A Comparison of Naïve Bayes and Logistic Regression Applied to the UCI Bank Marketing Data Set

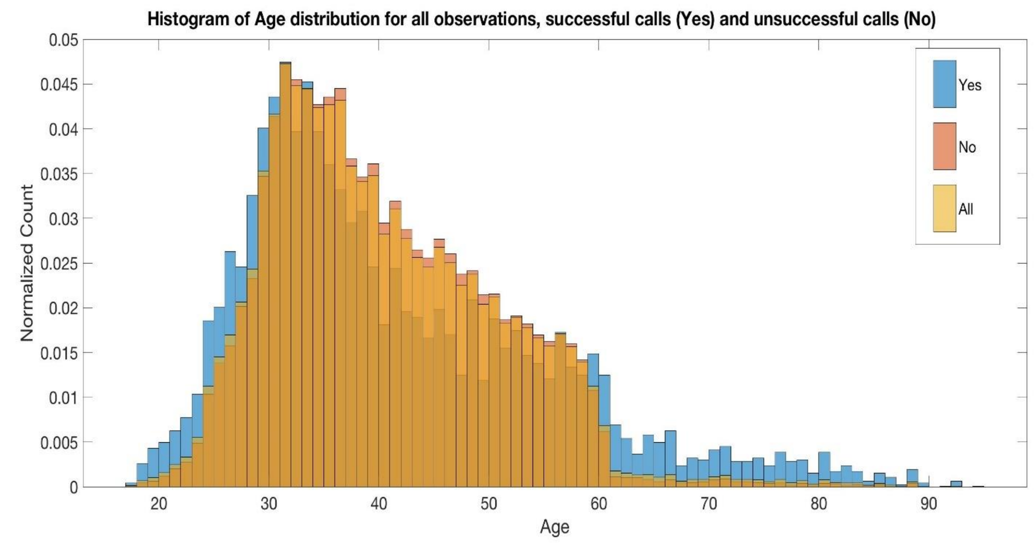
Toby Staines and Patrick Horgan

Description and motivation of the problem

- Predict the comparative success of a bank's marketing calls based on the results of a previous marketing campaign, enabling better allocation of resources in future campaigns
- We will undertake this work using two different interpretable machine learning techniques: Naïve Bayes and Logistic Regression, in order to compare their effectiveness

Dataset Summary and Initial Analysis

Feature	Min	Max	Median	Mean	Std	Skewnes
Age	17	98	38	40.0241	10.4212	0.7
Campaign	1	56	2	2.5676	2.77	4.7
Pdays	0	999	999	962.4755	186.9109	-4.
Previous	0	7	0	0.173	0.4949	3.8
EmpVarRate	-3.4	1.4	1.1	0.0819	1.571	-0.7
ConsPriceIdx	92.201	94.767	93.749	93.5757	0.5788	-0.2
ConsConfIdx	-50.8	-26.9	-41.8	-40.5026	4.6282	0.3
Euribor3m	0.634	5.045	4.857	3.6213	1.7344	-0.7
NrEmployed	4.96E+03	5.23E+03	5191	5.17E+03	72.2515	-1.0



- Dataset: Bank Marketing from UCI
- Training set: 28832 observations
- Validation set: 6178 observations
- Test set: 6178 observations
- The dataset has 20 predictors – 10 numeric, 10 categorical
- The ratio of Yes to No for the classifier in the data is 0.11 to 0.89
- One indicator (Duration) was removed because it would be unknown for future calls so should not be used as a predictor
- Mean, median, min, max & skew were calculated for the 10 numeric predictors
- The numeric predictors were not good approximations of a normal distribution

An Introduction to the Two Techniques

Logistic regression

- In a binary classification problem, results are either 1 (event occurrence) or 0 (no event occurrence). A standard linear regression built on these results predicts results y < 0 and y > 1, both of which are not possible.
- Logistic Regression uses a link function to convert a standard linear regression to a sigmoid curve bounded by 0 and 1.
- For any input X this gives an output $0 \leq y \leq 1$, which represents the probability of occurrence of an event (in our case a customer agreeing to make a long-term deposit).
- A threshold probability is then set, with inputs giving a probability greater than the threshold classified as 1 and those with a probability below the threshold classified as 0.

