Life Insurance Assessment (Kaggle)

Danijel Kopčinović

Description

Prudential, a life insurance company, wants to make quicker and less expensive quotes (offers) for new and existing customers, based on the information they have or receive about the customers and the targeted insurance product. Problem link: https://www.kaggle.com/c/prudential-life-insurance-assessment.

Data

The data used to solve the problem consists of:

- Response variable, indicating customers insurance category (offer, quote).
- 126 different variables, related to different insurance product and customer characteristics.

```
##   [1] "Product_Info_1"      "Product_Info_2"      "Product_Info_3"
##   [4] "Product_Info_4"      "Product_Info_5"      "Product_Info_6"
##   [7] "Product_Info_7"      "Ins_Age"             "Ht"
##  [10] "Wt"                  "BMI"                 "Employment_Info_1"
##  [13] "Employment_Info_2"   "Employment_Info_3"   "Employment_Info_4"
##  [16] "Employment_Info_5"   "Employment_Info_6"   "InsuredInfo_1"
##  [19] "InsuredInfo_2"       "InsuredInfo_3"       "InsuredInfo_4"
##  [22] "InsuredInfo_5"       "InsuredInfo_6"       "InsuredInfo_7"
##  [25] "Insurance_History_1" "Insurance_History_2" "Insurance_History_3"
##  [28] "Insurance_History_4" "Insurance_History_5" "Insurance_History_7"
##  [31] "Insurance_History_8" "Insurance_History_9" "Family_Hist_1"
##  [34] "Family_Hist_2"       "Family_Hist_3"       "Family_Hist_4"
##  [37] "Family_Hist_5"       "Medical_History_1"   "Medical_History_2"
##  [40] "Medical_History_3"   "Medical_History_4"   "Medical_History_5"
##  [43] "Medical_History_6"   "Medical_History_7"   "Medical_History_8"
##  [46] "Medical_History_9"   "Medical_History_10"  "Medical_History_11"
##  [49] "Medical_History_12"  "Medical_History_13"  "Medical_History_14"
##  [52] "Medical_History_15"  "Medical_History_16"  "Medical_History_17"
##  [55] "Medical_History_18"  "Medical_History_19"  "Medical_History_20"
##  [58] "Medical_History_21"  "Medical_History_22"  "Medical_History_23"
##  [61] "Medical_History_24"  "Medical_History_25"  "Medical_History_26"
##  [64] "Medical_History_27"  "Medical_History_28"  "Medical_History_29"
##  [67] "Medical_History_30"  "Medical_History_31"  "Medical_History_32"
##  [70] "Medical_History_33"  "Medical_History_34"  "Medical_History_35"
##  [73] "Medical_History_36"  "Medical_History_37"  "Medical_History_38"
##  [76] "Medical_History_39"  "Medical_History_40"  "Medical_History_41"
##  [79] "Medical_Keyword_1"   "Medical_Keyword_2"   "Medical_Keyword_3"
##  [82] "Medical_Keyword_4"   "Medical_Keyword_5"   "Medical_Keyword_6"
##  [85] "Medical_Keyword_7"   "Medical_Keyword_8"   "Medical_Keyword_9"
##  [88] "Medical_Keyword_10"  "Medical_Keyword_11"  "Medical_Keyword_12"
##  [91] "Medical_Keyword_13"  "Medical_Keyword_14"  "Medical_Keyword_15"
##  [94] "Medical_Keyword_16"  "Medical_Keyword_17"  "Medical_Keyword_18"
##  [97] "Medical_Keyword_19"  "Medical_Keyword_20"  "Medical_Keyword_21"
## [100] "Medical_Keyword_22"  "Medical_Keyword_23"  "Medical_Keyword_24"
## [103] "Medical_Keyword_25"  "Medical_Keyword_26"  "Medical_Keyword_27"
```

```
## [106] "Medical_Keyword_28"  "Medical_Keyword_29"  "Medical_Keyword_30"
## [109] "Medical_Keyword_31"  "Medical_Keyword_32"  "Medical_Keyword_33"
## [112] "Medical_Keyword_34"  "Medical_Keyword_35"  "Medical_Keyword_36"
## [115] "Medical_Keyword_37"  "Medical_Keyword_38"  "Medical_Keyword_39"
## [118] "Medical_Keyword_40"  "Medical_Keyword_41"  "Medical_Keyword_42"
## [121] "Medical_Keyword_43"  "Medical_Keyword_44"  "Medical_Keyword_45"
## [124] "Medical_Keyword_46"  "Medical_Keyword_47"  "Medical_Keyword_48"
```

