Project #5 – Supervised Machine Learning

Project Description

Time to kick some machine learning ass. After this week's introduction to ML and specifically supervised ML you are now ready to take on your own Supervised ML Project. You will choose your own data set and conduct a full machine learning project!

Project Goals

- Grow your autonomy in the supervised ML code & workflow
- Continue to practice data cleaning, EDA & visualizations as it is an integral part of any data analysis project
- Practice relating a Supervised ML models' predictions to a problem they can help solve
- Practice clearly communicating the value of your analysis & code

Project Requirements

- Plan the project in Trello Board
- Choose/collect a data set
- Describe your data set and formulate a precise problem that you want to solve
- Determine if your ML problem is most suited to solve with **classification** or **regression**.
- Decide on a baseline that you are trying to beat with your model
- Put the project on your GitHub

Technical Requirements

- Data cleaning (if necessary)
- EDA
    - Visualize findings
- Define your target variable and your features (independent variables)
- Preprocess input data (if necessary)
    - Scale
    - Create dummies
    - Impute
- Train/Test split
- Model Selection
    - Try out at least 4 different models
    - Use K-fold cross validation
- Hyperparameter tuning
    - Use GridSearch with a parameter grid and K-fold cross validation
    - During this evaluation, show how you cope with over/underfitting
- 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 with relevant measures

Presentation

The presentation should take max 10 minutes

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

- Title of the project + Student name
- Clear description of the problem you were trying to solve
- Clear description of your data set
    - Visualize!
- Clear communication of your models' performance
    - Which metric(s) are the most important for you?
    - Relate to baseline
- Clear communication of how your models' predictions can create value
- Challenges
- Learnings / highlights

Schedule

The presentations will take place on **Monday next week** and you should hand in the project (slides, links etc) before presenting

Good Luck!!