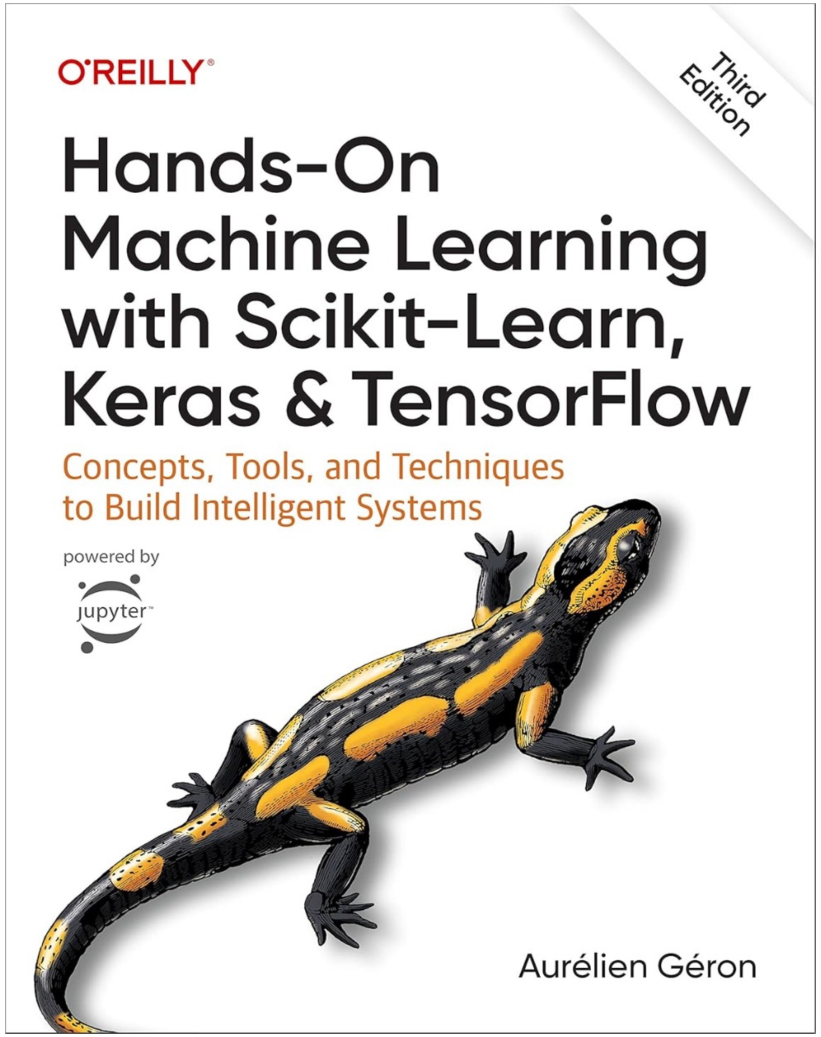
DS 5220
Supervised Machine Learning
Steve Morin

Semester Project Description

Introduction

This semester project is adapted from Geron's Chapter 2 - End-to-End Machine Learning Project.

I suggest you read that chapter to better understand the spirit of the semester project.

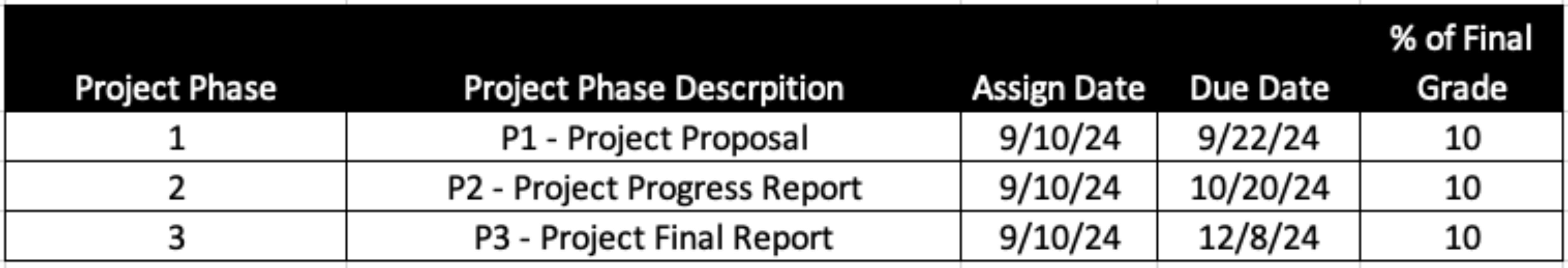

O'REILLY
Third Edition
Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow
Concepts, Tools, and Techniques to Build Intelligent Systems
powered by
jupyter
Aurélien Géron

