Linear Model

Author: Rajat Jain

Last Updated: 2018-04-24

Contents

Training & Test Data

We have split available usage data into training data (75% - 1987 records) and test data(25% - 663 records).

Summary of Training data

```
##  class     lr_cc_usage       lr_cl_usage       lr_mo_usage
##  0:1106   Min.   : 0.0000   Min.   :  0.000   Min.   : 0.0000
##  1: 881   1st Qu.: 0.0000   1st Qu.:  0.000   1st Qu.: 0.0000
##           Median : 0.0000   Median :  2.000   Median : 0.0000
##           Mean   : 0.3563   Mean   :  4.265   Mean   : 0.8938
##           3rd Qu.: 0.0000   3rd Qu.:  6.000   3rd Qu.: 0.0000
##           Max.   :20.0000   Max.   :185.000   Max.   :24.0000
##  storage_usage         ps_usage        stock_usage
##  Min.   :     0.0   Min.   :  0.000   Min.   :  0.000
##  1st Qu.:     0.0   1st Qu.:  0.000   1st Qu.:  0.000
##  Median :     0.0   Median :  3.000   Median :  0.000
##  Mean   :   255.3   Mean   :  4.703   Mean   :  1.099
##  3rd Qu.:     1.0   3rd Qu.:  6.000   3rd Qu.:  0.000
##  Max.   :107556.0   Max.   :182.000   Max.   :246.000
```

Summary of Test data

```
##           class      lr_cc_usage       lr_cl_usage      lr_mo_usage
##  OTHER       :394   Min.   : 0.0000   Min.   : 0.000   Min.   : 0.0000
##  PHOTOGRAPHER:269   1st Qu.: 0.0000   1st Qu.: 0.000   1st Qu.: 0.0000
##                     Median : 0.0000   Median : 2.000   Median : 0.0000
##                     Mean   : 0.3213   Mean   : 4.072   Mean   : 0.7587
##                     3rd Qu.: 0.0000   3rd Qu.: 6.000   3rd Qu.: 0.0000
##                     Max.   :22.0000   Max.   :81.000   Max.   :21.0000
##  storage_usage      ps_usage       stock_usage
##  Min.   :    0   Min.   : 0.000   Min.   : 0.0000
##  1st Qu.:    0   1st Qu.: 0.000   1st Qu.: 0.0000
##  Median :    0   Median : 2.000   Median : 0.0000
##  Mean   :  436   Mean   : 4.487   Mean   : 0.7104
##  3rd Qu.:    1   3rd Qu.: 6.000   3rd Qu.: 0.0000
##  Max.   :96273   Max.   :92.000   Max.   :48.0000
```

Training - Linear Model

Training a Logistic Regression model.

```
#logistic regression model
model <- glm(class ~ lr_cc_usage + lr_cl_usage + storage_usage + ps_usage + stock_usage,
             data = train, family = binomial("logit"))
```

Summarize trained model.

```
summary(model)
```

```
##
## Call:
## glm(formula = class ~ lr_cc_usage + lr_cl_usage + storage_usage +
##     ps_usage + stock_usage, family = binomial("logit"), data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0057  -1.0480  -0.9838   1.2696   2.0817
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -4.417e-01  6.305e-02  -7.006 2.44e-12 ***
## lr_cc_usage   -9.714e-02  3.498e-02  -2.777  0.00548 **
## lr_cl_usage    4.391e-02  7.838e-03   5.603 2.11e-08 ***
## storage_usage  1.997e-05  1.708e-05   1.169  0.24234
## ps_usage       1.296e-02  6.368e-03   2.035  0.04183 *
## stock_usage   -3.045e-03  5.586e-03  -0.545  0.58563
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2729.0  on 1986  degrees of freedom
## Residual deviance: 2674.1  on 1981  degrees of freedom
## AIC: 2686.1
##
## Number of Fisher Scoring iterations: 4
```

Prediction (Testing)

Once we have the model built on the training data, let's test in by predicting the output class on the test data.

```
pred <- predict(model, newdata=test, type = 'response')
pred.class <- ifelse(pred > 0.5, 'PHOTOGRAPHER', 'OTHER')
```

Performance

Based on the measure defined in the FPS, we will use classification accuracy as our performance measure.

Confusion Matrix

```
## Confusion Matrix and Statistics
##
##               Reference
## Prediction     OTHER PHOTOGRAPHER
##   OTHER          343          223
##   PHOTOGRAPHER    51           46
##
##                Accuracy : 0.5867
##                  95% CI : (0.5482, 0.6245)
##     No Information Rate : 0.5943
##     P-Value [Acc > NIR] : 0.6689
##
##                   Kappa : 0.0463
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.17100
##             Specificity : 0.87056
##          Pos Pred Value : 0.47423
##          Neg Pred Value : 0.60601
##              Prevalence : 0.40573
##          Detection Rate : 0.06938
##    Detection Prevalence : 0.14630
##       Balanced Accuracy : 0.52078
##
##        'Positive' Class : PHOTOGRAPHER
##
```

Accuracy

- Observed Accuracy : 58.67%
- Desired accuracy : 70%
- Performance is Not Satisfactory.