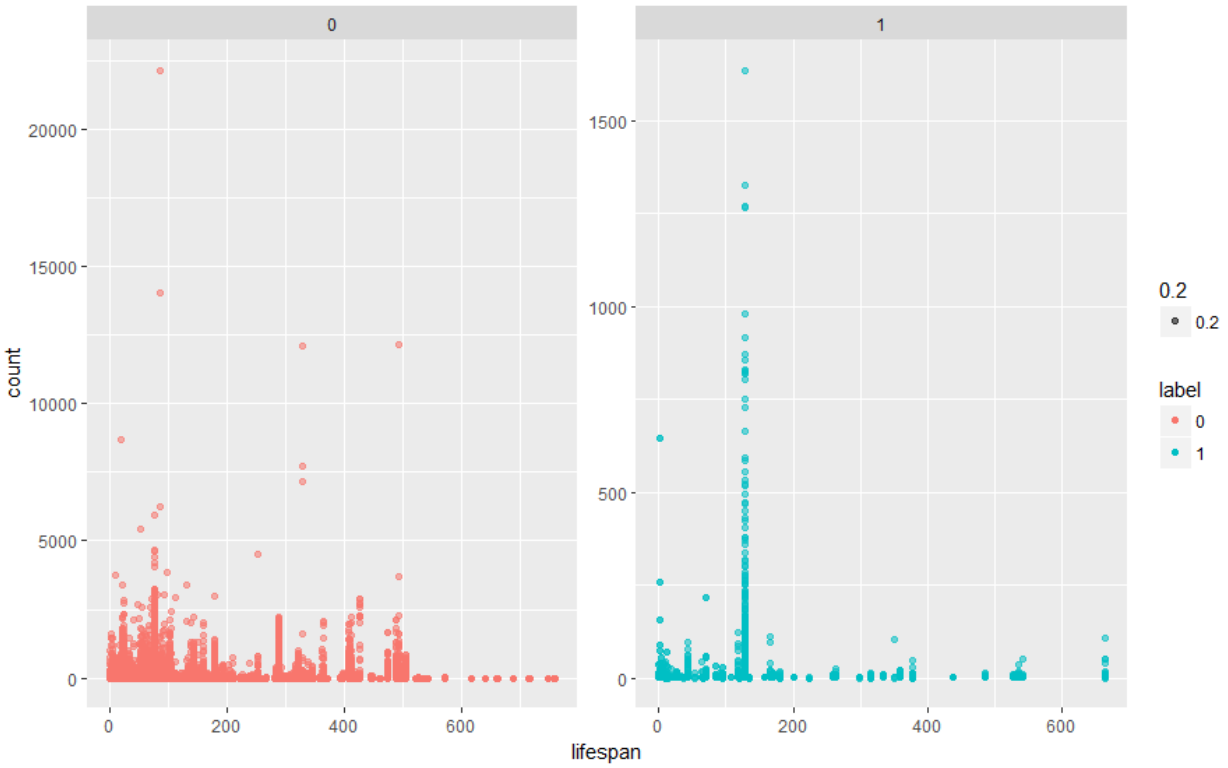
Ethereum blockchain account classification

by Marine Howard

I investigated the Ethereum blockchain dataset for a set of labeled accounts to predict the value of these account labels. Some fields are combined to get the total deposit value, total payout value and total deposit count; also, total participant rank by End date and Life span is calculated. I started with exploratory analysis and plotted each data set. The plot analysis showed that the count of transactions for label 1 accounts is higher than for label zero accounts; this provides a distinguishing feature between these two classes. (Graph 1).

