49-733 / 90-835 Designing Smart and Healthy Systems (Fall 2021)
HW #2: Modeling Patient Engagement
Due November 18, 2021

David Steier

The focus of this assignment is on modeling patient engagement, using the variables you created in HW#1. You will practice skills of clustering (unsupervised learning) and prediction (supervised learning). Each team will have access to all the care (outcome) events created by the class for use as target variables, and as well as all the predictor variables created by the class.

This is a group assignment for your team. Please submit one assignment for each team (and be sure to include your team name and list the team members on your submission). Files turned should one or more Jupyter notebooks. We will ask you to present the results of this HW in the midterm presentation on Nov 23.

Data

Once HW#1 is turned in, the results will be compiled and made available to you regarding the list of hypotheses turned in by each team , the outcome variables, and the predictor variables.

- `List of hypotheses`: Collected hypotheses, separated by group.
- Outcome variables
  - `Appointment_attending_info` Skipped, canceled, rescheduled, on-time
  - `Attend_all_buy` Attends multi-service appointment and buys from the clinic
  - `Attend_single_buy` Attends single service appointment and buys from the clinic
  - `Single_service` Attends single service appointment
  - `All_service` Attends multi-service appointment
  - `Buy_clinics` Buys something from clinic
  - `Buys_or_renews_membership`
  - `Overall_complains`, `frontdesk_complains`, `cv_complains`, `nutrition_complains`
  - `Cancels_membership` (Per year
  - `Subscriptions`
- Predictor variables
  - `Geography`
  - `Patient_demographics`
  - `Appointment_Service` : Number of appts where patient received single service that year and all service
  - `Psicol_features` : Structured fields pertaining to social support and patient psychology
  - `History`: Includes glucose, hemoglobin A1C, cholesterol, and creatinine levels
  - `History_hospital`
  - `Lab_findings_df` (IdProspectos only)
  - `Physical_examination_df` (IdProspectos only)
  - `Ndiet` nutritional diet of patients
  - `Neval` initial nutritional evaluation
  - `Nseg` nutritional followup of the patients
  - `Nplan` nutritional plan
  - `Nininicio` Initial nutritional survey
  - `Nref` Patient referrals
  - `Nret`: retinography (eye exam)

Tasks and Rubric for this Assignment

Each task is worth 6 points, plus 2 points for a timely submission, for a total of 20 points for the assignment.

1. Use the data to test the ten hypotheses created by one of the other groups. So for example, if the hypothesis is that presence of social support is associated with a decreased level of A1c, the team could test the correlation between `ApoyoSocialEmocional` in `Psicolfeatures` and `HBA1C` in `Medical_History`. Remember to match the timestamps as appropriate for the hypothesis you are testing and the availability in the data. Mapping from group numbers to hypotheses to test is as follows:
    a. Group 1 tests the hypotheses from Group 2
    b. Group 2 tests the hypotheses from Group 3
    c. Group 3 tests the hypotheses from Group 4
    d. Group 4 test the hypotheses from Group 1
2. Each group should implement a variation of the clustering method as described in class. That method described at a high level is
    a. Create a matrix in which each row is a patient and each column represents some time period in a sequence, where the first value in the sequence for a patient starts with their first appointment at CdA.
    b. Compute the value in each cell of the matrix i,j as the value of some outcome variable for patient i at time period j. So if the number of appointments scheduled is the outcome variable, the row for Patient A who had his or her first appointment in Jan 2014 might have the number of appointments for Jan 2014, Feb 2014, etc.
    c. Cluster patients by similar sequences of engagement using k-medoids. Similarity is measured by pairwise distance between patient engagement vectors

    Each group should implement a different variation of the method as follows,
    1. Group 1: Use **k = 3**, time period as **month**, and the engagement level as **highest** level of engagement in the month (this was the method implemented by Deloitte in 2018)
    2. Group 2: Use **k = 3**, time period as **quarter** (three months) and the engagement level as **highest** level of engagement in the quarter
    3. Group 3: Use **k = 3**, time period as **month**, the engagement level as **average** level of engagement in the month
    4. Group 3: Use **k = 4**, time period as **quarter**, the engagement level as **average** level of engagement in the month

    Visualize the clusters, and describe the differences. Evaluate the cluster separation using an internal metric such as the silhouette coefficient (youmay use others, see https://en.wikipedia.org/wiki/Cluster_analysis#Internal_evaluation)
    For ranking or average levels of engagement, use the scale as described in class. So if a patient attends a single service appointment (level 8) and complains (level 5), their highest level of engagement is 8, and their average engagement is 6.5.
    1. Cancels Membership
    2. No Engagement (absence of any records for the patient during the time period)
    3. Skips Appointment
    4. Cancels Appointment
    5. Complains
    6. Reschedules Appointment
    7. Buys from Clinic
    8. Attends Single-Service Appointment
    9. Attends Single-Service Appointment and Buys from Clinic
    10. Attends All-Service Appointment
    11. Attends All-Service Appointment and Buys from Clinic
    12. Buys or Renews Membership
3. Each group should develop two models that use predictor variables for a patient in one time period to predict the engagement level the patient will fall into the next time period. The time periods and method of calculating engagement levels will follow the mappings for groups in the previous task
    - Group 1: Based on the predictive variables in **month** i, predict the **highest** level of engagement in the time period in month i+1

- Group 2: Based on the predictive variables in **quarter** i, predict the **highest** level of engagement in quarter i+1
- Group 3: Based on the predictive variables in **month i**, predict the **average** level of engagement in month i+1
- Group 4: Based on the predictive variables in **quarter** i, predict the **average** level of engagement in month i+1

Each of the two models should be a **prediction** model to predict the engagement level (value between 1 and 12. In both cases, you should list the independent and dependent variables, show an appropriate visualization (scatterplot, ROC curve, confusion matrix, etc.) and metric (r-square, AUC, f1-micro-average, etc.) to evaluate your model, and analyze the errors to describe in which situations your model is most likely to make a mistake.