Machine Learning Engineer Nanodegree

Capstone Proposal

Mario Rugeles October 21, 2018

Proposal

Domain Background

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

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

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 government 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 volume 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 assessment. But the volume of cases can eventually exceed experts’s capacity to respond quickly with the right decision.

A dataset from the “right of request” for the Healthcare system is available for the public. This dataset contains information related with the patient and a feature indicating if the patient’s life is at risk. This dataset is suitable for a classification model than can predict if new patients rising new “right of request” have their life at risk automatically saving time for both the government and de patient.

Every record provides information about patient condition, demographic information and the company that supplies Healthcare services.

Datasets and Inputs

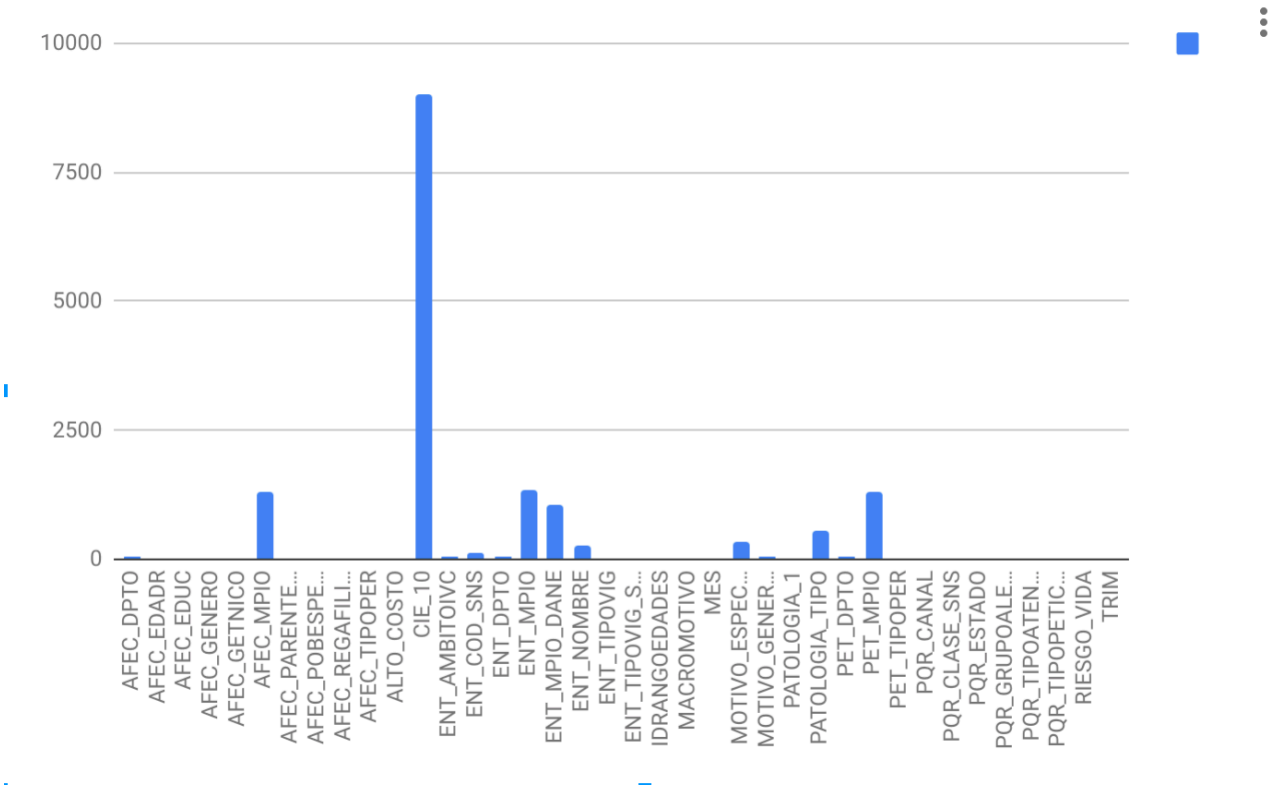
Colombia's government has a website offering a wide range datasets open for the public. In this database, Ministry of Health has published a series of datasets build from PQRD (Petitions, Complains, Claims and Demands from acronyms in Spanish)[3].