Graph 1

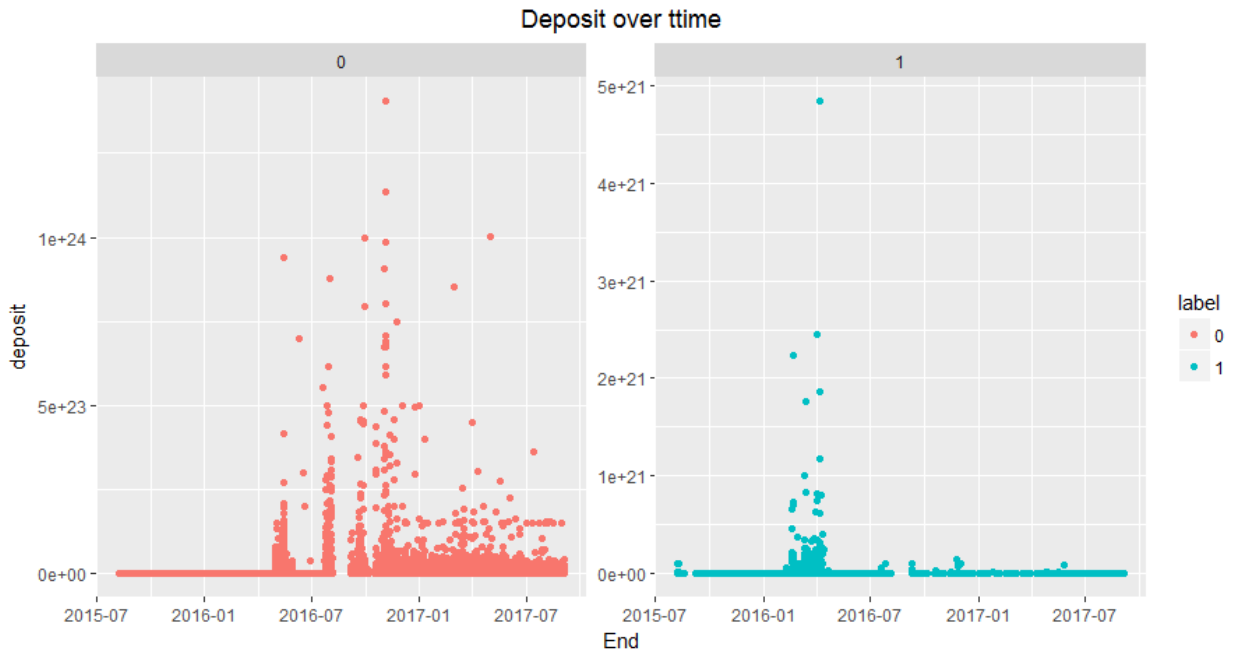


Looking at Graphs 2 and 3, I noticed a correlation between total deposit and payout; the calculated correlation came out relatively high and is significant. The observed correlation is higher for label 1 than label 0 accounts. (0.51 mean for label 1 and 0.28 mean for label 0). Therefore, I added the correlation to the list of features for the model development.

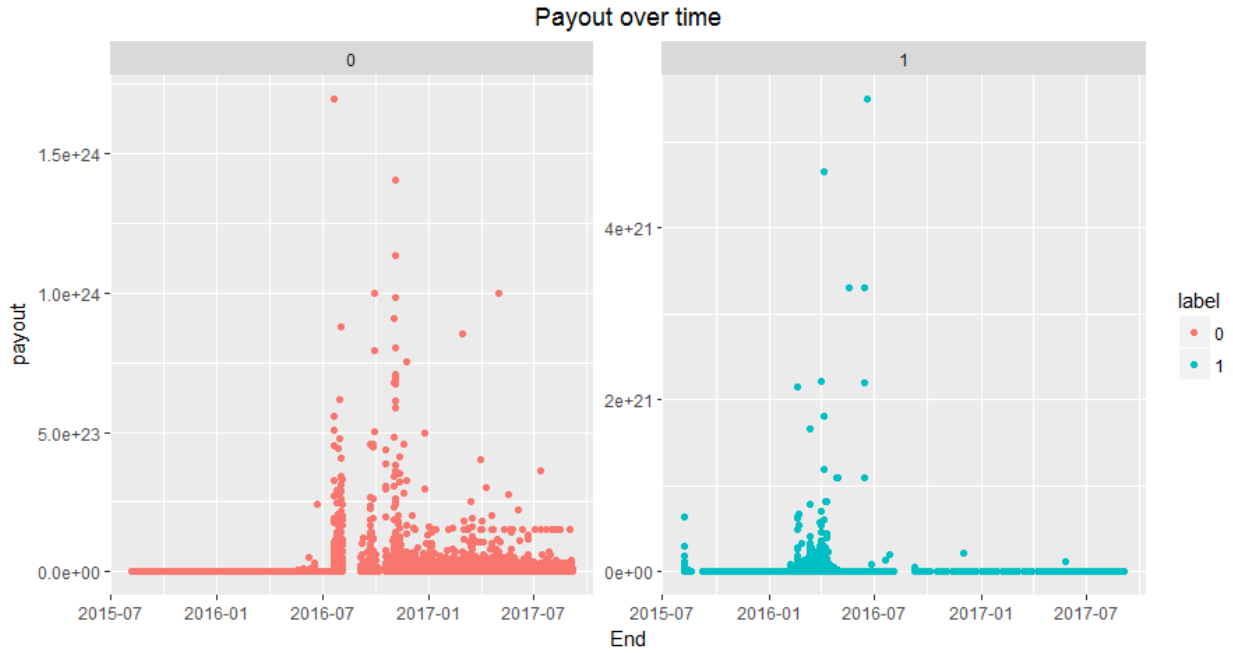
Ethereum blockchain account classification

