Logistic Regression Report – PIMA Indian Diabetes Dataset

Introduction and Dataset EDA:

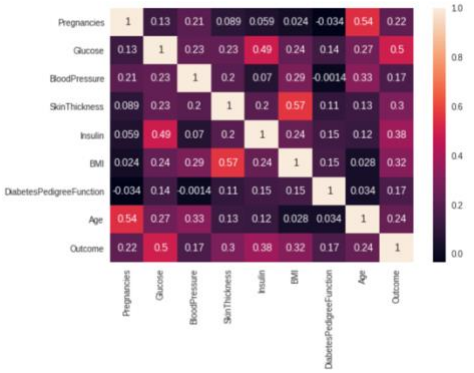
Diabetes mellitus is a metabolic disorder where the blood sugar levels are higher than normal for prolonged periods of time. Diabetes is caused either due to the insufficient production of insulin in the body or due to improper response of the body's cells to Insulin. Studies have estimated that the ratio of incidence of diabetes in the US alone is as high as one in three for men, and two in five amongst women. With such alarming numbers, it has become a pressing issue in the health care industry to rightly identify the factors that contribute to the occurrence of Diabetes in people.

To study this phenomenon closely, we chose to this dataset of the PIMA Indians which consists of diagnostic measurements of 768 female patients, all of whom are at least 21 years old, and 268 of which were diagnosed with Diabetes. The response variable, Outcome, which needs to be predicted is a binary classifier that indicates if the person was diagnosed with Diabetes or not (1 for yes, 0 for no). Other Information available includes 8 variables, such as, Age, Number of Pregnancies, Glucose, Insulin, Blood Pressure, Skin Thickness and BMI. More detailed description about the variables is listed in the table below.

At first glance, the dataset appeared to be clean. On deeper analysis, the dataset revealed many abnormal values for biological measures. Variables such as Skin Thickness and Glucose had 227 and 374 zero-values respectively. The missing values in the dataset constituted to about 30% of the observations. Since all of the variables are integers, we proceeded to impute the missing values for each of these variables with the respective mean values.

After having cleaned the dataset of all the null values, we conducted a basic analysis to understand the data distribution. The table on the left gives us a brief insight into the data. Along with this, the heatmap gives us the correlation between the different variables and the outcome column. Looking at the last column on the right, we can notice that the variables 'Glucose', 'Insulin' and 'BMI' have a strong correlation with the response variable. Researching about diabetes, we understand that Glucose and Insulin are the two vital factors that play an important role in keeping the sugar levels in your body under control, hence the high correlations only validate this fact. As for BMI, you can see that the mean value recorded amongst the 768 instances is 32.4 (18.5 and 24.9 – you're in the healthy weight range), which clearly lies on the higher end of the spectrum indicating that majority of the population from the dataset can be considered obese.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	121.677083	72.389323	29.089844	141.753906	32.434635	0.471876	33.240885	0.348958
std	3.369578	30.464161	12.106039	8.890820	89.100847	6.880498	0.331329	11.760232	0.476951
min	0.000000	44.000000	24.000000	7.000000	14.000000	18.200000	0.078000	21.000000	0.000000
25%	1.000000	99.750000	64.000000	25.000000	102.500000	27.500000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	28.000000	102.500000	32.050000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	169.500000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000



Logistic Regression and Interpretations:

Having concluded with data engineering and visualization, we then moved on to perform several logistic regressions runs with our dataset. We had setup our model to run at 95% confidence interval. In our first run, we notice that the p-values for variables 'Blood Pressure' and 'Age' were 0.6634 and 0.1711 respectively. Since these values were higher than our significant value of 0.05, it meant they were candidates for exclusion from the dataset. After removing these variables, we rerun the model to find results with improved accuracy and all of the variables having significant values. The variables that this renewed

Regression Coefficients	Coefficient	Standard Error	Wald Value	p-Value	Lower Limit	Upper Limit	Exp(Coef)
Constant	-8.922569679	0.977246788	-9.13031364	< 0.0001	-10.83797338	-7.007165974	0.000133345
Pregnancies	0.125308928	0.037894249	3.306805931	0.0009	0.0510362	0.199581656	1.133498568
Glucose	0.028623662	0.004561299	6.275330987	< 0.0001	0.019683516	0.037563809	1.029037256
BloodPressure	0.002679152	0.010447264	0.256445272	0.7976	-0.017797487	0.02315579	1.002682744
SkinThickness	0.036179332	0.016345787	2.213373536	0.0269	0.00414159	0.068217075	1.036841769
Insulin	0.004429031	0.001731299	2.558213063	0.0105	0.001035686	0.007822377	1.004438854
BMI	0.057624577	0.021059852	2.736228977	0.0062	0.016347267	0.098901886	1.059317228
DiabetesPedigreeFunction	0.79046906	0.347104936	2.277320137	0.0228	0.110143386	1.470794733	2.204430193
Age	0.000651419	0.011584585	0.056231548	0.9552	-0.022054367	0.023357205	1.000651631

Classification Matrix	1	0	Percent Correct
1	161	107	60.07%
0	62	438	87.60%

Summary Classification	Percent
Correct	77.99%
Base	65.10%
Improvement	36.94%

model included are: 'Pregnancies', 'Glucose', 'Skin Thickness', 'Insulin', 'BMI' and 'DiabetesPedigreeFunction'. The results are shown in the table above. From these results we can see that all of the variables are statistically significant in predicting whether a particular individual will be diabetic or not. Looking at the coefficients, we notice that variables 'Pregnancies', 'DiabetesPedgreeFunction' and 'BMI' have comparatively higher contributions towards a person being diabetic. As one unit increase in these variables, the odds ratio increases by the particular coefficient and would highly impact the response variable. The factor "Blood Pressure" has the least impact on the response. For inference, Women with higher number of pregnancies are more susceptible to being diabetic that not. Similarly, with BMI, in agreement with the data visualization report, the higher coefficient only indicates that individuals who are more obese have higher chances of being diabetic than those who have lower BMI index. I guess it's no surprise that if one is unfit and is less active, they are subjected to more sickness overall than in just this particular case.

From the above classification matrix, we can observe that our model's prediction accuracy is 60.07% amongst those who are diagnosed to be diabetic and 87.60% amongst those who are non-diabetic giving our model an overall accuracy of 77.99%. A type I error was made on 62 individual women who were wrongly classified to be diabetic whereas a type II error was made on 107 women who were not diabetic but were predicted to positive. In total 599 correct prediction were made.

The final logistic regression of the model (after deleting insignificant values) is as follows:

$$Log[p/(1-p)] = -9.221 + 0.137*Pregnancies + 0.031*Glucose + 0.038*SkinThickness + 0.005*Insulin + 0.055*BMI + 0.806*DiabetesPedigreeFunction$$

Conclusion:

We performed data analysis and constructed a logistic regression model to find information on the interaction between the different variable and its effects on the outcome. Based on the results of our model, it is good to see that our model predicted BMI and Pregnancies as significant variables in influencing the outcome, after all this is exactly what we have known from the medical studies. In another observations, we would usually identify Glucose to be a very important determinant of diabetes in real world and also since the correlation of Glucose is higher than that of BMI or Pregnancies, then how come our model shows a fairly low coefficient levels for Glucose. Having said that, we need to consider that the mean age of the women in our dataset was 33. One can say that at 33 years of age a person would have pretty healthy glucose level as compared those at higher ages. In addition, at the model level, the correlation captured by Glucose could also have been captured by some other variable, whereas the information captured by Pregnancies could not have been captured by other variables. Therefore, we can say that our model has done a good job with the given dataset.