Final Project for CIS 4321

Libin Varughese

Department of Computer Information Systems, California State Polytechnic University, Pomona

CIS 4321: Data Mining

Dr. Fadi Batarseh

May 12, 2023

Introduction

I am interested in how we can use technology to better understand the human body and more accurately diagnose diseases and other illnesses. By using modern computing technology and predictive models, we can vastly improve the medical industry by increasing the amount of preventative treatment for patients who are likely to acquire or potentially suffer from diseases such as cancer, diabetes, heart disease, and many more. The reason I am so interested in this domain and my personal motivation for this project is because I have family (all my direct family) is involved in the medical industry in one way or another, whether it be patient care or health administration, and I have family members who are diagnosed with the diseases I mentioned earlier that could have been possibly avoided if they received preventative treatment.

My framing question is essentially, “ How accurately can predictive models diagnose disease/illness?” The results that I am going to present based on this question are important to note because if the results show that predictive models can accurately predict and/or diagnose illnesses, then the technology should get more support and acceptance throughout the medical field to assist medical professionals and healthcare providers with providing preventative care. By providing preventative care instead of performing costly treatments, patients can save money and will not have to be burdened by expensive medical bills and healthcare providers can lower their costs and be viewed as not only more profitable, but as caring to their customers. This new view of them will allow them to reduce customer/patient churn rate, while also increase the number of customers they treat as their treatment options will be seen as more customer oriented than simply a way to make profit.

The stakeholders that are involved in this are healthcare providers, medical professionals, insurance companies, and customers. Healthcare providers are important, as mentioned earlier, as they can reduce their costs by providing cheaper preventative care rather than expensive treatments. The preventative treatments can also reduce the number of patient complaints and lawsuits due to mishaps, which will not only save the provider money, but allow them to both uphold and bolster their reputation in the eyes of the public. The providers are also the ones that would be the ones to implement the predictive models and the systems required to run them throughout their organization, so doctors and other medical professionals would access the systems through the provider. In this way, the healthcare provider is possibly the most important key stakeholder in this subject.

The insurance companies are a key stakeholder as the majority of people in the U.S. have some sort of health insurance that they rely on whenever they go for some sort of medical checkup or procedure. The companies can save money due to preventative care usually being cheaper than treatments, so if more patients are diagnosed earlier and make insurance claims for preventative care rather than treatment, the insurance company can spend less on covering the patient's bill. In some cases, preventative care is unclaimable, so if the patient is told to practice preventative care, then the company would not even have to cover anything other than the checkup/consultation fee the patient acquired, which is usually quite cheap.

The medical professionals and customers are both key stakeholders as they are the ones who are both using and directly benefit from the predictive models being implemented. Doctors can reduce the number of misdiagnoses they are responsible for, protecting them from lawsuits and the possibility of losing their license to practice, and they can treat more patients than they usually can, increasing the number of checkup/consultation fees they can collect, which benefits them, the healthcare provider, and the insurance companies. Patients/customers can get more accurate diagnoses and preventative care options that can help them not only save money as they will not have to go for treatments but can live healthier and longer lives due to the fact that they found out about their illnesses ahead of time and took steps to prevent it from severely affecting them. Customers in this way benefit the most from the models' adoption and become a key stakeholder.

Problem Statement

The problem I wish to explore is how can diseases such as cancer, diabetes, strokes, and many others be more accurately diagnosed? Many people every year die from these and other diseases and two of the reasons why this happens is either due to misdiagnosis from medical professionals and/or late diagnosis, which has led to patients being declared untreatable as the disease has either spread too far or has caused irreparable damage. This is something that I believe that predictive models can help solve and leads to my analytical question.

My analytical question is "How accurately can predictive models predict/diagnose a disease such as diabetes with health data collected from patients such as BMI, age, glucose levels, previous disease diagnoses?" This question is important as mentioned earlier because if the models are shown to have a high rate of accuracy in diagnosis of diseases, then the models can be implemented into the medical industry to assist in patient care, preventative care, and help all the involved key stakeholders save money that would be wasted on treatments and misdiagnoses due to human error.

