ETC3250 PROJECT: ANALYTICS IN BANKING

PRESENTATION BY:

DURKESWARI MOHAN	27675785
QIAN ZHI LAI	27556263
SAMANTHA KAR-KEI LOOI	26965607
DENNY BARING	26941910
MURPHY GUO	26879662
KYLE PUAH WAI HENG	27004783

OBJECTIVE

- Accurately predict the probability a client would subscribe to a bank term deposit, using the given predictors in our dataset
- Reduce Log loss function

$$Log\ loss = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{p}_i) + (1 - y_i)\log(1 - \hat{p}_i)],$$

Where:

- n is the number of data points in the test set
- $\hat{p}_i$ is the predicted probability
- $y_i \in \{0,1\}$
- $log()$ is the natural (base e) logarithm

DATA CLEANING

"UNKNOWN" VALUES

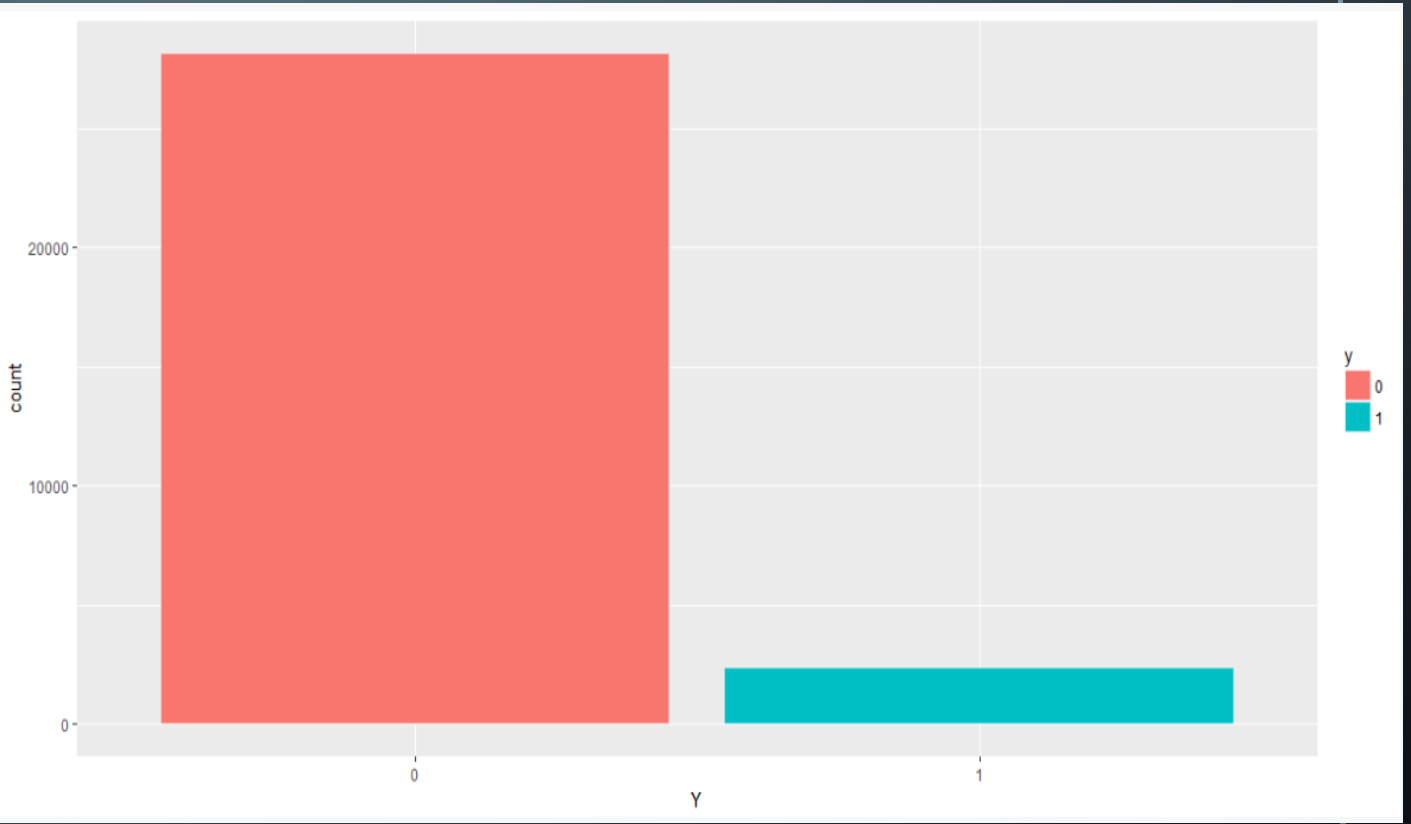
```
      age            campaign          pdays          previous
 Min.   :18.00   Min.   : 1.000   Min.   :  0.0   Min.   :0.00000
 1st Qu.:33.00   1st Qu.: 1.000   1st Qu.:999.0   1st Qu.:0.00000
 Median :39.00   Median : 2.000   Median :999.0   Median :0.00000
 Mean   :40.19   Mean   : 2.727   Mean   :995.5   Mean   :0.04886
 3rd Qu.:47.00   3rd Qu.: 3.000   3rd Qu.:999.0   3rd Qu.:0.00000
 Max.   :95.00   Max.   :56.000   Max.   :999.0   Max.   :3.00000

     marital                 job          default         housing
 divorced: 3550   admin.      :7496   no     :22901   no     :14133
 married :19146   bluecollar  :7054   unknown: 7532   unknown:  733
 single  : 7689   technician  :5280   yes    :    3   yes    :15570
 unknown :   51   services    :2999
                  management  :2233
                  entrepreneur:1175
                  (Other)     :4199

     loan                      edu            month       day_of_week
 no     :25093   university.degree  :8847   may    :7769   fri:5793
 unknown:  733   high.school        :6846   jul    :6685   mon:6226
 yes    : 4610   9                  :4541   aug    :5175   thu:6466
                 professional.course:3986   jun    :4374   tue:5915
                 4                  :3199   nov    :3616   wed:6036
                 6                  :1746   apr    :2458
                 (Other)            :1271   (Other): 359
      poutcome          contact       y
 failure    : 1320   cellular :16537   0:28094
 nonexistent:29013   telephone:13899   1: 2342
 success    :  103
```

