Springboard Data Science Course

Data Science Capstone Project 1

Orthopedic Biomechanical Features

Michelle Ide - 2/1/2021

~~DATA and ANALYSIS~~

The 'Orthopedic Biomechanical Features' project can be found at Kaggle's machine learning website. This capstone project utilizes one data file to classify patient x-ray results into Normal or Abnormal results. Below is a description of the data, analysis, and deliverables. Additional details about the project can be found in the *"Capstone 1 Project Proposal"*.

Data Source & Details

Data obtained from UCI Machine Learning Repository at the following link: http://archive.ics.uci.edu/ml/datasets/Vertebral+Column#.
(Dr. Henrique da Mota during medical residence period in the Group of Applied Research in Orthopaedics (GARA) of the Centre Medicao-Chirurgical de Radaptation des Massues, Lyon, France)

- Total 310 records: 210 Abnormal, 100 Normal
- 1 Target, binomial string 'Abnormal' or 'Normal'
- 6 quantitative Features with their new column names

Feature	Column
1) Pelvic Incidence	INCIDENCE
2) Pelvic tilt	TILT
3) Lumbar lordosis angle	ANGLE
4) Sacral slope	SLOPE
5) Pelvic radius	RADIUS
6) Grade of spondylolisthesis	DEGREE

Data Cleaning

Steps followed:

- ***Search for empty values:*** No empty values were found
- ***Review for uniqueness:*** Non-unique values were found as expected for this type of data, nothing highly unusual
- ***Target preparation:*** The string class was encoded for modeling purposes
- ***Feature preparation:***
    - Quantitative features were clean of strings or unusual text
    - Outliers were researched using Tukey's method and visual inspection to remove values >2.5 times the nearest value, this resulted in removal of a single record.

Analysis

Statistical Values

The feature 'degree' has a large variance as seen in the table below, distribution of the data demonstrated this larger value is due specifically to 'Abnormal' degree results.

	incidence	tilt	angle	slope	radius	degree
count	309.000000	309.000000	309.000000	309.000000	309.000000	309.000000
mean	60.272260	17.572396	51.942408	42.699864	117.953762	25.027289
std	16.804832	10.010988	18.583057	12.676949	13.326194	30.234211
min	26.147921	-6.554948	14.000000	13.366931	70.082575	-11.058179
25%	46.426366	10.688698	37.000000	33.340707	110.709912	1.594748
50%	58.599529	16.417462	49.775534	42.373594	118.343321	11.463223
75%	72.643850	22.181798	63.000000	52.549422	125.480174	40.880923
max	118.144655	49.431864	125.742385	79.695154	163.071041	148.753711

Correlations

Features were tested for correlations >90% but all were below the threshold. If interested in additional feature selection, results showed a high correlation between tilt and incidence.

Hypothesis Testing

Hypothesis statement: Features (a & b) with an alpha >0.05% do not demonstrate a statistically significant difference and therefore do not contribute to the target classification of Abnormal/Normal.

For this single target, multi-feature supervised learning project, a student's ttest was performed, testing the hypothesis on each feature. All failed the hypothesis test, validating their contribution to the labeling of target values and therefore were included in models.

Deliverables

309 records are included in final clean results: 209 Abnormal, 100 Normal

For use in machine learning models, the following variables were created and stored as dataframes:

- X.csv (features-only):
- Y.csv (target-only, *encoded*) Y.csv
- data.csv (full dataset *encoded*) data.csv