Pros

- Computationally very fast, with expense increasing linearly with increasing dimensionality.
- Highly interpretable.

Cons

- Not as accurate as more complex models such as Neural Nets and Support Vector Machines.

Naïve Bayes

- Naïve Bayes is a family of probabilistic classification algorithms based on Bayes Theorem, all sharing the common principle that every predictor being classified is independent of any other predictor. They work by calculating the probabilities for each predictor conditional on the class value, then using the product rule to obtain a joint conditional probability for the predictors, and using Bayes rule to derive conditional probabilities for the class. The class with the highest probability is then selected.

Pros

- Naïve Bayes have been shown to work surprisingly well for classification despite its simplicity, in particular in text classification problems (3).
- It also only needs a small number of observations to estimate the parameters needed for classification and is an interpretable classification method.

Cons

- Generally it predicts less accurately than other models (1).
- Naïve Bayes works by assuming independence between predictors which may not be true.

Hypothesis Statement

We expect Logistic Regression to correctly classify more observations than Naïve Bayes (2), though some situations have been noted in which Naïve Bayes performs better than Logistic Regression (3).

Moro et al (1) looked at the performance of Neural Networks, Support Vector Machines, Decision Trees and Logistic Regression with this dataset. We chose to compare one of these, Logistic Regression, to another thought to have similar advantages and disadvantages, Naïve Bayes.

We expect both models to perform reasonably well, though not as well as Neural Networks or Support Vector Machines, which outperformed Logistic Regression in the Moro et al paper.

Training and evaluation methodology

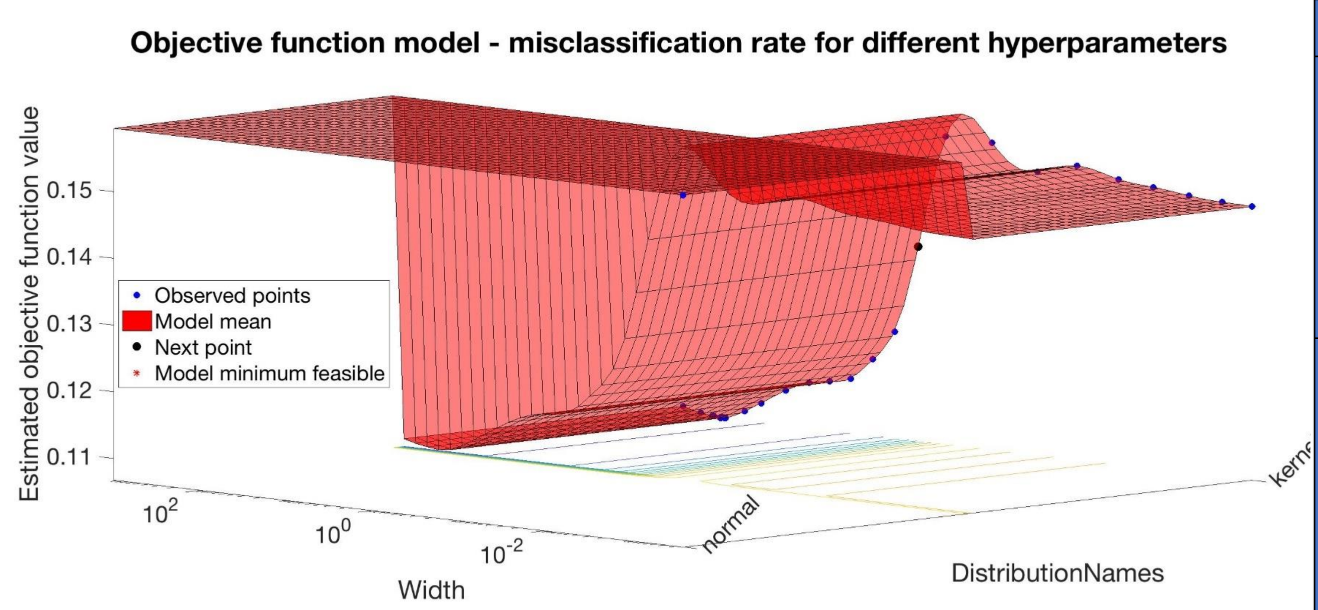
Of the four datasets available on UCI, we chose the dataset nearest to that used in Moro et al. for comparability. With a relatively large dataset, we choose to split into training, validation and test data with random sampling and a 70:15:15 ratio, rather than cross validate. Data was sampled at the outset and used in both techniques, ensuring faire evaluation of each model. For both methods the 10 numeric predictors were binned to create categorical predictors. These bins were chosen by looking at the basic statistics and distributions of the predictors. Evaluation measures used are the area under a Receiver Operating Characteristic curve (AUC), accuracy (proportion of observations correctly classified) and sensitivity (proportion of positive outcomes correctly identified)

