- **Business Problem:**

  A Private Bank offers variety of products like Checking / Savings accounts, investment products, credit products etc. The bank intends to cross-sell products to its existing customer base via different channels like emails, telephone calls, recommendations on online banking interfaces like mobile apps etc. As a part of current business opportunity, bank plans on identifying the potential customers from the list of eligible customers.

- **Background/History:**

  The Financial institutions / banks offer variety of products like bank accounts, credit / lending products, payment methods, investments to generate revenues. Credit cards being one of the widely used products, stand very good chance to continue garnering incremental revenues via increased usage at the various locations as the bank is able to facilitate the payments to merchants / service providers on behalf of card customers and able to charge the fees on such payments.

  Keeping this in mind, the bank is looking to follow the customer leads who would be potentially interested in opening a new credit card account. These customers are existing customers with the bank who are currently utilizing the other services like checking / savings accounts, other credit products etc. Bank has different features like age, gender, channel of connecting with customers, vintage, existing credit product usage, average account balance, active usage etc. as a part of this opportunity, we are looking to predict the potential leads based on various factors so that credit card offerings can be extended to the most interested and eligible customers who are more likely to open the credit card account.

- **<u>Data Explanation (Data Prep/Data Dictionary/etc):</u>**

Following features are present in the datasets:

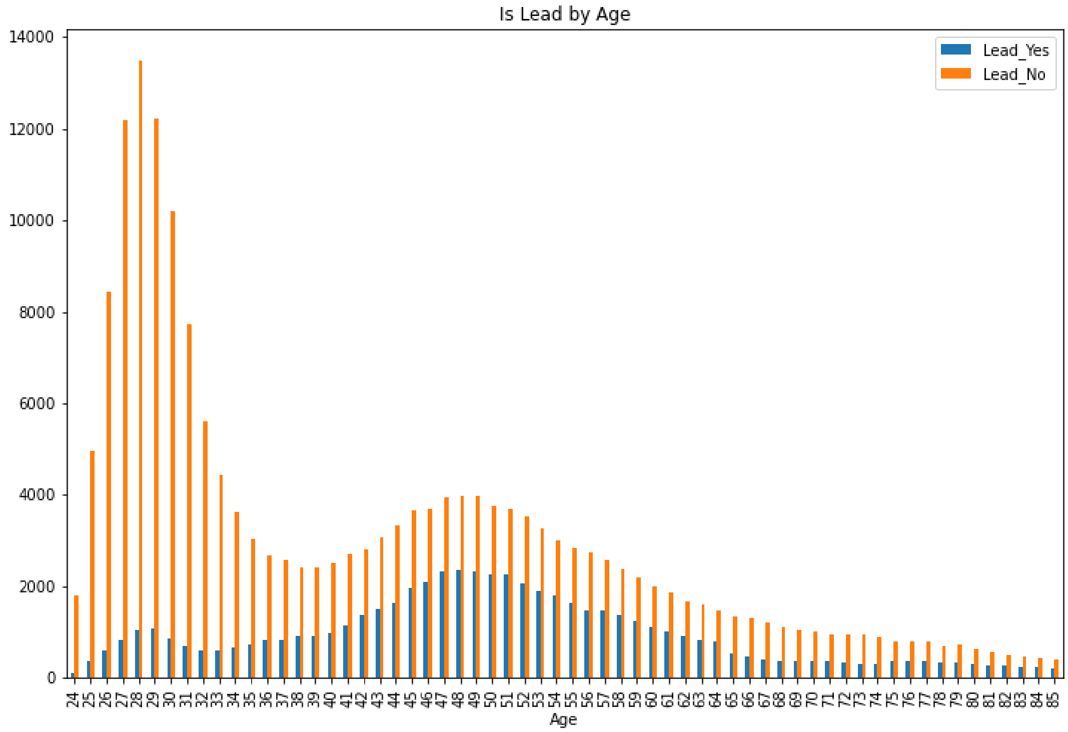
- ID : Customer ID
- Gender : Male / Female
- Age : Age in years
- Region Code : Represents the Region where the customer resides
- Occupation : Occupation status e.g. Salaried, Self-employed, Entrepreneur, Other etc.
- Channel Code : Method of connecting with the customers for credit card product offers
- Vintage : Vintage for the Customers (in months)
- Credit Product : If the customer has any active product (home loan, credit card etc.)
- Avg Account Balance : Average account balance for the past 12 months
- Is active : If the customer is active in last 3 months
- Is lead : (Target variable) if the customer is interested

Training dataset has 245725 records. Out of these, Credit_Product column is the only one which has NULL values for 29325 records in it. We will be dropping those records as a part of data preparation. Test dataset has 105312 records in it and similar to train dataset has Credit_Product column with 12522 records with NULL values. We will be dropping these records for analysis and predictions as well.

