Eric Chao, Camila Djurinsky Zapolski, Eric Powers

April 30th, 2024

Written Report - Osteoporosis

Project Inspiration:

For our final project, we wanted to explore if it is possible to predict osteoporosis using a variety of different linear algebra and machine learning concepts we have learned throughout the semester. The dataset used was obtained from Kaggle and contains 13 columns of information regarding an individual's health status (age, gender, alcohol consumption, etc.), which are our predictor variables. Additionally, each individual within this dataset has provided information on whether they have osteoporosis or not, which is our response variable. We chose this topic because we were all interested in the field of bioinformatics and how machine learning algorithms can be used to help advance medical research through the development of predictive modeling.

Linear Algebra Concepts:

The application of Linear Algebra concepts learned throughout this course is a critical element to our project. The list of concepts used in our project include:

- Classification
    - Random Forest, Gradient Boosting, Logistic Regression, K-Nearest Neighbors, Support Vector Machine
- Clustering
    - K-Means Clustering
- Dimensionality Reduction
    - Principal Component Analysis

Please refer to our appendix for definitions of each concept and how it applies to our project.

Methodology:

The motivation behind this project was to mimic a real-world Data Science project to provide us with a glimpse of the type of work that our group members may see in the future. The methodology that we followed in setting up this project was to follow a general Data Science and Machine Learning pipeline from getting the data to interpreting results that can provide actionable insights. Our process is broken down into five stages:

1. Dataset Selection and Wrangling
2. Visualize and Understand
3. Model Training
4. Model Enhancement / Feature Engineering
5. Insights and Model Deployment

These five stages provided our group with guidance in the appropriate steps needed to transform a brand new dataset into creating a Machine Learning model that can provide value on our topic of predicting for osteoporosis.

Dataset Selection and Wrangling:

The Osteoporosis dataset from Kaggle has 13 columns of categorical information regarding an individual's health status. In order to implement a variety of different machine learning tools, we had to encode the dataset. This means assigning numerical representations of the data in order for the algorithms to be able to conduct the necessary calculations. For example with decision trees classification, the algorithm must make a decision by choosing one specific category represented by a numerical value. This concept is applicable with gradient boosting and random forests as they use decision trees represented by encoded categories as well. By encoding every column, it ensures that problems with these three models along with other potential algorithms such as clustering or regression can be avoided.

Visualize and Understand:

Before creating any models, it is important to do some general exploratory data analysis to better understand the dataset. For each variable, a histogram was created to understand the distribution of ages captured within our dataset and the categories that each individual fell under. In regards to age, most individuals were between 18 and 40 years old, but there were still many between 40 and 90 years. For the categorical variables, there was practically an even distribution of each category per column of our dataset. This provided us with confidence that no predictor variable may be the cause of a biased prediction outcome and impacting any models that may be developed. In any data science project, we are typically faced with null values and outliers that must be dealt with. However, after our exploratory data analysis, we were fortunate to find out that there were no null values or outliers and that no further data cleansing needed to occur.

Model Training:

The model development process required the application of classification and clustering Linear Algebra concepts on our dataset to further understand its fundamental principles and why a certain methodology performs better than the other. With the classification methods, we can see that the methodologies that utilized decision trees in the background performed more accurately than the other algorithms. This could be attributed to the categorical nature of the dataset, making decision trees more equipped for its analysis. For our clustering methods, we wanted to test if it was rational to use unsupervised learning on a fully labeled dataset to cluster together individuals who do and do not have osteoporosis. Unsurprisingly, our accuracy did not compare to all the other clustering algorithms, highlighting the difference in use cases between an unsupervised and supervised model. Ultimately, it was discovered that the Gradient Boosting Classifier had the best overall accuracy score of 91.58% among all the model implementations.