- Dummy variables on all factor variables
- 14 variables -> 53 variables

IMBALANCED DATA

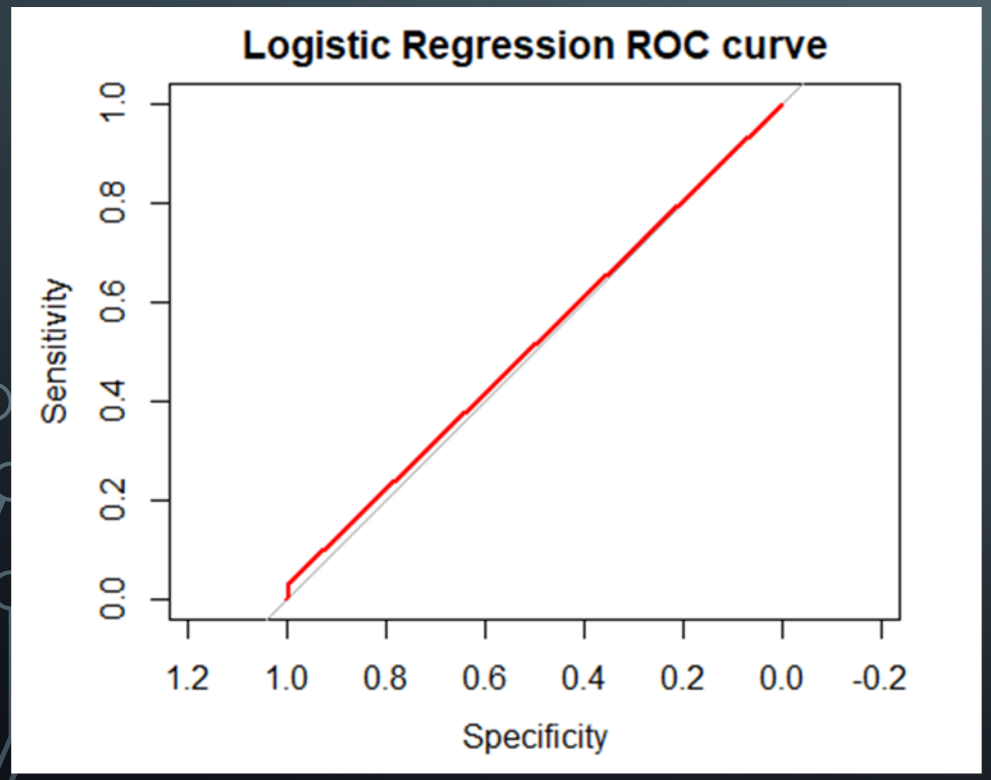


- 1:12 ratio
- Confusion matrix may not be ideal – results in misleading predictions
- ROC (Receiver Operating Characteristic) curve and AUC (Area under the Curve)

