Sales and Revenue Predictive Model

Jessiedee Mark B. Gingo, Data Scientist candidate, Home Credit PH

Business Understanding

Given the data set of the bank's clients, we want a model that predicts clients who are most likely to take an offer and predict its revenue. This is to be used in a marketing campaign for mutual funds, credit cards, and consumer loans products.

Data Cleaning

The Excel file has five sheets. Those are variable descriptions, social demography, account balances, behavior such as cash inflow and outflow, and sales revenues, which are converted to pandas dataframe. Inconsistencies are checked and fixed. Missing data are in the behavior data set, which has 1,587 out of 1,615 total clients. Sheets are merged and missing values are handled accordingly.

Feature Engineering

This is a feature selection process of which variables of high correlation are selected. Scikit-learn Logistic Regression and Recursive Feature Elimination (RFE) are used for models which have categorical output variables, i.e. sales. Linear Regression and RFE are used for models which have continuous output variables, i.e. revenues. The number of selected features are tuned to get the most accurate result.

Predictive Modeling

The data has 6 responses or outputs, i.e. mutual funds, credit cards, consumer loans sales, and mutual funds, credit cards and consumer loans revenue. The sales will have classification predictive models and revenue will have continuous predictive models. The train and test set are 60% and 40% of the total data, respectively. The models are created using the train set, and the customers will be predicted using the test set.

Sales Predictive Model

Scikit-learn RandomForestClassifier machine learning algorithm is used to train the model, and its hyperparameters are cross validated to get the most accurate results.

Revenue Predictive Model

XGBoostRegressor, the most accurate machine learning algorithm for continuous variables, is used to train the revenue predictive models. Parameters are also tuned to get the least Mean Absolute Error (MAE).

Answers

1. Which clients are more likely to buy consumer loans?

   ```
   list(consumer_loans.Client)
   [1207, 1510, 164, 532, 1455, 1393, 796, 1448, 1587, 1229, 392, 133, 1218, 411, 4,
   1589, 217, 803, 1588, 183]
   ```

2. Which clients are more likely to buy credit cards?

   ```
   list(credit_cards.Client)
   [851, 668, 978, 197, 352, 243, 104, 951, 1076, 1262, 715, 389, 1487, 592, 886, 145,
   701, 382, 535, 1227, 506, 587, 1419, 727, 640, 340, 1449, 1260, 151, 1424, 7, 119, 579,
   1008, 1120, 545, 835, 1278, 1304, 781, 783, 818, 299, 1057, 769, 1021, 1614, 289,
   1107, 223, 367, 1097, 813, 760, 331, 1331, 1074, 1443]
   ```

3. Which clients are more likely to buy a mutual fund?

   ```
   list(mutual_funds.Client)
   [354, 866, 910, 786, 1435, 1187, 109, 64, 1119, 779, 940, 968, 1578, 476, 484, 1129,
   246, 830, 353, 759, 1416, 386]
   ```

4. Which clients are to be targeted with which offer? General description.

   The clients' offers are identified based on the probability result of every sales category, (x>0.50 is 1, x<0.50 is 0). The mean is computed, the top 100 clients are sorted, and the maximum value of probability in the sales categories of each client is identified, thus the offer.

5. What would be the expected revenue based on your strategy?

   ```
   result.revenue.sum()
   EUR 1,033.48
   ```

6. How well do your models perform?

   The performance of the models is measured using the accuracy score sales predictive model, and MAE for revenue predictive model. Here as follows:

Sales Predictive Model Accuracy (ideal: high)			**Revenue Predictive Model MAE score (ideal: low)**		
Mutual Funds	Credit Cards	Consumer Loans	Mutual Funds	Credit Cards	Consumer Loans
81.48%	76.95%	74.48%	11.47	17.02	6.85

7. How did you ensure the models would continue performing well once applied to the 40% of clients?

   The quality performance is ensured because the models are trained out of the separate train set, and the test set would not affect the model. Also, the variables or features are checked for possible data leakage, or variables that would have an effect while training the model and make it very accurate on the training, but will be a failure in the test set or deployment.