PREDICTING DIABETES

Simple Analysis

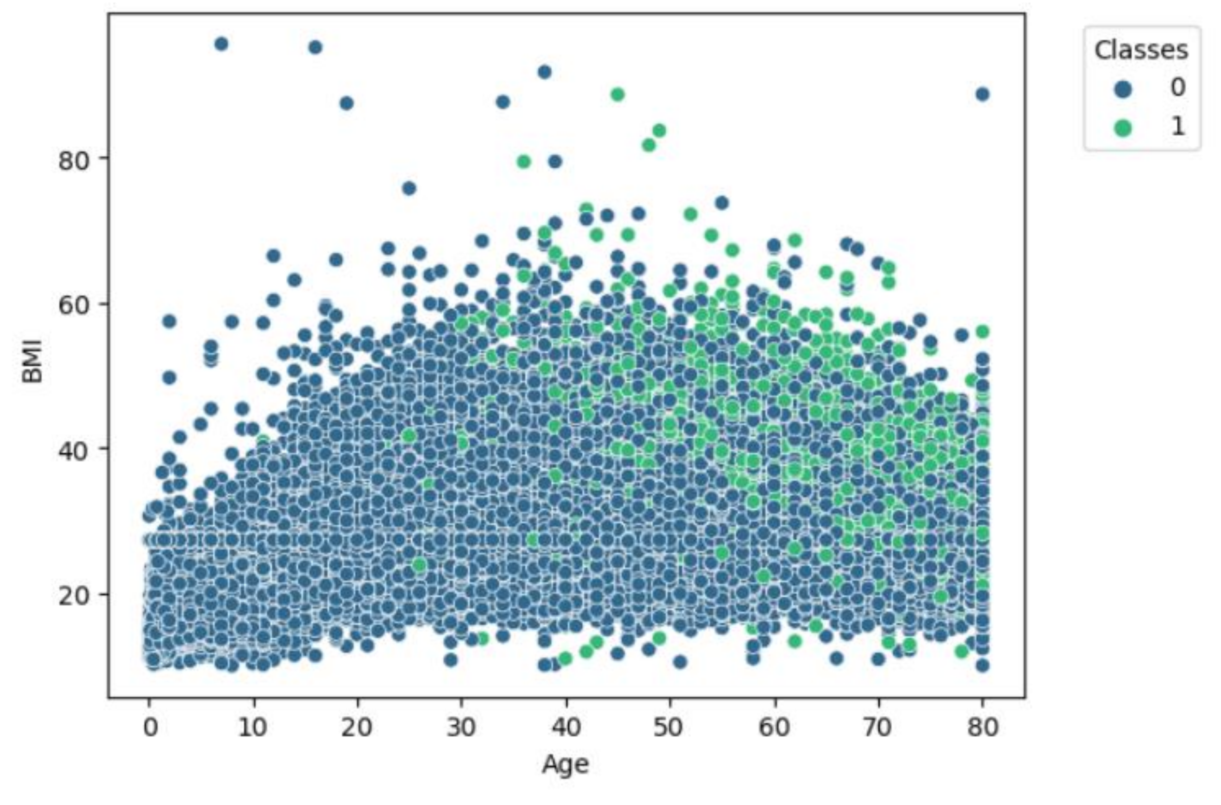
By

Nathan Notaras

RESULTS SUMMARY

- While a logistic regression model initially achieved an accuracy of 90%, the GAM model elevated this to an impressive 97%. This substantial improvement demonstrates GAM's enhanced ability to capture intricate patterns and relationships within the data that a linear model might overlook. Additionally, the fit of the model improved by 35%, as indicated by a Pseudo R-squared increase from 54% to 73%. This enhancement in fit reflects GAM's superior capacity to interpret and leverage feature significance, thereby providing a more accurate and comprehensive understanding of the dataset.
- This brief analysis underscores the significant advantages of using Generalized Additive Models (GAMs) over simple linear models, such as logistic regression, especially for complex datasets. In this case, the application of GAMs, which involve smoothing transformations for each feature, has markedly improved the model's performance.

