Machine Learning Engineer Nanodegree

Capstone Proposal

Churn Modeling

- Stiven López Giraldo
- Mayo 11, 2020

Proposal

Domain Background

Churn analysis is one of the most important problems for different companies, from sectors such as telecommunications to banking. A proper analysis together with a well-defined strategy allows to increase the user retention rate, and not only that, it can also help to focus on improving the user experience of those who are more likely to migrate to another campaign. Since users are a valuable asset for different companies and the constant creation of financial services from different companies in the sector, the need for customer retention increases.

Problem Statement

The bank's objective is to find out which users are most likely to leave the company. Certain user characteristics are calculated in order to generate that probability, and depending on each user, a business action is taken to retain this user.

Dataset and Inputs

The following kaggle data is used for the project. The data set contains sociodemographic information, credit rating, and even an estimate of your income, assuming that your salary is the only income the user has.

The data set has 13 variables in total including the target variable, this tells us if a user left the bank (closed his account) or if he continues to be a customer. In addition, the data set has 10,000 records, so it may not be a sufficiently representative sample depending on the number of users the bank has, however, it is an interesting data set because it brings information on user behavior.

Solution Statement

The solution to the problem of interest is to build a classification model that generates the probability that a user will leave the company, with this probability the users in the area in charge will create ranges and define actions for the different users and encourage them to don't leave the company. The model will be an Inference Pipeline trained in Amazon

SageMaker, this will be made up of a first container that does all the preprocessing of the data, and the second a container that houses an XGBoost for training.

The first container will perform transformations such as: select columns, convert categorical variables to dummy variables, among others. Scikit-learn transformers and custom pandas will be built.

The second container will receive the ready data from the first container, and will train the XGBoost to later make inferences.

The goal of developing such a model is to avoid data leakage problems, and to have the same transformations in training data as in new data.

Benchmark Model

The benchmark chosen is to infer for each new user the most frequent value of the users who are in our sample, that is, for each new user it is predicted that he will remain in the bank and that he does not have a high probability of leaving.

If we do this on the test data, we get the following metrics:

- AUC : 0.5
- Accuracy: 0.8
- Precision: 0
- Recall: 0
- F1 Score: 0

Evaluation Metrics

Our problem presents a class imbalance, although it is not as serious as certain problems that are usually found. In our case we have around 80% who are still clients and the remaining 20% who closed the account.

To evaluate performance, the F1 Score is used, which is the weighted average of precision and recall. Therefore, this score takes into account both false positives and false negatives and is useful for problems of unequal class distribution.

Project Design

Work flow:

- Download the data from S3, and break it into training, validation and test sets, and upload it back to S3 to train the model.
- Generate descriptive statistics and visualizations that allow a better understanding of the data.
- Build the custom transformers and the script that runs on the first container.
- Build the XGBoost estimator and debugger rules to generate more interpretability of the model.
- Deploy the endpoint and evaluate the latency with which it generates the inference.
- Send the results.