Springboard Data Science Course

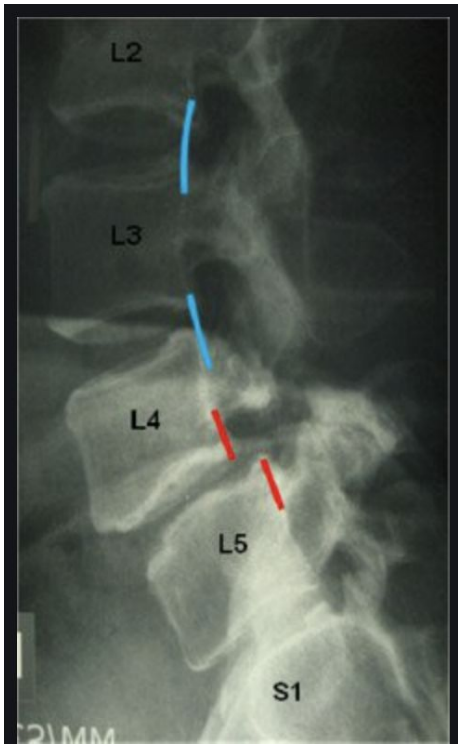
Data Science Capstone Project 1

Orthopedic Biomechanical Features

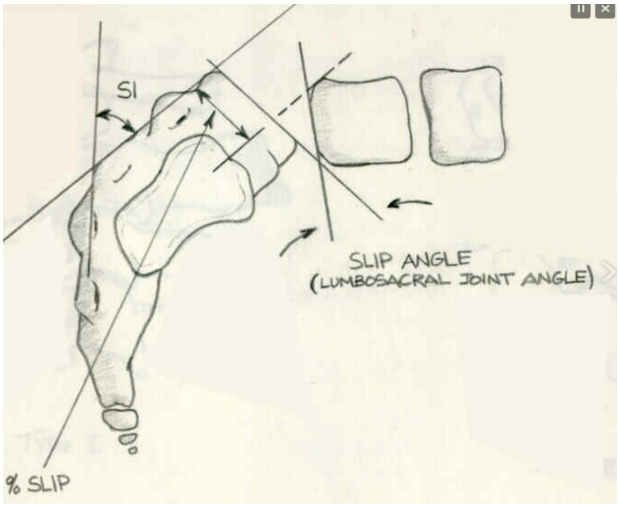
Michelle Ide - 2/1/2021

Symptoms for Spondylolisthesis are very common making early diagnosis and treatment difficult. They include:

- Muscle spasms in the hamstring.
- Back **stiffness**.
- Difficulty walking or standing for long periods.
- **Pain** when bending over.
- Numbness, weakness or tingling in the foot.

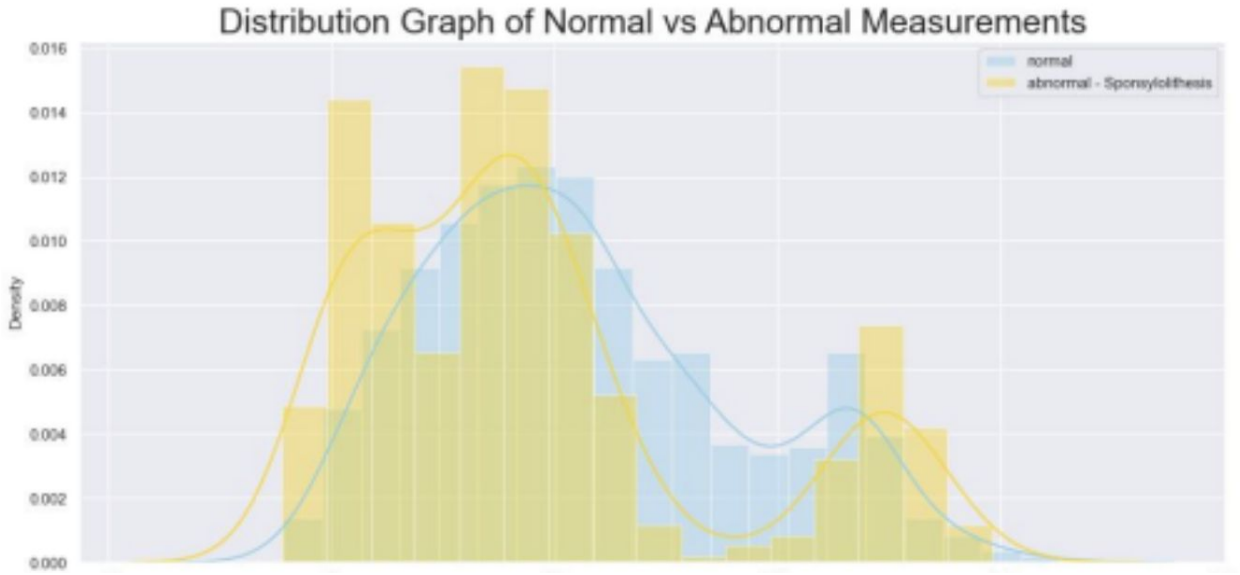


Spondylolisthesis is a medical term for abnormal alignment of the spine and has many causes but the outcomes are continued deterioration and increasing debilitation without treatment. Early treatment provides life-long improved outcomes.



