

Logistic Regression to Predict Diabetes Using Multiple Model Evaluation Techniques

Diksha Dinesh

RV College of Engineering, Bangalore, India

Abstract

Diabetes Mellitus, known popularly as Diabetes, is a metabolic disease that results in the rise of glucose levels in ones body. Diabetes is influenced by factors such as Insulin, Age and Blood Pressure. In this study, using the Pima Indians Diabetes Dataset, Logistic Regression is performed to predict the possibility of occurrence of Diabetes. Logistic Regression is implemented when the dependent variable(s) are categorical instead of numerical. The model used here is a Logit model which is obtained by performing logarithmic operations upon the Logistic model. A rigorous procedure of data cleaning and outlier elimination results in a model with high accuracy. A logistic model, which is classically used for binary dependent variables, is developed. Confusion matrices assist in the determination of model accuracy. Varying the number of parameters taken into consideration, the effect on AIC is noted. The models performance is further examined by implementing three techniques : AUC and ROC, KS statistics and Gain chart, and finally, Lift test. Percentage change in outcome is determined for each of the independent variables. The gains chart shows that 100% of the target values are covered by the first 80% of the target values alone. the lift chart depicts that 70% of the model records, account for 2.7 time the total targets found by selecting 70% of a file that doesn't have a model.

Keywords

Regression, Receiver Output Characteristics, Lift , Kolmogorov-Smirnov, Akaike's Information Criterion, Logistic, Confusion Matrix.

I. Introduction

Diabetes is a chronic, metabolic disease, characteristic of high glucose levels, which inevitably damage other organs such as the heart, eyes and kidneys. When the body fails to produce sufficient insulin, type 2 diabetes occurs, which is symbolic of all countries regardless of their economic status. Type 1 diabetes is also referred to as juvenile as it is chronic and insulin dependent. The globally accepted target is to hinder the rise of obesity as well as diabetes within the coming five years. Despite the alarming number of deaths due to Covid-19, it is in no way comparable to the leading cause of death as of date, which is Diabetes. 422 million people worldwide suffer from this condition and approximately 1.6 million deaths are entirely due to Diabetes.

Data Science provides cutting-edge techniques that allow predictive analysis, that result in models that can control such distressing numbers, by simply predicting future outcomes, giving the concerned professionals the time they need to prevent/treat the issue. Regressions are used to accurately predict the relationship between a desired outcome and its supposed causes. Logistic Regressions deal with scenarios consisting of a binary outcome, in this study, the outcomes are whether the person will have diabetes or not. Confusion matrices are used to compute the accuracy of the model, however for the overall evaluation of the model, is done by calculating the AUC (Area Under Curve) for an ROC (Receive Output Characteristics) curve. This is supported

by measuring the KS statistic and plotting a Gains chart, which deals with cumulative percentages used to determine number of targets satisfied in a given decile. Finally, the Lift chart effectively determines the efficiency of the predictive model with respect to results obtained without a model as such.

This study proposes a detailed approach to assess the performance of any predictive model. Section II will include a survey of related works by fellow researchers/students who have worked with logistic regression and diabetes prediction. Section III will provide the sequential flow of the whole prediction model, followed my Section IV which will state all results found during the course of this work. Finally, Section V will conclude the work done, and provide future scope.

II. Literature Survey

The authors in [1] improve upon logistic regression and K means clustering techniques by using Principal Component Analysis. A total of 25 more data points were classified accurately, increasing the logistic regression accuracy by 1.98%. the model successfully predicts diabetes based on health data obtained electronically [2]. focuses on classification algorithms such as Decision Tree, SVM and Naïve Bayes, and compares parameters such as precision, accuracy, f-measure and recall. The best amongst the three mentioned algorithms turns out to be Naïve Bayes, with an accuracy of 76.3%. The authors in [3] use optimal feature selection to design a predictive model that uses decision tree and random forest algorithms to yield an accuracy of 98.20% and 98.00% respectively. [4] uses Principal Component Analysis (PCA) and minimum redundancy maximum relevance to reduce dimensionality, in collaboration with Random Forest, to obtain a high accuracy of 80.04%. [5] proposes a system achieved a high classification result of 98.7%, using the Decision Tree classification approach. The authors in [6] use a combination of Machine and Deep Learning techniques to design a model that used classifiers that were rarely used in articles published in the last six years. They found that the accuracy was enhanced when these classifiers were used, obtaining an accuracy of 68%-74%.

III. Methodology

This sections will provide a detailed analysis of the processes carried out in the model developed. The stages include : Data Preprocessing, Checking for monotonicity, Logistic model creation, Confusion matric, AUC and ROC, KS statistic and Gains chart and Lift chart. The flowchart for the proposed model is depicted in fig. 1.