by Marine Howard

Graph 2



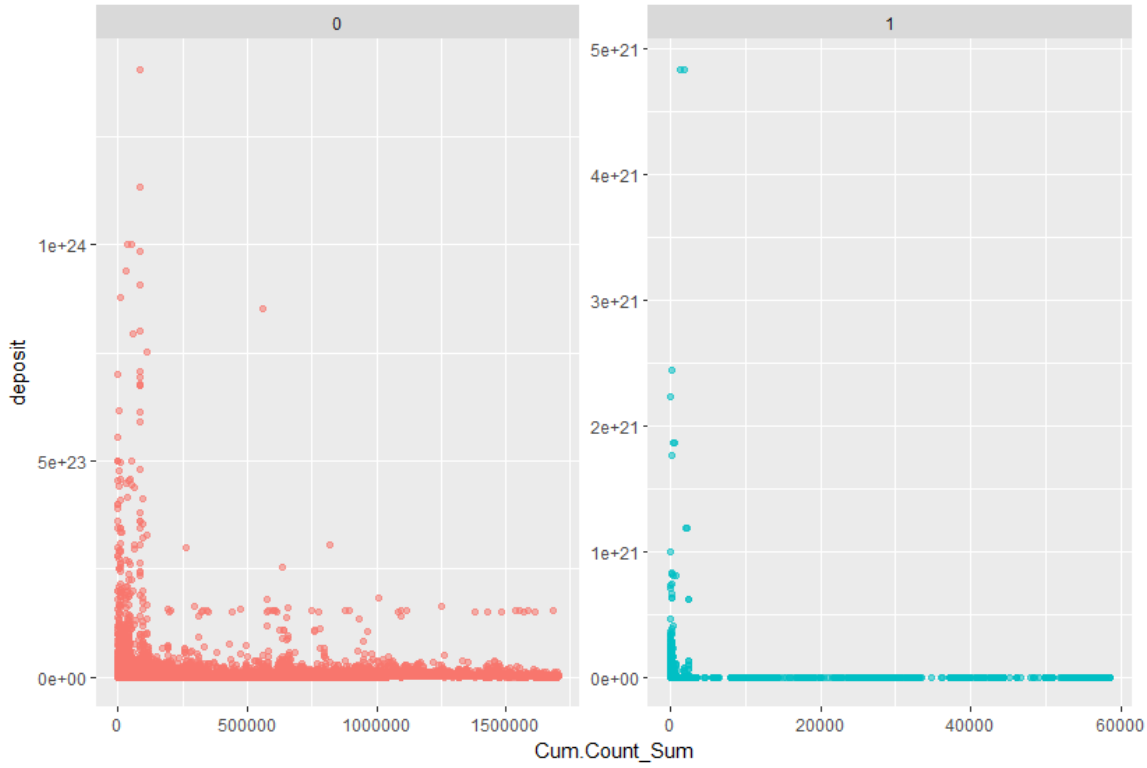
Graph 3



Investigating the data structure, I observed that only 1% of the data is labeled as 1, such a low percentage means that the data is highly unbalanced; however, this 1% is partially due to the fact that label 0 accounts have a higher number of recordings. One way to fix unbalanced data is to create new samplings of the data (under-sample, over-sample, synthesize new data, etc). The sampling techniques for time series data are tricky, and the models I reviewed didn't produce much accuracy; in addition, the

programs to train such models ran for hours at a time. Aggregation was used to solve the unbalanced data problems. In theory, machine learning techniques would have been a better method, but I didn't have the available computational power to try this method.

