SUPERVISED LEARNING IN BILLIONAIRES' AGE PREDICTION

Jiawei Xu, Iván López Anca

Statistical Learning.
Bachelor in Data Science and Engineering.

Access to our raw dataset. (https://www.kaggle.com/datasets/nelgiriyewithana/billionaires-statistics-dataset)

Iván López Anca,
Jiawei Xu.

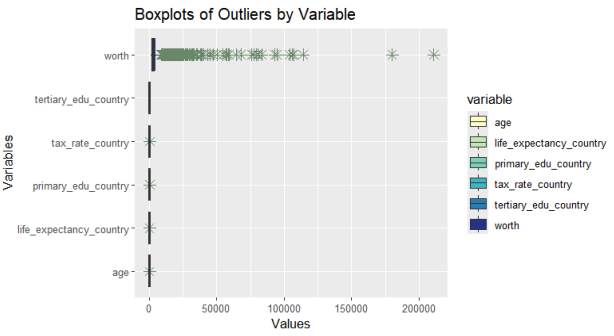
MOTIVATION

This project aims to predict whether a billionaire is an Adult or a Senior (over 65 years old). The goal is to build a model that predicts these age groups while identifying key factors like self-made status, industry, and economic indicators. This research holds real-world value for economists and investors seeking to understand the life paths and societal contributions of influential billionaires.

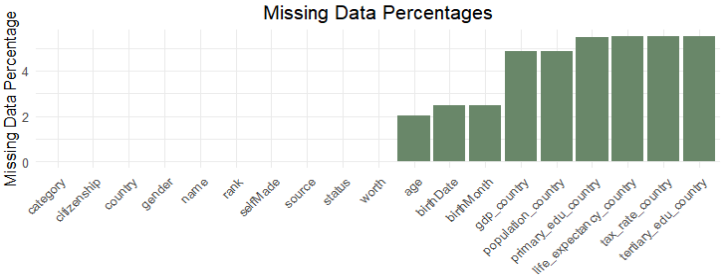
DATA PREPROCESSING & VISUALIZATION

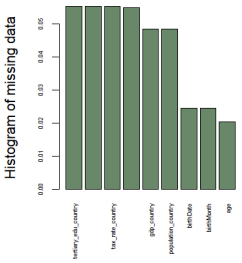
To begin with, we eliminated variables that appeared to be irrelevant to our analysis or redundant due to other available variables. For instance, we removed variables like birthDay and the latitude and longitude of the country. Additionally, we renamed some variables to simplify their long names. We then proceeded to ensure the quality of the data.

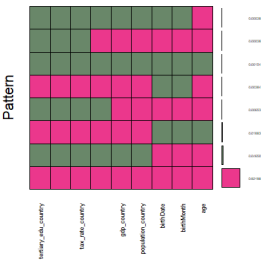
First, we checked for duplicates, confirming that none were present. Next, we addressed outliers using the 3-sigma method and boxplots.



Then, we handled missing values, starting by identifying hidden NAs represented as "". We visualized them with ggplot2, and used the VIM library to see patterns in the missing data.



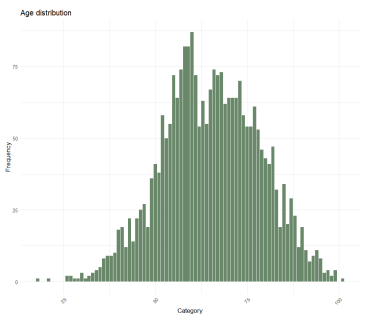


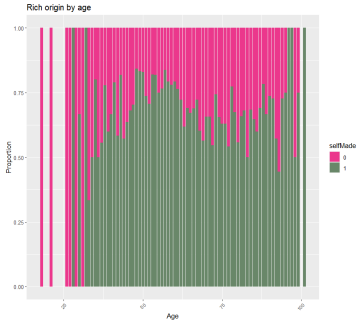


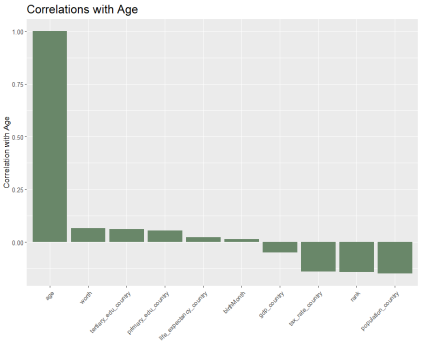
We then addressed these missing values using the MICE library, performing multivariate imputation with a random forest.

To conclude, we carried out some feature engineering, converting categorical variables into factors and scaling certain data for improved visualization outcomes.