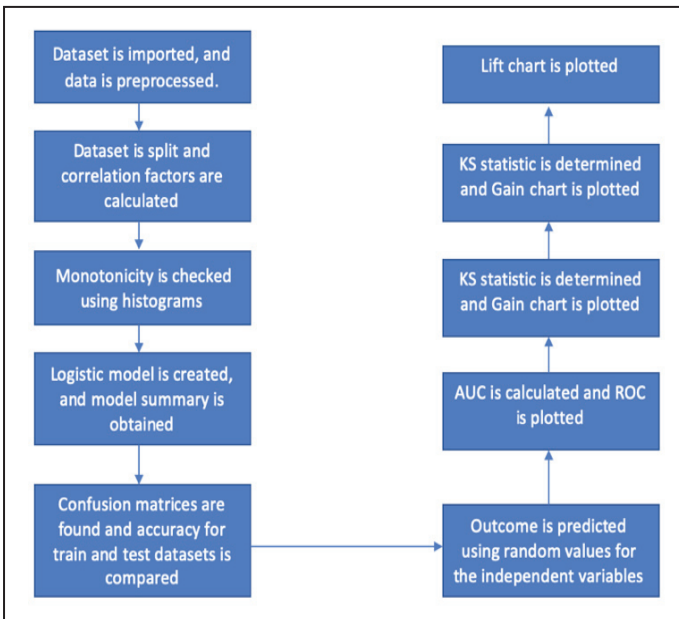


Fig. 1: Methodology flowchart

A. Toolkit adopted

Anaconda is a free open-source platform that assists all python/R Data Science and Machine Learning operations. Jupyter Notebook is a service provided by Anaconda, which allows the creations and modification of documents that contain live code, equations etc. this project is carried out using Jupyter Notebook.

A combination of modules are necessary are used, most valuable of which is Statsmodels. It provides the classes and functions required to implement different statistical models. Along with several other modules, a more accurate prediction model is developed.

B. Importing Dataset and Primary Description

The dataset used for this work is the Pima Indians Diabetes Dataset. It tabulates data obtained from a group of 768 females, portraying 8 characteristics for each of them, which are : Number of Pregnancies, Glucose Levels, Blood Pressure, Skin Thickness, Insulin levels, Body Mass Index, Diabetes Pedigree Function and Age. This dataset is imported and described to look for obvious anomalies. This description of the data can be seen in Fig. 2.

	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	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	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

Fig. 2: Dataset Description

From the description above, it can be noted that the following attributes have a minimum value of 0, which is humanly impossible : Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI. These inaccuracies may be due to errors during data collection, amongst other reasons. To overcome such inadequacies, data preprocessing must be carried out.

C. Data Preprocessing

Data preprocessing is performed to eradicate outliers if any and substitute all misplaced 0 values.

1. Removing Misplaced Null Values

The elimination of all misplaced zero values, i.e., the zeros in Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI, must be accounted for before fitting any statistical model. The steps carried out to do the same are :

- A copy of the dataset without any zeroes in its for each attribute is created.
- The mean for that particular attribute is calculated.
- All zeroes pertaining to that attribute are substituted with the mean value.

For example, when all the rows of the dataset, where the value of glucose was 0, were removed, the mean was found to be 121.6867. this value is put in place of all zeros in the Glucose column in the original dataset.

The same is repeated until the data description shows a minimum value of 0 only for the Pregnancies and Outcome columns.

2. Eliminating outliers

The outliers are determined using each attributes probability distribution function plot. The ditplot function is used to plot the PDFs. The steps carried out are :

- PDF for each attribute (excluding outcome) are plotted.
- outliers from the plot are identified, by looking for elongated tails, or values far from the mean.
- These outliers by are eliminated by considering a certain percentile of the total data available.

For example, Fig. 3 shows the PDF plot for the Diabetes Pedigree Function, where significant outliers were found to the right of the mean value. These were eliminated by removing all but 0.99 percentile of the entries.

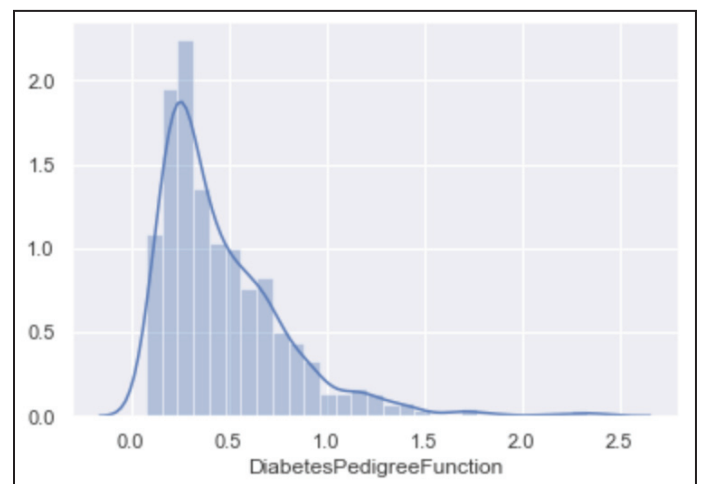


Fig. 3: PDF for Diabetes Pedigree Function

This elimination ensures that the mean is more centrally located. The same procedure is repeated for all independent attributes. The final dataset obtained is accurately preprocessed, and has 63 lesser entries than the original dataset. This new dataset is used for future operations.

