CAPSTONE PROJECT – ADVANCED STATISTICAL MODELLING

**Predicting The Onset of Diabetes – An investigation into the optimal statistical model for predicting the onset of gestational diabetes**

**William Belcher**

Gestational diabetes mellitus (GDM) is a subtype of diabetes diagnosed during the second part of pregnancy, gestation, and remains until the baby is born. Currently, all women are screened for GDM during their routine 24-to-28-week check-up [https://www.pregnancybirthbaby.org.au/gestational-diabetes]. This test is quite lengthy, taking a couple of hours. With the rapid increase in computing power and use of machine learning models in a clinical environment, it now seems plausible for a time-efficient screening process to be implemented for GDM. In this paper, the methods of Logistic Regression

**Introduction**

Gestation diabetes mellitus (GDM) is a form of diabetes that occurs during pregnancy and persists until after the birth of the child. [https://www.diabetesaustralia.com.au/about-diabetes/gestational-diabetes/] Diagnosis typically happens in around the 24th to 28th week of the pregnancy. GDM can be attributed to the increased insulin resistance caused by the hormone blocking nature of the placenta. During pregnancy, the need for insulin can be as high as 2 to 3 times higher than normal. If a woman already suffered from insulin resistance, the pancreas may be unable to cope with the heightened demand, leading to higher blood glucose levels and a diagnosis of GDM. Currently, women are referred to an oral glucose tolerance test at a pathology lab to determine if they have GDM. This test requires fasting from the previous night and can take between one to two hours. With the swift increase in computation and of machine learning models in the clinical setting, it now seems plausible for a statistical model to assist in the early detection and diagnosis of gestational diabetes mellitus. The early detection of GDM enables a management plan to be developed and enacted earlier, minimising the impact of the condition on the expecting mother

**Data**

The Diabetes data set, originally from the National Institute of Diabetes and Digestive and Kidney Diseases, contains several medical predictors and one target variable [https://www.kaggle.com/uciml/pima-indians-diabetes-database]. The response variable classified the patients into two classes: diabetic and non-diabetic. The medical predictors were eight different risk factors associated with GDM, being: number of pregnancies the patient has had, the plasma glucose concentration of two hours in an oral glucose tolerance test, diastolic blood pressure, triceps skin fold thickness, two-hour serum insulin, body mass index, diabetes pedigree function and age as in Table 1. These variables had many types, being either numerically discrete or continuous and the response variable being a binary 1 or 0, with a 1 representing the patient being diabetic. The dataset set is a subset of a larger database, with each observation being taken from a female patient over the age of 20 and of Pima Indian heritage. The following table details all 9 variables:

|  |  |
| --- | --- |
| Variable | Description |
| Pregnancies | Number of pregnancies (discrete) |
| Glucose | Plasma glucose concentration (mg/Dl) |
| BloodPressure | Diastolic blood pressure (mm Hg) |
| SkinThickness | Triceps skin fold thickness (mm) |
| Insulin | 2-Hour serum insulin (mu U/ml) |
| BMI | Body mass index (kg / |
| DiabetesPedigreeFunction | Diabetes pedigree function |
| Age | Age (years) |
| Outcome | Response (binary) |

Table 1: List of variables and their associated types

In total, there were 768 observations. It is worth noting not much information about the DiabetesPedigreeFunction field was supplied. However, for this report it is assumed that this function returns some information based on the family’s history of gestational diabetes.

*Data Pre-processing*

Data pre-processing is a technique in which the raw, imported data is transformed into a more meaningful and usable format. A summary of the dataset showed that minimum value for: Glucose, BloodPressure, SkinThickness, Insulin and BMI, was zero. It is unreasonable to have a zero value for these fields so the rows containing zeros in these columns were removed. Missing value imputation was not utilised in order to not introduce unnecessary variance into the models. The size of the cleaned data that was then utilised in the construction of the models contained 9 columns, 8 being predictors and 1 being a response with a total of 392 observations.

