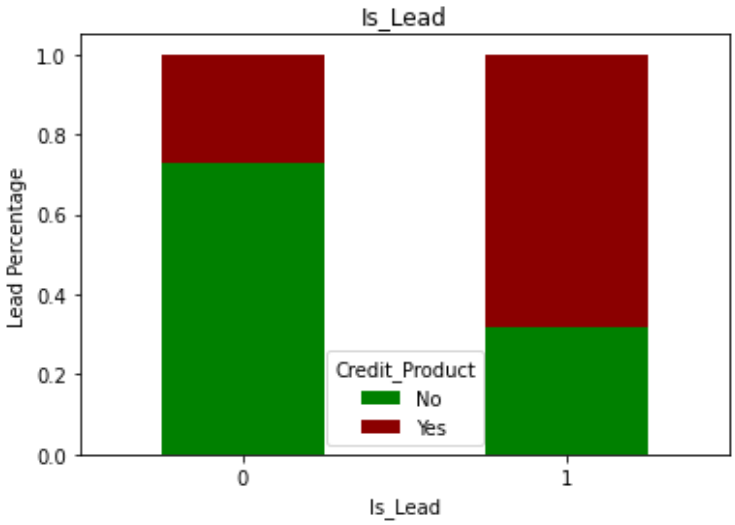
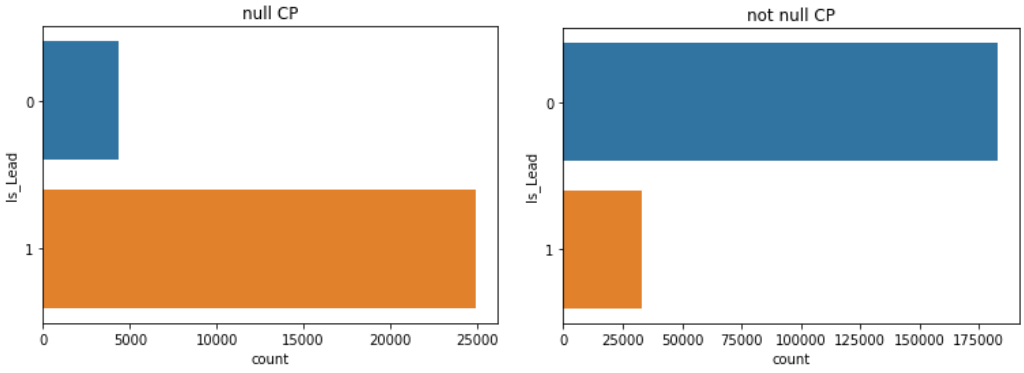
Credit Card Lead Prediction

1. **EDA**
    a. Imbalanced class label was found through countplot of the dependent variable-**'Is_Lead'**
    b. Credit_Product has missing values.
    c. **85%** of Missing Credit_Product records are interested in credit card.

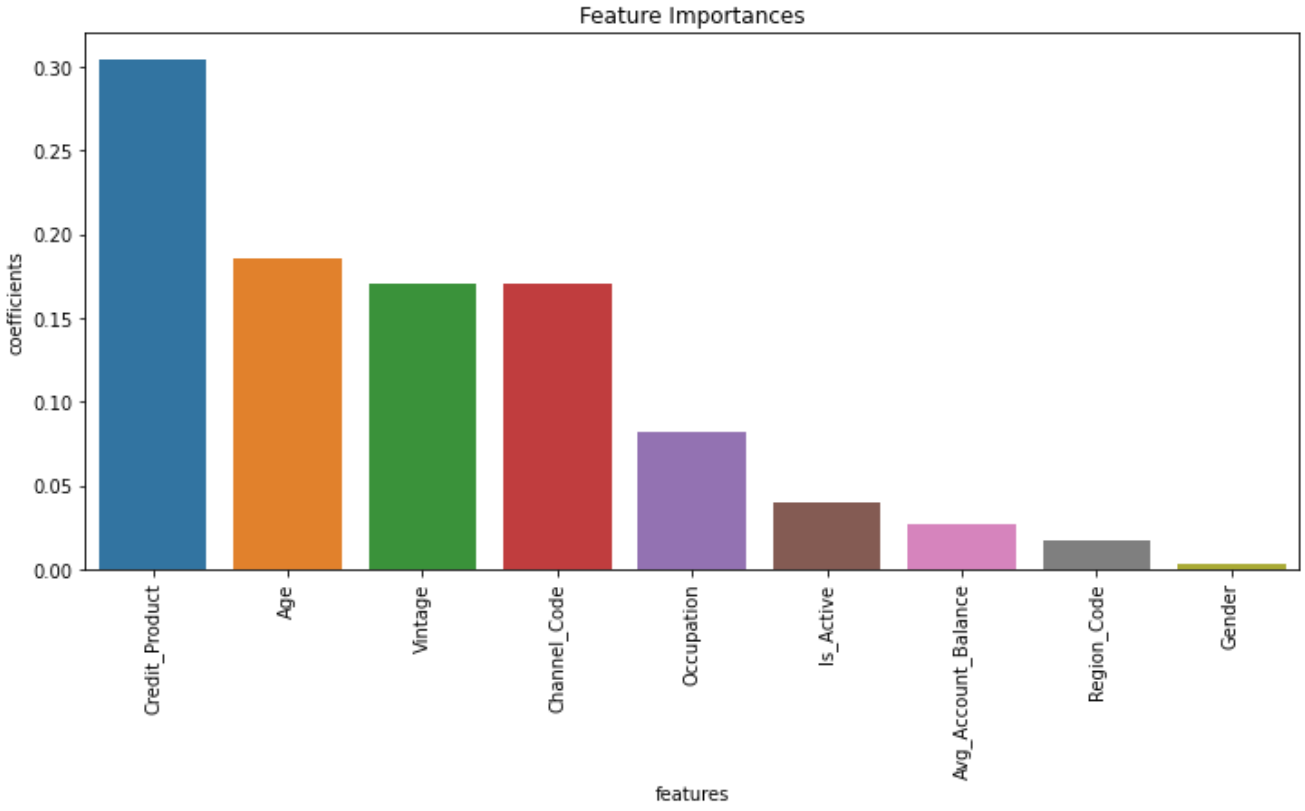




    d. **Credit_Product** is an important feature to predict Is_Lead based on it's correlation coefficient. Thus imputation is highly necessary.
    e. Customers above **Age 43** are interested in credit card, thus, age is also an important feature.
    f. Through pairplots, the data cannot be used for clustering.
    g. PCA and LDA couldn't help in feature extraction.

h. Many columns are categorical. Remaining columns can be converted to categorical by taking ranges(age, vintage, Avg_Account_Balance). On experimenting this point, no difference in model performance was observed.

2. **Preprocessing:**
    a. Identify and convert all string categorical variables into numerical categorical ones using **LabelEncoder**
    b. Removing missing values gave less **auc.**
    c. Credit_Product is imputed using IterativeImputer and observed better auc score.
    d. Feature importance is found with a base random forest model.
    e. Smote method is applied to handle class imbalance which has increased the roc_auc_score.



3. **Model**
    a. Bayes classifier and logistic was expected to perform better, but overfitting was observed.
    b. LGBM, Gradient descent and ada boost gave equivalent results.
    c. Random Forest model is finalised and trained and tested with intuitive parameters.

d. Auc score obtained = .91