D. Splitting the dataset

The preprocessed dataset is split into two datasets, namely, the Training dataset and the Testing dataset. The first set is used to

train the actual model, while the latter helps confirm the model is indeed accurate. An 80:20 split is performed, hence assigning 552 rows to the Train set and 216 rows to the Test set.

E. Correlation Factor Determination

Correlation factors are useful in assessing the effect one variable has on another. In this work, it is essential to know which of the attributes has the most effect on the Outcome. The correlation factors for each of the attributes are calculated, a heatmap as seen in Fig. 4 is produced.

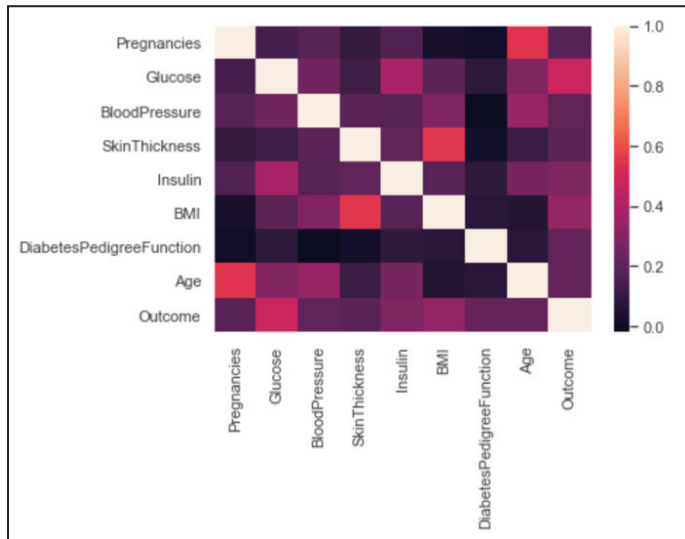


Fig. 4: Heatmap depicting correlation factors

It can be inferred from the heatmap that, the outcome is most dependent on Glucose levels, BMI, and insulin levels.

F. Monotonicity check

It is essential to determine the relationship between each of the attributes and the outcome. While some attributes portray decreasing trends with respect to increase in its value, some attributes portray a normal distribution. According to the data used, Blood Pressure shows a fairly normal distribution (Fig. 5), while Age shows a decreasing trend (Fig. 6).

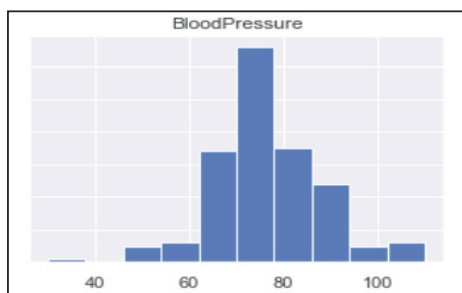


Fig. 5: Blood Pressure Plot

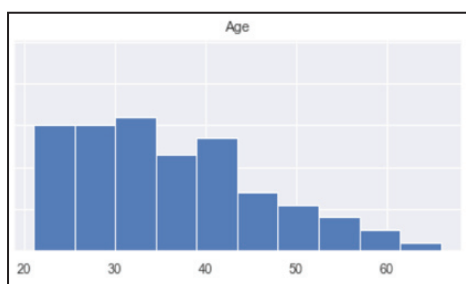


Fig. 6: Age Plot

G. Logistic Model

Logistic regressions are used when the dependent variable is categorical and binary. This work requires the prediction of whether a woman, given her medical details, will have diabetes or not, hence making it categorical and binary. The general equation of a Logistic Regression (1), shows that the right-hand side of the model is an exponent, making it generally computationally inefficient. To overcome this, both sides of the equation undergo logarithmic operations to produce (2). Hence, a Logistic Regression is commonly referred to as a Logit Regression.

$$\frac{p(X)}{1-p(X)} = e^{(B_0+B_1X_1+\dots+B_kX_k)} \tag{1}$$

$$\log(\text{odds}) = B_0 + B_1X_1 + \dots + B_kX_k \tag{2}$$

In equation (2), ‘odds’ refers to the ratio of an event happening, to the event not happening.

The model summary for the Train set is obtained as shown in Figure 7. McFadden’s pseudo-R-squared is used to compare variations of the same model, as is most favorable within the 0.2-0.4 range. The summary clearly denoted the dependent variable, and the model implemented. The Log-Likelihood is a negative quantity, and should be as high as possible. The Log-Likelihood-Null value is the LL value of a model with no independent variables, and can be regarded as the benchmark for the worst model possible. The LL Ratio p-value specifies how different the model is from its worst case scenario.