**Methods**

*Binary Logistic Regression*

Binary Logistic Regression models the relationship between a binary response and its predictors. More specifically, there exists a linear combination of variables that predicts the log-odds of the probability of an event from a logistic model. Let be the binary outcome where if the patient is diabetic and  if the patient does not suffer from GDM, with all observations being independent. Additionally, let be the set of explanatory variables, which can be of any form. In this case, is the probability of the patient being diabetic with the following logistic function

Thus,

From this function, .

*Maximum Likelihood Estimation*

Due to the nature of generalised linear models, and more specifically binary logistic regression which assumes the response comes from a binomial distribution, least squares regression cannot be employed and is instead substituted for maximum likelihood estimation (MLE). MLE attempts to maximize the value of the following equation using iterative methods

*Likelihood Ratio Test*

The likelihood ratio test assesses the goodness of fit between two statistical models, based on the ratio of their likelihoods. The effect of can be assessed by setting in the first model and letting in the second. The likelihood ratio statistic can the be calculated with the following formula

Where and are the log-likelihoods of model 0 and model 1, respectively.

*Model Selection Criteria*

Akaike’s Information Criteria (AIC) was selected as the measure of model fit. AIC estimates the distance between the true likelihood function of the data and the fitted likelihood function of the model plus a constant which penalises model complexity. AIC is defined by

Where is the number of parameters in model .

*Cross-fold Validation*

Cross-fold validation is employed as a method for estimating the model test error rate. This approach divides the data into groups of equal size in a random manner. To start, the first group is removed from the training data, leaving the model to be trained on the remaining groups. The Mean Square Error (MSE) is then calculated using the observations from the left-out group. This process is repeated until all groups have been utilised as the testing set, leaving a vector of MSEs, . The k-fold cross validation error is then calculated by finding the mean of these values:

*Performance metrics*

The result of a model is measured in terms of its accuracy. The values are obtained by using the generated models to predict the outcome on the test data set and making note of the resulting confusion matrix.

