Data challenge - Summary note

Raphaël Pesah

Description of the pipeline

(executable ipynb file available here)

- Preprocessing:
  - I kept all features except “MI”, I gave the value “False” to all NA in “multiple_full_time_jobs” and “combined_multiple_jobs”.
  - Since I’m using the HistGradientBoostingRegressor (dataset with many observations compared to the number of features), I didn’t need to scale numerical features and I used the “OrdinalEncoder” for categorical features.

- I compare two approaches, with regard to the way to tune the hyperparameters:

  - A randomized search method: the hyperparameters investigated here are max_leaf_nodes ([2, 5, 10, 20, 50, 100]) and the learning rate (loguniform(0.01, 1)). The values allowing the best predictions can be seen in Table 1 or in the linked file.

  - A grid-search method: the hyperparameters investigated here are max_depth ([3, 8]), max_leaf_nodes ([15, 31]) and, learning_rate": [0.1, 1]. The values allowing the best predictions can be seen in Table 2 or in the linked file.

	param_histgradientboostingregressor__learning_rate	param_histgradientboostingregressor__max_leaf_nodes	split0_test_score
0	0.125207	100	-5585.899178
1	0.488126	20	-6220.483257
5	0.845913	10	-6595.150396
3	0.176656	10	-7153.734201
2	0.122961	5	-9125.635232

Table 1: Best values for the hyperparameters with randomized-search

	param_histgradientboostingregressor__learning_rate	param_histgradientboostingregressor__max_depth	param_histgradientboostingregressor__max_leaf_nodes	split0_test_score
3	0.1	8	31	0.845413
7	1	8	31	0.819423
6	1	8	15	0.795300
2	0.1	8	15	0.825877
4	1	3	15	0.797735

Table 2: Best values for the hyperparameters with grid-search

Accuracy achieved

I performed inner and then outer cross-validation for both hyperparameters tuning methods.

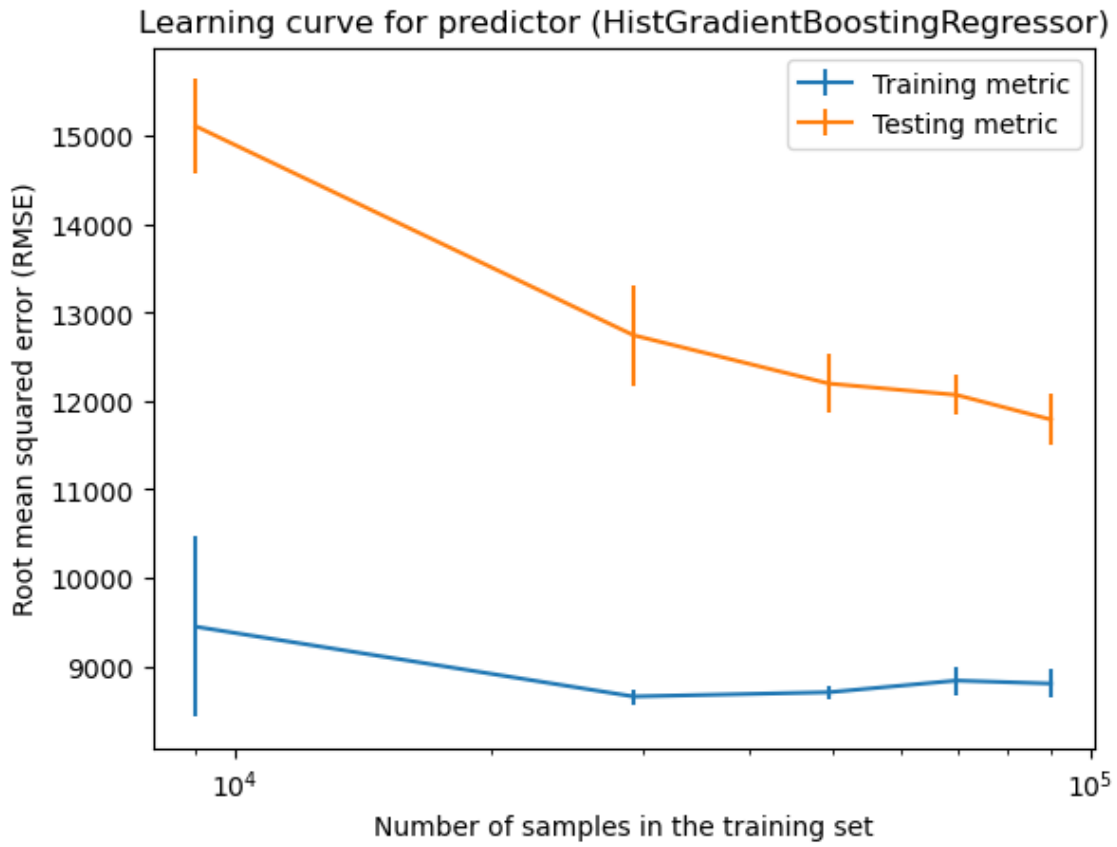
With the randomized-search, I used a ShuffleSplit strategy for the outer cross-validation and I achieved an **absolute mean cross-validated testing error of $5845** with a standard deviation of $109.