Model:	Logit	Pseudo R-squared:	0.297
Dependent Variable:	Outcome	AIC:	510.3617
Date:	2020-07-28 15:10	BIC:	549.1836
No. Observations:	552	Log-Likelihood:	-246.18
Df Model:	8	LL-Null:	-349.95
Df Residuals:	543	LLR p-value:	1.6404e-40
Converged:	1.0000	Scale:	1.0000
No. Iterations:	7.0000		

	Coef.	Std.Err.	z	P> z	[0.025	0.975]
const	-10.5379	1.0448	-10.0861	0.0000	-12.5857	-8.4901
Glucose	0.0360	0.0047	7.7279	0.0000	0.0269	0.0452
BMI	0.1051	0.0219	4.8008	0.0000	0.0622	0.1481
Pregnancies	0.1052	0.0400	2.6286	0.0086	0.0268	0.1837
BloodPressure	0.0022	0.0104	0.2106	0.8332	-0.0182	0.0226
DiabetesPedigreeFunction	1.5989	0.3845	4.1585	0.0000	0.8453	2.3524
Insulin	0.0041	0.0027	1.5088	0.1314	-0.0012	0.0094
SkinThickness	-0.0024	0.0163	-0.1492	0.8814	-0.0344	0.0295
Age	0.0031	0.0123	0.2543	0.7992	-0.0209	0.0272

Fig. 7: Logistic Regression Model Summary

Attributes having a p-value lesser than 0.05 are considered to be most significant. The coefficient of each attribute can explain the change in outcome with respect to a certain percentage change in that attribute.

It is found that the model characteristics are the same for both the Train and Test datasets.

H. Confusion Matrix

Confusion matrices are capable of efficiently describing the performance of a classification model based on its data when true values are known. the general format of a confusion matrix is shown in Fig. 8.

	<i>Class 1 Predicted</i>	<i>Class 2 Predicted</i>
<i>Class 1 Actual</i>	TP	FN
<i>Class 2 Actual</i>	FP	TN

Fig. 8: Confusion Matrix

The confusion matrix when found for the test data, revealed that it had an accuracy of a little lesser than that of the train data, which is typical of all statistical models.

I. Prediction with assumed cutoff

The logit model, which has been fit with all the attributes, can be used to predict the outcome given a random value (within reason), for each of the attributes. The outcome is very rarely either 0 or 1. A threshold of 0.5 is assumed, an outcome ≥ 0.5 is considered as 1 (Diabetic), else is 0 (Non-Diabetic). this however, is inaccurate, as the cut off cannot be assumed. To accurately determine a cutoff value, AUC must be computed.

J. AUC and ROC

AUC-ROC curves measure the performance of a classification model at multiple threshold settings. The Area Under Curve value specified how good the model is at separating outcomes, i.e, how good the model is able to separate diabetic from non-diabetic patients. Higher the AUC value, better is the model at distinguishing between 1s and 0s.

The Receiver Output Characteristics Curve is plotted with True Positive Rate against False Positive rate. The farther the ROC curve is from the 45 degree line, better is the model. A curve coinciding with the 45 degree line indicates it is incapable of distinguishing between outcomes. Fig. 9 shows an ROC curve.

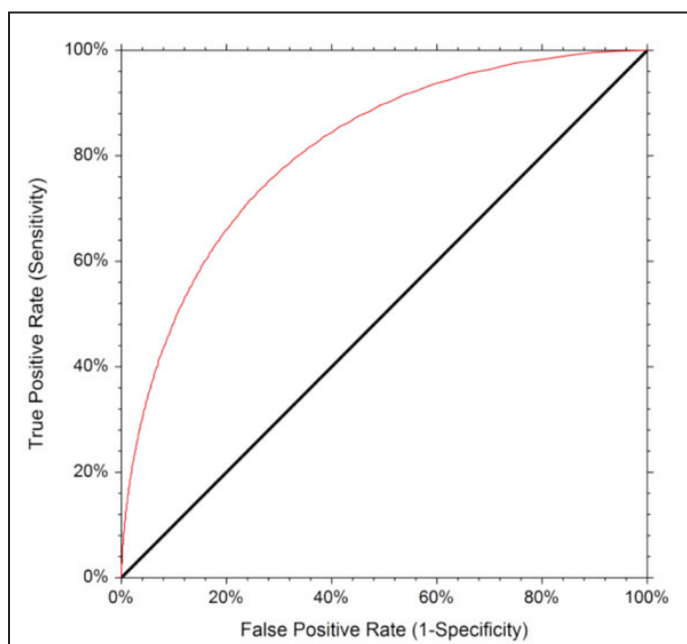


Fig. 9: ROC curve

The cutoff of the model is decided by determining the value of sensitivity when the difference between False Positive and True Positive is the lowest.

K. KS statistic and Gain Chart

The KS statistic and Gain chart are another method implemented to check whether the model can separate outcomes accurately.

1. KS statistic

The predicted probability for each data point is arranged in descending order, and divided into ten categories/deciles. A pivot table is created to assist the calculation of the KS statistic, which shows the number of entries in each decile (Fig. 10).

Decile	Defaulter_Count	Non-Defaulter_Count	max_score	min_score	Total_Female
10	18	4	0.962140	0.808947	22
9	18	3	0.804448	0.668967	21
8	11	11	0.659798	0.468632	22
7	9	12	0.465729	0.371161	21
6	7	15	0.367290	0.276587	22
5	7	14	0.272168	0.220210	21
4	2	20	0.218712	0.156427	22
3	1	20	0.154876	0.122692	21
2	0	22	0.119476	0.094504	22
1	0	22	0.093031	0.011194	22

Fig. 10: Pivot Table

The rate of defaulters and non-defaulters per decile is calculated, using which the KS statistics are found. The KS statistics are the difference between the cumulative defaulter and cumulative non-defaulter rates. The maximum value of KS statistic is determined, and assigned to the model. The decile the KS statistic belongs to is of importance. There should be a very small difference between the KS statistic of the train and test datasets.

