ANLY 530 Late Fall 2019

Course Project Instructions

The purpose of the project is to learn how to formulate a problem statement or research question, determine how to best find a solution to the stated problem or answer to the research question, do that and then develop a final written report and presentation. The project is team-based. Individual grades will include points for how well they contributed to the team effort.

The course project has two deliverables:

1. Project presentation
2. Final report

Each of these deliverables will be described in the paragraphs below.

Final presentations are due on February 19th 2019 and final report based on the discussions on presentation day will be due by February 21st 2019 (noon).

Competitiveness, market share, professional development and personal support to community action, health, culture, education and sport, are linked to a promising new market. Coupled with the development of organizations, the pressure to achieve goals more audacious, employees increasingly overwhelmed, they end up buying some disturbance in the health-related type of labor activity. The objective of this project is to apply some machine learning algorithms in the prediction of absenteeism at work. The database is the information collected records of absenteeism from work during the period of July/07 to July/2010 in a Courier company. Absences certified with the International Classification of Diseases were stratified into 21 categories, the data were tabulated and stored in two datasets (training and testing set).

Attribute Information:

1. Individual identification (ID)

2. Reason for absence (ICD).

Absences attested by the International Code of Diseases (ICD) stratified into 21 categories (I to XXI) as follows:

I Certain infectious and parasitic diseases

II Neoplasms

III Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism

IV Endocrine, nutritional and metabolic diseases

V Mental and behavioural disorders

VI Diseases of the nervous system

VII Diseases of the eye and adnexa

VIII Diseases of the ear and mastoid process

IX Diseases of the circulatory system

X Diseases of the respiratory system

XI Diseases of the digestive system

XII Diseases of the skin and subcutaneous tissue

XIII Diseases of the musculoskeletal system and connective tissue

XIV Diseases of the genitourinary system

XV Pregnancy, childbirth and the puerperium

XVI Certain conditions originating in the perinatal period

XVII Congenital malformations, deformations and chromosomal abnormalities

XVIII Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified

XIX Injury, poisoning and certain other consequences of external causes

XX External causes of morbidity and mortality

XXI Factors influencing health status and contact with health services.

And 7 categories without (CID) patient follow-up (22), medical consultation (23), blood donation (24), laboratory examination (25), unjustified absence (26), physiotherapy (27), dental consultation (28).

3. Month of absence

4. Day of the week (Monday (2), Tuesday (3), Wednesday (4), Thursday (5), Friday (6))

5. Seasons (summer (1), autumn (2), winter (3), spring (4))

6. Transportation expense

7. Distance from Residence to Work (kilometers)

8. Service time

9. Age

10. Work load Average/day

11. Hit target

12. Disciplinary failure (yes=1; no=0)

13. Education (high school (1), graduate (2), postgraduate (3), master and doctor (4))

14. Son (number of children)

15. Social drinker (yes=1; no=0)

16. Social smoker (yes=1; no=0)

17. Pet (number of pet)

18. Weight

19. Height

20. Body mass index

21. Absenteeism time in hours (target)

More about data data:

ID {31.0, 27.0, 19.0, 30.0, 7.0, 20.0, 24.0, 32.0, 3.0, 33.0, 26.0, 29.0, 18.0, 25.0, 17.0, 14.0, 16.0, 23.0, 2.0, 21.0, 36.0, 15.0, 22.0, 5.0, 12.0, 9.0, 6.0, 34.0, 10.0, 28.0, 13.0, 11.0, 1.0, 4.0, 8.0, 35.0}

Reason_for_absence {17.0, 3.0, 15.0, 4.0, 21.0, 2.0, 9.0, 24.0, 18.0, 1.0, 12.0, 5.0, 16.0, 7.0, 27.0, 25.0, 8.0, 10.0, 26.0, 19.0, 28.0, 6.0, 23.0, 22.0, 13.0, 14.0, 11.0, 0.0}

Month_of_absence REAL

Day_of_the_week {5.0, 2.0, 3.0, 4.0, 6.0}

Seasons {4.0, 1.0, 2.0, 3.0}

Transportation_expense REAL

Distance_from_Residence_to_Work REAL

Service_time INTEGER

Age INTEGER

Work_load_Average/day_ REAL

Hit_target REAL

Disciplinary_failure {1.0, 0.0}

Education REAL

Son REAL

Social_drinker {1.0, 0.0}

Social_smoker {1.0, 0.0}

Pet REAL

Weight REAL

Height REAL

Body_mass_index REAL

Absenteeism_time_in_hours REAL

Your job is to design a machine learning algorithm which tends to predict the absenteeism in hours.

Task 1- Mandatory

Since the target variable is continuous, you should break it to some smaller sub groups:

Group 0: Number of hours=0

Group 1: 0 < Number of hours <= 6

Group 2: Number of hours > 6

Task 2- Optional (extra credit)

Predict the number of hours of absence without converting it to categorical variable (consider the continuous value)

Final Project Report

The final report and presentation should cover virtually everything about the project. It should cover the situation, problem or challenge that required attention, the relevant background, related work, data, and technical details of the analysis, conclusions and possible directions for future work. It is recognized that not all of the following sections will pertain to each report. However, it is strongly recommended that these

section topics be used as a guideline for your final project reports. Final presentations can follow your final report in text and graphical content.

- Introduction, motivation and general description of the situation, problem or challenge.
    - What is the situation, problem or challenge you are addressing?
    - What preliminary examination leads you to believe analytics could help?
    - What are the shortcomings of the current work/analysis that analytics could help with?
- Related work.
    - Provide a thorough background for the project; e.g. about the situation, problem or challenge, about other companies that have undergone similar situations, problems or challenges and how they handled them or did not, etc.
    - How does this project relate to other work that has been done on this situation, problem or challenge?
- Data
    - Give a complete description of the data you use during the project, including any you reject.
        - Provide a detailed description of your data.
        - Provide any exploratory data analyses you complete.
- Technical Approach
    - Give a detailed description of the process for your entire project.
    - Given a detailed description of your approach to the algorithm you have proposed. You do not have to describe well known approaches themselves, e.g. linear regression. You do have to describe how you applied the approach you used.
- Test and evaluation
    - Describe how you test your approach to ensure that it is valid.
    - Discuss the validity of your approach.
    - Describe how you will evaluate your results and/or conclusions including any specific metrics, output data, completed analyses, etc.
    - Discuss the baseline you will use to compare your results to.
    - Discuss how well your approach worked to address the situation or challenge, solve the problem or answer the research question.
    - Discuss any potential future work. For example, if you were not able to resolve the situation or problem or answer the research question what will it take to do so? What else needs to be done?
    - Evaluate and report whether or not someone unfamiliar with your work could accurately replicate it.
- Written work and Presentation Style
    - Written work will be graded using the rubric provided.
    - Presentation style will be graded on comprehensiveness and inclusiveness, as well as using the rubric provided.
- The final report should follow the guidelines provided by the APA or IEEE.