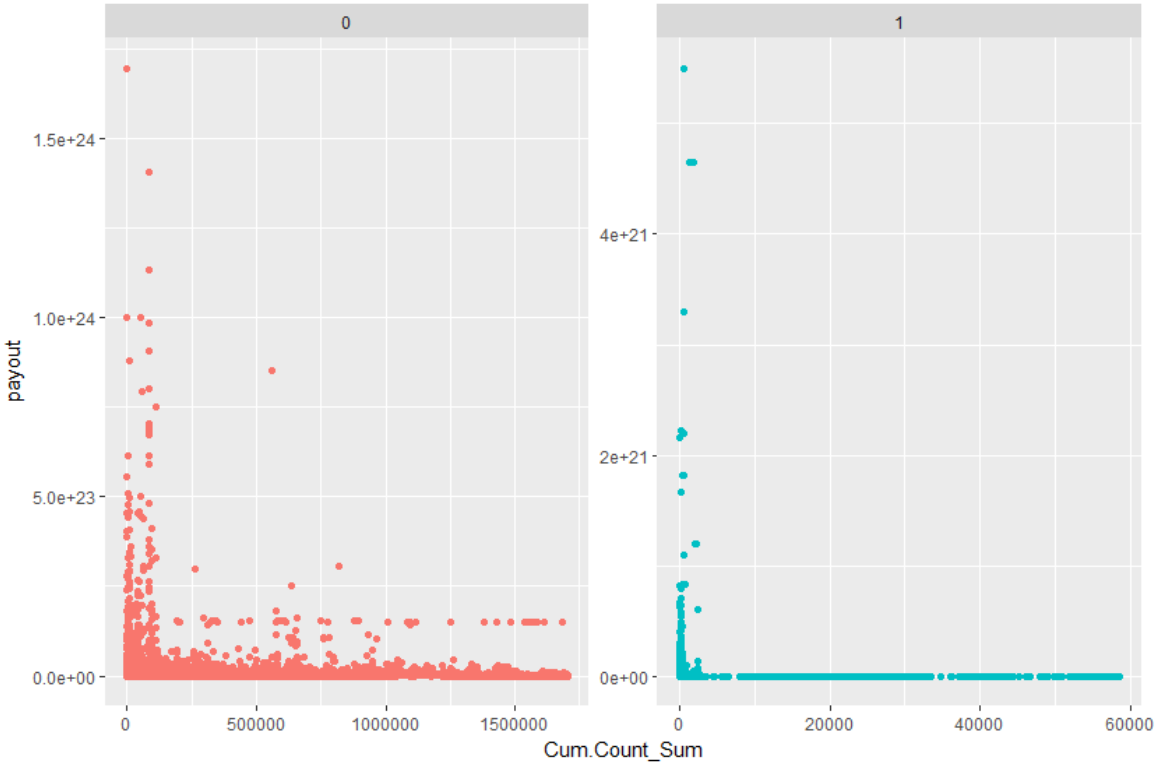
Graph 4



Ethereum blockchain account classification

by Marine Howard

Graph 5



In graph 4 and 5 deposit and payout for label 0 accounts becomes 0 as rank goes up but for label 1 there is a relatively constant non-zero deposit value and payout.

After noticing that the distribution of transactions was different for the labeled accounts, I added entropy of deposit values as an additional aggregated feature to prevent an excessive loss of information.

I created a new feature weighted deposit by finding the product of deposit and rank.

With these potential features I fit a linear model with different combination of features which however doesn't yield high predictive power.

In the end, I explored Gradient Boosted Models (GBM) with different features and tuned a number of parameters. I started with 5 features: deposit, payout, count, life span and entropy, which took me to 88% accuracy. Since deposit and payout are correlated, it made sense to remove one of them. In my most accurate model, I kept payout. In addition, I correlated deposit and payout, which drove down the accuracy in training, but overall test accuracy went up to 92%.

Ethereum blockchain account classification

by Marine Howard

I fit a gbm (gradient boosted model) to a subset of the data (training data) to generate a list describing how each variable reduced the squared error. The following table and Graph 4 shows the importance of the features. Table 1 has Confusion Matrix for the Model.

```
                              var rel.inf
count                   count 24.53
payout                 payout 22.43
average_balance average_balance 17.41
corr                     corr 14.16
lifespan             lifespan 11.70
entrop_dep         entrop_dep  9.78
```

