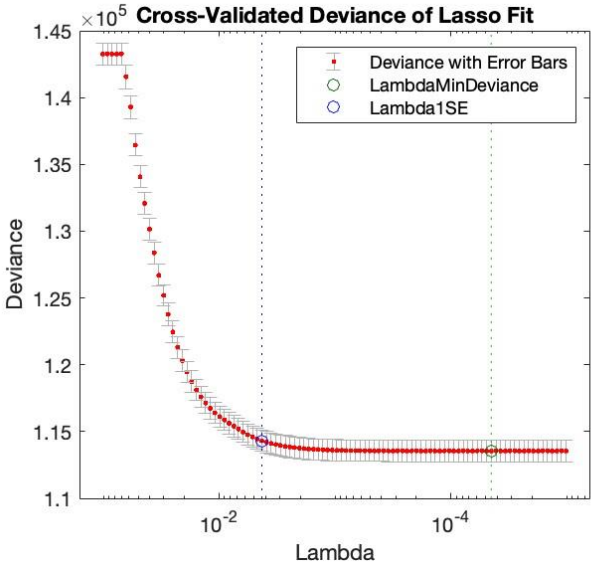
Glossary

BRFSS	Behavioral Risk Factor Surveillance System
CDC	Centers for Disease Control
L1 & L2	Lasso Regularisation and Ridge Regularisation are techniques in machine learning to prevent overfitting and enhance the performance of models in logistic regression or linear regression.
MSE rule	Selects λ minimising Mean Square Error during cross-validation.
1SE rule	Chooses a simpler model within one standard error of the minimum cross-validated error.
ANOVA	Analysis of Variance is a statistical method used to analyse the differences among group means in a sample.
SMOTE	The Synthetic Minority Oversampling Technique is an oversampling technique to deal with imbalanced datasets.

Intermediate results

The choice of features:

In Logistic Regression, I used a cross-validated lasso regularization with 3-fold and chose two sets of parameters with MSE rule and 1SE rule[1,2]. The results of cross-validation deviance of Lasso fit are shown in Figure. The MSE rule means it has the minimum mean squared error across the cross-validated fold is indicated by the green circle in Figure, and the 1SE rule means we choose the simplest model within one standard error of the minimum cross-validated error is indicated by the blue circle in the Figure.



In Naïve Bayes, I used the Chi-squared test and ANOVA, with their p-values to determine the importance of features. First, I chose a set of parameters with p-values less than 0.0024, since there are 21 features in total. Second, I chose 15 features with the smallest p-values as a set of parameters, with the same number of features in the Logistic Regression second model.

I also considered using p-values equal to zero as a selection criterion, but it eliminated too many factors. To ensure a fair comparison with Logistic Regression, I decided not to use it in the project. The training results will be presented in the implementation details section below.

Reference:

1. Tibshirani R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B (Methodological)*. 1996;58(1):267-288. http://www.jstor.org/stable/2346178

2. Zhou H, Seker V, Downar T. Enhanced Lasso Regularization-Based Self-Adaptive Feature Selection Algorithm for the High-Dimensional Uncertainty Quantification of TREAT Transient Test Modeling. *Nucl Technol*. 2020;206(6):839-861. doi:10.1080/00295450.2020.1746620

Implementation details

The choice of cutoff point in Logistic Regression models:

In Logistic Regression, predictions are presented as probability, and a cutoff point is needed to classify outcomes. Typically, the cutoff point is set at 0.5. However, for imbalanced targets, different approaches are considered. Initially, I considered using the Youden index, which maximises the combination of sensitivity and specificity and is commonly used in medical data. But with 10-fold cross-validation, deciding on the model for this calculation remains uncertain. In the end, I chose 0.14 as a cutoff based on the original target distribution.

Even when using different cutoff points, the values with matrices change significantly, but the overall patterns remain consistent. Interestingly, at the 0.5 cutoff point, the F1-score value in Model 1 performs better. However, just like I presented in the poster, this difference could be random, and with fewer features might also prevent overfitting.

The following table presents how the matrices would change with different cutoff points.

	Cutoff = 0.14			*Cutoff = 0.5*		
Results	Training		Testing	Training		Testing
Logistic Regression	Model 1	Model 2	Model 2	Model 1	Model 2	Model 2
Training errors	0.2701	0.2702	-	0.1361	0.1362	-
F1-score	0.4415	0.4416	0.4411	0.2392	0.2384	0.2406
Precision	0.3099	0.3099	0.3101	0.5389	0.5346	0.5251
Recall	0.7677	0.7679	0.7639	0.1538	0.1534	0.1561
Accuracy	0.7298	0.7298	0.7293	0.8639	0.8636	0.8623

The training results for Naïve Bayes models:

The table below shows the training results of Naïve Bayes with 21, 15, and 12 features, using p-values to assess variable importance. Overall, the model with 21 features still has the best performance. However, it is worth applying various feature selection techniques to observe how different models will perform.

	Naïve Bayes		
Training results	Model 1	Model 2	*Model 3*
Number of features	21	15	12
Mean Training errors	0.1843	0.1842	0.1827
Mean Validated F1-score	0.4309	0.4302	0.4283
Mean Validated Precision	0.3776	0.3777	0.3792
Mean Validated Recall	0.5020	0.4997	0.4925
Mean Validated Accuracy	0.8156	0.8159	0.8172
Mean Validated AUC	0.8086	0.8075	0.8070
Training Time (s)	2.6534	1.9106	1.6897