Project Definition

COMP 6981

Fall 2023

Objective

In this project, you will practice data preparation in the context of predictive modeling. The **due date** of the project is **Nov 20th** (11:59 PM). Submit your file in the "Assignments Folder" (under "Assessment") of Brightspace.

Project outline

The project can be implemented either in **R** or **Python**. Pick a dataset in form of tabular data, and pick a target/response variable (for prediction) in the dataset. Depending on the measurement level of the target variable, your problem would be either a regression or classification. Now construct your model based on a model type, and take note of its predictive performance called the **bottom-line**.

At each **step**, through applying a data preparation technique, you need to **improve upon your last achieved** predictive performance (starting with the bottom-line). In doing so, you apply the data preparation technique of choice on your dataset, re-model your data based on the new dataset, and then measure the new predictive performance and compare it with the previous one. If an improvement is achieved replace the new dataset with the old one and continue the process. You need to improve upon the predictive performance **five times.**

Data preparation technique(s) which is done between two measurements of predictive performance, constitutes a **single step** in this project. For example, if you decide to drop one variable and then measure the performance, then dropping a single variable would constitute a single step. If you decide to drop a set of variables instead and measure the performance, then dropping several values constitute a single step. Or as another example, you may decide to drop certain outliers first and then then rest, in which case you are implementing two steps.

Complementary notes:

1. You may use any dataset.

2. You may use any regression/classification model type except for Neural Networks; but you need to **stick with your selected model type,** throughout the project.

3. You may use any values for hyper-parameters in your model; but you need to **stick with all the parameter values** throughout the project.

4. You may use any measure for predictive performance; but you need to **stick with the measure** of your choice, throughout the project.

5. Improvements upon a predictive performance will **not** be counted if it takes only a single step. Therefore, you will have to improve upon the predictive performance **again.**

6. Once you have gone through 15 data preparation steps, you may stop. There will be **no marks deducted** at this point, if you have **not been able** to improve upon the predictive performance (even once).

7. You may use **any data preparation technique** discussed in this course. Your are also **allowed** to employ data preparation techniques which are not covered in the course.

8. **Leave explanation** in the notebook surrounding the following points:

    (a) The dataset you have chosen, its variables, in particular the target variable.

    (b) Your thinking process or the rationale behind each data preparation step (include visualization, summary statistics, etc, where applicable).

    (c) At each step, if your data preparation fails, try to speculate why the technique you have chosen has not resulted in improvement.

9. **Bonus mark:** If you extend your data preparation process to **25 steps,** you will be entitled to **3 extra marks.**

Project delivery format

You must complete your project in a Python/R **Notebook** (code + comments + explanations + plots). And then convert the final Notebook to a **PDF** file. Therefore, the project submission is **exclusively** in PDF format. Here are some useful resources in this regard:

1. Google Colab: resource 1, resource 2, resource 3. Exporting a Notebook to PDF in Google Colab: resource 1, resource 2.

2. Jupyter: resource 1, resource 2, resource 3, resource 4. Exporting a Jupyter Notebook to PDF: resource 1, resource 2.

3. R Studio: resource 1, resource 2. Exporting a Notebook to PDF in R Studio: resource.

Some useful resources

Here are some useful **Python** resources relating to this project:

- General predictive modeling: resource 1, resource 2, resource 3.
- Pandas: resource 1, resrouce 2.
- Scikit-Learn: resource 1, resource 2.
- Matplotlib: resource 1, resource 2.
- NumPy: resource 1, resource 2.

Here are some resources on **R**:

- General: resource 1, resource 2, resource 3, resource 4.
- Predictive modeling: resource 1, resource 2.
- **Base R** provides many data preparation techniques. Nonetheless, you may want to check out the main R repository, known as CRAN, where there exist tons of libraries which extend the capabilities of Base R (including data preparation capabilities).