Diabetes Prediction

Grant Corkren

**Preface:**

This project was originally done during my master’s program with the help of Will Davis but the code was lost. I have redone it in python to showcase my capabilities in Python and added some models that are not included within the SAS enterprise miner. This Word document in combination with the Python code is the entirety of my reasoning of the steps I took in order to obtain the final/production-level model.

**Abstract:**

This project is done to prevent untreated diabetes from going undetected by using the same survey questionnaire that we use for the predictive models. This could be beneficial to individuals, health insurance companies, and care providers. The end resulting model will predict whether or not someone has diabetes.

**Tools:**

This project was done in Python and utilized pandas, seaborn, scikit learn, and statsmodel libraries. All random states were set to 1 for reproducibility.

**Data:**

The dataset included prediabetes, which was not included in the modeling process, so a new dataset set was made.

**Class imbalance**:

First and foremost, the dataset provided has a class imbalance between having diabetes and not having diabetes. The dataset was first split into two sections. This was done to obtain a dataset that aligns with the proportion of diabetes responses in the original dataset which will be used to evaluate how the models do with the original dataset. The other section of the split was sampled at a fifty-fifty split of having diabetes and not having diabetes. The resulting split included all responses that had diabetes and a random sampling of those that did not have diabetes.

**Sampling and train test split:**

The data was then split again into a training and test split where 70% went to training and 30% to testing. The data was then split again into the independent variables and the dependent variables.

**Correlation between variables:**

Multicollinearity affects some model's performance such as the logistic regression, so an analysis of correlation was done. The correlation of independent variables was somewhat high for a few of the variables, but not high enough to justify removing the variables from the modeling process.

**Models:**

The variance inflation factor was checked to remove the variables before applying the logistic regression. Hyperparameter tuning was done for the neural network model. All of the models performed poorly even after adjusting for prior probability.

**Evaluation:**

All the models performed poorly on the equally weighted sample. Another round of analysis was performed with a simple random sample of the original dataset and clustering was done to create a new variable. Logistic regression was tried along with a deep neural network and categorical boosting. The performance of these additional models is nearly the same as just saying that no one has diabetes which is a model with 86% accuracy. The conclusion is that the survey needs better questions that give more insight into whether someone has diabetes.

**Appendix**

50/50 split testing sample

A screen shot of a black and white screen

Description automatically generated

Original proportion testing sample

A screenshot of a black screen

Description automatically generated