IA & SHS: Data challenge

April 16, 2023

Inputs: You will work on a data set containing compensation of state employees. The aim is to predict the annual compensation from information such as their gender, employment date, the agency they belong to, grades and so on. Specifically, you have at hand a dataset to train your model (`train.csv`) with the true labels, and a test dataset (`test.csv`) without the labels, on which you are asked to predict the annual compensation.

Link to download the data – `http://bitly.ws/D6bk`

You you will be asked to:

1. **[10 points]** Predict the annual compensation on the test set using a Machine-Learning pipeline, as taught in class. Compare at least two approaches seen in class, taking into consideration the preprocessing techniques and hyperparameter selection. Evaluating how different preprocessing techniques or hyperparameter selections affect the predictive performance counts as distinct methods tested.

2. **[6 points]** Improve on this minimal pipeline, deep-diving into one question of your interest. This question should be **yours** and can relate to anything. Here are example of questions you can take, but feel free to find yours:

   - Showing how the training sample size changes the learning capacities, proposing as an output a plot with the RMSE as a function of n.
   - Investigating variable importances.
   - Going beyond the mean error (RMSE) to observe how the model predicts extreme values.

3. **[4 points]** Estimate the effect of gender on salaries. Imagine you are a civil servant statistician in charge of the data analysis. You are asked by a journalist the question: *is there an effect of the gender on the annual compensation?* What answer would you provide? You can use the matching approach seen in class, or any other approach you think is helpful.

Bonus: An additional point will be awarded for the top three results in predicting the salary on the hold-out set. **Before** May 13, at noon you will send at `benedicte.colnet@inria.fr` the following documents,

1. The predicted values for the annual compensation. Please send it within the test data set with an extra column called `ANNUAL.prediction` in a `.csv` file called name_surname_predictions.csv;

2. A summary note called name_surname_summary.pdf explaining your approach. For this task, you are not required to focus on formatting, and the report can be concise (maximum of 2 pages). Please mention the minimal pipeline you have proposed, the accuracy achieved, the additional question you chose to address, and your proposed approach for answering the question related to gender.

Furthermore, on May 15, you will be asked to present your approach and methodology. You will prepare the slides for a **5 minutes** speech, followed by 5 minutes of questions from the class and teachers. The idea is to focus on your original approach, rather than explaining which covariates are in the data set as everyone is supposed to have discovered them.

We stress that you will have to **stick to a 5 minutes presentation**.

You are free to use the programming language and tools that work best for you.