Analysis

By inspecting the data, we see that a lot of entries (observations/variables) are undefined (NA).

```
##  Product_Info_3  Employment_Info_1 Employment_Info_2 Employment_Info_4
##  Min.   : 1.00   Min.   :0.00000   Min.   : 1.000    Min.   :0.000
##  1st Qu.:26.00   1st Qu.:0.03500   1st Qu.: 9.000    1st Qu.:0.000
##  Median :26.00   Median :0.06000   Median : 9.000    Median :0.000
##  Mean   :24.42   Mean   :0.07758   Mean   : 8.642    Mean   :0.006
##  3rd Qu.:26.00   3rd Qu.:0.10000   3rd Qu.: 9.000    3rd Qu.:0.000
##  Max.   :38.00   Max.   :1.00000   Max.   :38.000    Max.   :1.000
##                  NA's   :19                          NA's   :6779
##  Employment_Info_6 InsuredInfo_3    Insurance_History_5 Family_Hist_2
##  Min.   :0.000     Min.   : 1.000   Min.   :0.000       Min.   :0.000
##  1st Qu.:0.060     1st Qu.: 3.000   1st Qu.:0.000       1st Qu.:0.362
##  Median :0.250     Median : 6.000   Median :0.001       Median :0.464
##  Mean   :0.361     Mean   : 5.836   Mean   :0.002       Mean   :0.475
##  3rd Qu.:0.550     3rd Qu.: 8.000   3rd Qu.:0.002       3rd Qu.:0.580
##  Max.   :1.000     Max.   :11.000   Max.   :1.000       Max.   :1.000
##  NA's   :10854                      NA's   :25396       NA's   :28656
##  Family_Hist_3   Family_Hist_4   Family_Hist_5   Medical_History_1
##  Min.   :0.00    Min.   :0.000   Min.   :0.00    Min.   :  0.000
##  1st Qu.:0.40    1st Qu.:0.324   1st Qu.:0.40    1st Qu.:  2.000
##  Median :0.52    Median :0.423   Median :0.51    Median :  4.000
##  Mean   :0.50    Mean   :0.445   Mean   :0.48    Mean   :  7.962
##  3rd Qu.:0.60    3rd Qu.:0.563   3rd Qu.:0.58    3rd Qu.:  9.000
##  Max.   :1.00    Max.   :0.944   Max.   :1.00    Max.   :240.000
##  NA's   :34241   NA's   :19184   NA's   :41811   NA's   :8889
##  Medical_History_2 Medical_History_10 Medical_History_15
##  Min.   :  1       Min.   :  0.0      Min.   :  0.0
##  1st Qu.:112       1st Qu.:  8.0      1st Qu.: 17.0
##  Median :162       Median :229.0      Median :117.0
##  Mean   :254       Mean   :141.1      Mean   :123.8
##  3rd Qu.:418       3rd Qu.:240.0      3rd Qu.:240.0
##  Max.   :648       Max.   :240.0      Max.   :240.0
##                    NA's   :58824      NA's   :44596
##  Medical_History_24 Medical_History_32
##  Min.   :  0.00     Min.   :  0.00
##  1st Qu.:  1.00     1st Qu.:  0.00
##  Median :  8.00     Median :  0.00
##  Mean   : 50.64     Mean   : 11.97
##  3rd Qu.: 64.00     3rd Qu.:  2.00
##  Max.   :240.00     Max.   :240.00
##  NA's   :55580      NA's   :58274
```

The undefined values can either be removed from the dataset or somehow estimated. Since this data has a