The type of data analytics approach that I will be using to support my analytical question is classification as the model is trying to predict one of two possible classes that the patient belongs to. Those two classes are if they have diabetes (1), or they do not have diabetes (0). Classification is about placing predicted and future values into predefined classes, which is very useful in the medical field as in the case of disease, one either has it or does not. There is no other option, so classification will help support my analytical question.

Understanding the Dataset

My dataset, in CSV format, is from Kaggle.com and it is titled “Diabetes prediction dataset.” The target variable of interest is the column labeled ‘diabetes,’ which is a binary class label and is an integer type with the only values being 0 (meaning negative for diabetes) and 1 (meaning positive for diabetes. The dataset contains no missing values and consists of 100,001 rows/instances full of values from 9 different columns (diabetes being the 9th column). The first attribute is gender, which is a string type and is either male or female. The second attribute is age, which is an integer type. The third attribute is hypertension, which is a binary class label, an integer type, and it states if the patient has been (signified by a 1) or has not been (signified by a 0) diagnosed with hypertension. The fourth attribute is heart disease, which is also a binary class label, an integer type, and it states if the patient has been (signified by a 1) or has not been (signified by a 0) diagnosed with heart disease. The fifth attribute is smoking history, which is a string type and provides a simple detailed explanation of if the patient has a history of smoking and if they do they, what is their current status (current, former, never, no info, ever). The sixth attribute is BMI, which is a float type and stands for Body Mass Index. The seventh attribute is HbA1c level, which is a float type and is the average amount of glucose in your blood from the last two to three months measured in millimoles per mole. The eighth attribute is blood glucose level, which is an integer type, and is the amount of glucose (sugar) in someone’s blood in milligrams per deciliter.

The challenge with my dataset is regarding the smoking history attribute column. The data has many varied values depending on the patient’s history with smoking such as current, former, never, etc. Though the gender column is also in a string format, it only has two values, male and female, and that can be changed to 0 (for male) and 1 (for female). I can do something similar for the different values for smoking history, but I am unsure if that is necessary. The column may not be relevant at all or changing the values to integers may cause issues when inputting the data into a predictive model. I will take the chance and remove the column from the dataset the models will use as I believe that is the best decision in this situation.

I will also be sampling the data for my analysis as I doubt my computer can handle such a large dataset. I will use 10,099 rows of the dataset and I believe this will not only run on my computer, but still be enough data to provide accurate results. This new sampled dataset will be called “Diabetes prediction dataset sample” and will exclude the smoking history data column.

Data Analytics

The first method I used was Decision Trees for my diabetes dataset. Decision Trees work by using a series of if-then rules to make decisions on how to split the data it is working with. It starts with a root node, then branches off into decision nodes until it reaches a prediction or classification, which is referred to as a leaf node. By visualizing the entire tree, one can see where the rules the tree has are applied and which variables are seen as the most important. It can also work with any number of variables, or predictors in our case, and the models are simple to understand and analyze. It was well suited for my data because decision trees can be used for classification and the target variable in my data was a binary variable (one of two predefined classes), which is well suited for decision trees. The way I applied the decision tree was I first split the data into an ‘X’ feature matrix containing the predictor variables and a ‘y’ target variable; then I split the data into training (70%) and testing (30%) sets. After this, I fitted the training data to the decision tree and then I plotted the tree and produced a complex tree model. I later used GridSearchCV to prune the tree and have it produce the optimal parameters and performance metrics. I then replotted the decision tree after visualizing in a bar graph which features were seen as the most important to view the optimization that took place and which rules were classified as the most important in making decisions.