Graph 6

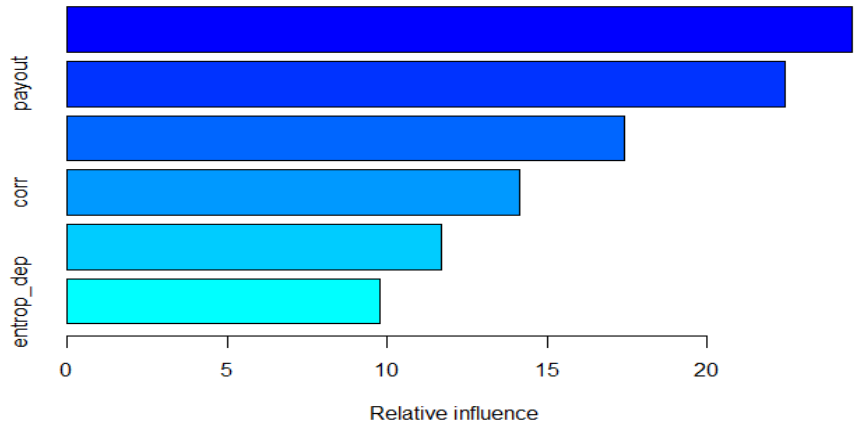


Table 1 Confusion Matrix

Train set	Test Set
Reference Prediction no yes no 295 15 yes 16 74 Accuracy : 0.9225 95% CI : (0.8918, 0.9467) No Information Rate : 0.7775 P-Value [Acc > NIR] : 6.52e-15 Kappa : 0.7769 Mcnemar's Test P-Value : 1 Sensitivity : 0.9486 Specificity : 0.8315 Pos Pred Value : 0.9516 Neg Pred Value : 0.8222 Prevalence : 0.7775 Detection Rate : 0.7375 Detection Prevalence : 0.7750	Reference Prediction no yes no 189 9 yes 10 35 Accuracy : 0.9218 95% CI : (0.8806, 0.9523) No Information Rate : 0.8189 P-Value [Acc > NIR] : 3.932e-06 Kappa : 0.7387 Mcnemar's Test P-Value : 1 Sensitivity : 0.9497 Specificity : 0.7955 Pos Pred Value : 0.9545 Neg Pred Value : 0.7778 Prevalence : 0.8189 Detection Rate : 0.7778 Detection Prevalence : 0.8148

Ethereum blockchain account classification

by Marine Howard

Balanced Accuracy : 0.8900 'Positive' Class : no	Balanced Accuracy : 0.8726 'Positive' Class : no

The results of the Gradient Boosting Model iterations are in the Appendix.

Appendix

Used Generalized Boosted Models

```
Stochastic Gradient Boosting

400 samples
  5 predictor
  2 classes: 'no', 'yes'

Pre-processing: centered (5), scaled (5)
Resampling: Cross-Validated (10 fold, repeated 5 times)
Summary of sample sizes: 360, 360, 360, 360, 360, 360, ...
Resampling results across tuning parameters:

  interaction.depth  n.trees  ROC        Sens       Spec
  1                   50      0.8845520  0.9761290  0.4666667
  1                  100      0.8858781  0.9464516  0.5511111
  1                  150      0.8856631  0.9432258  0.5711111
  2                   50      0.8896057  0.9458065  0.5466667
  2                  100      0.8877419  0.9387097  0.5800000
  2                  150      0.8845161  0.9316129  0.5777778
  3                   50      0.8944444  0.9458065  0.5688889
  3                  100      0.8913620  0.9335484  0.5800000
  3                  150      0.8869892  0.9251613  0.5800000

Tuning parameter 'shrinkage' was held constant at a value of 0.1
Tuning
 parameter 'n.minobsinnode' was held constant at a value of 10
ROC was used to select the optimal model using the largest value.
The final values used for the model were n.trees = 50, interaction.depth = 3,
shrinkage = 0.1 and n.minobsinnode = 10.
```

Confusion Matrix and Statistics

Train	Test
Reference Prediction no yes no 280 12 yes 30 78 Accuracy : 0.895 95% CI : (0.8607, 0.9233) No Information Rate : 0.775 P-Value [Acc > NIR] : 3.554e-10 Kappa : 0.7189 Mcnemar's Test P-Value : 0.008712	Reference Prediction no yes no 175 13 yes 25 30 Accuracy : 0.8436 95% CI : (0.7917, 0.8869) No Information Rate : 0.823 P-Value [Acc > NIR] : 0.22704 Kappa : 0.5161 Mcnemar's Test P-Value : 0.07435

```
            Sensitivity : 0.9032
            Specificity : 0.8667
         Pos Pred Value : 0.9589
         Neg Pred Value : 0.7222
             Prevalence : 0.7750
         Detection Rate : 0.7000
   Detection Prevalence : 0.7300
      Balanced Accuracy : 0.8849

       'Positive' Class : no
```

```
            Sensitivity : 0.8750
            Specificity : 0.6977
         Pos Pred Value : 0.9309
         Neg Pred Value : 0.5455
             Prevalence : 0.8230
         Detection Rate : 0.7202
   Detection Prevalence : 0.7737
      Balanced Accuracy : 0.7863

       'Positive' Class : no
```

```
          Reference
Prediction  no yes
       no  313  26
       yes   6  55

               Accuracy : 0.92
                 95% CI : (0.8889, 0.
9446)
    No Information Rate : 0.7975
    P-Value [Acc > NIR] : 1.296e-11

                  Kappa : 0.7272
 Mcnemar's Test P-Value : 0.0007829

            Sensitivity : 0.9812
            Specificity : 0.6790
         Pos Pred Value : 0.9233
         Neg Pred Value : 0.9016
             Prevalence : 0.7975
         Detection Rate : 0.7825
   Detection Prevalence : 0.8475
      Balanced Accuracy : 0.8301

       'Positive' Class : no
```