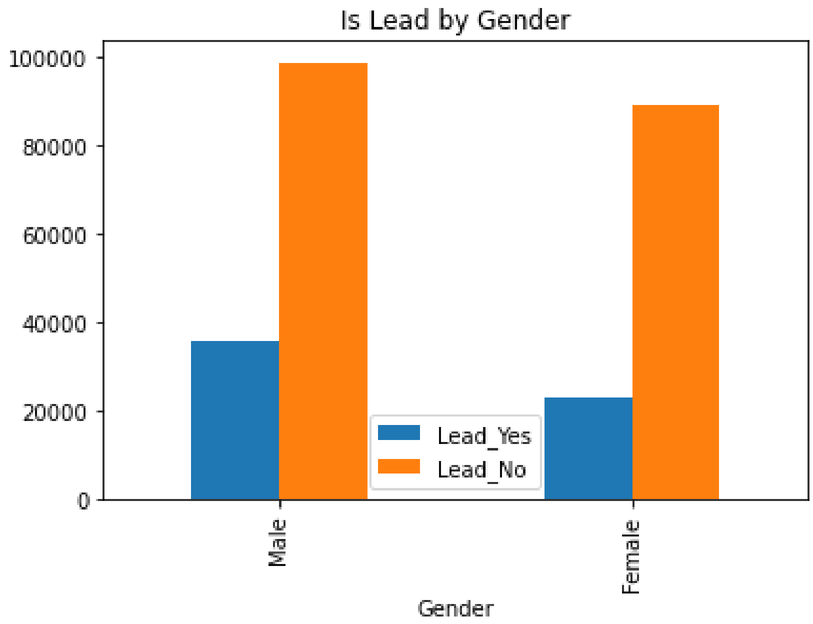
Also, the overall data is imbalanced in terms of Is_Lead column (Target variable) values whereas only 33313 records having value of ‘1.0’ and all the remaining 183087 records with ‘0.0’ as value. So, during data clean up / preparation, we will be creating a balanced training dataset with equal number of data records in each category (0.0 / 1.0). This will ensure that our predictions are unbiased and free of apparent picture of higher accuracy due to such imbalanced data records.

- **Analysis & Methods:**

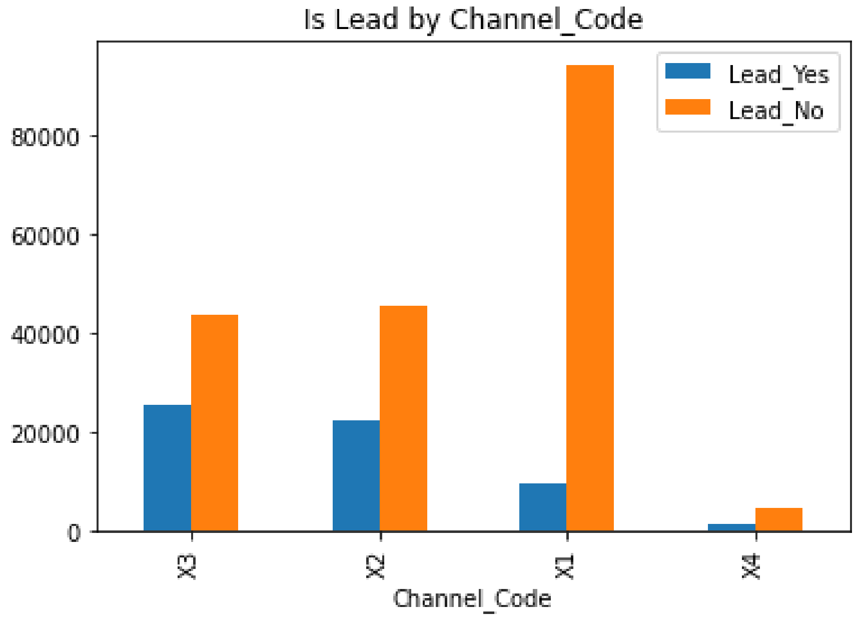
Age distribution indicates fair amount of positive leads (Is_Lead as 1.0) for age groups 27 – 30 years and higher percentage of (Is_Lead as 1.0) in the age group 42 – 58 years. The younger customers tend to have lower positive leads.