Looking further into the Gradient Boosting model, it had the best precision — the proportion of observed prediction over truly expected — of all the algorithms, with 86% precision in predicting that an individual does not have osteoporosis and 100% precision in predicting that an individual has osteoporosis. None of the algorithms achieved a precision and accuracy score as significant as the Gradient Boosting Classifier. The closest results were those of the K-Nearest Neighbors algorithm, with 86.48% accuracy, 80% precision in predicting individuals without osteoporosis, and 97% precision in predicting individuals with osteoporosis. Thus, the Gradient Boosting is clearly the best algorithm for this dataset.

Model Enhancement / Feature Engineering

After identifying that Gradient Boosting is the best predictive model, we wanted to identify ways to simplify our model without reducing its overall accuracy. Given the nature of our dataset, many of the common approaches towards dimensionality reduction could not be used. The first method we tested was creating a correlation matrix to identify highly correlated variables between predictor variables. Knowing that correlation is often derived from a scatterplot of points from continuous variables, it was difficult to identify any correlations. The next method we tested was using Principal Component Analysis discussed in class. Knowing that in theory, principal components are derived by finding linear combinations between predictor variables, it was evident it was not possible with categorical variables. Our final approach towards enhancing our model was to use the feature importance function provided in the random forest model. This function provides modelers with the most important variables in predicting for osteoporosis. Since gradient boosting models utilize the concept of decision trees like a random forest, it seemed relevant to us to use this feature. It was evident that our continuous variable age had significant importance in predicting osteoporosis. We decided to eliminate the five least important variables to create a simplified model. The result produced an accuracy score nearly identical to our thirteen variable model.

Insights and Model Development

At the conclusion of our analysis and model development, we have found that the gradient boosting model performs the best at predicting for osteoporosis. More impressively, we were able to eliminate five variables from our original model and maintain an almost identical prediction accuracy. This project highlights the potential capabilities of predictive modeling and linear

algebra in the healthcare field. With this model, there are a variety of next steps that can be taken. In regards to helping society, the vision for a model like this could be to help identify, warn, and provide resources to individuals who may unknowingly have or will have osteoporosis. Creating a website or application where users can answer questions pertaining to the predictor variables to generate a prediction using our model can be an immense resource for health professionals.

Behind the scenes, there are always improvements that can be made to our model. It was evident that there were a lack of continuous variables within our dataset. Finding additional continuous variables such as height, weight, blood pressure, etc. could prove to be helpful in further predicting for osteoporosis. Additionally, the use of other dimensionality reduction methods can be used to enhance our models. For example, the concept of Multiple Correspondence Analysis was discovered when finding an alternative for PCA, but with categorical variables. The possibilities of a project like this in healthcare truly are endless and hopefully the understanding of these linear algebra and data science concepts can lead to exciting innovations for the future.

Appendix: Linear Algebra Concepts

- Random Forest Classifier

Random Forest Classifiers construct many individual decision trees at training. Predictions from all these trees are put together to make the final prediction. Random Forest classifiers are able to tackle classification and regression problems. Within the internal workings of the Random Forests, the algorithm uses criteria to partition the data through the use of linear algebra calculations like the dot product to determine the optimal way of partitioning based on the features of the specific dataset.

- Gradient Boosting Classifier

Gradient Boosting algorithms optimize the loss function by iteratively fitting weak learners (usually decision trees) to the residuals of the previous iteration. Gradient Boosting relies on gradients, vectors, and matrices within its algorithm for proper optimization. This optimization involves constructing trees that minimize the loss function, involving feature transformations. One can visualize the training of gradient boosting models as matrix operations, with each row being the data sample and each column being the tree prediction.

- Logistic Regression

Logistic Regression is a linear classification algorithm that models the relationship between independent variables — age, gender, race, ethnicity, etc — and a binary dependent variable — having osteoporosis (1 yes, 0 no). The algorithm makes predictions on the data based on a linear combination of the input features, multiplied by their corresponding coefficients, and added together (a.k.a a dot product between the feature vector and the parameter vector). Logistic Regression can also be visualized using matrix notation, with the rows being the sample data and the columns being their respective features (the parameter vector). Logistic Regression uses a cost

function that measures the difference between the predicted probabilities and the actual labels (predicted vs true). Finally, Logistic Regression uses a Sigmoid function to transform the output of the linear combination into a value between 0 and 1, which is the output that represents the probability of belonging to the "positive" class — having osteoporosis.