With the grid-search, I used a KFold strategy for the outer cross-validation and I achieved an **absolute mean cross-validated testing error of $6695** with a standard deviation of $145.

So I chose the first model for the prediction on the hold-out dataset.

Additional question I chose to address

The figure below, a plot with the RMSE as a function of the sample size n, shows how the training sample size changes the learning capacities. This model used here is the HistGradientBoostingRegressor before the tuning of the hyperparameters. We clearly see that the testing error continuously decreases as the sample size increases. However, for the training error, we see stagnation and even a small increases when n is large, which may lead us to this that the model starts to underfit at some point.



Approach for answering the question related to gender

In order to know the effect of gender on the annual compensation, we need to find a way to check if the differences we may observe are causal or not. But since we only have observational data, we will thus use a matching approach (executable R file available here).

Here we consider that the gender variable (SEX) is the treatment variable and that the annual compensation (ANNUAL) is the outcome: we used propensity score matching to estimate the average marginal effect of the “treatment” (being a woman) on annual compensation, accounting for confounding by the included covariates.

1) We matched the “treated” group (i.e. women) and the “control” group (i.e. men) using a nearest neighbor matching on several covariates likely related to the level of compensation. The covariates we think are relevant and included are : the age in months (AGY), the ethnicity (RACE), the type of employment (EMPTYPE), the year of hiring (HIREDT converted to a year), the RATE of employment (i.e. 0 is 40h, and e.g. 0.75 is three quarter of this), the number of hours worked per week (HRSWKD, although we realized it’s redundant with RATE) and the class code of the specific job position (JOBCLASS).

2) After matching, all standardized mean differences were below 0.15, indicating relatively good balance (see Figures 1 and 2 for rough illustrations). We had to use a random sample of 20,000 individuals (out of 112 582) for the matching due to computational constraints, and 2776 from the control group were not matched.

3) To estimate the treatment effect and its standard error, we fit a linear regression model with annual compensation as the outcome and the treatment, covariates, and their interaction as predictors and included the full matching weights in the estimation. The lm() function was used to fit the outcome, and we then performed a g-computation to estimate the average treatment effect in the treated (ATT). The estimated effect was $748 (SE = 136, $p < 0.001$), indicating that the average “effect” of being a woman is to decrease annual compensation.

So we can conclude that there is a causal effect of gender on the annual compensation, that disadvantages women.