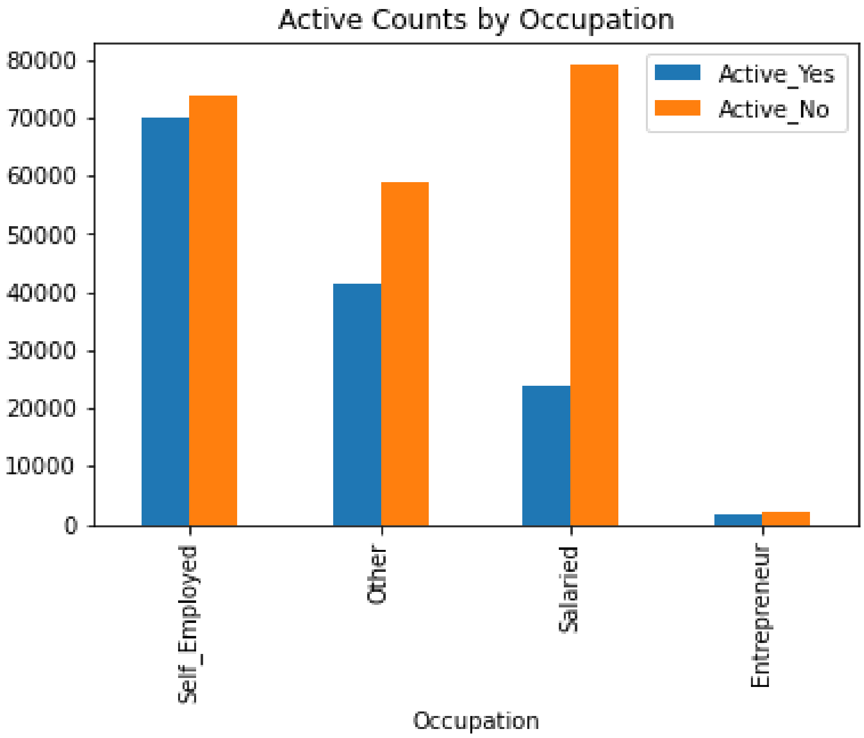
There slightly higher number of Male customers with Is_Lead as 1.0 compared to Female customers.



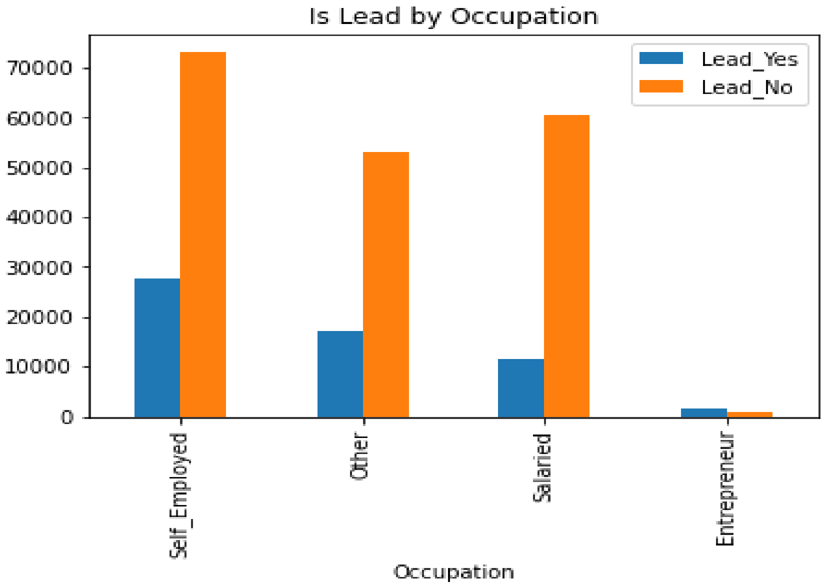
Channel code X3 and X2 have higher number of positive leads (Is_Lead = 1.0)



