Approach

To solve the given problem, I followed the following approach:

- Understand the data.
- Data Pre-processing and Feature Engineering
- Data Modeling

Data Understanding:

We are given an insurance policy dataset and we need to find out that a customer is interested in this policy or not. For that we were given customer demographics, information regarding holding policies of the customer and recommended policy information.

So after analyzing the data, I have found out:

- Upper_Age and Lower_Age are multi-collinear to each other. So I have dropped the Lower_Age feature.
- Age is directly proportional to policy premium.
- For individual type of insurance has same Upper and Lower Age and for Joint type, different Upper and Lower Age. So we can say that, for individual we are having duplicate data.
- Policy duration range is 1 – 14+
- The dataset has target variable value 0 as 73% and 1 as 23%. So it is an Imbalanced Dataset.

Data Pre-processing and Feature Engineering:

- **Pre-processing:**

  At first I divided the data into two parts. Numerical and Categorical. After that I checked for missing values in the dataset. Heath_Indicator, Holding_Policy_Duration and Holding_Policy_Type have missing values. All are categorical in nature. We cannot impute those missing values with mode value because that will give different inference. We also cannot delete those values as that will lead to information loss. So I imputed those missing values using KNNImputer.

- **Feature Engineering:**

  After analyzing the dataset, I have found that, most of the records are of age between 18 – 30 and > 50. So I created a new column age and divided the age records into three categories.
  - 1 for age between 18 and 30
  - 2 for age between 31 and 50
  - 3 for age greater than 50

  Other than that, I also analyzed that, most of the insurance are of less than 10 years. So I created a new column and divided the insurance duration into three categories:

  - 1 for duration between 1 and 5
  - 2 for duration between 6 and 10
  - 3 for duration greater than 10

After data pre-processing and feature engineering, I have done OneHotEncoding for all the categorical feature and created the final dataset for model building.

Model Building:

As I already mentioned that It is a imbalanced dataset that's why I tried OverSampling method to make it Balanced Dataset. After that I divided the data into train and test (70:30 ratio) for model building.

As It is a classification problem, at first I tried Logistic Regression for model building. But it did not give a good roc score because insurance premium column had some outliers. So after that I tried Decision Tree Classifier and got roc_score 0.83 which is good. To make it better, I applied Ensemble Techniques. For Bagging I tried Random Forest Classifier and Extra Trees Classifier and for Boosting, I applied AdaBoost, GradientBoost, XGBoost and LightGBM.

Among all the algorithms, Random Forest Classifier and LightGBM Classifier gave the best result. So I have done hyper parameter tuning for both and got more better roc score.

After that I applied these two algorithms in Test Dataset and got better result for Tuned LightGBM Classifier. And Tuned LightGBM Classifier is my final model with roc score 0.61.