Working with Real Data

Popular data repositories:

- OpenML.org (https://openml.org)
- Kaggle.com (https://www.kaggle.com/datasets)
- PaperWithCode (https://paperswithcode.com/datasets)
- UC Irvine Machine Learning Repository (https://archive.ics.uci.edu/datasets)
- Amazon’s AWS datasets (https://registry.opendata.aws)
- TensorFlow datasets (https://www.tensorflow.org/datasets)

Project Phases, Due Dates and Team/Group Requirements



Project Phase	Project Phase Descrpition	Assign Date	Due Date	% of Final Grade
1	P1 - Project Proposal	9/10/24	9/22/24	10
2	P2 - Project Progress Report	9/10/24	10/20/24	10
3	P3 - Project Final Report	9/10/24	12/8/24	10

Teams/groups are allowed. There is a maximum of 3 people per team/group.

All members of a team must submit identical submissions on the Canvas to receive credit.

Main Steps

I. Frame the problem and look at the big picture.

II. Get the data.

Minimum requirements for Project Phase 1
(Steps I and II)

III. Explore the data to get insights (EDA).

IV. Prepare the data to better expose the underlying data patterns to machine learning algorithms.

Minimum requirements for Project Phase 2
(Steps III and IV)

V. Explore many different models and short list the best ones.

VI. Fine-tune your model.

VII. Present your solution.

Minimum requirements for Project Phase 3
(Steps V, VI and VII)

Project Deliverables

Project Phase	Steps	Deliverables
1	I and II	A PowerPoint document submitted as a .pdf file that addresses all the points in these steps.
2	III and IV	Step III: • One Jupyter notebook (eda.ipynb and eda.html) dedicated to data exploration as described in this document. • All .py modules that are imported by the notebooks(1). • The .yml for the environment in which the data exploration was completed. • A PowerPoint document submitted as a .pdf file that describes the findings. Step IV: • One Jupyter notebook (prep.ipynb and prep.html) dedicated to data preparation as described in this document. • All .py modules that are imported by the notebooks(1). • The .yml for the environment in which the data preparation was completed. • A PowerPoint document submitted as a .pdf file that describes the data preparation.

Notes:

1. Include a .zip of the utils directory that includes any .py files that are imported by the .ipynb.
2. Do not combine .ipynb files.

Project Deliverables (continued)

Project Phase	Steps	Deliverables
3	V, VI and VII	Step V: • Jupyter notebook(s) (model_exp_x.ipynb and model_exp_x.html) dedicated to “quick and dirty” model exploration as described in this document. • All .py modules that are imported by the notebooks(2). • The .yml for the environment in which the model exploration was completed. • A PowerPoint document submitted as a .pdf file that describes the model exploration and provides a short list of candidate models. Step VI: • Jupyter notebook(s) (model_ft_x.ipynb and model_ft_x.html) dedicated to fine tuning a short list of candidate models as described in this document. • All .py modules that are imported by the notebooks(2). • The .yml for the environment in which the model fine tuning was completed. • A PowerPoint document submitted as a .pdf file that describes the model fine tuning and indicates the best model. Step VII: • A PowerPoint document submitted as a .pdf file that addresses all the points in this step.

Notes:

1. A file name that ends in a _x is meant to enable multiple notebooks in a submission. For example, for Step VI, one might submit the following: model_ft_1.ipynb and model_ft_2.ipynb.
2. Include a .zip of the utils directory that includes any .py files that are imported by the .ipynb.
3. Do not combine .ipynb files.
4. You may combine the .pdf files if each step above is clearly delineated by inserting a title page indicating the step number for each step.

I. Frame the Problem and Look at the Big Picture

1. Define the objective of the project in business terms.

2. How will your solution be used?

3. How should performance be measured?

4. Is the performance measure aligned with the business objective?

5. What would be the minimum performance needed to reach the business objective?

6. List the assumptions you have made so far.

7. Verify assumptions if possible.

II. Get the Data

1. Document where you got the data by providing a URL.

2. Get the data.

3. Convert the data to a format you can easily manipulate (without changing the data itself).

4. Perform a train/test split. Save the test set to a file.

   Put it aside and don’t use it until it is time to measure the generalization performance of your selected model.