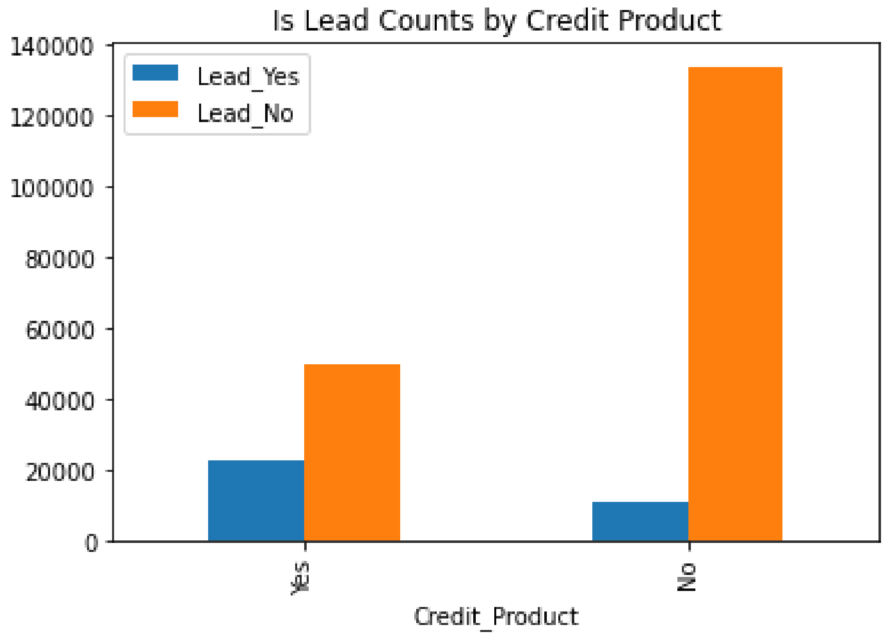
Higher number of Self_Employed and Other occupation along with Salaried customers have Active_Yes indicators



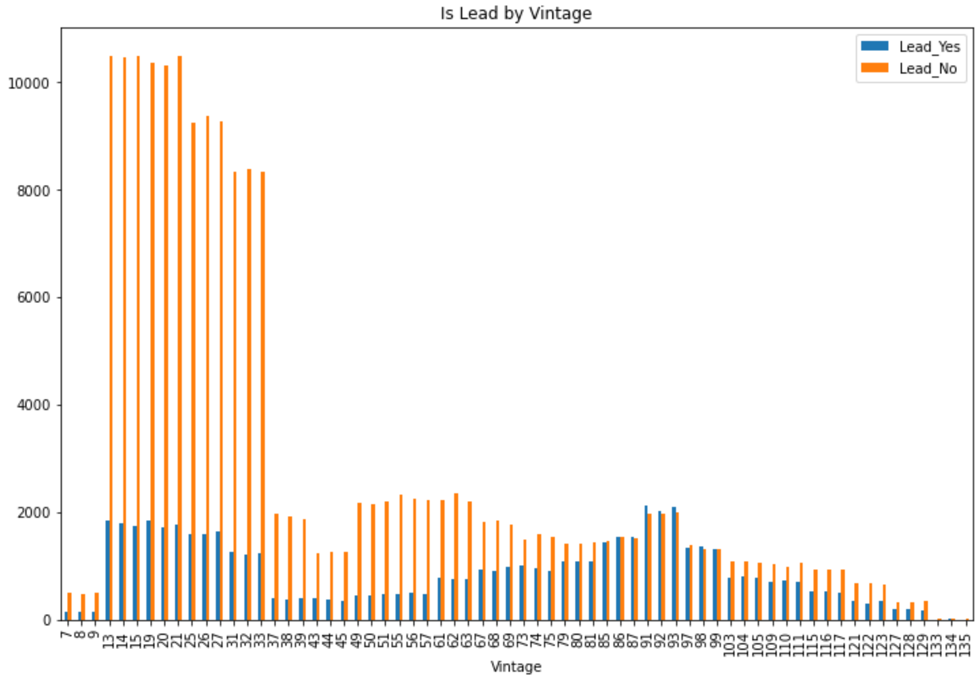
Higher number of Self employed customers are positive leads (Is_Lead as 1.0).