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 life is at risk.

The final dataset will have around 50 features, most of them categorical, some of these features are categories with a description (feature code / feature description) this descriptions can be discarded, so only the codes will be kept. As there are many category features, a category reduction will be required. Features are also highly unbalanced.

The dataset has around 2 million of records, so it provides a quite large sample for a model to learn.



Features Description

Feature	Classes	Description
AFEC_DPTO	39	Patient's state
AFEC_EDADR	11	Patient's age range

AFEC_EDUC	10	Patien's Education Level
AFEC_GENERO	4	Patient's Genre
AFEC_GETNICO	10	Patient's etnic origin
AFEC_MPIO	1299	Patient's City/Town
AFEC_PARENTESCO	18	Parentesc relation with the patient (Who raise de PQRS)
AFEC_POBESPECIAL	10	Is the patient from a especial vulnerable group of the population?
AFEC_REGAFILIACION	7	What kind of healtcare is afiliated to?
AFEC_TIPOPER	4	Is a person or a entity?
ALTO_COSTO	23	Is a hight cost healt issue?
CIE_10	9029	Diagnostic code (ICD-10 in english)
COD_MACROMOT	8	Reason for raising PQRSD - Macro Category Code
COD_MOTESP	323	Reason for raising PQRSD - Especific Category Code
COD_MOTGEN	43	Reason for raising PQRSD - Generic Category Code
ENT_AMBITOIVC	33	???
ENT_COD_DEPTO	35	Healthcare provider's State Code
ENT_COD_MPIO	1119	Healthcare provider's City / Town Code
ENT_COD_SNS	107	Healthcare provider's Sanitary Code
ENT_DPTO	36	Healthcare provider's State Name
ENT_MPIO	1349	Healthcare provider's City / Town Name
ENT_MPIO_DANE	1037	Healthcare provider's City / Town Name
ENT_NOMBRE	250	Healthcare provider's Name
ENT_TIPOVIG	30	Healthcare provider's Category
ENT_TIPOVIG_SNS	13	Healthcare provider's Category Sanitary Code
IDPATOLOGIA_2	457	Patology Id 2
IDRANGOEDADES	10	Age range Id
ID_MES	201178	record Id

MACROMOTIVO	9	Reason for raising PQRSD - Macro Category Description
MES	12	Month number (1 -12)
MOTIVO_ESPECIFICO	334	Reason for raising PQRSD - Especific Category Description
MOTIVO_GENERAL	42	Reason for raising PQRSD - Generic Category Description
PATOLOGIA_1	23	Patology description
PATOLOGIA_TIPO	553	Patology type description
PET_COD_DEPTO	34	Petition's State Id
PET_DPTO	38	Petition's State Name
PET_MPIO	1290	Petition's City/Town Name
PET_TIPOPER	3	Petition Entity Name (company or person)
PQR_CANAL	12	How the petition was raised (phone call, interner, etc)
PQR_CLASE_SNS	2	What kind of petition is?
PQR_ESTADO	10	Petiton's Curren State
PQR_GRUPOALERTA	4	Petiton's Priority
PQR_TIPOATENCION	8	Petition's administrative type
PQR_TIPOPETICION	5	Petition's category
RIESGO_VIDA	2	It's patient's life at risk
TRIM	4	N/A

Solution Statement

For this project, a classification model will be builded to predict whether 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

A Support Vector Machine model for classification will be build to compare the project results. SVM provides probability classification that fits to the project's unbalanced data nature.

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 analyze columns datasets for category features may be helpful in this step. Once every categorical feature 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

The 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.

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 general flow for the construction of the pipeline will be:

- Preprocess / Balance dataset
- Ensemble method: A selected algorithms for categorization will be trained and the model with best performance will be selected.
- Once the algorithm is selected, a parameter grid search is applied to tune the model
- Next step is to perform feature engineering over the dataset, train the selected algorithm with the reduced dataset and compare with the original model and assess which model works best.

- Eval final model by comparing with benchmark

The step of "Ensemble method" is about selecting the best algorithm. Given the case no model has good performance for the task, Model Stacking will serve as a backup plan .

References

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