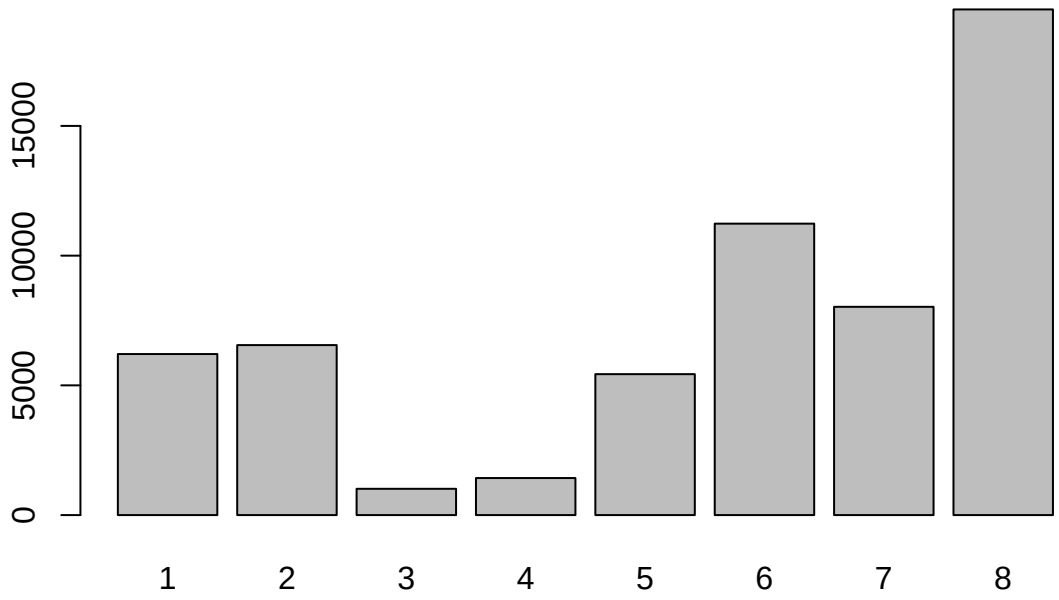
lot of undefined values, we cannot remove it, therefore we make simple assessments:

- For a categorical variable, we set the assessed value to the most frequent value.
- For a numerical value, we set the assessed value to 0.

We can also see that some variables are way out of range compared to other variables, so we "stratify" all the variables with a $log1p(x) = log(x + 1)$ to make the data more adjusted for the modelling.

```
##  Product_Info_3   Employment_Info_1 Employment_Info_2 Employment_Info_4
##  Min.   :0.6931   Min.   :0.00000   Min.   :0.6931    Min.   :0.000000
##  1st Qu.:3.2958   1st Qu.:0.03440   1st Qu.:2.3026    1st Qu.:0.000000
##  Median :3.2958   Median :0.05827   Median :2.3026    Median :0.000000
##  Mean   :3.2036   Mean   :0.07226   Mean   :2.1141    Mean   :0.005181
##  3rd Qu.:3.2958   3rd Qu.:0.09531   3rd Qu.:2.3026    3rd Qu.:0.000000
##  Max.   :3.6636   Max.   :0.69315   Max.   :3.6636    Max.   :0.693147
##  Employment_Info_6 InsuredInfo_3    Insurance_History_5 Family_Hist_2
##  Min.   :0.0000    Min.   :0.6931   Min.   :0.0000000   Min.   :0.0000
##  1st Qu.:0.0000    1st Qu.:1.3863   1st Qu.:0.0000000   1st Qu.:0.0000
##  Median :0.1398    Median :1.9459   Median :0.0001667   Median :0.1967
##  Mean   :0.2273    Mean   :1.8351   Mean   :0.0009804   Mean   :0.1981
##  3rd Qu.:0.4055    3rd Qu.:2.1972   3rd Qu.:0.0012659   3rd Qu.:0.3810
##  Max.   :0.6931    Max.   :2.4849   Max.   :0.6931472   Max.   :0.6931
##  Family_Hist_3    Family_Hist_4    Family_Hist_5    Medical_History_1
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.6931
##  Median :0.0000   Median :0.2806   Median :0.0000   Median :1.3863
##  Mean   :0.1691   Mean   :0.2448   Mean   :0.1157   Mean   :1.4159
##  3rd Qu.:0.3923   3rd Qu.:0.4008   3rd Qu.:0.2988   3rd Qu.:2.1972
##  Max.   :0.6931   Max.   :0.6646   Max.   :0.6931   Max.   :5.4848
##  Medical_History_2 Medical_History_10 Medical_History_15
##  Min.   :0.6931    Min.   :0.00000    Min.   :0.000
##  1st Qu.:4.7274    1st Qu.:0.00000    1st Qu.:0.000
##  Median :5.0938    Median :0.00000    Median :0.000
##  Mean   :5.1721    Mean   :0.03661    Mean   :0.969
##  3rd Qu.:6.0379    3rd Qu.:0.00000    3rd Qu.:0.000
##  Max.   :6.4754    Max.   :5.48480    Max.   :5.485
##  Medical_History_24 Medical_History_32
##  Min.   :0.0000     Min.   :0.00000
##  1st Qu.:0.0000     1st Qu.:0.00000
##  Median :0.0000     Median :0.00000
##  Mean   :0.1556     Mean   :0.01516
##  3rd Qu.:0.0000     3rd Qu.:0.00000
##  Max.   :5.4848     Max.   :5.48480
```