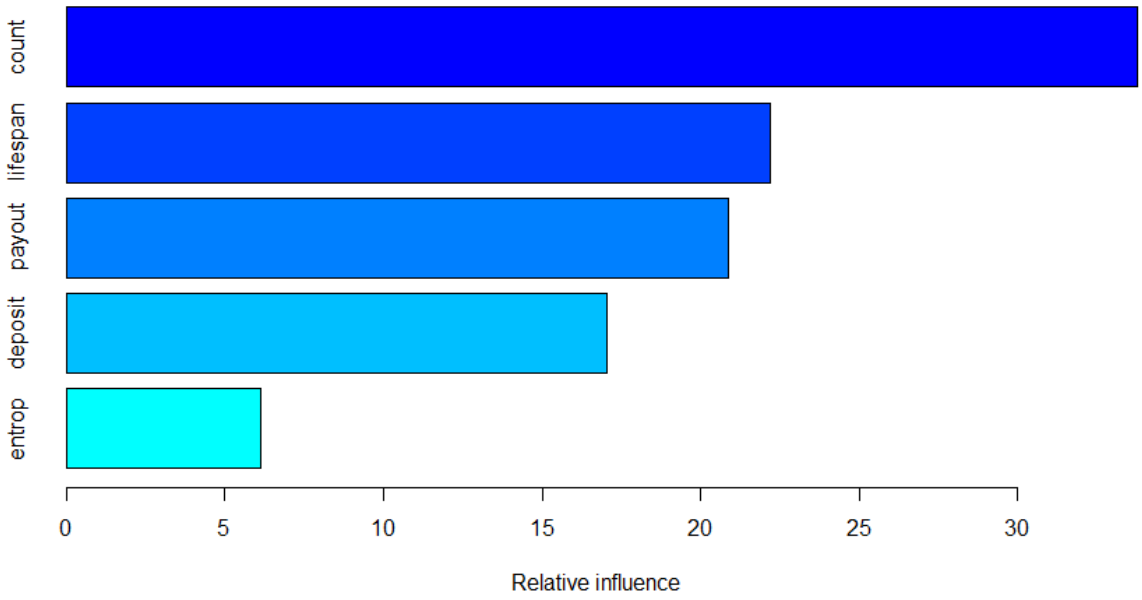
```
  Reference
Prediction  no yes
       no  180  22
       yes  11  30

               Accuracy : 0.8642
                 95% CI : (0.8146, 0.9046)
    No Information Rate : 0.786
    P-Value [Acc > NIR] : 0.001227

                  Kappa : 0.5626
 Mcnemar's Test P-Value : 0.081723

            Sensitivity : 0.9424
            Specificity : 0.5769
         Pos Pred Value : 0.8911
         Neg Pred Value : 0.7317
             Prevalence : 0.7860
         Detection Rate : 0.7407
   Detection Prevalence : 0.8313
      Balanced Accuracy : 0.7597

       'Positive' Class : no


>
```

XGboost

```
          Reference
Prediction   0   1
         0 320  11
         1   4  65

               Accuracy : 0.9625
                 95% CI : (0.9389, 0.
9789)
    No Information Rate : 0.81
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8737
 Mcnemar's Test P-Value : 0.1213

            Sensitivity : 0.9877
            Specificity : 0.8553
         Pos Pred Value : 0.9668
         Neg Pred Value : 0.9420
             Prevalence : 0.8100
         Detection Rate : 0.8000
   Detection Prevalence : 0.8275
      Balanced Accuracy : 0.9215
```

```
Reference
Prediction   0   1
         0 178  24
         1   8  33

               Accuracy : 0.8683
                 95% CI : (0.8192, 0.
9082)
    No Information Rate : 0.7654
    P-Value [Acc > NIR] : 4.232e-05

                  Kappa : 0.5937
 Mcnemar's Test P-Value : 0.00801

            Sensitivity : 0.9570
            Specificity : 0.5789
         Pos Pred Value : 0.8812
         Neg Pred Value : 0.8049
             Prevalence : 0.7654
         Detection Rate : 0.7325
   Detection Prevalence : 0.8313
      Balanced Accuracy : 0.7680
```