2. Gain chart

The Gain chart uses the data used by the KS statistic evaluation, with some additions. It is necessary to calculate the cumulative of the default percentages, for both train and test datasets. The gain at a given decile is the ratio of cumulative number of targets up to that decile to the total number of targets in the complete dataset. An example of a Gain chart can be seen in Fig 11.

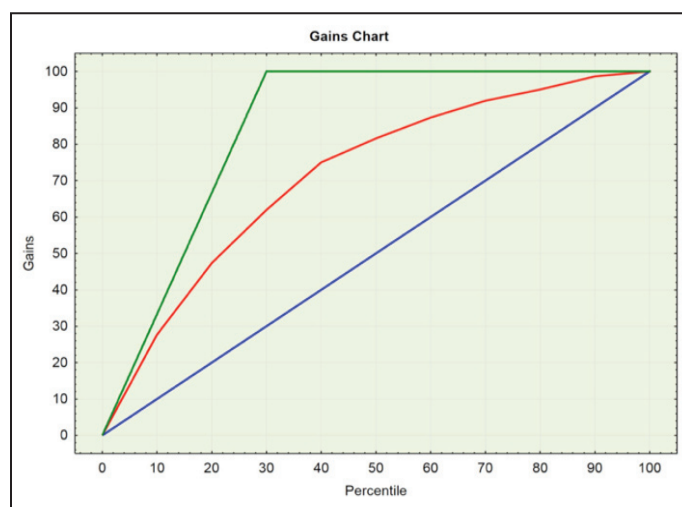


Fig. 11: Gain Chart

L. Lift test

Gain and Lift charts successfully evaluate the performance of the classification model. The data used by the gain chart, along with the ratio of the cumulative defaulter percentage to the baseline

percentage, is required to plot the Lift chart. An example for a Lift chart is shown in Fig. 12.

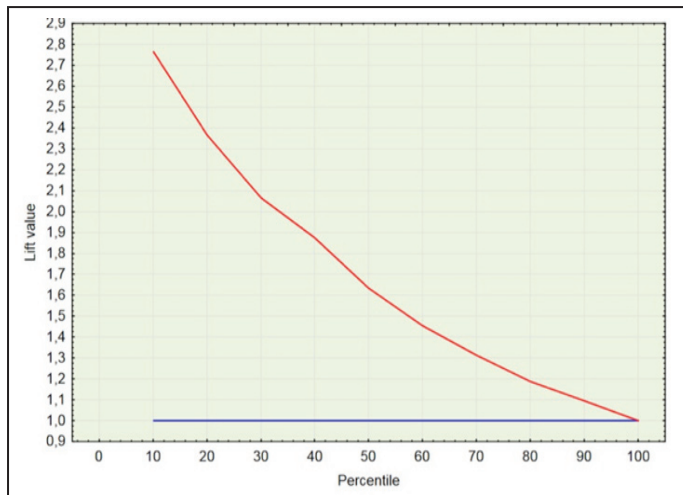


Fig. 12: Lift chart

Lift charts can help understand how much better a model can perform compared to not having a model. It is the ratio of the actual gain percentage to the expected percentage for a given decile.

IV. Results and Discussion

This section will discuss the results obtained during the course of the entire project.

A. Results from Logit summary

McFadden’s pseudo-R-squared is a measure of the performance of a model, and should ideally be in the range of [0.2,0.4]. The pseudo-R-squared value obtained for this predictive model is 0.297, as shown in Figure 7, which is well within the ideal range.

1. AIC Parameter Variation

Akaike’s Information Criterion is another basis for measuring the quality of the model. Lower the AIC score better is the model. The addition of attributes to the any model, improves its prediction accuracy. Table 1 depicts the decrease in value of the AIC score with the addition of parameters.

Table 1: AIC parameter with Respect to Variables

Independent variable	AIC
Glucose	570.6051
Glucose + BMI	531.7738
Glucose + BMI + Insulin	528.9344
Glucose + BMI + Insulin + Age	526.8819
Glucose + BMI + Insulin + Age + Skin Thickness	515.4470
Glucose + BMI + Insulin + Age + Skin Thickness + Blood Pressure	513.4550
Glucose + BMI + Insulin + Age + Skin Thickness + Blood Pressure + Diabetes Pedigree Function	511.5125
Glucose + BMI + Insulin + Age + Skin Thickness + Blood Pressure + Diabetes Pedigree Function + Pregnancy	510.3617

The AIC score of the final model, including all attributes is 510.3617.