For visualization, we plotted some interesting plots about the variable age, which is going to be our target variable to understand the relationship with other variables.







Iván López Anca,
Jiawei Xu.

CLASSIFICATION - PREDICTION

Regression analysis

First we did both a simple regression model with the most correlated variable with our target, age, which is worth. As there wasn't a linear relationship, we had to use the exponential and then we predicted it. Knowing that the squared correlation between the actual age and the predictions is 0.0039

```
From the model we can see that:
-Residual standard error: 0.2102 (2536 degrees of freedom).
-Multiple R-squared: 0.00369, Adjusted R-squared: 0.003297.
-F-statistic: 9.393 on 1 and 2536 DF, p-value = 0.002201.
```

Then we did a multiple regression model with all the numerical and few-level factor variables:

```
Residual standard error: 0.1984; R^2 = 0.1235, Adjusted R^2 = 0.1119.
F-statistic: 10.69 on 33 and 2504 DF, p-value < 2.2e-16 (model significant).
```

and we predicted it, receiving that the model explains about 12.35% of the variance in the log-transformed age variable.

Classification analysis

First we splitted the data into 70% for the train test and 30% for the test set.

What we have done for every model is create a train and predict for the age in the test set

Then we trained a Naive Bayes model whose accuracy was of the 55% and mispredicted many values for adult and seniors. Then we plotted the probability of the model and the ROC curve receiving a value of 0.613 as Area Under the Curve.

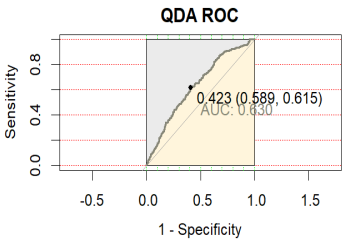
This values seem to us low, so we decided to train many other models with other techniques.

So we proceeded with Linear Discriminant Analysis (LDA), creating the model and receiving many significant predictors as rank or worth, and then we predicting the model, having as results:

```
          Reference
Prediction Adult Senior
    Adult    209    138
    Senior   178    236
> accuracy_lda
 Accuracy
0.5847569
```

We also plot the Roc Curve, receiving as AUC 0.626

We were again shocked by this low values, but we kept training:



Now we trained a Quadratic Discriminant Analysis (QDA) having as before significant differences in features like worth and rank., making before the prediction and having as before so many mispredicted values for seniors and adults and an accuracy of 59.39%, the best for now

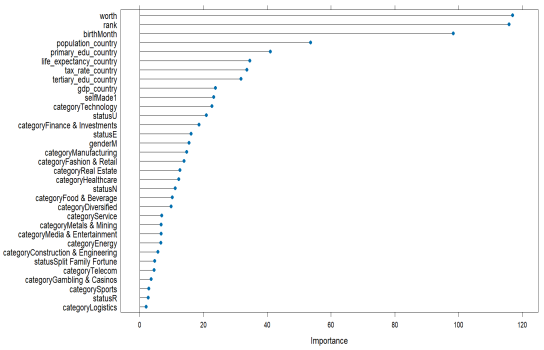
After this we plotted the ROC curve and as the image shows, the AUC was also the best for now, with a value of 0.630

We realized that this was something happening with the performance, but we had already set balanced classes and eliminated the noise, so we kept trying with other training models, in fact with machine learning techniques with their hyperparameters selection.

The first one we used was decision trees, with randomly chosen hyperparameters, we plotted it and we received an accuracy of 61.62%, still low but better than before. So we tried to construct other trees with cross validation for hyperparameters selection, however even the

accuracy was higher, it was that hard to deal with the tree, that we chose to predict with the one we selected before

Now we proceeded with a random forest. Having the highest accuracy for the moment as well as the best predictions for the moment. We tried to set other hyperparameters with cross validation to see if we could even improve it more. However, we didn`t, so we kept the first random forest as our best predictor.



We also checked for variable importance receiving almost the same results as before:

With Simple Gradient Boosting we had to retransform the values to 0 and 1, and same as before, the model predictions were very similar to the previous ones

However XGBoost was very tricky to build, as it was like the optimization of the hyperparameters of before with cross validation with five folds, having received as final hyperparameters:

```
Aggregating results
Selecting tuning parameters
Fitting nrounds = 500, max_depth = 2, eta = 0.01, gamma = 1, colsample_bytree = 0.3, min_child
_weight = 1, subsample = 1 on full training set
```

We also plotted the importance of the variables, having very different results as before. As now our most important feature is the population followed by the educations

The prediction was of approximately 62% of accuracy so we are going to still trust the random forest, although this model tends to be very accurate

