KBC BANK

Data Science : Direct Marketing Optimization

Folder : KBC.zip
Sub-folders : /home/pgoyal/Projects/kbc/
Contents :
1. IPython Notebooks and Helper Python scripts
2. Train Dataset
3. In-between datasets created
4. ./model/ folder : Model objects
5. ./model/gs_files : Grid Search files generated

Final Customer List : /home/pgoyal/Projects/kbc/Final_Target_List_Approach2.csv
Result:

- Which clients have higher propensity to buy consumer loan?
  Kindly Refer: /home/pgoyal/Projects/kbc/cl_approach2_Test_classif_preds.csv

*The file has the list of customer IDs along with the predicted values in the priority order

- Which clients have higher propensity to buy credit card?
  Kindly Refer: /home/pgoyal/Projects/kbc/cc_approach2_Test_classif_preds.csv

*The file has the list of customer IDs along with the predicted values in the priority order

- Which clients have higher propensity to buy mutual fund?
  Kindly Refer: /home/pgoyal/Projects/kbc/mf_approach2_Test_classif_preds.csv

*The file has the list of customer IDs along with the predicted values in the priority order

- Which clients are to be targeted with which offer? General description.
  Kindly Refer: /home/pgoyal/Projects/kbc/Final_Target_List_Approach2.csv

* The file has the file list of the customer IDs along with their predicted Revenue and the concerned product (Mutual Fund/Credit Card/Consumer Loan). The lit has been prepared for the Top-100 customers to be targeted in the Marketing campaign

- What would be the expected revenue based on your strategy?

Kindly Refer: /home/pgoyal/Projects/kbc/5_Customer_Shortlisting_Approach_1_&_Approach_2.ipynb
*** Expected Revenue : 1338.12**
* The above file has the final steps of the selection of the top-100 for both adopted approaches, and the final Approach-2 results are selected for the

Steps Adopted:
1. Creation of analytical dataset (both training and targeting sets) :
   a. Given dataset was separated between train and test datasets
   b. Feature creation and processing was done for all the 3 products
   c. Target variable (Binary and Revenue values) added for different modeling purposes
2. Develop 3 propensity models (consumer loan, credit card, mutual fund) using training data set

Approach-1

* All the entries from the Train Dataset are taken along with their Target Revenues
* Regression Model has been buiilt using the XGBoost library and a thorough Grid Search has been done
* To ensure the generalization being introduced by the model, the Train dataset has been split into 2 parts (80% and 20%) :
* a) 1st model is trained on the 80% dataset and validated on 20% dataset, and further tuned to optimize the hyper-parameters.

* b) The above selected hyper-parameters are further used for model trained on the 100% Train dataset and results in the final model
* The final model is used to score on the Test dataset.

Other considerations:
* The target variable with the binary value of 0 and 1 for the persons with Revenue or without Revenue has been ignored, and only the variable with the actual revenue has been taken into account in lieu of it. The same exercise has been performed for the 3 flags : Mutual Fund, Consumer Loan, and Credit Card, resulting in 3 separate final models

Approach-2 (Selected Approach)
Step-1. Shortlisting Potential customers by Classifier model
Step-2. Calculating Probability Threshold for the Step-1 customers
Step-3. Regression model trained on ONLY all those customers with Target=1 (or Revenue>0)
Step-4. (Part-a) Scoring on Test data using Step-1 Classifier Model and using the Probability Threshold calculated in Step-2 to further shortlist highly potential customers
Step-4. (Part-b) Scoring on Step-4a customers using Step-3 Regression Model and predicting the amount of Revenue the highly potential customers will bring

* All the entries from the Train Dataset are taken along with their Target Revenues > 0
* Classifier Model has been built using the XGBoost library and with thorough Grid Search and 3 different sets of models are prepared, which are further compared on the basis of LogLoss value on validation dataset, and the best model is selected
* To ensure the generalization being introduced by the model, the Train dataset has been split into 2 parts (80% and 20%) :
* a) 1st model is trained on the 80% dataset and validated on 20% dataset
* b) The best selected hyper-parameters are further used for model trained on the 100% Train dataset and results in the final model of Approach-2
* c) Depending upon the best possible probability threshold a CutOff has been found
* The final model from above step is used to score on the Test dataset, and further Probability Threshold has been put
* Regression Model has been built using the XGBoost library and with 3 different set of hyper-parameters from the Approach-1, for the concerned flag

Other considerations:
* The target variable with the binary value of 0 and 1 for the persons with Revenue or without Revenue has been taken into account in Step-3 for shortlisting, and the variable with the actual revenue has been further taken into account for the regression model purpose.
** The same exercise has been performed for the 3 flags : Mutual Fund, Consumer Loan, and Credit Card, resulting in 3 separate final models for Classification and 3 separate final models for Regression

3. Optimize targeting clients with the direct marketing offer to maximize the revenue
Scoring : Shortlisting Top-100 Customers for the Marketing offer :
Assumption :
1. Finally 100 customers are shortlisted for Marketing offers (This has been kept as variable :))
2. No constraints on the minimum number of offers for any of the Product in the campaign
3. Maximization of Revenue has been taken into consideration

Steps Followed (Common for both Approach 1 and Approach 2 results) :
1. Loading the final revenue predictions files for 3 cases (Mutual Fund, Consumer Loan, Credit Card), which has Test Dataset entries sorted on the basis of Revenue in decreasing order
2. Shortlisting the Top-102 entries for each of the 3 cases (As total 100 customers are supposed to be shortlisted finally, so the top 34 entries are selected from each case, by taking the worst case scenario into consideration) : Resulting in 102x3 (=306) entries in the dataframe after combining
3. Selecting Top-100 customers overall with the aim of maximizing the Revenue