FCIM.FIA Autumn 2024
Lab: Learning

Handed out: Tuesday, October 22, 2024

Get a good job with more pay and you're okay!

In such a mesmerizing place as Luna-City, the economy is carefully engineered to ensure fairness and prosperity for all its citizens. Each employee's compensation is not just a matter of their job title but is also influenced by a variety of factors that reflect their contributions to society.

To establish an order and to ensure the fairness of it, the Economic Council was created. They used an extremely powerful tool, known as Lunar Sheets, to carefully document all data required to compute the salary for a certain citizen. However, while it was nice to have it all gathered in one place, Lunar Sheets was kind of weak in computations and much more work was done to correct results provided by it.

Recently, anomalies in the salary computation have become public, sparking concern among the city's workers. These discrepancies have led to questions about whether all factors are being accurately accounted for, or if there are hidden biases in the current system. It emerged in increasing doubts towards the Economic Council!

Desperate and lost, the Economic Council asked your company for you specifically to be assigned to their request. Your new challenge is to develop a fair and accurate system that will help compute citizen's salaries (yours inclusively!). You need to consider how each factor influences the overall pay and how any change in these variables might impact salary computations. The economy of a whole colony is in your hands!

General Guidelines

- Submit your solution as a .zip archive, containing .ipynb and/or .py code files, and a PDF report describing what you have implemented, on ELSE.
- Do NOT host your solution in public repositories (e.g. Github etc). You can use private repositories if you need to.
- **Plagiarism is NOT tolerated!**

Grading Policy

Task 1 Import the provided data. Make a detailed dataset analysis and present some statistics in the form of Matplotlib or Seaborn visualisations. Pre-process the data if necessary. **(2p.)**

Task 2 Based on the analysis from task 1, perform Feature Selection. You should state which columns you will be using for the final predictions and show why you are choosing the respective columns. Perform Linear Regression on your train set, using the selected features. **(1p.)**

Task 3 Train at least two new Linear Regression models. You can use one of the following: *Ridge*, *Lasso*, *Elastic Net Regularization*, *Ordinary Least Squares*, *Least-angle Regression (LAR)*. **(2p.)**

Task 4 Show the performance of your models. You can use one or multiple metrics that you consider more suitable for your case (e.g. MAE, MSE, RMSE, etc.) and explain why you decided to use them. Make conclusions on the obtained results and the performance of each of the developed models. **(1p.)**

Task 5 Cluster your data and show a visual representation of it (you can use K-means, for example, or another algorithm you consider suitable). **(2p.)**
Note: when clustering, you should eliminate the target column from the dataset.

Task 6 Draw conclusions on the obtained clusters. Analyze the values predicted by your best Linear Regression model and resultant clusters. **(1p.)**

Report & Presentation Clear explanations, report formatting, code quality, comments in the code, docstrings, visualisations if relevant etc. **(1p.)**

Good Luck!