MODEL ESTIMATION

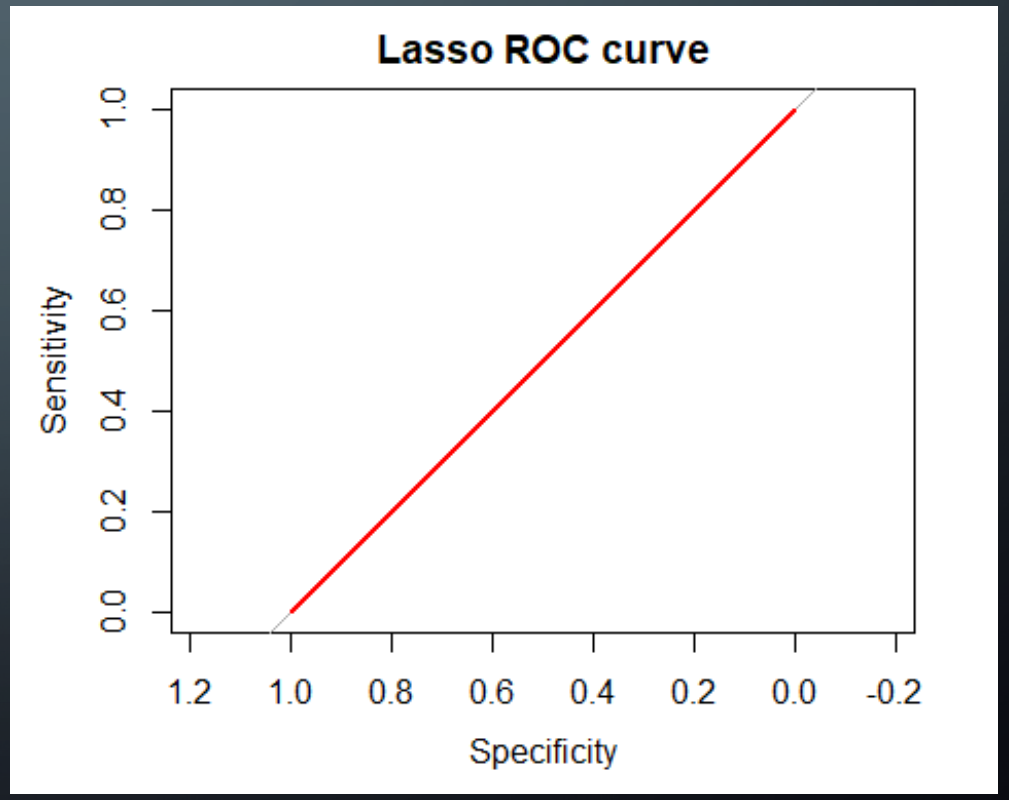
LOGISTIC REGRESSION

Actual	Prediction	
	0	1
0	28035	59
1	2266	76



LASSO MODEL

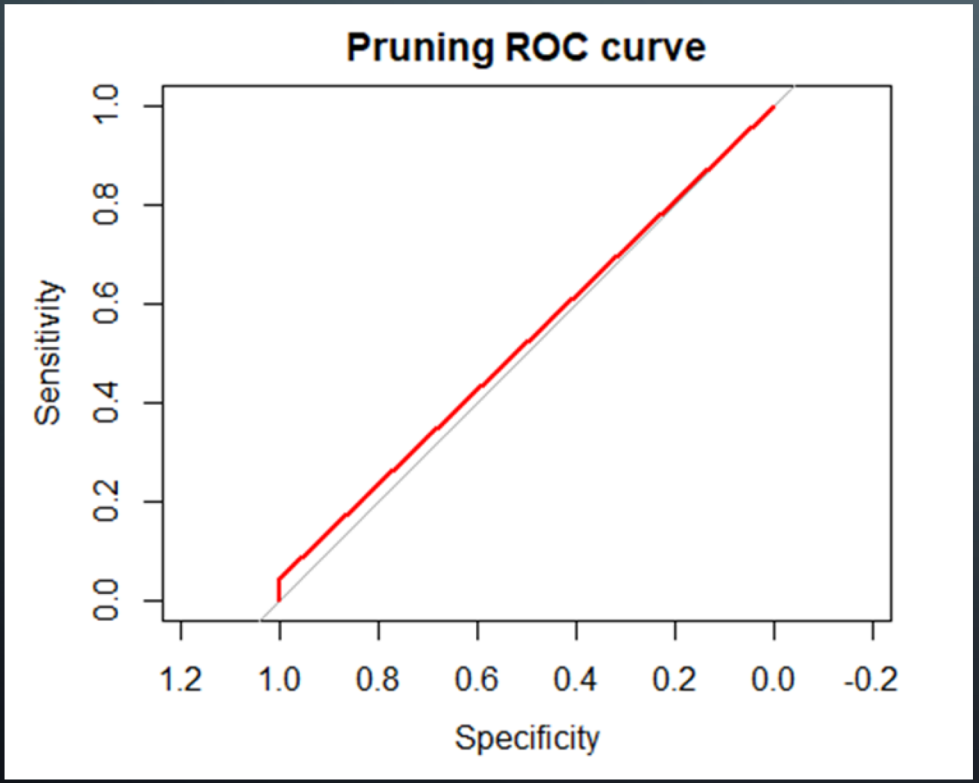
Actual	Prediction	
	0	1
0	28094	-
1	2342	-