The second method I used was Logistic Regression. Logistic Regression is a linear model that is applied to classification tasks where a discrete value needs to be predicted. It maps a continuous value into probability, which can then be applied to the prediction a binary label. It is used to describe the relationship between a target variable and a set of predictor variables and to create a model for interpretation of that relationship. The way I applied logistic regression to my data was I first scaled the training and testing data, after importing the StandardScaler from sklearn, that I had made earlier before implementing the decision tree method. After that, I imported the logistic regression model and fitted it with my scaled training data and the target variable training data. Then I created a prediction variable for my target variable to use in a classification report imported from sklearn and compared the actual testing data values with the predicted ones to view the model's accuracy, recall, precision, and f1-score. To improve the accuracy/effectiveness of the model, I used GridSearchCV again and fitted it with the training data. I created a second prediction variable based on the grid search and printed out another classification report to compare the performance of the logistic regression model before and after the grid search.

The last method I used was Neural Networks. Neural Networks is based on the cerebral cortex of the human brain and is a "Data processing system consisting of a large number of simple, highly interconnected processing elements (artificial neurons)" (Tsoukalas and Uhrig, 1997). They can be used for both regression and classification, which means it will work for my data task and they “combine the predictor variables’ information to capture the complicated relationships between the predictors and between the predictors and target variable” (Module 11 Colab on Neural Networks). “A neural network is similar to linear regression, but neural networks model a more complex relationship and more nonlinear relationship” (Module 11 Colab on Neural Networks. The way I applied neural networks was I created a new scaler to scale the data that I partitioned earlier into training and testing sets and fit the scaler with our predictors’ training data. I then scaled the predictors’ training and testing data, instantiated a neural network after importing it from sklearn, and created another grid search for the neural network’s parameters. The next step was to fit the grid search variable, review its results in a data frame, creating another prediction variable and analyze its performance with a classification report, ROC curve, and an AUC score.

Results

My results have shown that the Neural Network model provides the highest accuracy and was able to model the data the best with minimal runtime. The Decision Tree took the longest to run fully, especially with the grid search, but I believe that is due to the large number of rows (10,099 rows) from the sampled dataset and the limitations stemming from my laptop. The Logistic Regression model got high accuracy of 96%, but that is the lowest of the three models.

However, when I compared the classification models against each other using the classification model evaluation Python script, the Logistic Regression model had the highest accuracy, the best Roc Curve, and highest AUC score compared to the Neural Network and Decision Tree. This might be due to these models being compared without grid search optimization, so the results may be varied if they were compared post-optimization. I believe that this means that classification algorithms/models that are based on or similar to linear models or linear regression produce more accurate predictions based on binary labels and predictors that are all numeric. It could also be due to the fact that my dataset does not have many predictor variables, only seven total, so if I had maybe double or even triple the amount, the accuracy of the models may skew more towards decision trees rather than the other two. If I had included the smoking history data column into my models, it may have also varied the results as that variable had multiple class labels for each different patient history provided. I removed it as I believed it would mess with the functionality of the models as I was not accustomed to working with such varied class labels in a numeric format. It should also be noted that the logistic regression's AUC score and accuracy is not much higher than the neural network's AUC score and accuracy with the difference only being .002 for AUC score and .002 for the accuracy. This could be a random occurrence that might change if I ran the models again or due to the fact that the models were not optimized with the grid search.

Conclusion

My analysis has shown that predictive models can quite accurately diagnose disease/illnesses as all the models, after some optimization, have accuracies of above 95%. This means that predictive models should be integrated into the healthcare system to increase the number of correct diagnoses and amount of preventative care, so that all the key stakeholders can both save money and cut down costs on expensive, later-stage medical treatments for diseases that can be prevented or caught in the early stages.

The analysis answered in the way that it did (suggesting that logistic regression was the best model when compared to the others) most likely due to the nature of the data, which dealt with a binary label and few numeric predictors, so linear models and linear regression like models would possibly be the best-case models in these situations. If the data contained more varied data types, such as the smoking history data column, the analysis may have shown that one of the other two models would have proven to be more accurate.

For future analysis, I should include the smoking history data column after converting it from an object data type to an integer data type. This would allow me to answer the questions of “How would the inclusion of the smoking history data column affect the results/accuracy of the models?” and “Is the smoking history data column a necessary predictor variable to include for diabetes predictions/diagnoses?”

Important note: In order to keep this report concise, I have not included any images of visualizations or other graphics. To see the visualizations, look at the Python file (.ipynb) with the same name.