Machine Learning Engineer Nanodegree

Capstone Proposal

Manuel Maqueda Vinas July 10th, 2020

Proposal

Pulmonary Fibrosis Progression (Kaggle Competition)

Domain Background

Pulmonary fibrosis, a disorder with no known cause and no known cure, created by scarring of the lung. Outcomes of the disease can vary from long-term stability to rapid deterioration, but doctors are not easily able to understand the exact outcome in advance.

In addition some of the current methods implies long times and substantial effort in order to produce an accurate prognose. This increases patients anxiety and delays the application of the right mitigation action.

Problem Statement

The goal of this exercise is to help doctors with predictions of pulmonary fibrosis progressions using CT scan images, metadata and FVC (force vital capacity) data as an input.

Datasets and Inputs

The datasets are provided by the Open Source Imaging Consortium.

The input data corresponds with a baseline CT scan on Week 0 and a series of FVC measurements that corresponds to follow-up visits of the patients in the next 1-2 years. Note also that some of the week's visit number can be also negative in the case that the patient's data had been acquired before the baseline CT scan visit.

Input Data fields

- Patient- a unique Id for each patient (also the name of the patient's DICOM folder)
- Weeks- the relative number of weeks pre/post the baseline CT (may be negative)
- FVC - the recorded lung capacity in ml
- Percent- a computed field which approximates the patient's FVC as a percent of the typical FVC for a person of similar characteristics
- Age
- Sex
- SmokingStatus

CT scan

A baseline CT scan obtained in week 0. This image is stored in a DICOM subfolder with the Patient unique ID.

Size of the dataset

- 176 unique patients in the train dataset (including 176 DICOM folders)
- 5 patients in the tests dataset (including 5 DICOM respective folders)

Data characteristics

Input data can be split in two different types: tabular data related with patient's visits and a DICOM image obtained from CT scan.

In terms of the tabular data some of the characteristics that we will be exploring are:

- Distribution of FVC over age, sex and smoking status.
- Distribution of smoking status over age and sex

In terms of DICOM images: * Compare images related with the highest and lowest FVC measurements. * Experiment with pixel and tissue densities and relate with FVC, sex, age and smoking status. * Explore DICOM metadata and find out if there is any other relevant property that can be used.

Solution Statement

The solution aims to predict the pulmonary fibrosis progression for the last 3 weeks of the patients' visit. In particular we will attempt to predict the FVC measurements for those visits but also a confidence metric (standard deviation), which should reflect both accuracy and certainty of the predictions.

Benchmark Model

There are other existing solutions that attempts to resolve this problem as part of the Kaggle competition. The ambition of the ML model that I will build as part of this project is contribute and improve other models as part of the competition.

I am also going to benchmark our model against a couple of existing Kaggle models that use Quantile regression. * Notebook 1 * Notebook 2

Evaluation Metrics

The objective of the competition is to obtain good predictions for the follow-up patient's FVC measurements. In particular, the evaluation metric for those measurements is a modified version of the Laplace Log Likelihood. Below it is shown the respective formula:

$$\Delta = \min\left(\| FVC_{ture} - FVC_{predicted} \| , 1000 \right)$$

$$f_{metric} = -\frac{\sqrt{2}\Delta}{\sigma_{clipped}} - \ln\left(\sqrt{2}\sigma_{clipped}\right).$$

Project Design

The project will be structured in 3 subsections: exploration data analysis, modelling and model evaluation.

The exploration and data analysis aims to understand the nature of the tabular and image data. I will be looking at the correlation between the given tabular data and also I will be experimenting to extract some properties from the CT scan images (such as fibrosis tissue density and air volume in the chest).

With that information I will be defining a regression model to predict FVC for the subsequent patient's visits. For that I will be optimising the model using the Laplace Log Likelihood metric.

At the end I will be training the model using the train dataset and validating the results using the available public test dataset.

I also expect to do a couple of iterations when modelling to define a final one to be submitted in both, as part of the nanodegree project and in the Kaggle competition.

Reference

- Kaggle Competition
- Scikit Linear Models
- EDA - 1
- EDA - 2
- EDA - 3