Approach Document

Steps:

1. **Clean the train data. Notice that there are good number of "NAN" for "Credit_Product" feature.**
2. **Label all NaNs to "Unknown" literal string.**
3. **Using the following columns Gender, Age, Region_Code, Occupation, Channel_Code and Vintage I have created a new column known as "new_cluster"**
4. **Then encoded this "new_cluster" column to get category codes.**
5. **The aim was to replace each of the "Unknown" values of "Credit_Product" column with the mode of the cluster which they closely belong to.**
6. **The whole aim of dividing the whole dataset was to club closely related subsets into their respective cluster.**
7. **Then, aim was to find the mode within the cluster and replace the values of all the "Unknown" with the cluster mode.**
8. **This has helped reduced the number of "Unknown" if not eliminated them completely.**
9. **Next aim was to quantize (turn the categorical feature values to numerical values).**
10. **For quantization, I first did One-Hot encoding.**
11. **Then, executed XGBoostClassifier to generate the coefficients for each of the encoded features.**
12. **Once the coefficients were found, I am replacing each of the categorical feature values with their coefficient values.**
    Example, if coefficient if "Female" Gender was 0.25732, then in original train dataset, all the "Female" values will be replaced with 0.25732.
13. **This quantization has been done for all the categorical features in the original dataset.**
14. **After the above exercise, I am left with all NUMERICAL features (keeping aside, Is_Lead and ID columns).**
15. **I am then normalizing them all and bringing them between 0 and 1.**
16. **After the normalization process, I am training my dataset using some classifiers.**
17. **Tried classifiers were LogisticRegressionCV, RidgerRegressionCV, AdaBoostingClassifier, RandomForestClassifier, DecisionTreeClassifier etc.**
18. **For test data, I quantize and normalize my test data features and apply the PREDICT method of the model.**