PRELIMINARY FINDINGS



- A scatter plot of Age vs BMI for original data shows no real linear pattern a standard linear model could be used to predict diabetes.
- The green dots show some negative linear trend for higher values of age and BMI for people with diabetes.
- The blue dots show no trend for people without diabetes.

LOGISTIC RESULTS (LINEAR MODEL)

- Pseudo-R2 was only 0.54.
- More than half the features were not significant (p-value>0.05).
- This linear model failed to capture significance of most of the features.

Logit Regression Results

Dep. Variable:	diabetes	**No. Observations:**	79988
Model:	Logit	**Df Residuals:**	79973
Method:	MLE	**Df Model:**	14
Date:	Tue, 23 Jul 2024	**Pseudo R-squ.:**	0.5432
Time:	20:34:45	**Log-Likelihood:**	-10595.
converged:	True	**LL-Null:**	-23193.
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-11.7939	2.58e+06	-4.56e-06	1.000	-5.07e+06	5.07e+06
year	-0.0316	0.019	-1.704	0.088	-0.068	0.005
gender	0.2890	0.037	7.845	0.000	0.217	0.361
age	0.0454	0.001	40.435	0.000	0.043	0.048
location	-0.0011	0.001	-0.919	0.358	-0.004	0.001
race:AfricanAmerican	-2.2694	2.58e+06	-8.78e-07	1.000	-5.07e+06	5.07e+06
race:Asian	-2.3039	2.58e+06	-8.91e-07	1.000	-5.07e+06	5.07e+06
race:Caucasian	-2.4019	2.58e+06	-9.29e-07	1.000	-5.07e+06	5.07e+06
race:Hispanic	-2.3968	2.58e+06	-9.27e-07	1.000	-5.07e+06	5.07e+06
race:Other	-2.4219	2.58e+06	-9.37e-07	1.000	-5.07e+06	5.07e+06
hypertension	0.7741	0.049	15.832	0.000	0.678	0.870
heart_disease	0.7754	0.063	12.362	0.000	0.652	0.898
smoking_history	0.0900	0.010	8.596	0.000	0.070	0.111
bmi	0.0010	2.82e-05	33.867	0.000	0.001	0.001
hbA1c_level	0.4525	0.006	70.151	0.000	0.440	0.465
blood_glucose_level	0.3186	0.005	59.451	0.000	0.308	0.329

LOGISTIC RESULTS (GAM MODEL)

- Pseudo-R2 increased to 0.73.
- Every feature now is significant (p-value<0.05).
- This GAM model was successful in capturing significances of all of the features.

LogisticGAM

Distribution:	BinomialDist	Effective DoF:	73.0026
Link Function:	LogitLink	Log Likelihood:	-6328.2693
Number of Samples:	79988	AIC:	12802.5439
		AICc:	12802.6828
		UBRE:	2.1608
		Scale:	1.0
		Pseudo R-Squared:	0.7271

Feature Function	Lambda	Rank	EDoF	P > x	Sig. Code
s(0)	[9.1]	20	6.1	8.97e-01	
s(1)	[9.1]	20	1.8	1.11e-08	***
s(2)	[9.1]	20	11.0	0.00e+00	***
s(3)	[9.1]	20	14.4	7.31e-01	
s(4)	[9.1]	20	1.0	1.63e-11	***
s(5)	[9.1]	20	1.0	4.81e-14	***
s(6)	[9.1]	20	1.0	0.00e+00	***
s(7)	[9.1]	20	1.0	2.22e-16	***
s(8)	[9.1]	20	0.0	0.00e+00	***
s(9)	[9.1]	20	1.0	0.00e+00	***
s(10)	[9.1]	20	1.0	0.00e+00	***
s(11)	[9.1]	20	5.0	0.00e+00	***
s(12)	[9.1]	20	12.7	0.00e+00	***
s(13)	[9.1]	20	7.2	0.00e+00	***
s(14)	[9.1]	20	8.8	0.00e+00	***
intercept		1	0.0	3.28e-02	*