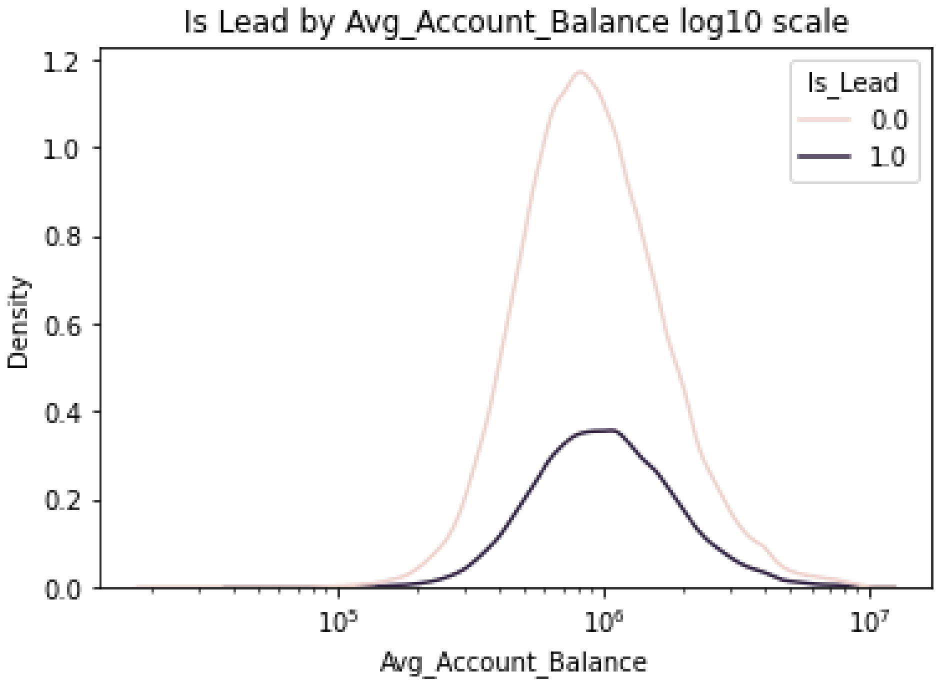
Higher number of customers with existing Credit Products are found to be positive leads (1.0).



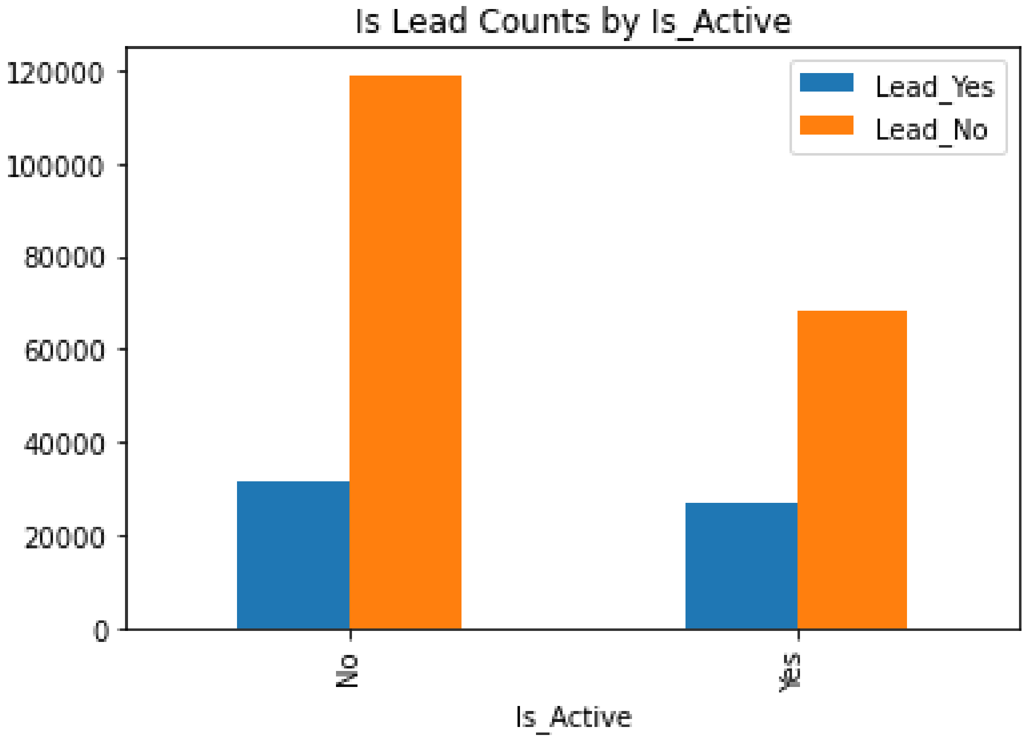
For Vintage, customers in the range of 13 moths to 33 months as well as 81 months to 99 moths have higher positive leads (1.0).