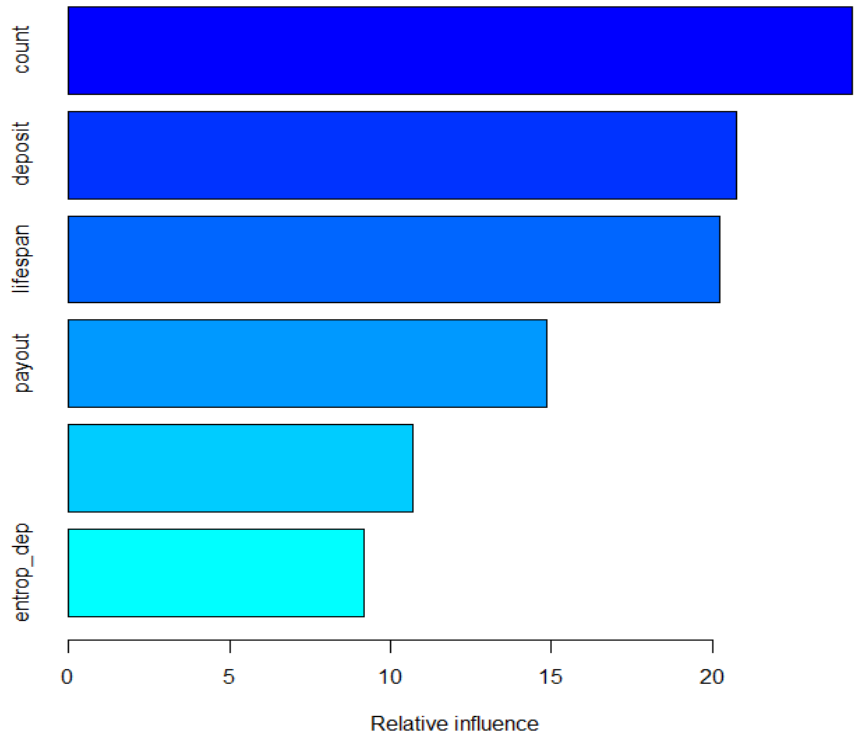
'Positive' Class : 0	'Positive' Class : 0




```
  Feature        Gain      Cover Frequency
1:        lifespan 0.28363029 0.25777238 0.2359551
2:          payout 0.21745732 0.09961687 0.1235955
3:           count 0.16522877 0.28077212 0.2359551
4: average_balance 0.12613386 0.13963677 0.1573034
5:         deposit 0.12425415 0.13711716 0.1198502
6:      entrop_dep 0.08329562 0.08508471 0.1273408
```

Ethereum blockchain account classification

by Marine Howard




```
          Reference
Prediction   0   1
         0 313   4
         1   0  83

               Accuracy : 0.99
                 95% CI : (0.9746, 0.
9973)
    No Information Rate : 0.7825
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9701
 Mcnemar's Test P-Value : 0.1336

            Sensitivity : 1.0000
            Specificity : 0.9540
         Pos Pred Value : 0.9874
         Neg Pred Value : 1.0000
             Prevalence : 0.7825
         Detection Rate : 0.7825
   Detection Prevalence : 0.7925
      Balanced Accuracy : 0.9770

       'Positive' Class : 0
```

```
          Reference
Prediction   0   1
         0 187  17
         1  10  29

               Accuracy : 0.8889
                 95% CI : (0.8425, 0.
9255)
    No Information Rate : 0.8107
    P-Value [Acc > NIR] : 0.0006776

                  Kappa : 0.6156
 Mcnemar's Test P-Value : 0.2482131

            Sensitivity : 0.9492
            Specificity : 0.6304
         Pos Pred Value : 0.9167
         Neg Pred Value : 0.7436
             Prevalence : 0.8107
         Detection Rate : 0.7695
   Detection Prevalence : 0.8395
      Balanced Accuracy : 0.7898

       'Positive' Class : 0
```

The last model is somewhat overfit but it also provides the good Accuracy on the test set. I can probably leave payout out of the model with the same accuracy.

Ethereum blockchain account classification

by Marine Howard

After adding the Pearson correlation between deposit and payout as a feature. The Accuracy on the test set is 92% even though training set accuracy is also only 92%.

Confusion Matrix and Statistics

Train set	Test Set
Reference Prediction no yes no 295 15 yes 16 74	Reference Prediction no yes no 189 9 yes 10 35
Accuracy : 0.9225 95% CI : (0.8918, 0.9467) No Information Rate : 0.7775 P-Value [Acc > NIR] : 6.52e-15	Accuracy : 0.9218 95% CI : (0.8806, 0.9523) No Information Rate : 0.8189 P-Value [Acc > NIR] : 3.932e-06
Kappa : 0.7769 Mcnemar's Test P-Value : 1	Kappa : 0.7387 Mcnemar's Test P-Value : 1
Sensitivity : 0.9486 Specificity : 0.8315 Pos Pred Value : 0.9516 Neg Pred Value : 0.8222 Prevalence : 0.7775 Detection Rate : 0.7375 Detection Prevalence : 0.7750 Balanced Accuracy : 0.8900	Sensitivity : 0.9497 Specificity : 0.7955 Pos Pred Value : 0.9545 Neg Pred Value : 0.7778 Prevalence : 0.8189 Detection Rate : 0.7778 Detection Prevalence : 0.8148 Balanced Accuracy : 0.8726
'Positive' Class : no	'Positive' Class : no

	var	rel.inf
count	count	24.53
payout	payout	22.43
average_balance	average_balance	17.41
corr	corr	14.16
lifespan	lifespan	11.70
entrop_dep	entrop_dep	9.78

