Machine Learning Engineer Nanodegree

Capstone Proposal

Mario Rugeles

October 10, 2018

Proposal

Domain Background

The “right of request” is a legal mechanism included in the Colombian Constitution of 1991 designed to allow efficient comunication between a citizen and the Goverment. This mechanism is one of the most fundamental democratic tools in Colombia as it allow all citizens, but most importanlty, the most vulnerable to ensure the Goverment is fulfilling his obligations with the people. Once a “right of request” is required by a citizen, the goverement has from 10 to 30 days to resolve the case.

This in practice, means that the Goverment needs to ensure it has the procedures and infrastructure to resolve a right of request as soon as posible.

Problem Statement

One of the basics rights of every citizen is the access to the health care system. Although Colombia faces big challenges to ensure access to the health care system for everyone, it’s recognized as one of the best in LATAM [1]. The Goverment of Colombia has been creating online resources to allow citizens to raise complaints or ask for information in all ministeries and public entities[2], and it’s expected that the volumen of “right of request” will increase as more citizens has access to the internet. The Ministry of Health receives all kinds of right of request and must find a way to decide which cases have more priority as many of these cases are about citizens with life at risk.

Deciding when a person’s life is at risk may depend on both technical and legal grounds, and it require experts in the field (medicine and law) to make the right assesment. But the volumen of cases can eventually exceed experts’s capacity to respond quickly with the right desition.

Datasets and Inputs

Colombia’s Goverment has a website offering a wide range datasets open for the public. In this database, Ministry of Health has published a series of datasets build from PQRD (Petitions, Complains, Claims and Demands from acronyms in spanish)[3]. These datasets have information of “right of request” raised by citizens regarding issues with the Health Care System.

There’s a dataset per year, and every record has information about the patient, demographics and administrative information related with the

patient. The features include a field indicating if the patient’s lifes is at risk.

Solution Statement

For this project, a classification model will be builded to predict wheter a “right of request” (PQRD) could imply if a patient’s life is at risk. As the data is from more than one dataset, it is required to find the common features between them. The final dataset will have around 50 features, most of them categorical, so a process of dimensionality reduction will need to be done. New PQRD’s are added every year, so a process to upgrade de model should be considered. Also, every year dataset has around 600K records so model time creation should be take into account, at least deciding how important is it, and being the case, proposing which strategies could be implemented in a next project version.

Benchmark Model

The PQRD database has information from 2013 to 2017, so it gives a good chance to test the model with new information. A first model will be created for the first year an evaluated first with its testing and validation sets and then with the dataset of the next year and so on. The model will be finally created with data up to 2016, and the final evaluation will be done with dataset from 2017.

Evaluation Metrics

The model must have a hight recall metrics because it needs to avoid false negatives, as we don’t want people with life at risk as not in risk. A initial beta of 2 will be used in the F-beta score metrics.

Project Design

Data preprocessing

The first part of the project will be to unify the features from the datasets, all datasets from 2013 to 2017 have pretty much the same features, but every year a few ones may be added or removed, so it’s required to ensure that all datasets have the same features. The second part is that the datasets have mostly categorical features, that implies high dimensionality for categories with high cardinality, so supervised ratio and weight of evidence should be considered in every case, some utility methods to analyse columns datasets for category features may be helpful in this step. Once every categorycal fearture has been processed with one of the strategies (one hot encoding, supervised ratio or weight of evidence), a value normalization will be applied to the dataset.

Data analysis

De second part is about finding anomalies in the data, is there noise in the dataset?, it has skewed features?, in every case appropiate procesing should be applied.

The dataset has about 50 features, so dimensionality reduction needs to

be applied. PCA will be used in this step.

Building the model

Next is the model construction, this is a Supervised Learning project for categorization, so several algorithms will be evaluated, a pipeline will be implemented for this step to measure which algorithm work best. An analysis from the results with the selected model will be documented. After a model is selected, a tuning process will be done to improve the model's performance using grid search.

General model pipeline

The model will be build incrementally, that is: a model will be created with data from 2013, tested and the updated with data from 2014, and so on. A versionig approach with git will be used, as the project is intented to work in a production environment, so a very basic deployment flow will be used.

In this final section, summarize a theoretical workflow for approaching a solution given the problem. Provide thorough discussion for what strategies you may consider employing, what analysis of the data might be required before being used, or which algorithms will be considered for your implementation. The workflow and discussion that you provide should align with the qualities of the previous sections. Additionally, you are encouraged to include small visualizations, pseudocode, or diagrams to aid in describing the project design, but it is not required. The discussion should clearly outline your intended workflow of the capstone project.

References

1. **Health Care in Colombia: Top Quality and Affordable**, **The healthcare system in Colombia**
2. **Datos Abiertos, Colombia Digital**
3. **Health Monistry PQRDs**