FINAL ASSIGNMENT

Karla Guadalupe Sam Millan
Statistical Learning
10/06/2022

ADCTL

Data Split & Feature Selection

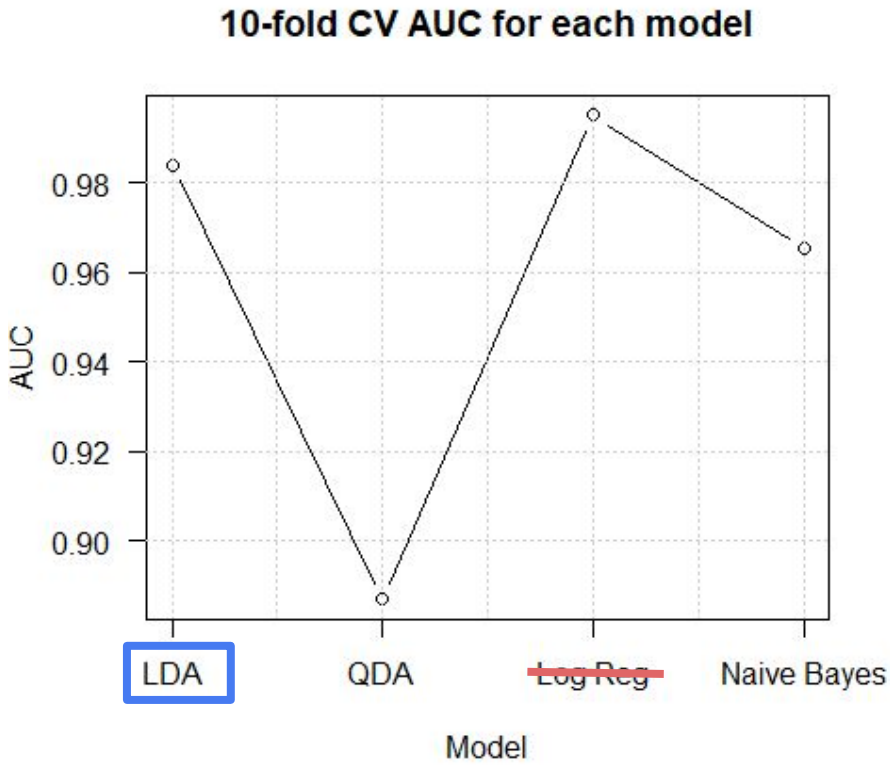
- The data was split at the beginning, saving the first **80%** of the rows for training, and **20%** for validation.
- For feature selection, as there were too many variables in these challenges to use stepwise selection, and since ridge regression doesn't set the values to 0 and it is therefore not easily interpretable, it was decided to use **lasso** with the **caret library**
  - **Lambda** was fine tuned, with values ranging from 10^{10} - 10^{-2} and a length of 100 and using a **k-fold CV of k=10**
  - Its **best value** was determined to be **0.01**
  - The result of this was the **38 most important variables**, shown as follows:

```
 [1] "Supp_Motor_Area_L"   "OFClat_R"            "Hippocampus_L"       "Hippocampus_R"
 [5] "Amygdala_R"          "Occipital_Mid_R"     "Angular_L"           "Caudate_L"
 [9] "Heschl_L"            "Temporal_Pole_Sup_R" "Temporal_Mid_L"      "Temporal_Pole_Mid_R"
[13] "Cerebelum_3_L"       "Cerebelum_10_L"      "Vermis_1_2"          "Vermis_7"
[17] "Vermis_10"           "ABCA7"               "AGTRAP"              "C1orf63"
[21] "C6orf115"            "CBL"                 "CETN2"               "CYTH1"
[25] "DCUN1D1"             "DNAJC7"              "F13A1"               "FKBP5"
[29] "FNIP1"               "FTHL8"               "IQGAP1"              "LCOR"
[33] "NACA"                "RGS19"               "RPL36AL"             "TACC3"
[37] "TRIM41"              "VCAN"
```

ADCTL

Models AUC and MCC

- Again, using **caret**, different models were tried with the train set from the split , which were: LDA, QDA, Logistic Regression and Naive Bayes
- **Logistic Regression is discarded** as it doesn't converge, due to the high number of predictors.
- Among the rest, LDA is the best performing one, with a **ROC = 0.984**
  - QDA ROC = 0.887
  - Naive Bayes ROC = 0.9655
- The LDA model is used to create the prediction probabilities and the classes with both sets of the data split.
- **AUC** & **MCC** are computed with both sets of the data split (training and validation)
  - AUC = 0.9998 & 0.9925
  - MCC = 0.985 & 0.83
- The final LDA model is trained on the whole dataset and used for the test predictions.



