inspect and Clean the data
Handling null values and other special attributes
Treating data imbalance
2
EDA
Outlier Handling
Data preparation , creation of dummies
Split Training data and Test Data
Scale the data set
Build Logistic Regression model
Iterate the model build
3
Arrive at final model
Optimum Probability threshold
Evaluation Metric
Test the model on test data set
Add the lead score
Extract list of hot leads

STAGE -1

Stage 1 – Data Understanding

- The Data in the form of csv file
- It has 37 features and each feature has 9240 entries
- The data has categorical variables, continuous variables, of demography, Metric scores like activity score, profile score
- The data has answers to subjective questions, mode of contact , response to email/call etc
- It has details of source/origin of the lead etc
- It has a feature **“Converted”** which is the target variable here 0-not converted ; 1-converted

Handling “Select” and Null values

- “Select” is as good as null; because the user has not chosen any option in drop down of the online form; it’s the default tab that has been carried into our data
- There were features with the level “Select” those were imputed with null values
- Features having null values > 40 % are discarded
- Another attribute “Not Specified” was created for null values for “City” and “Specialization” feature
- Imputed the null values with mode for categorical variable and median for continuous variable

STAGE -1

Handling Data Imbalance

- There were quite a few features having data imbalance ex: “Do not email", "Do not call” etc

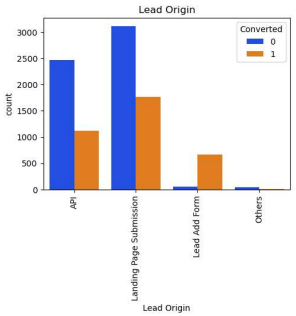
- It is important to handle these features, we decided to eliminate these feature columns, as these features would impact our prediction to sway in one direction

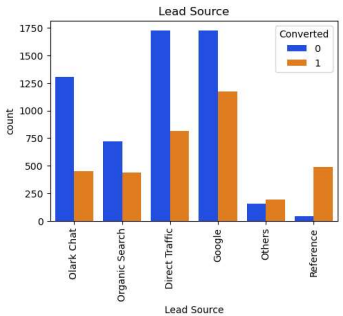
- There are feature like “Last Notable activity” and “Tags”, these are the responses collected after the marketing executives already contacted the leads; these kind of data which deals with customers/Students and not the leads are not taken for analysis

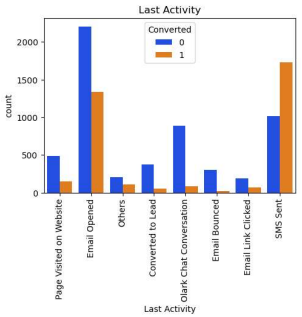
STAGE -2

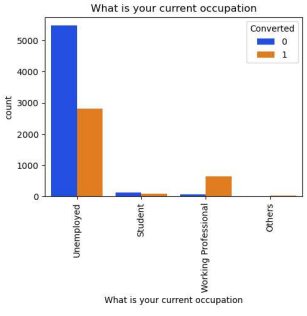
EDA to First ML Model

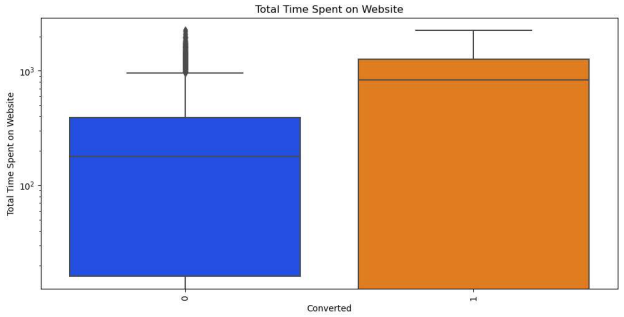
- At the start of EDA we had 11 independent variables
- EDA inference were made for lead orgin,lead source,last activity and Occupation
- Outlier were capped for 99 % quantile; Dummies were created for categorical variables
- Training/Test split in 70-30 proportion
- First Logistic regression model was built and P value and VIF were observed
- Model was iterated using RFE once to bring down the number of features from 33 to 20
- Later it was manual elimination , while monitoring the p < 5% and VIF < 5%











STAGE -3

Final ML Model to “Hot” Leads

- Final Logistic regression model had 15 features
- Area under ROC curve was 0.88 , which is high and preferred
- Optimum threshold probability was arrived through intersection of sensitivity, specificity and accuracy curve and it came to around 0.35
- Accuracy of Train data set was 81% and Accuracy of Test data set was 80 %, indicating a good prediction
- Finally we assigned the “lead score” for each lead which is the conversion probability in percentage