MODEL ESTIMATION

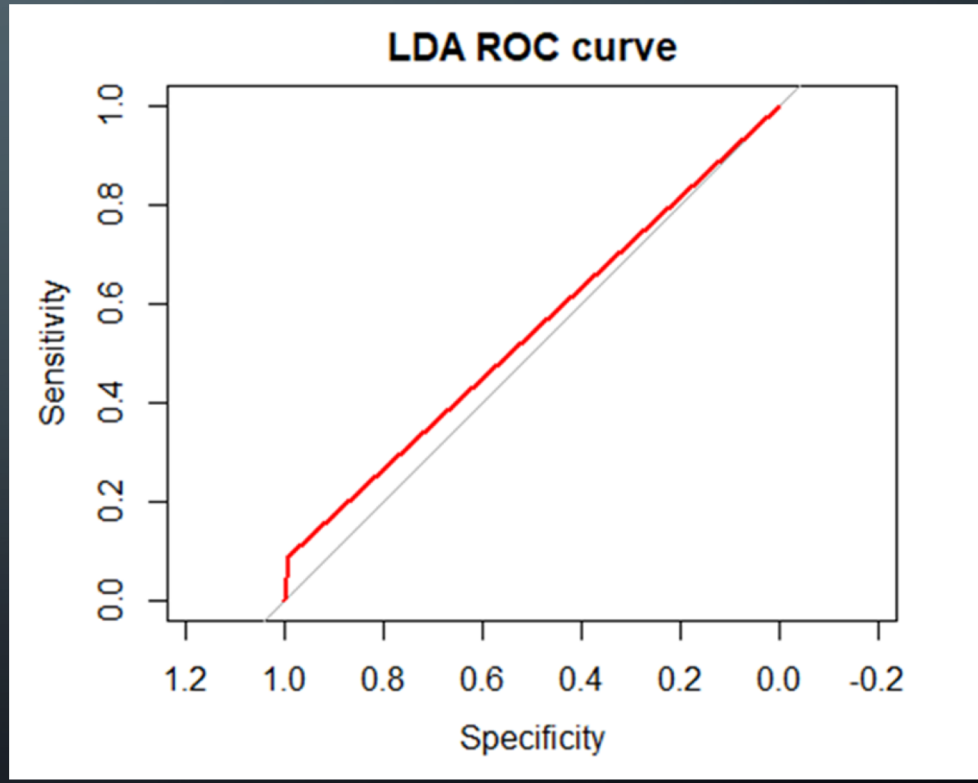
PRUNED TREE

Actual	Prediction	
	0	1
0	28046	48
1	2241	101



LDA MODEL

Actual	Prediction	
	0	1
0	27845	249
1	2135	207



RESULTS AND FINDINGS

CONFUSION MATRIX

Model	Prediction Accuracy
Logistic Regression	0.9236
Lasso	0.9231
LDA	0.9217
Pruned tree	0.9248

AREA UNDER THE ROC CURVE

Model	AUC
Logistic Regression	0.5152
Lasso	0.5
LDA	0.5398
Classification tree	0.5207

DISCUSSION

- Aim to minimise the log loss function was not satisfied with LDA or Classification tree due to the nature of its prediction by giving a binary outcome instead of probability of an outcome occurring
- Since the objective was to minimise the log loss function we chose Logistic regression as our model to predict on the test set

CONCLUSION & RECOMMENDATION

- Logistic regression chosen as final model
- A different way to tackle imbalanced data
  - Resampling method
- Prediction type for classification tree
  - Changing the argument *type = "class"* to *type = "vector"* in the *predict* command