Summary report-

1) We started with understanding the data by checking the data types of all columns and then a basic clean up.
2) Then we saw a lot of categorical fields that had a value 'Select', which looks like some kind of default value. This could be treated as Null so we replace all cells having 'Select' with Null.
3) Then we checked the missing values in all fields. We categories fields in three different pattern- i) Columns having %nulls < 25, ii) columns having %nulls between 25 and 40, and iii) columns having %null >40.
    a. For columns having %nulls > 40 → We simply dropped all such fields.
    b. For columns having %nulls between 25 and 40 → All of them were categorical variables. And, we treated Null as an information. We filled Null with 'NA'.
    c. For columns having %nulls < 25 → We treated such columns by filling nulls with mode/50th percentile value.
4) We then checked the outliers in the numeric columns.
5) We started analyzing data by performing EDA. Found that most of the yes-no fields are singular in nature, i.e. they mostly contains only one value.
6) We then wrapped all the transformation steps into functions, which could help us to transform train and test data as per our need with a very minimal code required.
7) We used LabelEncoder() to encode categorical variables.
8) We then performed model training. We used statsmodel's Logit class to train the model.
9) By checking the p-score and VIF, we performed feature selection.
10) We then finally performed an exercise to find the optimal cutoff point.