Stochastic Gradient Boosting

400 samples
  6 predictor
  2 classes: 'no', 'yes'

Pre-processing: centered (6), scaled (6)
Resampling: Cross-Validated (10 fold, repeated 5 times)
Summary of sample sizes: 360, 361, 360, 360, 359, 361, ...
Addtional sampling using SMOTE prior to pre-processing

Resampling results across tuning parameters:

interaction.depth	n.trees	ROC	Sens	Spec
1	50	0.8633707	0.8938710	0.6719444
1	100	0.8752786	0.9200403	0.6866667
1	150	0.8817524	0.9322177	0.6888889
2	50	0.8879767	0.9206653	0.6972222
2	100	0.8928017	0.9283669	0.7091667

```
  2                   150      0.8998353  0.9271774  0.7136111
  3                    50      0.8917773  0.9361089  0.6930556
  3                   100      0.8985481  0.9412500  0.7047222
  3                   150      0.9031188  0.9368145  0.7019444

Tuning parameter 'shrinkage' was held constant at a value of 0.1
Tuning
 parameter 'n.minobsinnode' was held constant at a value of 10
ROC was used to select the optimal model using the largest value.
The final values used for the model were n.trees = 150, interaction.depth = 3
, shrinkage =
 0.1 and n.minobsinnode = 10.

Area under the curve: 0.9764
```

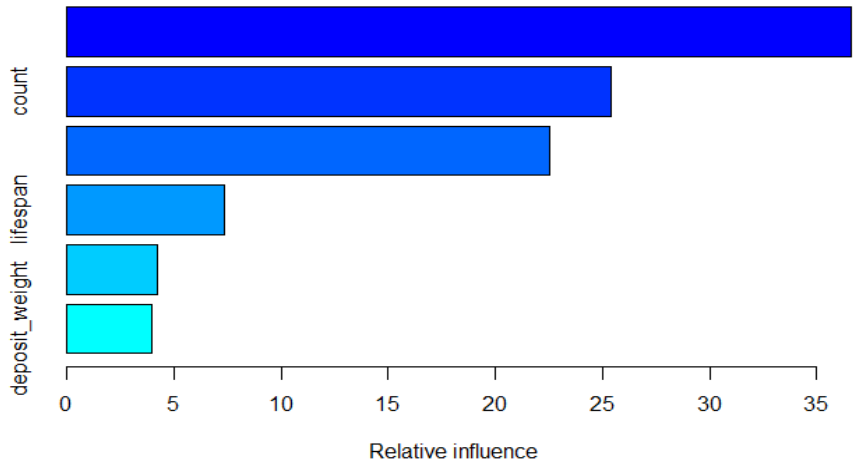
Confusion Matrix and Statistics

```
          Reference
Prediction  no yes
       no  307  17
       yes   6  70

               Accuracy : 0.9425
                 95% CI : (0.915, 0.9
632)
    No Information Rate : 0.7825
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.823
 Mcnemar's Test P-Value : 0.03706

            Sensitivity : 0.9808
            Specificity : 0.8046
         Pos Pred Value : 0.9475
         Neg Pred Value : 0.9211
             Prevalence : 0.7825
         Detection Rate : 0.7675
   Detection Prevalence : 0.8100
      Balanced Accuracy : 0.8927

       'Positive' Class : no
```

```
          Reference
Prediction  no yes
       no  178  16
       yes  19  30

               Accuracy : 0.856
                 95% CI : (0.8054, 0.
8976)
    No Information Rate : 0.8107
    P-Value [Acc > NIR] : 0.03942

                  Kappa : 0.5422
 Mcnemar's Test P-Value : 0.73532

            Sensitivity : 0.9036
            Specificity : 0.6522
         Pos Pred Value : 0.9175
         Neg Pred Value : 0.6122
             Prevalence : 0.8107
         Detection Rate : 0.7325
   Detection Prevalence : 0.7984
      Balanced Accuracy : 0.7779

       'Positive' Class : no
```

Ethereum blockchain account classification

by Marine Howard



	var	rel.inf
payout	payout	36.607246
count	count	25.408465
corr	corr	22.508487
lifespan	lifespan	7.327994
entrop_dep	entrop_dep	4.197963
deposit_weight	deposit_weight	3.949845