Choice of parameters and experimental results

Naïve Bayes

The initial model was created using the default prior probability, a normal distribution and bins of equal sizes for the numerical predictors. Then models were creating using specific individual bin sizes determined by examination of the numerical predictors to improve the model. Next the prior probability was adjusted to see the change in performance of the model.

The best model was created using hyperparameter optimization by attempting to minimise the cross-validation loss for the model by altering the distribution, kernel and width (see graph on right). Applying prior probability or binning did not improve this model.



		Training	Validation	Test
Naïve Bayes	AUC	0.75	0.75	0.76
	Accuracy	0.89	0.90	0.89
	Sensitivity	0.30	0.31	0.27
Logistic Regression	AUC	0.80	0.80	0.80
	Accuracy	0.90	0.91	0.90
	Sensitivity	0.25	0.27	0.23

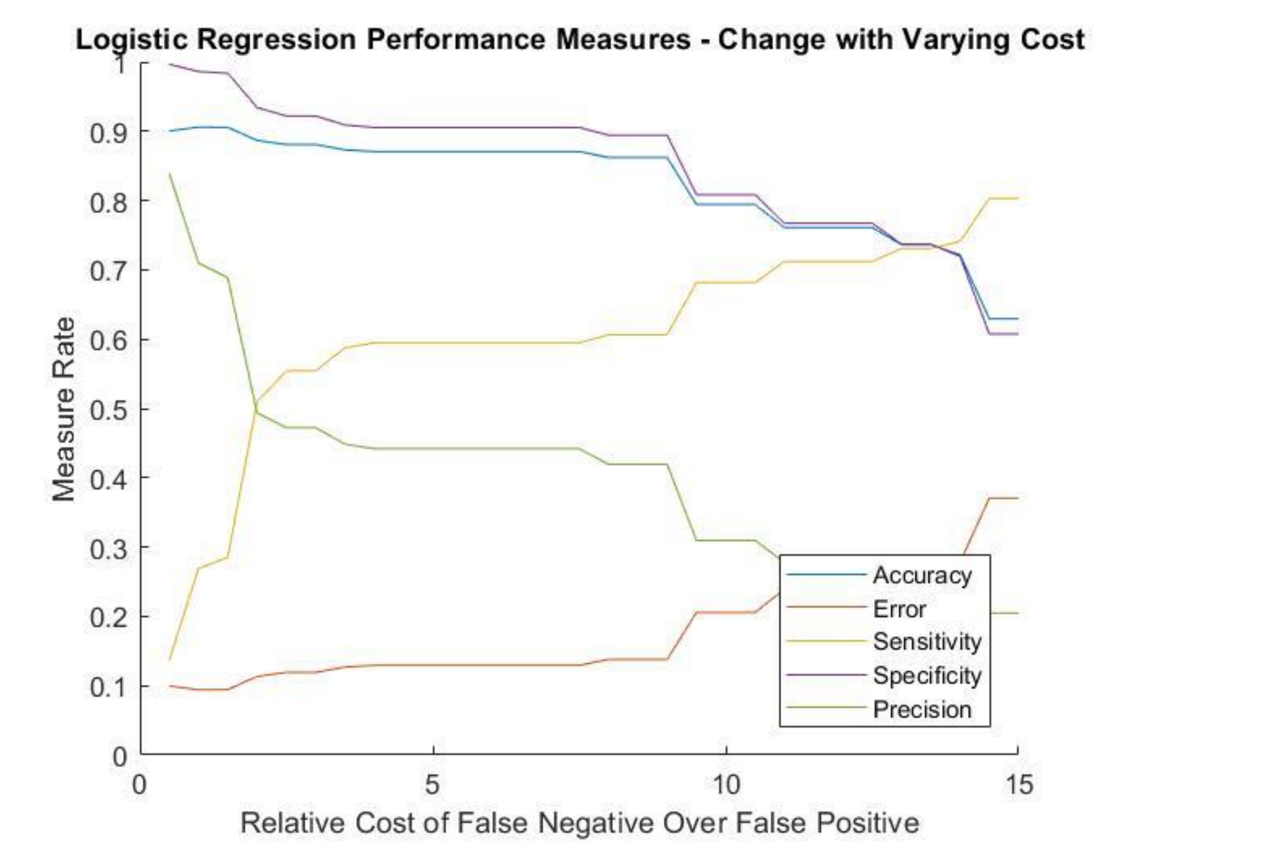
Logistic Regression

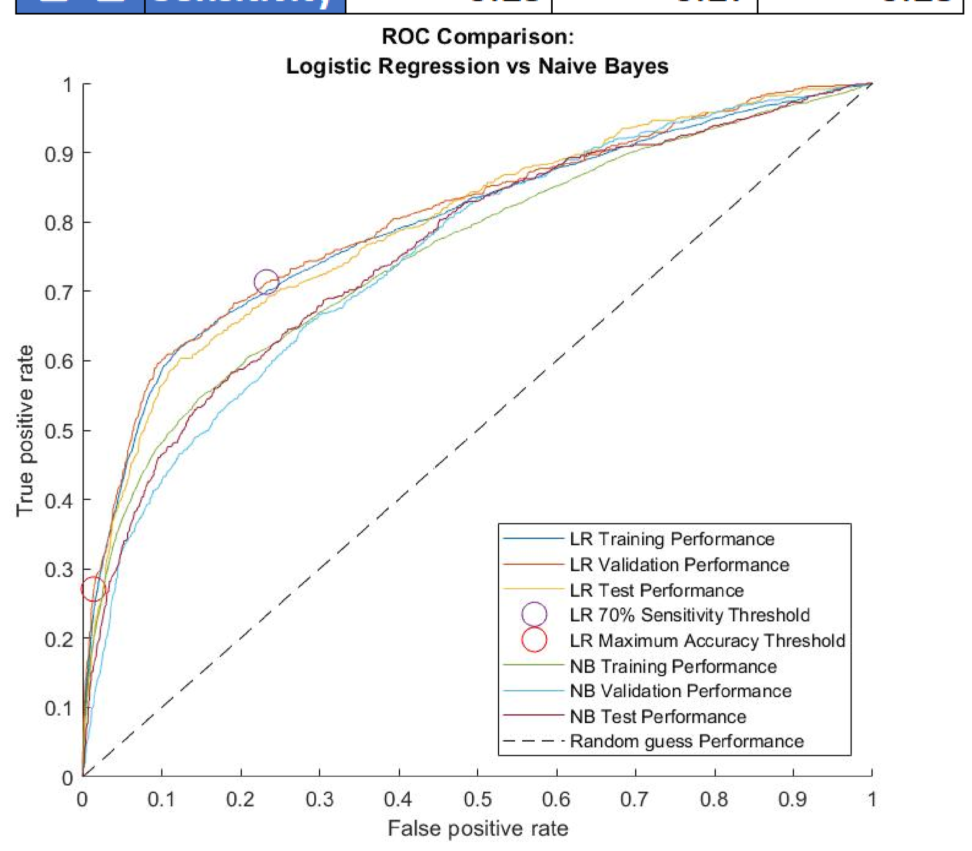
The optimal coefficients for the Logistic Regression are determined automatically by maximum likelihood estimation, so there is little requirement for parameter selection prior to creating the model.

