Medumo - Identifying Patient Engagement Patterns

Group #8, Progress Report 2

Zhonghao Guo, Zihao Yuan, Tlanzhi Wu, Jialiang Shi
{ gzh1994, yuan1z, tianzhi, markshi } @bu.edu

1. Project Description

This project aims to build a model to predict whether a customer will show up to an appointment or not based on demographics characteristics (age, gender, etc), time-related factors, engagement channels (mobile, web, email, etc.), engagement events (notifications, web-clicks, etc).

The main question to be answered is “Which factors (and how much) affect patients that don’t show up to an appointment”. This is critical to determining whether a patient’s medical care will be successful or not.

Other strategic questions to be answered includes:
(1) “Are demographic variables indicative of patients more/less likely to cancel appointments at the last minute”
(2) “What other factors are responsible for driving patient engagement?”

2. Proposed Work

2.1 Proposed Phases of Work

There are roughly four major phases of this project:

(1) Data Representation
(2) Comparative Analysis Research on Models
(3) Models Building and Finetuning
(4) Feature Importance Analysis

2.2 Models

(1) Logistic Regression
(2) Support Vector Machine
(3) Decision Tree (Random Forest)
(4) Neural Networks

3. Dataset Overview

3.1 Size

The size of training set is 1996, and test set size is 222. Even though the size of training and test set is not very large, the number engagement events associated with each training sample is relatively large, which includes over 170,000 events records in total, and about 88 events for each training sample on average.

3.2 Data Imbalance

There is major data imbalance problem in the training set that about 90% of training samples are labeled negative (showed up), and only 10% are labeled positive (cancellations within 3 days of colonoscopies and no show).

3.3 Missing Data

Engagement Events		Enrollment	
Hospital Id	0.000	Hospital Id	0.000
Patient Id	0.000	Patient Id	0.000
Event Date	0.000	Registration Date	0.000
Event Time	0.000	Procedure Date	0.000
Event Name	0.000	Email	0.000
Message Id	0.620	SMS	0.000
Module Id	0.740	Date of Birth	0.298
Event_Desc	0.063	Gender	0.469

Figure 1. Missing Data

4. Data Representation

4.1 Data Preprocessing

Most engagement event data is hard to use, e.g ***User clicking on button***, ***Viewing Page***, etc. We put aside those incomprehensible data for now, and keep ***Notification_Sent*** and its corresponding ***Message_ID***. We may want to utilize all the data in the future, but for now, we plan to focus more the models instead of data processing.

The Enrollment Data (collated final).csv, train.csv and test.csv are three main data sources. There are repetitive data in Enrollment Data, so we deleted it manually.

We merged train.csv with other information in enrollment data.csv and finally get 1996 valid training

samples and 222 test samples, with 11 features including Patient_Id, Hospital_Id, Registration_Date, Precedure_Date, Waiting_Time, Email, SMS, Date_of_Birth, Gender, Message_Id, No_Show/ LateCancel Flag. Sample data is shown in Figure 2.

	Patient Id	Hospital Id	Registration Date	Procedure Date	Waiting Time
2	87255	19	2018/8/2	2018/8/22	1
3	85877	19	2018/8/1	2018/8/29	5
4	85281	19	2018/8/1	2018/8/9	1
5	85267	19	2018/7/27	2018/8/22	8
6	84052	19	2018/7/27	2018/8/23	6
7	84041	19	2018/7/27	2018/8/2	12
8	82100	19	2018/7/26	2018/8/1	8
9	81574	19	2018/7/25	2018/8/28	20
10	81564	19	2018/7/25	2018/8/22	19

(Continued)

Email	SMS	Date of Birth	Gender	Message_Id	No Show/LateCancel Flag
0	0	1950/1/27	Female	7400	0
1	1	1955/2/17	Male	7398	0
0	1	1962/6/14	Female	7397	1
0	1	1948/1/18	Male	7399	0
1	1	1967/1/8	Female	7398	0
1	1	1959/11/22	Male	7400	0
1	1	1943/2/22	Male	7398	1
0	0	1970/4/18	Female	7397	0
0	1	1957/6/21	Male	7399	0

Figure 2. Pre-processed Sample Data

4.2 Data Cleaning

As we can see the Missing Data, shown in Figure 1, for the enrollment information, 29.8% ***Date of Birth*** is missing, and 46.9% ***Gender*** is missing. We know that these two features are determinant and can not be filled by any strategies. The current solution is just to discard them. Thus we have 964 training data and 208 testing data left. Any bool type parameters in dataset is converted to binary 0/1. ***Date of Birth*** and ***Procedure Date*** are combined and translated to a new feature **Age**. ***Registration Date*** and ***Procedure Date*** are converted to new feature Waiting Time. ***Patiend_Id*** are deleted.

5. Experiment Result

We have only tested SVM models for now, and will do hand on experiments on other models shortly.

5.1 Support Vector Classifier

There is a compromise between number of samples and number of features. We want to fully utilize all features for SVM classifier, but in sacrifice of over 1000 samples. We will run SVM on sparse data in the future. Currently, we have finished the following tasks using SVM:

(1) Data Standardization

We standardized dataset along all axis, and centralized to zero mean and rescaled to unit variance component wise.

(2) Feature Importance Analysis

Gives us the importance dict for each feature. ***Waiting Time*** and ***Age*** are two very significant features as we assumed before. Hospital Id are least important, and is most likely to be discarded when doing dimensionality reduction. But missing ***Gender*** and ***Date of Birth*** are still unsolved.

(3) Grid Search

Grid Search is a useful module in sklearn to do finetuning. Here we try different combinations:
{kernel:(linear, poly, rbf),
C:(1,2,3,4,5),
gamma:(1,0.5,0.25,0.125),
class_weight:({1:0,0:9},'balanced')}
By exhaustive searching all possibilities, the optimal params are returned and stored as *SVC_model.m*.

(4) Prediction

By making prediction on test set and calculating the accuracy, unexpectedly, when *class_weight* fixed to {1:0,0:9}, the training accuracy is about 0.89 and test accuracy is 1.0. We strongly suspect that data leakage happens, where there is causality between label and features.

However, When *class_weight* is 'balanced', the accuracy is approximately 0.75. A further exploration on unbalanced control parameter is necessary.

6. Evaluation Criterion

By trying out different models to classify whether a patients would show up or not, we hope to show that the best model can achieve accuracy, precision, and recall are all higher than the 70%.

7. Timeline and Leader

Task	Deadline	Lead
Exploratory Data Analysis	Nov.20	Jialiang Shi
Research on Models	Nov.29	Zhonghao Guo
Implement & Finetune Models	Dec.12	All
Feature Importance Analysis	Dec.25	Zihao Yuan
Report and Presentation	Dec.20	All

References

[1] Working with missing data, pandas 0.23.4 documentation.
[2] Donald B. Rubin. Inference and missing data. Biometrika, Volume 63, Issue 3, 1 December 1976, Pages 581–592.
[3] Bernhard Scholkopf, Alexander J. Smola. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press Cambridge, MA, USA.
[4] S. B. Kotsiantis, D. Kanellopoulos and P. E. Pintelas, Data Preprocessing for Supervised Learning. International Journal of Computer Science Volume 1, 2006.