2. Change in Outcome with respect to change in attribute

The Logit summary specifies the coefficients for the constant as well as each of the attributes. Using these values, the percent change in outcome can be assessed for unit change in each of the attributes having a p-value < 0.05.

$$\text{Log}(\text{odds}_1) = B_0 + B_1X_1 \tag{3}$$

$$\text{Log}(\text{odds}_2) = B_0 + B_1X_2 \tag{4}$$

$$\text{Log}(\text{odds}_2/\text{odds}_1) = B_1(X_2 - X_1) \tag{4)-(3)}$$

Using the above equations, the values in Table 2 are calculated.

Table 2: Change in outcome with respect to change in input variable

For unit change in	% change in outcome
Glucose	3.66
BMI	11.08
Number of Pregnancies	11.09

From the table above it can be inferred that:

- For unit change in Glucose levels, the odds of getting diabetes increases by 3.66%
- For unit change in BMI, the odds of getting diabetes increases by 11.08%
- For unit change in number of pregnancies, the odds of getting diabetes increases by 11.09%

Since the Diabetes Pedigree Function is a small value in the range of 0.07 – 1.5, change in Outcome for change in 0.01 in the Diabetes Pedigree Function is calculated. It is found that, when the Diabetes Pedigree Function increases by 0.01, the odds of getting Diabetes increases by 1.6%.

B. Results from Confusion Matrix

The Confusion Matrix is calculated for both the train and test data, and compared. The confusion matrix of the Train set is shown in Table 3, followed by the Confusion matrix for the test set, shown in Table 4.

Table 3: Confusion Matrix – Train dataset

	Predicted Non-Diabetic	Predicted Diabetic
Actual : Non-Diabetic	331	39
Actual : Diabetic	74	108

Table 4: Confusion Matrix – Train dataset

	Predicted Non-Diabetic	Predicted Diabetic
Actual : Non-Diabetic	126	17
Actual : Diabetic	28	45

The accuracy of the model is calculated from the confusion matrix by finding the ratio between the sum of True Positives and True Negatives to the total number of data points. The accuracies of the test and train data are depicted in Table 5.

Table 5: Test and Train accuracies

	Accuracy (%)
Train dataset	79.528
Test dataset	79.166

The Train dataset’s accuracy is slightly higher than that of the Test dataset’s accuracy, which is common in classification models.

C. Results from AUC and ROC curve

AUC measures the degree of separability. The value of AUC ranges from 0 to 1, where a AUC of 0 means the model has no power of separability, while a AUC of 1.0, means that the model is able to perfectly distinguish between outcomes.

The AUC of the model created in this project has a value 0.87, which means the model is able to distinguish between 87% of the cases. The ROC curve plotted is shown in Fig. 13.

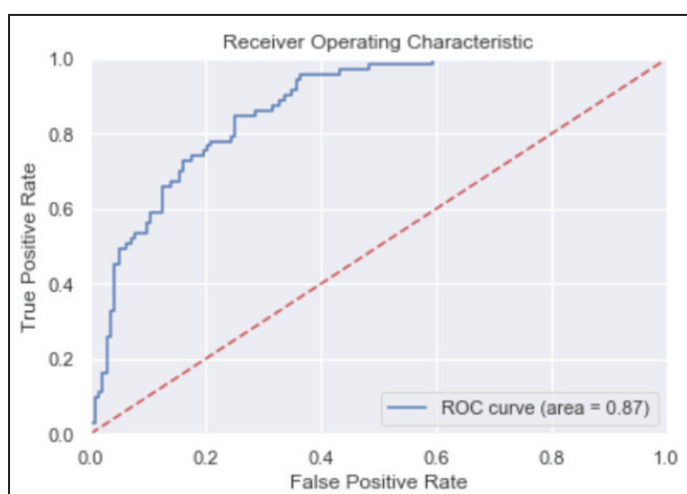


Fig. 13: ROC curve

Curves that reach the top-left corner indicate better performance. The points lying along the red dotted diagonal act as a baseline for the model the farther away the curve is from this baseline, better is the model. A curve that reaches true positive faster, is also said to have better performance. The minimum difference between TPR and FPR, i.e., sensitivity, is used to calculate the new cutoff.

The new cutoff is calculated to be 0.456 which increases the model accuracy from 79.166% to 79.629%. these results are tabulated in Table 6.

Table 6: Model Accuracy Comparison

	Accuracy (%)
Logit model	79.166
Logit model + AUC + ROC	79.629

Hence, the implementation of ROC curve and computation of AUC aided in making the model better at distinguishing between positive and negative outcomes.

D. Results from KS Test and Gain chart

1. Results from KS Test

KS statistic is calculated to distinguish between events and non-events. The KS statistic for the Train and Tests sets are determined.

Fig. 14 shows the defaults and non-default percentages, along with indication of the KS statistics and the decile it belongs to.

