Imbalanced Classification Project

Project Deliverables

You will be required to submit:

- A GitHub repository with your project written in Python or R.

Instructions

Problem Statement

Beta Bank customers are leaving: little by little, chipping away every month. The bankers figured out it's cheaper to save the existing customers rather than to attract new ones.

We need to predict whether a customer will leave the bank soon. You have the data on clients' past behavior and termination of contracts with the bank.

Build a model with the maximum possible F1 score. To pass the project, you need an F1 score of at least 0.59. Check the F1 for the test set.

Additionally, measure the AUC-ROC metric and compare it with the F1.

1. Download and prepare the data. Explain the procedure.
2. Examine the balance of classes. Train the model without taking into account the imbalance. Briefly describe your findings.
3. Improve the quality of the model. Make sure you use at least two approaches to fixing class imbalance. Use the training set to pick the best parameters. Train different models on training and validation sets. Find the best one. Briefly describe your findings.
4. Perform the final testing.

Data description

- Dataset URL (CSV File): https://bit.ly/2XZK7Bo

Data description

- Features
    - RowNumber — data string index
    - CustomerId — unique customer identifier
    - Surname — surname
    - CreditScore — credit score
    - Geography — country of residence
    - Gender — gender
    - Age — age
    - Tenure — period of maturation for a customer's fixed deposit (years)
    - Balance — account balance
    - NumOfProducts — number of banking products used by the customer
    - HasCrCard — customer has a credit card
    - IsActiveMember — customer's activeness
    - EstimatedSalary — estimated salary
- Target
    - Exited — customer has left

Project evaluation

We've put together the evaluation criteria for the project. Read this carefully before moving on to the task.

- How did you prepare the data for training? Have you processed all of the feature types?
- Have you explained the preprocessing steps well enough?
- How did you investigate the balance of classes?
- Did you study the model without taking into account the imbalance of classes?
- What are your findings of the task research?
- Have you correctly split the data into sets?
- How have you worked with the imbalance of classes?
- Did you use at least two techniques for imbalance fixing?
- Have you performed the model training, validation, and final testing correctly?
- How high is your F1 score?
- Did you examine the AUC-ROC values?
- Have you kept to the project structure and kept the code neat?

You can use the CRISP-DM methodology that we've used in the past, whose approach will help you address the above questions.