Credit Card Cross Sell

VARUN BHANDARI

Binary Classification

Problem to identify prospective customers of a bank's credit card.

JUPYTER NOTEBOOK AND WRITE-UP

Broad Overview

Step 1: EDA
Perform basic Exploratory Data Analysis to check for relationship among variables.

Step 2: Feature Selection
Select relevant features that would determine a person’s interest in credit card.

Step 3: Data Munging
Estimate missing values based on given data.

Step 4: Model Building and Prediction
Build model to estimate if a person is a prospective lead for buying credit card.

EDA

The variables that have been provided for the analysis are mostly categorical.

Variables such as gender, region, occupation and channel don’t appear to show a direct relationship, however they may be related in a more complex way.*

An analysis of average account balance with prospective lead data also doesn’t show a direct relationship since most of the people who are prospective customers are distributed throughout the account balance spectrum.*

However, for the missing data for Credit Product column there is a heavy bias towards prospective credit card customers.*

*-Graphs are available in the jupyter notebook enclosed herewith.

Feature Selection

The feature called ID is not relevant to the categorisation of data and has therefore been dropped.

For the purpose of model building some of the categorical features have been numerically encoded, this includes occupation, customer channel, gender and customer activity.

Later these variables have been converted to one hot encoded forms for use in the algorithm.

Variable region code is converted to a numerical form, since regions are generally allotted codes based on their geographical placement.

Variables vintage, age and average account balance have been used directly.

The variable for credit product has missing values and is therefore handled next.

Data Munging

The variable credit product shows whether a customer has an active credit product such as a loan or card.

There are however multiple missing values for this field. Among all of these missing values nearly 85 % are the people who are prospective customers for a credit card.

Therefore this is an important variable that will likely determine if a customer buys a credit card.

There is a marked increase in the percentage of people opting for credit cards when they own a credit product. Hence this can be safely assumed to be ‘Yes’.

The final model is built on this data.

Model Building and Prediction

The variables have not shown a direct correlation to the target variable, hence there may be some complex relationship and therefore a non linear model has been used.

A random forest model (101 trees) has been used to capture the model. To avoid overfitting training set is divided into a test and train set of 30-70 ratio and 10-fold cross validation is performed on the training set to check the results.

Later an xgboost model (101 trees) has been used to further improve the model. The xgboost model has also been built on 70% of training set along with 10-fold cross validation.

Finally, a lightgbm model is used. This gives the best result among all the three. This model has also been built on 70% training data and evaluated using 10-fold cross validation.

Final output is provided based on this model.