Decile	Defaulter_Count	Non-Defaulter_Count	max_score	min_score	Total_Female	Default_Rate	Default %	Non_Default %	ks_stats	max_ks
10	48	8	0.968859	0.738358	56	85.71%	26.37%	2.16%	24.21	
9	38	17	0.732509	0.588136	55	69.09%	20.88%	4.59%	40.50	
8	26	29	0.585821	0.425665	55	47.27%	14.29%	7.84%	46.94	
7	23	32	0.424503	0.327285	55	41.82%	12.64%	8.65%	50.93	*****
6	17	38	0.321767	0.263364	55	30.91%	9.34%	10.27%	50.00	
5	12	43	0.262569	0.196437	55	21.82%	6.59%	11.62%	44.97	
4	12	43	0.195947	0.140174	55	21.82%	6.59%	11.62%	39.95	
3	3	52	0.139504	0.101419	55	5.45%	1.65%	14.05%	27.54	
2	3	52	0.099840	0.059650	55	5.45%	1.65%	14.05%	15.14	
1	0	56	0.059386	0.016267	56	0.00%	0.00%	15.14%	0.00	

Fig. 14: KS statistic for Train dataset

The KS statistic is determined for the test dataset as well, and the results are tabulated in Table 7.

Table 7: KS Statistic Comparison

	Decile	KS Statistic
Train dataset	7	50.93
Test dataset	7	55.73

The table above shows that the KS statistic for the Train dataset is located in the 7th decile and is 50.93, while the KS statistic for the Test dataset, while in the same decile, is 55.73.

2. Results from Gain chart

The default cumulative percentages for train and test datasets are shown in Fig. 15.

Decile	default_cum%_train	Base %	default_cum%_test
10	26.37	10	24.66
9	47.25	20	49.32
8	61.54	30	64.38
7	74.18	40	76.71
6	83.52	50	86.30
5	90.11	60	95.89
4	96.70	70	98.63
3	98.35	80	100.00
2	100.00	90	100.00
1	100.00	100	100.00

Fig. 15: Default Cumulative Percentages

These cumulative percentages are used to plot the Gain chart, as shown in Fig. 16.

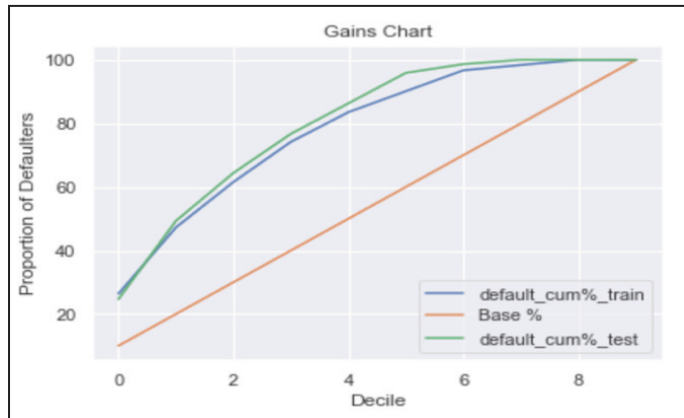


Fig. 16: Gain chart

The Gain chart provides the information tabulated in Table 8.

Table 8: Results from Gain chart

Percentile of data from Test and Train datasets	% targets in Train dataset	% targets in test dataset
20%	63	66
40%	84	86
60%	95	98
80%	100	100

From the table it can be inferred that :

- First 20% of the data covers 63% of the Train targets and 66% of the test targets.
- First 40% of the data covers 84% of the Train targets and 86% of the test targets.
- First 60% of the data covers 95% of the Train targets and 98% of the test targets.
- First 80% of the data covers 100% of the Train targets and 100% of the test targets.

E. Results from Lift test

To plot the Lift chart, Lift Train and Lift test must be calculated, which are obtained by dividing the default cumulative percentages by the baseline percentage. These values are depicted in Fig. 17.

Decile	lift_train	lift_test	Baseline
10	2.637000	2.466000	1
9	2.362500	2.466000	1
8	2.051333	2.146000	1
7	1.854500	1.917750	1
6	1.670400	1.726000	1
5	1.501833	1.598167	1
4	1.381429	1.409000	1
3	1.229375	1.250000	1
2	1.111111	1.111111	1
1	1.000000	1.000000	1

Fig. 17: Lift Train and Lift Test

The Lift chart is shown in Fig. 18.

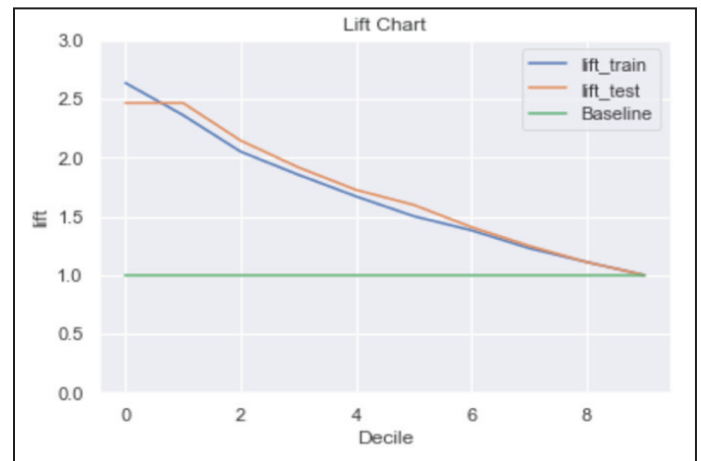


Fig. 18: Lift Chart

The Lift value interpreted from the Lift chart is tabulated in Table 9.

