Documentation

### 

1. **Approach**:

I started exploring the data and analyzing different variables, from the analysis I created two new features. One feature is by subtracting the Age with Vintage so as to get the idea, at what age customer met the Bank. Second, is by dividing the average account balance with tenure. Apart from these I have created feature by aggregation with respect to categorical values, and continuous variable. I used ensemble of two different algorithms I.e, LightGBM and LightAutoML.

1. **Data Preprocessing / Feature Generation**:
   1. After Visualization it was clear that the plot of Credit\_Product was similar to Is\_Lead w.r.t Avg\_account\_balance and Age.
   2. I Created different aggregation like mean, median and std for categorical w.r.t continuous variable.
   3. Also, created aggregations of Region Code with (Occupation, Channel Code) like mean, median and std w.r.t continuous variable.
   4. I created two new features:
      1. By Dividing Avg Account balance with tenure.
      2. By Subtracting Age with Tenure(converted into year).
2. **Key Observations/Trends**:
   1. The rejection chances are higher in the ages less than 30. The chances of getting a loan approved is higher in the age group of 40-50.
   2. The data are normally distributed around 0 for both the approved values for capital gain and capital loss.
3. **Model Selection**:
   1. I trained three models:
      1. LightGBM
      2. LightAutoML
   2. For model selection I created grid search, Bayesian pipelines and ROC-AUC score as metric to select the best model as was defined in the competition.
   3. I created validation set once by using the stratification sampling and other without stratification.
   4. Gridsearch and Bayesian itself uses a validation set for identifying best features.
   5. The best model was selected using the ROC-AUC score on the validation set.
   6. The final output is an ensemble of different algorithms like LightGBM and LightAutoML along with different features used while training them or hyper tuning them.
   7. The final result is calculated by taking mean of 10 different models trained on different parameters or features.