Project #5 – Supervised Machine Learning

Project Description

You will choose a new data set and do a full Supervised ML Project! Emphasis in this project is in the ML method and practising to do all steps with new data. You should therefore pick a dataset suitable for ML fast, so you can start working directly in the ML workflow.

Project Goals

- Grow your autonomy in the supervised ML code & workflow
- Practice relating a Supervised ML models' predictions to a problem they can help solve
- Practice clearly communicating the value of your analysis & code
- Put the project on GitHub

Project Guidelines

- Choose/collect a data set
- Describe your data set and formulate a precise problem that you want to solve
    - **Classification** or **Regression**?
- Plan the project in Trello Board

Technical Requirements

- Define your target variable and your features (independent variables)
- Define the metric you will optimise your model for
- Decide on a baseline that you are trying to beat with your model
- Train/Test split
- Preprocess input data (if necessary)
    - Scale
    - Create dummies

    - Impute
- Model Selection
    - Try out at least 4 different models
    - Use K-fold cross validation
- Hyperparameter tuning
    - Use GridSearchCV with a parameter grid and K-fold cross validation
- When you've decided on your final model, move on to predict test data
    - Do not modify the model after using it to predict test data!
- Evaluate final performance on your test data against the baseline

Presentation

The presentation should take max 5 minutes

The slides should include the following (not necessarily in this order):

- Title of the project + Student name
- Clear description of the problem you are trying to solve
- Clear description of your data set
- Clear communication of your models' performance
    - Relate to baseline
- Clear communication of how your models' predictions can create value
- Challenges
- Learnings / highlights

Schedule

The presentations will take place on …?

Good Luck!!