When we have cleared and prepared the data, we can build a model that we will use for later predictions. The model we build is based on the XG Boost algorithm for different machine learning purposes: regression, binary and multiclass classification.



Since our target variable ("Response"), that we will predict, has 8 different values, we will build a model for multiclass classification. This is one approach that doesn't have to bring the best results so in practice we always try many approaches to find the best performing one. Often it is a combination of a few models.

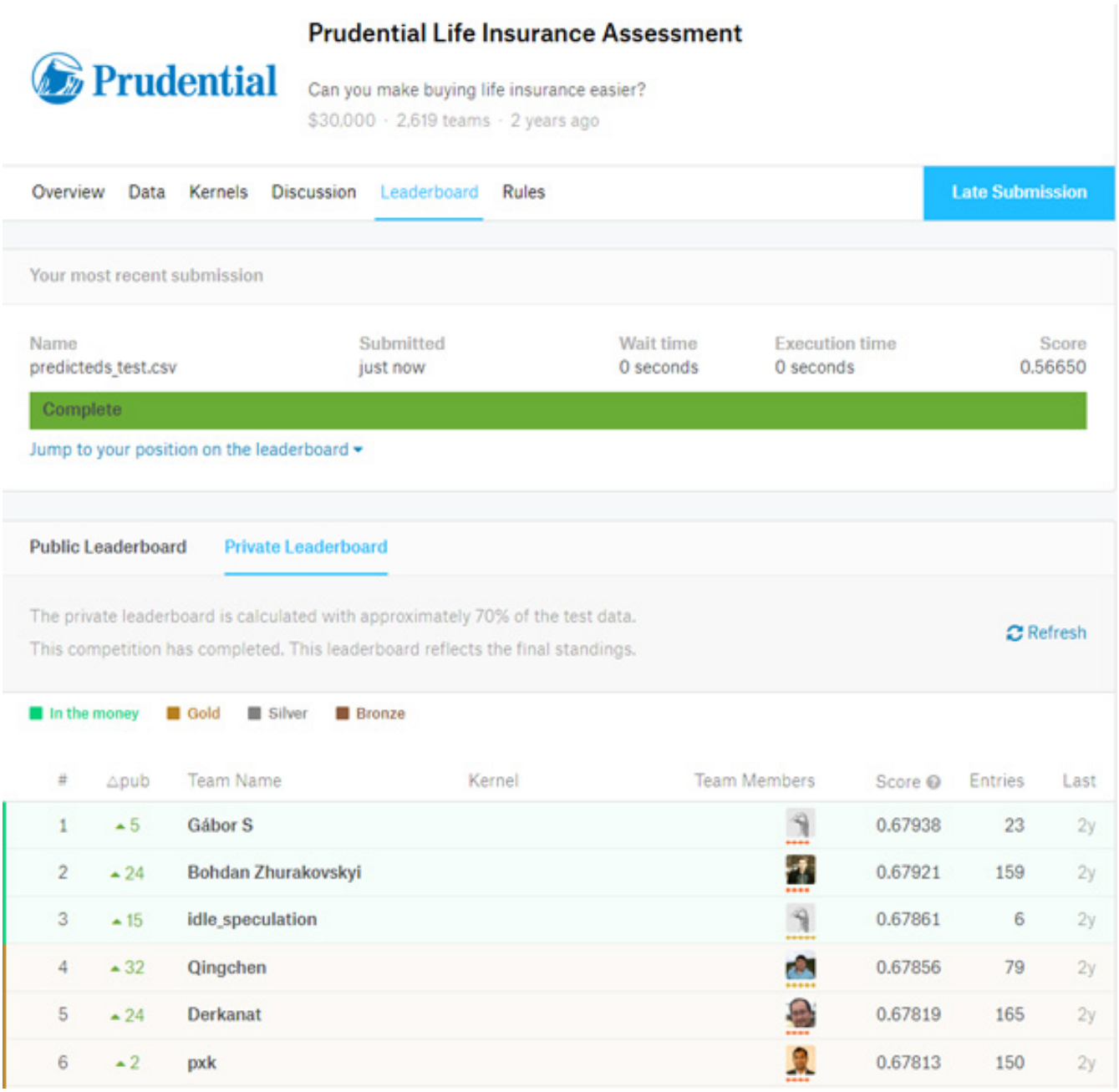
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    1    2    3    4    5    6    7    8
##          1  305  147    9   11   40   78   30   12
##          2  183  343   12    1  106   78   17   14
##          3   18   21   60    5   12    5    0    0
##          4   24   21   35  148    0   17    1    4
##          5   80  126   17    0  576   75    8    4
##          6  228  269   53   82  206 1214  246  120
##          7  132  120    5    5   70  306  693  166
##          8  257  257    4   34   98  506  589 3589
##
## Overall Statistics
##
##                Accuracy : 0.5826
##                  95% CI : (0.5737, 0.5915)
##     No Information Rate : 0.3287
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.4626
##  Mcnemar's Test P-Value : NA
##
```