The main parameter choice is in where to set the classification threshold to translate the output probabilities in to predicted classifications.

The simplest approach is to select the threshold which gives the highest overall accuracy, but this may not be appropriate, depending on domain context. Due to the skewed outcome ratio in our data the maximum accuracy approach greatly favours specificity over sensitivity (4).

Implementing a cost function on the validation data allowed us to vary the relative cost of false negative and false positive results, with the results illustrated in the graph on the right.





Analysis and critical evaluation of results

On total accuracy and AUC, Logistic Regression performed slightly better than Naïve Bayes, but on these measures, performance of the two models was similar.

For both methods there was low variance between the test, validation and test results. This may be due to the relatively large training dataset.

The individual predictor binning levels decreased the loss for the validation dataset of the Naïve Bayes model (from 0.195 to 0.182) though introducing the binning for the Logistic Regression had had very little effect. With the optimised hyperparameters for the Naïve Bayes model (which gave by far the lowest loss 0.103) the binning increased the loss, and adjusting the prior probability to reflect the whole data set rather than the training dataset made little difference. Some of the Naïve Bayes models created using optimised hyperparameters took significantly more time (the largest being 98.7 seconds objective runtime and the smallest under 1 second) than the initial Naïve Bayes models and the Logistic Regression models.

In the context of our data, it is likely that the bank would consider a false negative (missed sales opportunity) as significantly more costly than a false positive (a wasted phone call). Both models showed poor sensitivity initially, although Naïve Bayes's was higher. In order to account for this, we ran the Logistic Regression once for maximum accuracy and once with a minimum sensitivity of 70% in the validation set. This required a cost ratio of 11:1 and achieved an increase in sensitivity of 44.3%, with a loss in total accuracy of 14.5%.

Lessons learned and future work

For the Naïve Bayes model we could look at more sophisticated binning techniques and methods, rather than those created using only a visual analysis of the predictors (5).

Some of the predictors, such as pdays, contain data that could possibly benefit from further transformations or cleansing. There were a large proportion of observations that contained numeric data that did not reflect a true value, e.g. 999 standing for "not previously contacted", that could be dealt with more systematically.

The final Naïve Bayes model did not have any adjustments for a misclassification cost, so this may also be an area to explore further and compare with our results from the Logistic Regression.

Moro et al also carried out some further work, reducing the features even further, and it would be interesting to see how this would affect the performance of both of our models.

References:

[1] S. Moro, P. Cortez, and P. Rita, "A data-driven approach to predict the success of bank telemarketing," *Decis. Support Syst.*, vol. 62, pp. 22–31, 2014.

[2] T. K. Asociación Española para la Inteligencia Artificial., *Inteligencia artificial : revista iberoamericana de inteligencia artificial.*, vol. 18, no. 56. Asociación Española para la Inteligencia Artificial (AEPIA), 1997.

[3] A. Y. Ng and M. I. Jordan, "On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes." pp. 841–848, 2002.

[4] J. J. Chen, C. A. Tsai, H. Moon, H. Ahn, J. J. Young, and C. H. Chen, "Decision threshold adjustment in class prediction," *SAR QSAR Environ. Res.*, vol. 17, no. 3, pp. 337–352, 2006.

[5] F. Kaya, "Discretizing Continuous Features for Naive Bayes and C4.5 Classifiers," *Univ. Maryl. Publ.*, 2008.