Proper diagnosis requires x-rays containing various angles of the spine to correctly identify normal versus abnormal. Several angles are used and by simultaneously comparing, physicians can identify potential issues. Unfortunately, this is not as straightforward as it may seem, these angles do not have distinct cut-off values clearly demarking normal from

abnormal. By taking these in conjunction, angles often overlap for normal and abnormal. For example, below is a distribution of abnormal versus normal measurements for the "degree" of the spine.



In yellow we see abnormal measurements overlap with normal. These types of comparisons and classifications can utilize machine learning to assist radiologists in determining proper labels, saving time and money resulting in improved outcomes for patients with faster results and treatment.

Project

This is a supervised classification project using quantitative biomechanical data taken from patient x-rays to predict results as either Normal or Abnormal. In this project abnormal indicates spondylolisthesis of the lumbar spine specifically.

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 for cleaning

- ***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

A dataframe stored as csv file contained the feature data "X" and encoded target data "Y" for use in machine learning models.

Machine Learning Modeling

Method

- Stratified Train-Test Split, 30% test

- Resampling: SMOTE, ADASYN
- Parameter Tuning: GridSearch with KFold split

Models

- Logistic Regression
- Gradient Boost
- Support Vector Machine
- Random Forest
- KNearest Neighbors
- Naive Bayes

Results

Detailed results for all models with both resampling methods are contained in the addendum at the end of this report.

Both sampling techniques performed similarly with Logistic Regression demonstrating the best performance. However, for weighting results to prevent false negatives, ADASYN resampling would provide better results.

Using SMOTE resampling, Logistic Regression had a 95% ROC AUC and Recall of 97% for Normal and 89% for Abnormal results. Therefore, 97 out of 100 normal results were accurately found and 89 out of every 100 abnormal results were found and properly classified.

With ADASYN, the Logistic Regression recall of 100% for Normal and 84% Abnormal provides important accuracy by determining 100 out of 100 normal results.

Of the remaining models, Random Forest was a close second to Logistic Regress however it performed similarly to SVM, and Gradient Boosting.

KNNeighbors and Naive Bayes are not competitive with the other models and not recommended for this purpose.

ADDENDUM

Summary of Accuracy Scores for both resampling methods:

SMOTE

	Trianing Score	ROC AUC	Recall - Abnormal	Recall - Normal	Precision - Abnormal	Precision - Normal	Weighted Average - Precision
Logistic Regression	63%	95%	89%	97%	98%	81%	93%
Gadient Boosting	41%%	96%	86%	87%	93%	74%	87%
SVM	0%	NA	86%	87%	93%	74%	87%
Random Forest	34%	95%	87%	87%	93%	76%	88%
KNeighbors	0%	89%	83%	80%	90%	69%	83%
Gaussian NB	0%	89%	76%	87%	92%	63%	83%

ADASYN

	Trianing Score	ROC AUC	Recall - Abnormal	Recall - Normal	Precision - Abnormal	Precision - Normal	Weighted Average - Precision
Logistic Regression	61%	96%	84%	100%	100%	75%	92%
Gadient Boosting	36%	94%	84%	77%	88%	70%	82%
SVM	0%	N/A	93%	70%	83%	87%	86%
Random Forest	5%	95%	83%	93%	96%	72%	88%
KNeighbors	0%	81%	83%	80%	90%	69%	83%
Gaussian NB	0%	86%	75%	87%	92%	62%	82%

CONFUSION MATRIX SUMMARY

SMOTE

		ABN	NOR
LR	ABN	56	1
	NOR	7	79
GB	ABN	54	4
	NOR	9	26
SVM	ABN	54	4
	NOR	9	26
RF	ABN	55	4
	NOR	8	26
KNN	ABN	52	6
	NOR	11	24
NB	ABN	48	4
	NOR	15	26

ADASYN

		ABN	NOR
LR	ABN	53	0
	NOR	10	30
GB	ABN	53	7
	NOR	10	23
SVM	ABN	52	4
	NOR	11	26
RF	ABN	52	2
	NOR	11	28
KNN	ABN	52	6
	NOR	11	24
NB	ABN	47	4
	NOR	16	26