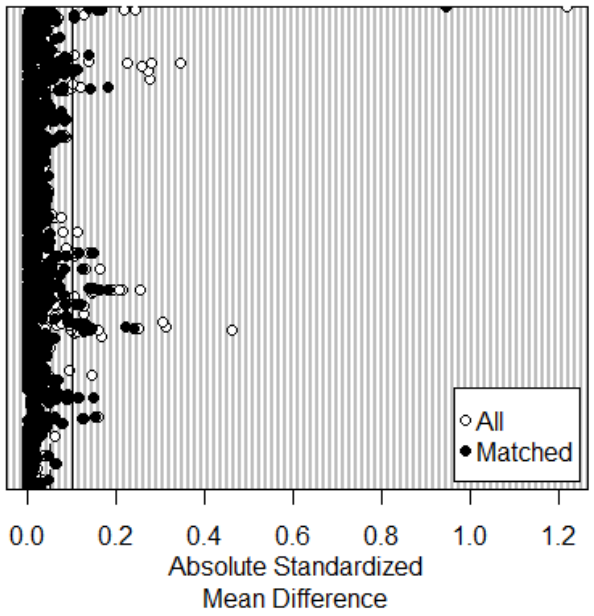


Figure 1: Rough Love plot (the variable labels were not easily readable)

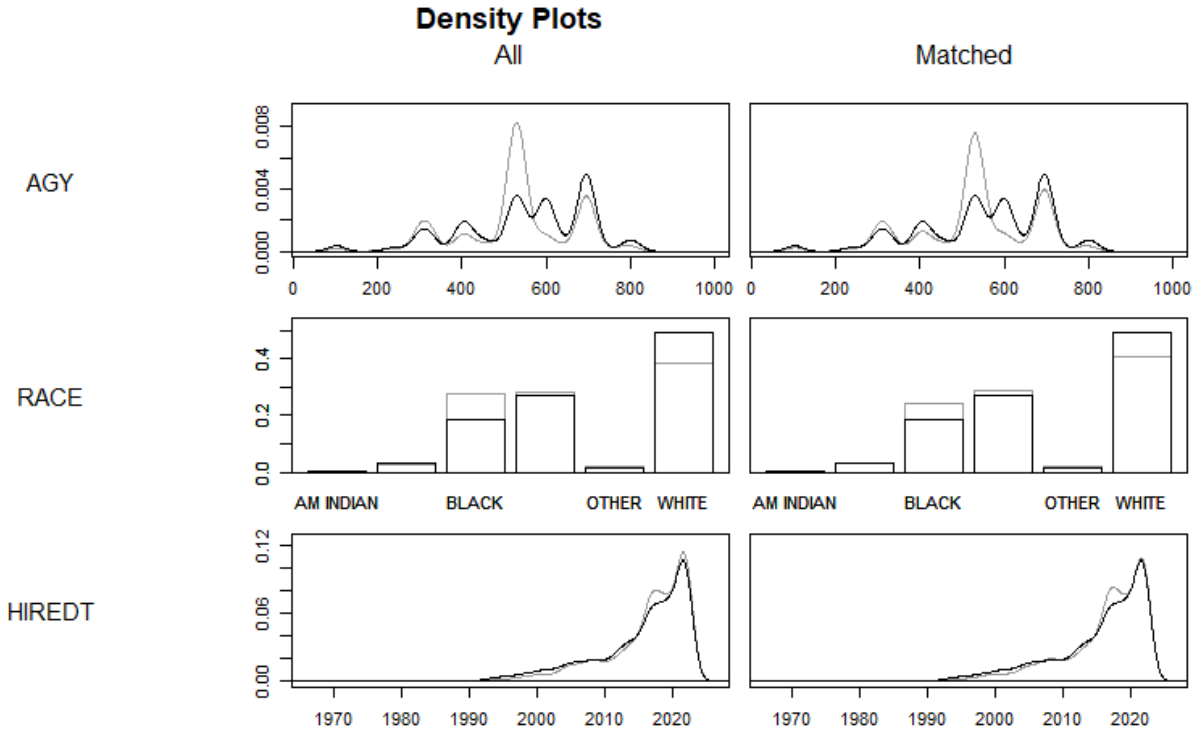


Figure 2: Density plot of some variables for illustration (age in months, ethnicity and date of hiring) - color black is for women and gray for men