We finally confirmed our theory that there was something wrong with the data, however we don't think we have the techniques to correct these mistakes. However, we tried a little bit of subsampling with rose but nothing special happened.

```
          Reference
Prediction Adult Senior
    Adult    215    111
    Senior   172    263
> accuracy = confusionMat
> accuracy
 Accuracy
0.6281209
>
```

Almost finished, we dealt with logistic regression, having as significant variables the ones in XGBoost, which surprised us a lot. We also computed the prediction, having the low value of 58% for accuracy as well . So we tried penalized logistic regression to see if there was any improvement as it is more accurate for this type of data considering all variables, which gave us even worse performance. We plot the ROC and an area of 0.626 was captured under the curve

We came to the conclusion that doing the penalized logistic regression with cross validation would lead into better results. However, as we had so many problems with the RStudio updates we decided not to capture and predict this models.

Finally, we ensembled a lot of this models to see if their combination could lead in the best predictor for this dataset. Nevertheless, it was not the case. Although it seem to be good:

Having finished we compared the ROCS stating, ordering them from the best to the worst in a similar way we are going to order the predictors in 4.

```
  auc.qda   auc.glm   auc.lda   auc.gnb
0.6296999 0.6257583 0.6255165 0.6125378
```

```
          Reference
Prediction Adult Senior
    Adult    210    119
    Senior   177    255
> ensemble_accuracy = con
> ensemble_accuracy
 Accuracy
0.6110381
```

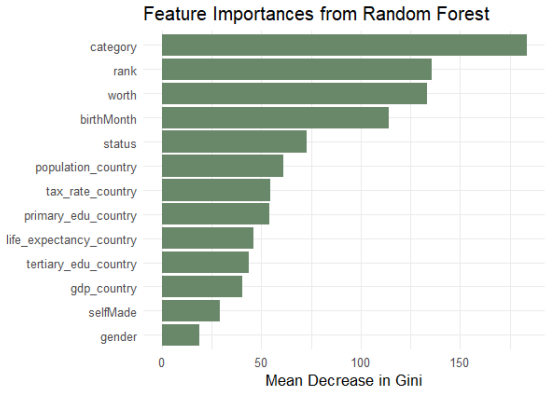
Then we made another predictor selecting the optimal threshold for classifying "Senior" vs. "Adult" based on logistic regression probabilities. By testing thresholds from 0.05 to 0.5, We calculate the profit for each threshold using confusion matrices. The threshold with a balance between the highest median profit and lowest variability (narrowest IQR) is chosen.Being the chosen option 0.5. Now, using this threshold in a logistic regression model to classify the testing set, we can see that the result shows a slight bias towards classifying as "Adult", having a positive profit margin.

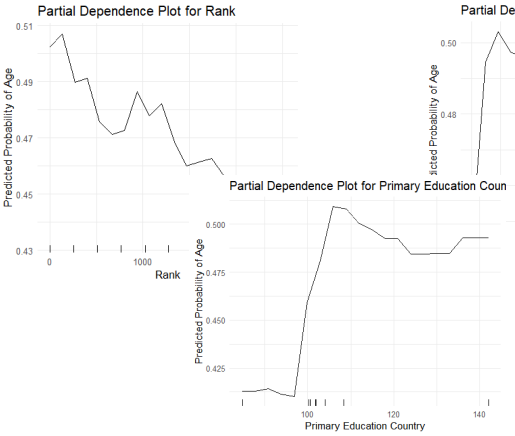
CLASSIFICATION - INTERPRETATION

Having age divided in two groups, Adults and Seniors, we started our analysis.

Random Forest

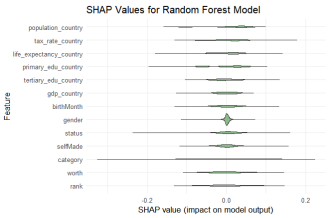
To begin with, we made a Feature Importance Plot and Partial Dependence Plots with a Random Forest. With this ones, we could conclude that the features that contribute the more are for example rank or worth, and using these for the Partial Dependence Plots, we can see how the higher the rank is, the higher the age will be, but that the worth doesn't really affect, which could be because the billionaires have similar worth values. Then, we saw that the Primary Education of the country does have an impact on the probability of being a senior, which could be related with the economic development and wealth accumulation that this education allowed.