- K-Nearest Neighbors

K-Nearest Neighbors algorithms use classification or regression to make predictions based on the similarity of instances. The similarity of data points is determined by distance metrics which involve vector operations. Additionally, the data points themselves are represented as feature vectors in order to be compared and analyzed. In some scenarios, K-Nearest Neighbors also uses dimensionality reduction (i.e. PCA and SVD) to preprocess the data.

- Decision Tree Classifier

Decision Tree Classifiers are a type of non-parametric, supervised learning method that we utilized for our Osteoporosis analysis. Behind the scenes, these classifiers pick features recursively to split on that maximize information gain at each step. The algorithm is able to compare the entropy of a parent node with all of its children as it navigates closer and closer to the leaf node (final decision). For our scenario, the leaf nodes at the end of the tree represent whether or not a person has osteoporosis.

- Support Vector Classification

Support Vector Classification (SVC) is a powerful tool that we used for our binary classification task. The goal of SVC is to identify the optimal hyperplane that divides classes in a feature space. The best hyperplane to divide our osteoporosis and non-osteoporosis classes was chosen by the algorithm with the help of linear algebra concepts like matrix inversion, matrix multiplication, and inner products. SVC maximizes the distance between the support vectors for a given class and the

hyperplane, which is often referred to as the margin. The reason for this is because, in general, the larger the margin the lower the generalization error of the classifier.

- K-Means Clustering

K-Means Clustering is an unsupervised learning method used to cluster together points based on similar traits. Each point is clustered into their respective group based on their distance to a corresponding centroid. Since our intention is to classify individuals based on whether they have osteoporosis or not, we wanted to observe whether or not these data points could be clustered into the two respective groups. Although clustering is not commonly used on data with defined labels, we wanted to be able to identify the discrepancies between clustering and classification methods.

- Principal Component Analysis on Categorical Data

Principal Component Analysis is a dimensionality reduction technique used to find the linear combination of our predictor variables into a set of linearly uncorrelated variables called principal components. The goal of this process is to explain as much of the variance explained in the original dataset using fewer “dimensions” or variables. In working with PCA in our osteoporosis dataset, it was observed that this dimensionality reduction technique does not work with categorical variables. In learning about the intricacies and details behind PCA, it was concluded that continuous predictor variables must be used for any process involving PCA.

Bibliography

1. "Learn." *Scikit*, scikit-learn.org/stable/. Accessed 25 Apr. 2024.
2. "What Is Random Forest?" *IBM*, www.ibm.com/topics/random-forest. Accessed 25 Apr. 2024.
3. Saini, Anshul. "Gradient Boosting Algorithm: A Complete Guide for Beginners." *AnalyticsVidhya*,10,Jan.2024,www.analyticsvidhya.com/blog/2021/09/gradient-boosting-algorithm-a-complete-guide-for-beginners/.
4. "What Is Logistic Regression?" *IBM*, www.ibm.com/topics/logistic-regression. Accessed 25 Apr. 2024.
5. "K-Nearest Neighbor(KNN) Algorithm." *GeeksforGeeks*, GeeksforGeeks, 25 Jan. 2024, www.geeksforgeeks.org/k-nearest-neighbours/.
6. "1.10. Decision Trees." *Scikit*, scikit-learn.org/stable/modules/tree.html. Accessed 25 Apr. 2024.
7. "1.4. Support Vector Machines." *Scikit*, scikit-learn.org/stable/modules/svm.html. Accessed 25 Apr. 2024.
8. "K Means." *CS221*, stanford.edu/~cpiech/cs221/handouts/kmeans.html. Accessed 25 Apr. 2024.
9. "A Step-by-Step Explanation of Principal Component Analysis (PCA)." *Built In*, builtin.com/data-science/step-step-explanation-principal-component-analysis. Accessed 25 Apr. 2024.