Problem Statement

Happy Customer Bank is a mid-sized private bank that deals in all kinds of banking products, like Savings accounts, Current accounts, investment products, credit products, among other offerings.

The bank also cross-sells products to its existing customers and to do so they use different kinds of communication like tele-calling, e-mails, recommendations on net banking, mobile banking, etc.

In this case, the Happy Customer Bank wants to cross sell its credit cards to its existing customers. The bank has identified a set of customers that are eligible for taking these credit cards.

Now, the bank is looking for your help in identifying customers that could show higher intent towards a recommended credit card, given:

- Customer details (gender, age, region etc.)
- Details of his/her relationship with the bank (Channel_Code,Vintage, 'Avg_Asset_Value etc.)

Methods Used

- Machine Learning
- Data Visualization
- Feature Engineering
- Predictive Modeling
- etc.

Technologies

- Python
- Pandas
- Scikit-learn
- Seaborn
- etc.

Project Description

Basically, the project includes 4 steps:

- Load and Import Data
- Make Plots for Exploratory Data Analysis
- Data Wrangling and Feature Engineering
- Model Building

Load data

The raw data is loaded.

Link - https://datahack.analyticsvidhya.com/contest/job-a-thon-2/?utm_source=datahack&utm_medium=Navbar&utm_campaign=Jobathon#ProblemStatement

EDA

Basically, I made histograms, bar charts, Boxplot, heat map and pair plot

Data Wrangling and Feature Engineering

I have changed missing values in "Credit_Product" feature into "others"

Removed the outliers from Avg Account balance feature.

Removed Id Feature from Data-frame as it has unique values.

Removed Region Code from data-frame as it has no major impact on Depentent feature.

Performed Lable encoding and one hot one hot encoding on data.

Model Building

Methods I used in the model building part include:

- SMOTE: The target variable “Is_Lead” is highly imbalanced. SMOTE generates the virtual training records by randomly selecting one or more of the k-nearest neighbors for each example in the minority class. After the oversampling process, the data is reconstructed.

- Decision tree: The decision true are generally in form of if-then-else statements. The deeper the tree, the more complex the rules and fitter the model. I tried decision tree model first because it's a very interpretable model and it's very transparent. If the company is going to show the workflow of the machine learning model to non-technical audience, decision tree may be a good choice. Decision tree model can be easily visualized.

- Ensemble Bagging : Bootstrap aggregating, also called bagging (from bootstrap aggregating), is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid overfitting.

- RandomizedSearchCV: Random Search is a hyperparameter tuning algorithm. It helped me select models by setting up a grid of hyperparameter values and selecting random combinations to train the model and score. This allows me to explicitly control the number of parameter combinations that are attempted. It’s more efficient than GridSearchCV and its chance of finding the optimal parameter are comparatively similar.