Table 9: Results from Lift Chart

	Lift value
Test dataset	2.637
Train dataset	2.466

From Table 9 it is inferred that, when selecting 100% of the data based on a model, it can be expected to find 2.637 times the targets for the Train set, and 2.466 times the Test set, found by randomly selecting a file without a model.

V. Conclusion and Future Scope

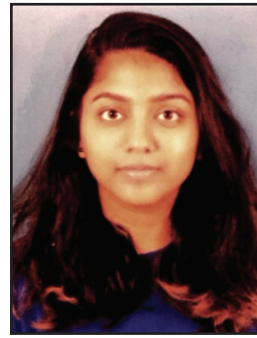
The main purpose of this work was to firstly, create a model that was capable of predicting whether a female patient, given her medical history, has or will have Diabetes. To accommodate the categorical nature of the data in the Pima Indians Diabetes Dataset, a logistic regression model was designed that was able to sufficiently predict outcomes. The accuracy of this model was improved upon by computing the AUC factor and plotting ROC. To determine how well the model was able to classify the outcomes, KS test was done and Gain chart was plotted, which yielded very interesting results, which could not have been obtained from a logistic regression model alone. As a sequel to the KS test and Gain chart, Lift test was also performed proving the ability of the model to predict accurately. The series of tests performed, help establish the validity of the model, as well as improve its accuracy, giving it an edge over other models.

This model can be used to predict the occurrence of any disease regardless of the number of classifiers. Future work would comprise of further improving the accuracy using classification algorithms and Deep Learning.

References

[1] Changsheng Zhua, Christian Uwa Idemudia, Wenfang Feng, “Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques”, in Informatics in Medicine Unlocked, April 2019.
 [2] Deepti Sisodia, Dilip Singh Sisodia, “Prediction of Diabetes using Classification Algorithms”, In International Conference

- on Computational Intelligence and Data Science (ICCIDS 2018).
- [3] N. Sneha, Tarun Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection", In Journal of Big Data, 2019.
 - [4] Quan Zou, Kaiyang Qu, Yamei Luo, Dehui Yin, Ying Ju, Hua Tang, "Predicting Diabetes Mellitus With Machine Learning Techniques", In Front. Genet. 9:515.
 - [5] Mustafa S. Kadhmi, Ikhlas Watan Ghindawi, Duaa Enteesha Mhawi, "An Accurate Diabetes Prediction System Based on K-means Clustering and Proposed Classification Approach", in International Journal of Applied Engineering Research ISSN 0973-4562, Vol. 13, No. 6, 2018.
 - [6] Souad Larabi-Marie-Sainte, Linah Aburahmah, Rana Almohaini, Tanzila Saba, "Current Techniques for Diabetes Prediction: Review and Case Study", In Journal of Applied Sciences, 2019.
 - [7] Sun, Y.L.; Zhang, D.L., "Machine Learning Techniques for Screening and Diagnosis of Diabetes: A Survey", In Teh. Vjesn. 2019.
 - [8] Aishwarya R., Gayathri P., Jaisankar N., "A Method for Classification Using Machine Learning Technique for Diabetes" in International Journal of Engineering and Technology(IJET) 5, 2013.
 - [9] Aljumah A.A., Ahamad M.G., Siddiqui M.K., "Application of data mining: Diabetes health care in young and old patients", in Journal of King Saud University - Computer and Information Sciences 25, 2013.
 - [10] Aiswarya Iyer, S. Jeyalatha, Ronak Sumbaly, "Diagnosis of diabetes using classification mining techniques", in International Journal of Data Mining & Knowledge Management Process, 2015.
 - [11] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., Chouvarda, I., "Machine Learning and Data Mining Methods in Diabetes Research" in Computational and Structural Biotechnology Journal, 2017.
 - [12] Kumar, P.S., Umatejaswi, V., "Diagnosing Diabetes using Data Mining Techniques", in International Journal of Scientific and Research Publication 2017.
 - [13] Sharief, A.A., Sheta, A., "Developing a Mathematical Model to Detect Diabetes Using Multigene Genetic Programming", in International Journal of Advanced Research in Artificial Intelligence (IJARAI), 2014.
 - [14] Priyam, A., Gupta, R., Rathee, A., Srivastava, S., "Comparative Analysis of Decision Tree Classification Algorithms", in International Journal of Current Engineering and Technology, 2013.
 - [15] Kumari, V.A., Chitra, R., "Classification Of Diabetes Disease Using Support Vector Machine", in International Journal of Engineering Research and Applications(IJERA), 2013.



Diksha Dinesh is currently pursuing her Bachelors in Technology from RV College of Engineering, in the Electronics and Communications branch. She has working on several projects during her time as a student, some of which include : developing a hybrid resource algorithm in virtual environment, automatic light intensity controller and front end for a vending machine web application. She has a keen interest in Data Science, which is why she took up this research topic. She will begin working as a Software Engineer soon.