Average account balance distribution on log10 scale is shown below amongst potential leads

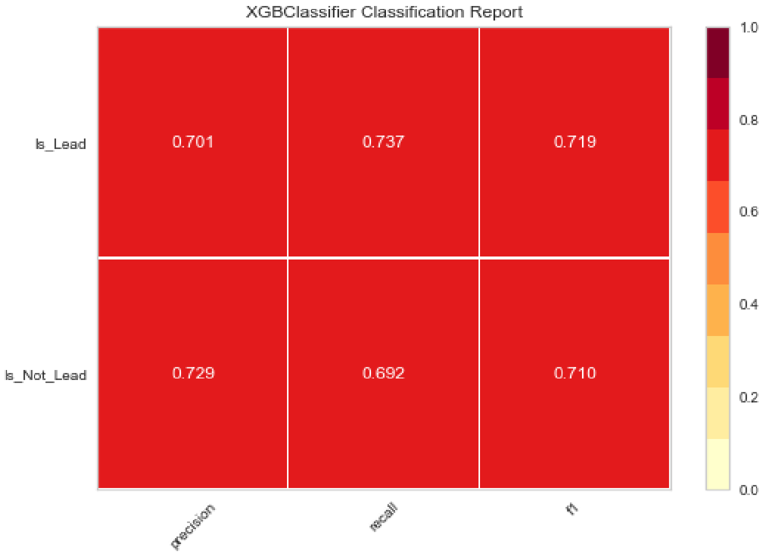


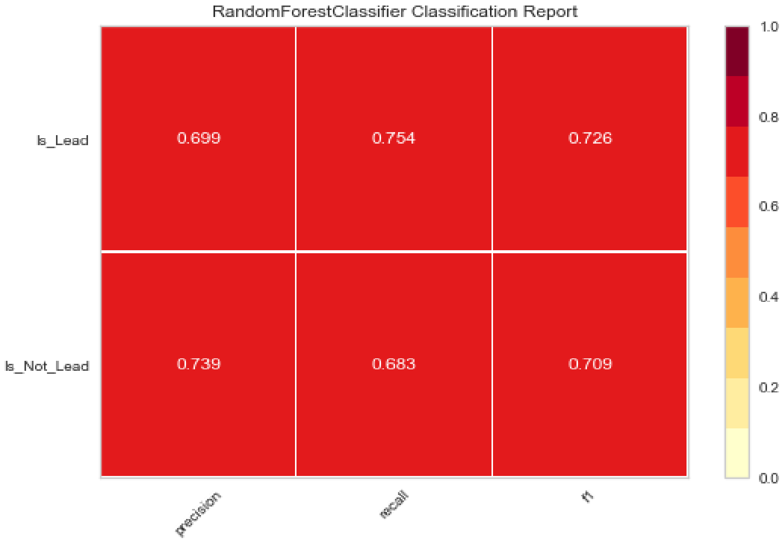
Active Customers have a Good ratio for being a positive potential leads as compared to Inactive customers looking at the below chart



