Simón Mijares

Udacity Nanodegree: Machine Learning Engineer Nanodegree

Domain: Telecommunications

April 2021

Capstone Proposal: Churn Predictor

Project Overview

Traditionally the base client grow rate was one of the main KPI in a service business, but currently is known that it does not matter if those clients will not stay long enough to develop some fidelity to your services.

For this reason, monitor your levels of churn (or attrition) is important, more in current times when the pandemic took most people out of their regular routine, struggling with income stability, heath issues among many others. Not only clients are having difficult times, but the companies are also having difficulties in different stages of its own flows. Transport, distribution, or the sole challenge of working remote to name a few examples.

The Telco's are one case, and it will be the case of study for this project. The possible variables of study are certainly multiples but all orbit around the same concept, technical variables, commercial variables, interaction variables and commitment with the services variables. These will be explained bellow.

As part of this work, I will try to obtain a model in Sagemaker capable of predict when a client will most probably churn, so the company could try to reach him and try to keep him happy.

Data set

To train the model we will be using the data set available in kaggel at *https://www.kaggle.com/barun2104/telecom-churn.*

In this dataset we have the following columns:

Churn: 1 if customer cancelled service, 0 if not. *This will be the variable to predict.*

AccountWeeks: number of weeks customer has had active account

ContractRenewal: 1 if customer recently renewed contract, 0 if not

DataPlan: 1 if customer has data plan, 0 if not

DataUsage: gigabytes of monthly data usage

CustServCalls: number of calls into customer service

DayMins: average daytime minutes per month

DayCalls: average number of daytime calls

MonthlyCharge: average monthly bill

OverageFee: largest overage fee in last 12 months

AMDWAN1080APRIL

Solution Statement

To solve the problem ML models will be trained and tested. Bias due to statistical imbalance will be taken in consideration, in this case recall will be maximize as priority without leaving out the precision.

We will base the development on the Fraud Detection case of use of the course since the application is remarkably similar on spirit. In the mentioned example the target was detecting fraudulent operations avoiding as much as possible detecting false negatives. In our case we need to detect which clients pretend leaving the company, avoiding false positives as well. In the figure bellow we could see a representation on how disproportionate predictions could be due to imbalance in the classes. Causing bias in your model.

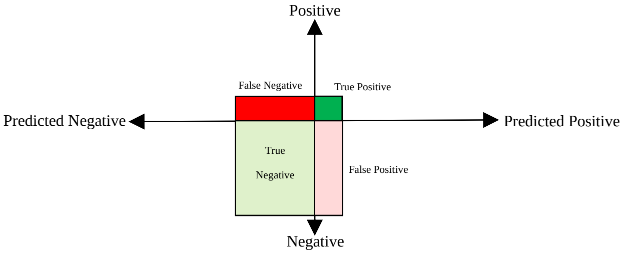

Fig 1. Imbalanced Training data and biased model

$$Recall = \frac{\Sigma\ \text{True positive}}{\Sigma\ \text{positive}}$$

$$Precision = \frac{\Sigma\ \text{True positive}}{\Sigma\ \text{predicted positive}}$$

Project Design

The first phase will be an exploratory analysis of the data. Show how are they composed, distributed, and interrelated. If any parameter has a high correlation with other, it could be ignored to optimize resources.

After that, the data will be group in training and test data. Assuring enough positive results in both groups.

Once the data is ready, we can use it to train the model. Then we will proceed to tune it optimizing for recall and comparing the results and getting to conclusions.

Finally, not part of the analysis, but after obtaining the results we will release the resources in Sagemaker to avoid additional costs.