   Do not look at it or analyze it to avoid data leakage through the data scientist.

   Provide a sample size and a file name in the .pdf.

III. Explore the Data

1. Create a copy of the data for exploration.

2. Create a dedicated Jupyter notebook to keep a record of your data exploration.

3. Identify the target attribute.

4. Study each attribute and its characteristics.

   The goal here is to prepare for setting the attribute configuration for machine learning.

   We are looking for each attribute's role in supervised machine learning. Possible roles are:

   - attributes with missingness above threshold

   - non machine learning attributes

   - attributes to exclude from machine learning

Consider reviewing

https://scikit-learn.org/stable/modules/feature_selection.html

for ideas on feature selection to aid in establishing the supervised machine learning configuration.

III. Explore the Data (continued)

5. Visualizations as needed.

6. Study the correlations and associations between attributes and between attributes and target.

   This might inform your attribute configuration.

7. Study how you would solve the problem manually.

8. Identify additional data that would be useful (go back to “Get the Data”).

   You are not required to get the additional data.

   You should give this some thought and document what data might add to your project.

III. Explore the Data (continued)

10. Document what you have learned in a PowerPoint slide deck.

    In addition to other content your documentation must include a table of attributes (rows) with the following columns:

    a. Attribute name
    b. Attribute type (nominal or numerical)
    c. % missing values
    d. Role in attribute configuration for machine learning.

IV. Prepare the Data

Notes:

- Work on copies of the data. Keep the original data intact.

- Implement all data transformations with reusability in mind. This is important for five reasons:

  - So you can easily prepare the data the next time you get a fresh data set.

  - So you can apply these transformations in future projects.

  - To clean and prepare the test set.

  - To clean and prepare new data instances once your solution is being used.

  - To make it easy to treat your preparation choices as hyperparameters.

sklearn pipelines will enable you to do this!

IV. Prepare the Data (continued)

Notes (continued):

- Create a dedicated Jupyter notebook to keep a record of your data preparation.

Data Preparation / Preprocessing

1. Pre Scikit-learn pipeline data preparation.

   This type of data preparation should be minimized because it does not allow us to benefit from the power of Scikit-learn pipelines.

   The types of preprocessing that might be completed here include:

   a. aggregating data objects

IV. Prepare the Data (continued)

2. Perform attribute selection by setting the Scikit-learn pipeline machine learning attribute configuration.

   Exclude:

   - attributes with missingness above threshold
   - non machine learning attributes
   - attributes to exclude from machine learning

   This step is informed by the eda performed earlier.

3. Perform attribute transformations:

   a. impute missing values.

   b. scale features

   c. discretize continuous features

V. Shortlist Promising Models

1. Train many default models from different categories (linear, ensemble, etc.) using default parameters.

2. Measure and compare their performance:

   a. For each model, use *k*-fold cross-validation and compute the mean and standard deviation of the performance measure on the *k* folds.

Notes:

- If the data is huge sample smaller training sets so you can train many models.

  Smaller data sets may penalize complex models such as large neural nets or random forests.

- Automate these steps.

V. Shortlist Promising Models (continued)

3. Analyze the most significant attributes for each algorithm.

4. Analyze the types of errors the models make:

   a. What data would a human have used to avoid these errors?

5. Perform a quick round of feature selection and engineering.

6. Perform one or two more quick iterations of the five previous steps.

7. Shortlist the top three to five most promising models, preferring models that make different types of errors.

VI. Fine-tune the System

1. Fine-tune the hyperparameters using cross validation.

   a. Treat your data transformation choices as hyperparameters.

   b. Use random search over grid search.

2. Try model ensemble methods.

   Combining your best models will often produce better results then running them individually.

3. Once you are confident about your final model, measure its performance on the test set to estimate generalization.

Notes:

- Use as much data as possible for these steps.

- Try to automate these steps as much as possible.

- Don’t tweak your model after measuring the generalization error.

  Tweaking your model at this point would cause you to overfit the test set.

VII. Document Your Solution

1. Document what you have done in a PowerPoint slide deck.

   a. Highlight the big picture first.

2. Explain why your solution achieves the business objective.

3. Present interesting points you noticed along the way:

   a. Describe what worked and what did not work.

   b. List the assumptions and your systems limitations.

4. Ensure your key findings are communicated through visualizations and easy to remember statements.

VII. Document Your Solution (continued)

6. Include a flow chart of the machine learning process in an appendix of the document.