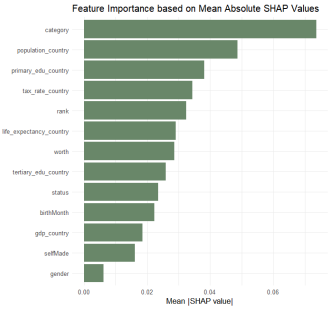


SHAP Values

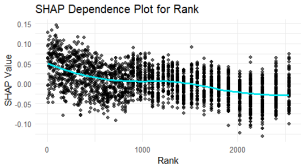
Then, we computed SHAP values, which show how each feature contributes to the model output. Wider distributions indicate more variability in feature impact, and the center around zero, such as in gender, suggest a neutral effect.

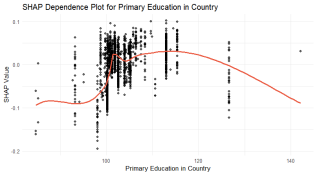


Then, we computed the Feature Importance Plot for this SHAP values, where we can see that rank and worth still have great impact, but category is the one that has the greatest impact.



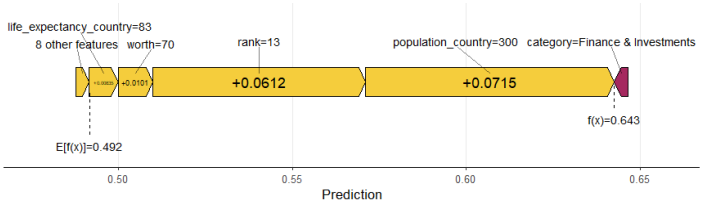
Finally, we calculated some Dependence plots, which show that the SHAP values decrease as the rank value increases, and that the relation with population is non-linear, getting the same results as with the RF.





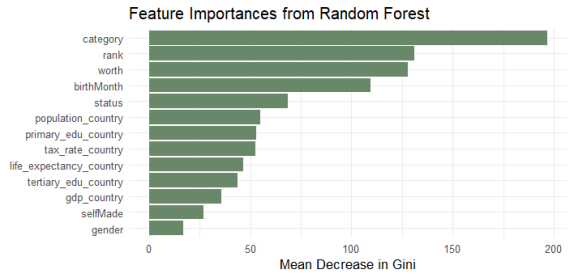
LightGBM

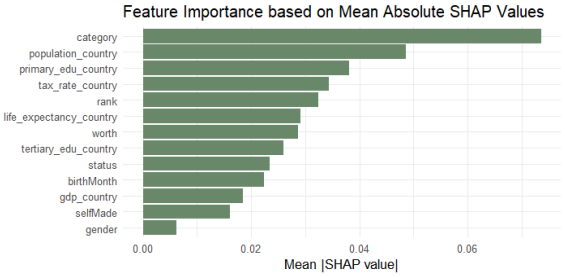
We trained a LightGBM model and created a new observation for a billionaire to predict his age, interpreting the result and understanding the impact of each variable on the prediction. After analyzing the model with Shapley values, we understood that features such as rank and worth have a positive impact on the billionaire being classified as "Senior," while features such as the education levels in the country push the prediction toward "Adult." Then we visualized the contribution of each feature using a waterfall chart and saw how each feature influenced the final prediction.



FEATURE SELECTION

So given the 2 graphs from the previous analysis:





We can see that the most important features are category, rank, worth, population_country, the primary education and the rate. Also, features such as gender, selfMade of birthMonth are irrelevant. Therefore, the features we should select are overall the ones mentioned at the beginning.

COMPARISON OF PREDICTOR SETS

In our work, we were continuously comparing the performances of our models, seeing that although the accuracies were close, there were differences. As we expected, for such complex data, more complex methods like Random Forest and XGBoost showed superior performance in terms of accuracy and AUC, with Random Forest standing out at 0.63 accuracy. Simpler models like Naive Bayes and Penalized Logistic Regression performed poorly, with accuracies of around 0.55, suggesting they struggle with more complex data.

However, the accuracy was lower than expected due to limitations in resources and time, indicating that better resources could improve the model's performance.

naive

Prediction	Reference Adult	Reference Senior
Adult	228	177
Senior	159	197

lda

Prediction	Reference Adult	Reference Senior
Adult	209	138
Senior	178	236

qda

Prediction	Reference Adult	Reference Senior
Adult	234	156
Senior	153	218

dt

Prediction	Reference Adult	Reference Senior
Adult	217	122
Senior	170	252

rf

Prediction	Reference Adult	Reference Senior
Adult	214	108
Senior	173	266

gbm

Prediction	Reference Adult	Reference Senior
Adult	213	121
Senior	174	253

xbst

Prediction	Reference Adult	Reference Senior
Adult	203	97
Senior	184	277

lgr

Prediction	Reference Adult	Reference Senior
Adult	210	139
Senior	177	235