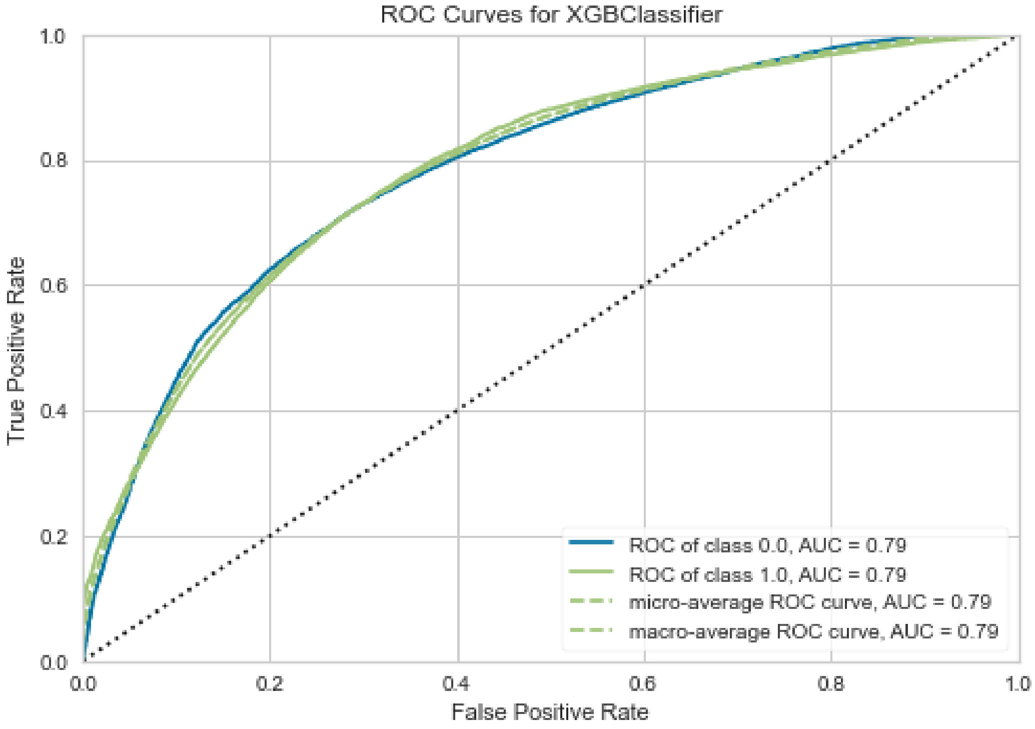
Models training and evaluation:

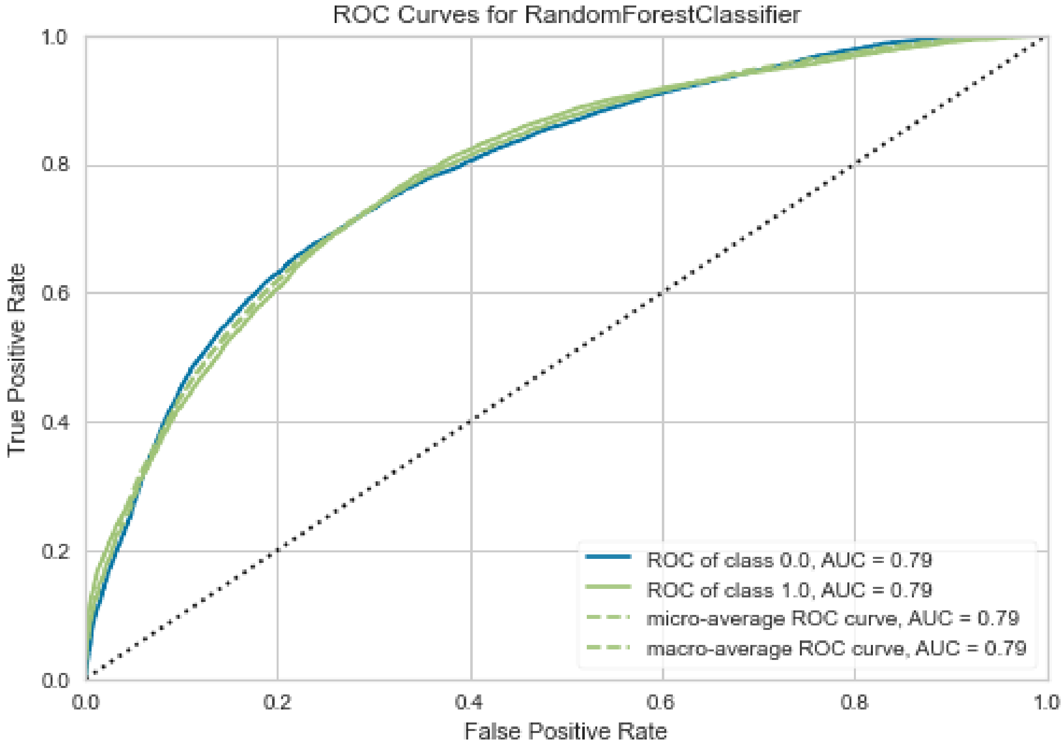
After splitting the training data for train and evaluate processes, we analyzed and evaluated multiple models like Logistic Regression, Random Forest Classifier, Gaussian Naïve Bayes, Gradient Boosting Classifier, Ada Boost Classifier, XG Boosting classifier. The Top 2 performing ML algorithms were Random Forest Classifier and XG Boost classifier. Below are the classification report and ROC curve details for these classifiers. For the evaluation of other classification methods, please refer the jupyter notebook **Credit_Card_Lead_Prediction_Model_Train.ipynb**.