```
## Statistics by Class:
##
##                      Class: 1 Class: 2 Class: 3 Class: 4 Class: 5 Class: 6
## Sensitivity           0.24857  0.26304 0.307692  0.51748  0.51986   0.5327
## Specificity           0.96934  0.96118 0.994785  0.99121  0.97125   0.8748
## Pos Pred Value        0.48259  0.45491 0.495868  0.59200  0.65011   0.5021
## Neg Pred Value        0.91812  0.91372 0.988531  0.98815  0.95166   0.8876
## Prevalence            0.10318  0.10965 0.016398  0.02405  0.09317   0.1916
## Detection Rate        0.02565  0.02884 0.005045  0.01245  0.04844   0.1021
## Detection Prevalence  0.05314  0.06340 0.010175  0.02102  0.07450   0.2033
## Balanced Accuracy     0.60896  0.61211 0.651239  0.75435  0.74555   0.7037
##                      Class: 7 Class: 8
## Sensitivity           0.43750   0.9181
## Specificity           0.92200   0.7814
## Pos Pred Value        0.46293   0.6729
## Neg Pred Value        0.91429   0.9512
## Prevalence            0.13320   0.3287
## Detection Rate        0.05827   0.3018
## Detection Prevalence  0.12588   0.4485
## Balanced Accuracy     0.67975   0.8498
```

We see that our model provides a solid base for correctly predicting the response variable class with overall accuracy higher than 58%.