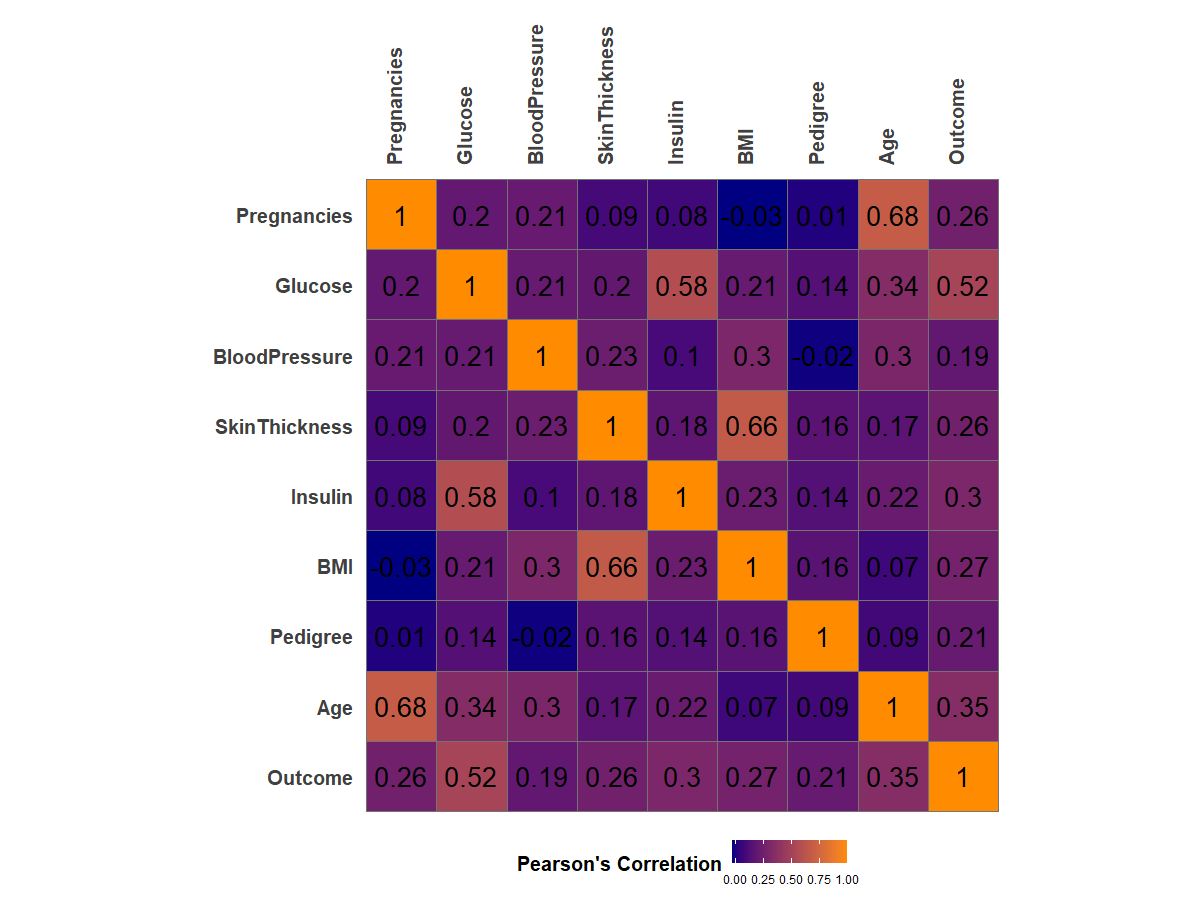
|  |  |  |
| --- | --- | --- |
|  | Predicted Outcome | |
| Actual Outcome | True Positive | False Positive |
| False Negative | True Negative |

Table 2: Confusion Matrix

With accuracy being calculated in the following way:

**Results and Discussion**

Figure **X** is the Pearson’s Correlation heat map, which visually illustrates the correlation between pairs of predictors.

****

The correlation between pregnancy and age (0.68), BMI and skin thickness (0.66), insulin and glucose (0.58) are considered high as they have a correlation over 0.5 and are significantly larger than other pairwise correlations. This suggest these three pairs of predictors are correlated.

VIF, interaction plots. “however when comparing models that included the interaction terms, these interactions were found to be non-significant”

In total, nine models were fit utilising best subset selection. Best subset selection was employed due to the low number of predictors in the data set. Appendix **X** illustrates the predictors found to be significant for each model. From the AIC value, it can be seen that a model with four predictors proved to fit best, with these predictors being: glucose, BMI, DiabetesPedigreeFunction and Age.

**Conclusion**

**References**

* [4] RStudio Team (2020). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA <http://www.rstudio.com/>

Mode: Desktop. Version 1.25042

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| # Of parameters / predictors | 1 | 2 | 3 | 4\* | 5 | 6 | 7 | 8 | 9 |
| Intercept |  |  |  |  |  |  |  |  |  |
| Pregnancies |  |  |  |  |  |  |  |  |  |
| Glucose |  |  |  |  |  |  |  |  |  |
| BloodPressure |  |  |  |  |  |  |  |  |  |
| SkinThickness |  |  |  |  |  |  |  |  |  |
| Insulin |  |  |  |  |  |  |  |  |  |
| BMI |  |  |  |  |  |  |  |  |  |
| DiabetesPedigreeFunction |  |  |  |  |  |  |  |  |  |
| Age |  |  |  |  |  |  |  |  |  |
| loglikelihood | -195 | -155 | -148 | -144 | -140.8 | -140.4 | -140.2 | -140.09 | -140.03 |
| AIC | 390 | 312 | 301 | 294 | 289 | 290 | 292 | 294 | 296 |