Machine Learning Engineer Nanodegree

Capstone Proposal

Chun Zhu
February 9th, 2018

Proposal

Domain Background

Patients who do not show up for their scheduled appointments without cancelling nor advance notice is a severe issue that significantly affects clinic efficiency, patient care and healthcare costs. When no-show happens, the medical recource is underutilized. The doctor's valuable time is lost, and the other patients miss their timely treatment because the schedule is taken by the no-show patients. According to some paper, an average of 42% of appointments become no-shows (http://www.annfammed.org/content/2/6/541.full) and the waste helathcare costs can reach hundreds of thousands of dollars yearly in US. Hence, accurate prediction of no-show appointment is very important for both the medical resource and the patients. It is also a cornerstone for future no-show reduction strategy.

Some research papers related to this topic:

http://www.ijimai.org/journal/node/1623

dmkd.cs.vt.edu/papers/THSE15.pdf

Problem Statement

The problem is to predict whether a patient will show up in a medical appointment in public hospitals in Brazil based on some attributes of the patient and the appointment itself. This is a binary classification problem.

Datasets and Inputs

The dataset is from Kaggle

https://www.kaggle.com/joniarroba/noshowappointments/data

The fields of data are below:

- PatientId: Identification of a patient
- AppointmentID: Identification of each appointment
- Gender: Male or Female. In common sense, women takes more care of their health in comparison to man.
- ScheduledDay: The day someone called or registered the appointment, this is before appointment.
- AppointmentDay: The day of the actuall appointment, when they have to visit the doctor. The appointment day may affect the no-show result. Monday or Friday may not the same for the patient.
- Age: How old the patient is.
- Neighbourhood: the neighbourhood of the hospital in which the appointment is carried out. It is possible for patients to come from outside of the neighbourhood, or even the city.
- Scholarship: whether the patient is covered by Bolsa Família, a social welfare program in Brazil.Ture of False .
- Hipertension: True(1) or False(0)

- Diabetes: True(1) or False(0)
- Alcoholism: True(1) or False(0)
- Handicap: an integer ranging from 0 to 4, indicating the level of the handicap the patient is suffering from.
- SMS_received: whether an SMS messages was sent to the patient to remind him/her of the appointment.
- No-show: the target variable, Yes or No. will be transformed to 1 or 0

PatientId and AppointmentID will not be included in my model. Indeed, the appointment show-up history of the patient is important for prediction, but this dataset will be randomly split into 80% training and 20% testing sets (cross validation will be applied on the training set later). Some records in testing dataset may happen prior to training dataset. So including the patient appointment history is not a good idea for this project. In addition, I will add the waiting time as a new variable between scheduled day and AppointmentDay.

The dataset contains 110527 records/apporintments. Patients showed up in 88208 appointments - 79.8% of total appointments, while no show in 22319 appointments - 20.2% of total appointments. The classes are not balanced.

Solution Statement

Logistic regression, random forrest and support vector machine will be applied seperately on the datasets as classification algorithms. Cross-validation will be applied to evaluate best parameters. The training dataset includes both the features (input) and the output variable. From it, the supervised learning algorithms seek to learn the mapping function from the input to the output. The goal is to build a good model that can make predictions for the new dataset (test dataset).

Benchmark Model

The benchmark model will be a simple out-of-the-box verson of random forests trained on the same training data as the final solution. I will then fine tune the model and compare the performance of the 2 models.

Evaluation Metrics

F1-score will be another evaluatiion metric to quantify both the benchmark model and the solution model.

TP = true positive TN = true negative FP = false positive FN = false negative

precision = TP/(TP+FP) recall = TP/(TP+FN)

F1-score = 2 * precision * recall/(precision + recall)

F1 score is the harmonic average of the precision and recall, where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0.

Project Design

1. Explore the dataset, remove NAs and any unreasonable records, like age < 0, AppointmentDay is prior to ScheduledDay.
2. Add derived relevant features: waiting time between ScheduledDay and AppointmentDay, Day of the week for AppointmentDay.
3. Feature scaling.
4. Logistic regression, random forrest and support vector machine will be applied seperately on the datasets as classification algorithms. Cross-validation will be applied to evaluate best parameters.
5. Predict on the test dataset by the best classifer.

6. Calculate f1 score of the test results and compared with the benchmark model.