After building the model, we make predictions for the test data given in the task. This data was originally used for scoring the participants in the contest.

Our overall score is 0.56650 which is a solid result taking into consideration that the best results are around 0.69.



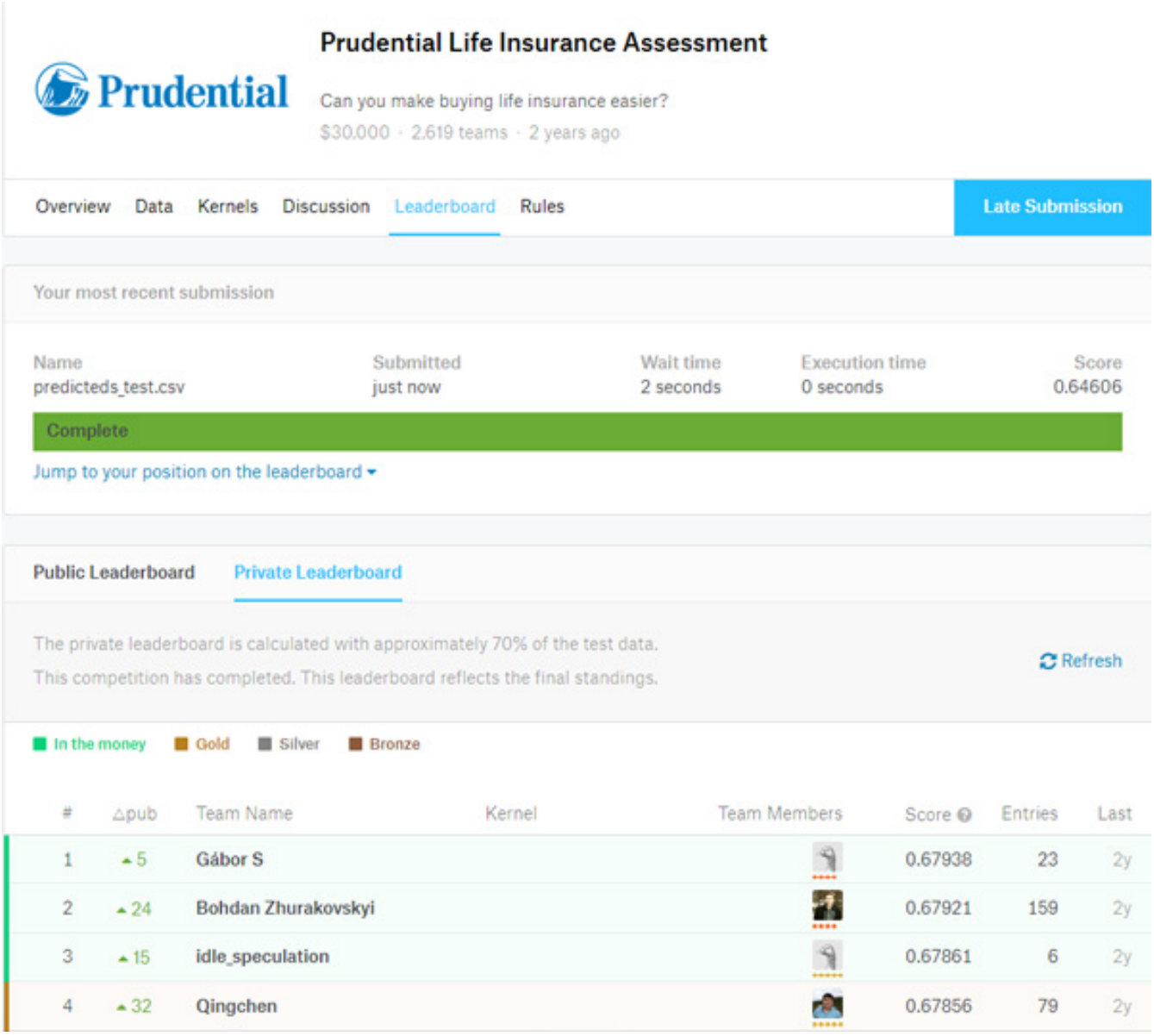
Improvements

The above result can certainly be improved with more insight into the data and certain well known tactics:

- Removing the variables that seem uncorrelated to the response variable.
- Combining and refactoring the variables to get new variables.
- Applying different machine learning and deep learning algorithms.
- Applying different methods to assess the missing data.
- Combining results from different approaches into one result.

And many others... The field is developing on almost daily basis and new approaches and algorithms are continuously giving new options and better results than previous ones.

With a different approach to modelling and predicting the response variable, we got a much better result.



Contact and Info

Danijel Kopčinović, IT Market

Mail: danijel.kopcinovic@itmarket.hr

Tel: +385956472127