ROC Curves:


ROC Curves for XGBClassifier
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC of class 0.0, AUC = 0.79
ROC of class 1.0, AUC = 0.79
micro-average ROC curve, AUC = 0.79
macro-average ROC curve, AUC = 0.79


ROC Curves for RandomForestClassifier
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC of class 0.0, AUC = 0.79
ROC of class 1.0, AUC = 0.79
micro-average ROC curve, AUC = 0.79
macro-average ROC curve, AUC = 0.79

- **Conclusion:**

Looking at the above analysis, we can notice that certain age groups along with communication channels, occupation, vintage and activeness of customers, existing credit products have varying degrees of effect on the prediction indicator of positive leads.

We noticed that XGBoost classifiers were best performing in terms of precision, recall, f1 scores and ROC curves for both the categories of Is_Lead target variable. Also, with Stratified K fold evaluation method, these produced accuracy scores of 74.6% and 71.7% respectively for train and evaluation datasets. Random Forest classifiers were a good close match on the similar parameters like precision, recall, f1 score and ROC curves. Also, they produced 74.2% and 72% accuracy on train and evaluation datasets using Stratified K fold method. Considering the future use cases of exposing the models to additional data on ongoing basis and potentially trying to utilize the model for other predictions, XGBoost model can be preferred since it does not change much even when the mixture of incoming / training data changes a lot or datasets are unbalanced since the Random Forest classifiers may tend to give more preference to data with higher presence in case of unbalanced datasets and thus susceptible to bias in predictions.

- **Assumptions:**

Assuming that given data has no outliers or incorrect values present in it, since many of the features are categorical in nature and numerical feature didn't appear to contain any out of the range / unexpected values.

- **Limitations / Challenges:**

The dataset available is imbalanced in terms of target variables values. If we use the training dataset as is, it results in lower accuracy / precision / recall for minority class (Is_Lead = 1.0).

So, appropriate measure of data clean-up is needed for balancing data. Though, dropping the excessive records from majority class (Is_Lead = 0.0) results in lower available data for training purposes and thus lowers the overall prediction power for the models. So, possibly higher number of records more training / test data with positive leads (Is_Lead = 1.0) value would be really useful in improving the model accuracy / precision.

- **Future Uses/Additional Applications:**

  The approach / models being created as a part of this opportunity can be partially utilized or concepts used for issuance of other credit products / loan products or investment products in future with some relevant data available in terms of target variables.

- **Recommendations:**

  The parameters observed to be more correlated to target variable should be appropriately used for model training and testing. We can normalize the data in Numerical feature columns and perform label encoding for the Categorical columns prior to feeding data for model training.

  I would also like to include other factors like Credit Scores / history, Annual Income and Existing Loan liabilities etc. for Risk assessments / mitigations as well as offering customized Credit Card Products as a next phase of the project.

  The recommended model for the prediction purpose for the given problem is XGBoost.

- **Implementation Plan:**

  Since the outcome variable would be Binary “Is_Lead” values of 1 or 0, we should be using machine learning models like Logistic regression, Random Forest Classifiers, Naïve Bayes, Gradient Boost and Extreme Gradient Boost algorithms for final predictions. We can train these models and evaluate the models based on parameters like precision / recall / accuracy, ROC charts and then generate the final predictions. We will be using the Stratified K-fold cross validations for model evaluation.

- **Ethical Assessment:**

  Factors like Age and Gender are present in the dataset. There are certain age groups which seemed to have a degree of influence on positive leads. However, as a ethical considerations to produce unbiased models, these factors should not be given weightage.

- **References** :

https://www.kaggle.com/sajidhussain3/jobathon-may-2021-credit-card-lead-prediction

https://medium.com/geekculture/xgboost-versus-random-forest-898e42870f30

https://www.educba.com/random-forest-vs-xgboost/