Justin Lewinski, Hudson Pak, Everett Prussak
Professor Doosti
MGSC 310-01
19 May 2023

Tree Model and Logistic Regression Model Predicting Diabetes in a Population

This executive summary provides an in-depth overview of a project focused on developing logistic regression and decision tree models for predicting the likelihood of diabetes based on various predictor variables. The objective of this project was to analyze a dataset containing demographic and clinical information and build predictive models to assist healthcare professionals in identifying individuals at risk of developing diabetes.

Problem Description and Motivation

Diabetes is a prevalent chronic disease that affects millions of people worldwide. Early identification and intervention can play a crucial role in managing the disease and preventing complications. Knowing the risk each individual is at for developing diabetes within a community can prove invaluable to getting them the resources and information they need. Predictive models can aid in identifying individuals who are at high risk of developing diabetes, allowing healthcare professionals to intervene with appropriate preventive measures.

The motivation behind this project is to use knowledge gained from working in class in order to build two predictive models that are useful and accurate. By identifying the key predictors and building these two models, healthcare professionals can enhance their ability to identify individuals who may benefit from early interventions, lifestyle changes, or further diagnostic tests. We explore in this project the application of both logistic regression and decision tree models to compare their performance in predicting diabetes.

Data Insights

This dataset contains information from an all female group of individuals, including demographic variables (age, gender, etc.) and clinical measurements (glucose levels, BMI, etc.). Exploratory data analysis revealed important insights into the dataset, facilitating further analysis and modeling. The key observations include: The dataset consists entirely of female individuals over the age of 21, with approximately 33.27% diagnosed with diabetes. The age distribution ranges from 21 to 81 years, with a mean age of 31.61 years. Glucose levels show significant

variation, with a mean of 121.03 mg/dL. BMI ranges from 18.2 to 67.10 kg/m², with a mean of 32.89 kg/m².

Building of Logistic Regression and Coefficient Analysis

A logistic regression model was developed to predict the likelihood of diabetes based on the available predictor variables. The dataset was divided into training and testing sets to evaluate the model's performance. The training split was 80% of the data, while the testing split was the remaining 20%. The logistic regression model incorporated all variables other than "outcome" as predictor variables to predict "outcome". Of the coefficients input into the model, there were 5 significant variables to a 95% confidence interval. Those variables were Glucose levels, BMI, DiabetesPedigreeFunction (A statistic developed in the medical field to calculate diabetes risk based on familial history), Age, and Age Squared, a variable manually added to the dataset by us. Of these variables, the one with the largest impact on the odds of having diabetes was determined to be DiabetesPedigreeFunction. For every standard deviation increase in DiabetesPedigreeFunction, the odds of having diabetes *decrease* by .3565%.

Logistic Regression Model Evaluation

After building the model and reviewing the original Train and Test scores, we found that the model was underfit, meaning the model was performing better on unseen data than it was on the training data. To counteract that, we decided to create the variable age squared, the age variable squared, to the dataset. This was to give the dataset more data to work with and help it to find the "outcome" variable. After doing this and comparing train and test scores again, we had shifted the model to actually be slightly *overfit*, which we were willing to accept due to how slightly the model was overfit. The logistic regression model achieved an accuracy of 77.8% on the training dataset and 78.5% on the testing dataset. Accuracy, sensitivity, and specificity were calculated to assess the model's performance in identifying individuals with and without diabetes. The sensitivity of the model, which measures the proportion of true positives correctly identified, was 84.35%. The specificity, measuring the proportion of true negatives correctly identified, was 74.46%.

Furthermore, an ROC curve analysis was conducted to evaluate the model's performance at different thresholds. The area under the curve (AUC) was calculated as a measure of the model's performance in distinguishing between individuals with and without diabetes. The

logistic regression model had an AUC value of 0.8711 for the training model, and .839 for the test, indicating a relatively strong performance in predicting correctly.

Decision Tree Model Development

In addition to logistic regression, a decision tree model was developed to predict diabetes based on the same set of predictor variables. As is common in decision tree models, we realized we had to trim our nodes down to a split that would reduce model complexity to a sufficient amount. After analysis, we found that it would be best to have 10 nodes at the end of our decision tree, as it was able to explain the entire model's complexity without being overly complex.

Decision Tree Model Evaluation

Upon building our decision tree, we found that glucose < 127.5 was the initial split, meaning it is the most important variable in the model. This was followed by Age < 28.5 and Glucose < 157.5. The decision tree model achieved an accuracy of 81.18% on the training dataset and 74.77% on the testing dataset. Similar to the logistic regression model, sensitivity and specificity were calculated to assess the performance of the decision tree model.

The sensitivity of the decision tree model was 78.23% for the train, and 74.77% for the test, indicating its ability to correctly identify true positives. The specificity was 82.73% for the train and 77.92% for the test, representing the model's ability to accurately identify true negatives. The confusion matrix provided insights into the number of true positives, true negatives, false positives, and false negatives. The decision tree model yielded an AUC value of 0.8048 for the training model, and .7229 for the test, indicating a strong discriminatory power.

Comparing Our Models

Comparing the logistic regression and decision tree models, both demonstrated reasonable predictive performance. The logistic regression model achieved slightly higher accuracy, sensitivity, and specificity compared to the decision tree model. The area under the ROC curve (AUC) for the decision tree model was 0.7229 for the testing model, suggesting good predictive power but slightly lower than that of the logistic regression model. The decision tree model's AUC value indicated its ability to distinguish between individuals with and without diabetes, albeit slightly less effectively than the logistic regression model.

Conclusion

The logistic regression model demonstrated good performance in predicting diabetes, with an accuracy of 80.2% and an AUC of 0.839. It exhibited strong sensitivity and specificity, indicating its ability to correctly classify individuals with and without diabetes. The decision tree model, while slightly less accurate than the logistic regression model, also provided reasonable performance with an accuracy of 74.77% and an AUC of 0.7229. It demonstrated good sensitivity and specificity, although all numbers were slightly lower compared to the logistic regression model. Based on the comparison of the two models, the logistic regression model appears to perform better in terms of this dataset because it achieved higher accuracy and precision.

These predictive models can assist healthcare professionals in identifying individuals at risk of developing diabetes, enabling early interventions, lifestyle modifications, or further diagnostic tests. It is important to note that the models should be validated using additional datasets and evaluated in a clinical setting before being implemented for any real world use. Overall, this project demonstrates the abilities of logistic regression and decision tree models in predicting diabetes and highlights the importance of using data analysis and statistical modeling techniques to enhance healthcare decision-making and improve patient outcomes.