ADMCI

Data Split & Feature Selection

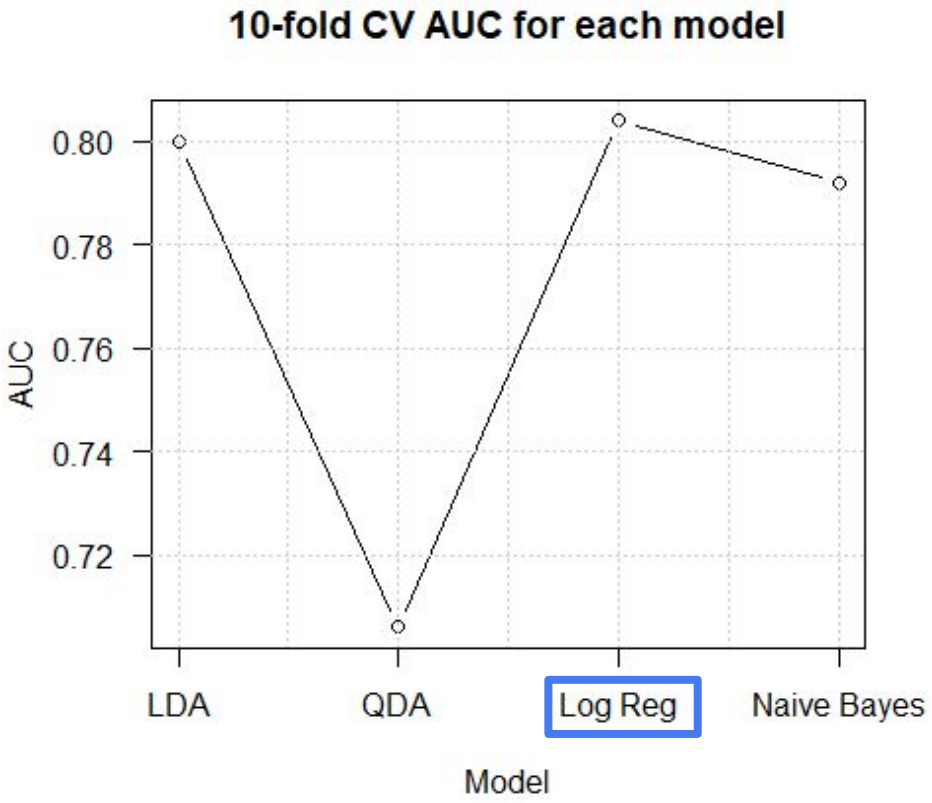
- The data was split at the beginning, saving the first **80%** of the rows for training, and **20%** for validation.
- For feature selection, it was again decided to do **lasso** with the **caret library**
- **Lambda** was fine tuned, with values ranging from 10^{10} - 10^{-2} and a length of 100 and using a **k-fold CV of k=10**
  - Its **best value** was determined to be **0.04037017**
  - The result of this was the **12 most important variables**, shown as follows:

```
 [1] "Left.Precentral.Gyrus"          "Left.Angular.Gyrus"
 [3] "Right.Angular.Gyrus"            "Left.Inferior.Occipital.Gyrus"
 [5] "Left.Middle.Temporal.Gyrus"     "Right.Middle.Temporal.Gyrus"
 [7] "Left.Caudate"                   "Right.Caudate"
 [9] "Left.Hippocampus"               "Right.Hippocampus"
[11] "ARPC5"                          "PIP5K2A"
```

ADMCI

Models AUC and MCC

- LDA, QDA, Logistic Regression and Naive Bayes were tried
- Among them, **logistic regression** is the best performing one, with a **ROC = 0.804**
  - LDA ROC = 0.800
  - QDA ROC = 0.706
  - Naive Bayes ROC = 0.792
- The Log Reg model is used to create the prediction probabilities and the classes with both sets of the data split.
- **AUC** & **MCC** are computed with both sets of the data split (training and validation)
  - AUC = 0.8746 & 0.92
  - MCC = 0.6506 & 0.5825
- The final model is trained on the whole dataset and used for the test predictions.



MCICTL

Data Split & Feature Selection

- The data was split at the beginning, saving the first **80%** of the rows for training, and **20%** for validation.
- For feature selection, it was again decided to do **lasso** with the **caret library,** for the same reasons
- **Lambda** was fine tuned, with values ranging from 10^{10} - 10^{-2} and a length of 100 and using a **k-fold CV of k=10**
  - Its **best value** was determined to be **0.01321941**
  - The result of this was the **55 most important variables**, shown as follows:

```
 [1] "Frontal_Inf_Oper_R"    "Frontal_Inf_Orb_2_L"   "Rolandic_Oper_L"      "Supp_Motor_Area_R"
 [5] "Frontal_Med_Orb_R"     "Insula_L"              "Cingulate_Post_L"     "ParaHippocampal_L"
 [9] "ParaHippocampal_R"     "Amygdala_R"            "Lingual_L"            "Lingual_R"
[13] "Fusiform_R"            "Parietal_Sup_R"        "Parietal_Inf_L"       "Angular_L"
[17] "Angular_R"             "Paracentral_Lobule_R"  "Caudate_R"            "Pallidum_R"
[21] "Thalamus_R"            "Heschl_R"              "Temporal_Pole_Sup_R"  "Cerebelum_Crus1_R"
[25] "Cerebelum_9_L"         "Cerebelum_10_R"        "AMY1C"                "APBB3"
[29] "ARHGAP4"               "CTSW"                  "CXCL16"               "DGKQ"
[33] "EDG4"                  "EXOC3"                 "FAM39DP"              "GPR97"
[37] "GSTM1"                 "HLA.H"                 "HSP90AB1"             "ITGAM"
[41] "KIAA0513"              "KLF6"                  "LOC728499"            "MICAL1"
[45] "MX1"                   "NDUFA1"                "PHACTR2"              "RPA3"
[49] "SELPLG"                "TGFBR3"                "TMEM8"                "TMEM86B"
[53] "TOP3A"                 "TTC15"                 "VWCE"
```

MCICTL

Models AUC and MCC

- LDA, QDA, Logistic Regression and Naive Bayes were tried
- **Logistic Regression is discarded** as it doesn't converge, due to the high number of predictors (55).
- Among the rest, **LDA** is the best performing one, with a **ROC = 0.9471**
  - QDA ROC = 0.5123
  - Naive Bayes ROC = 0.8538
- The LDA model is used to create the prediction probabilities and the classes with both sets of the data split.
- **AUC** & **MCC** are computed with both sets of the data split (training and validation)
  - AUC = 1 & 0.9375
  - MCC = 0.9712 & 0.5221
  - AUC doesn't vary much, but MCC does, suggesting overfitting, likely due to the high number of predictors.
- The final LDA model is trained on the whole dataset and used for the test predictions.