Phase 3 Project Launch

Machine Learning Classification

July 8, 2022

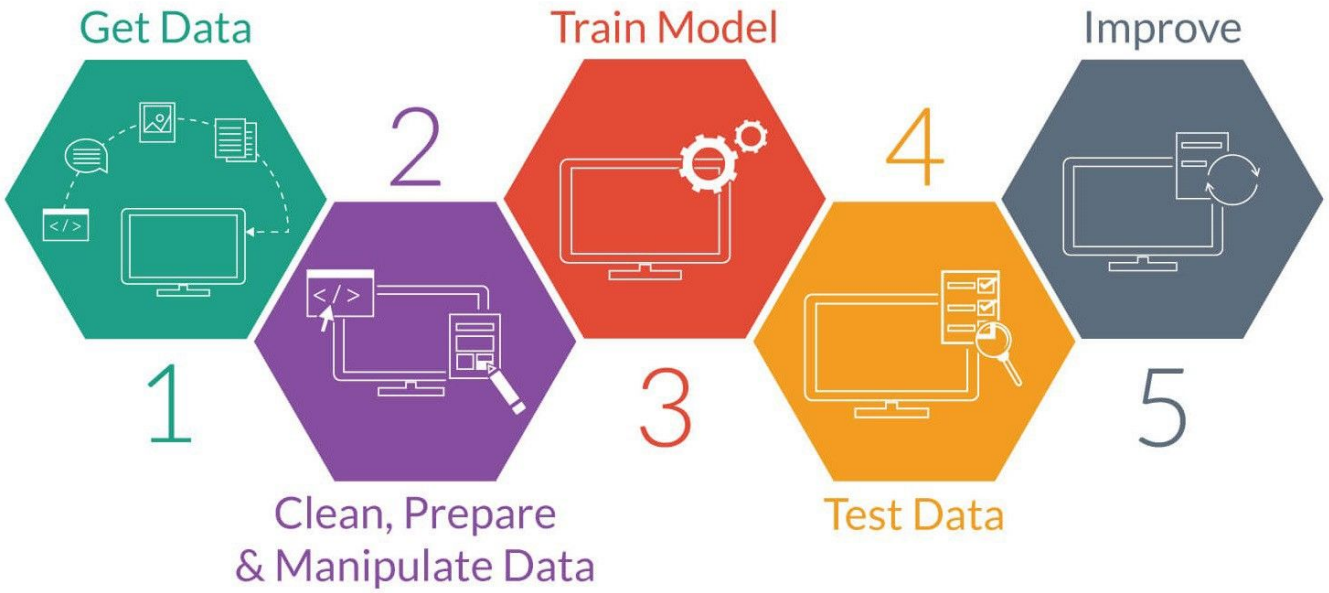
// FLATIRON SCHOOL

Classification Project

Project Assignment

- Phase 3 Project Rubric
  - Note: Rubric is similar, with a few changes
- Phase 3 Project Checklist

CRISP-DM

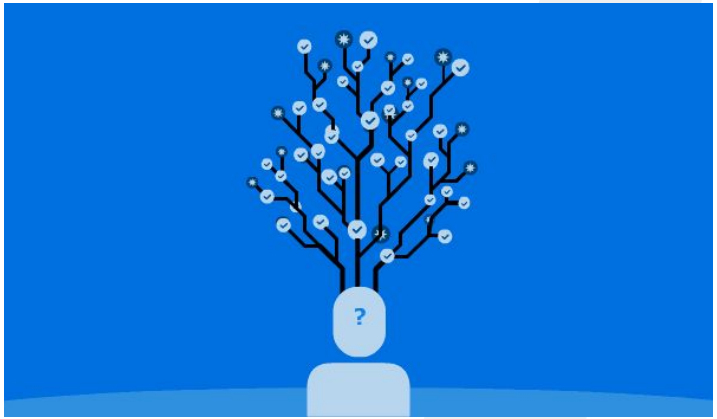

Get Data
Train Model
Improve
2
4
1
3
5
Clean, Prepare
& Manipulate Data
Test Data

Remember the Workflow Steps!

1. Identify your business problem/stakeholder
2. EDA/Data Prep (class imbalance, visualizations, summary stats, correlations)
3. Train/Test Split, Cross-Validation, Transformations (OneHotEncoder, StandardScaler, MinMaxScaler, Regularization, SimpleImputer, etc.)
4. Baseline Model (Dummy Classifier)
5. Evaluate Baseline
6. Run a model
7. Evaluate model - confusion matrix, log loss, accuracy, precision, recall, ROC-AUC, recall, etc. Talk about class imbalance
8. Decide if you need another model, then repeat steps 6 & 8

Business Understanding

- **Select a Stakeholder**
  - Discuss **cost** of different errors.
  - Decide on an appropriate **metric.**



- **How will your model be used?**
  - More interpretable or better performance?

Data Understanding/EDA

- Do not forget to describe your data! Where did it come from? How big is it? What features does it contain?
- How well does the sample represent the population?
- Thorough **EDA**! Address class imbalance, missing values, correlations. **Visualize!**
- Especially if your final model is less interpretable, demonstrating good EDA will help earn the **trust** of your stakeholder

Train/Test Split, Transformations, Cross-Validation

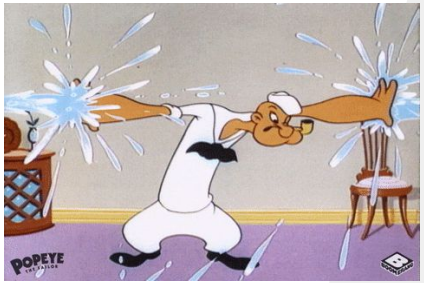
- Train-test-split BEFORE transforming anything
- **Fit and transform the training set, then <u>just transform</u>** the **test** set exactly how the **training** set was transformed
- Many of the same tools as Phase 2, but now with additional practice with cross-validation
- **Pipelines** aren't required, but are convenient!

Iterative Model Building:

Should you ever fit on the test set?

Iterative Model Building: Train Test Split

- Decide an approach to model validation
  - Untouched **hold out set** recommended
- **Cross-validation** is recommended for tuning
  - Minimize **data leakage**



Iterative Model Building: Use All the Models!

- **Apply multiple models!** Decide between model types based on **not only** the scores **but also** your business and data understanding!

- **Document** your iterative progress - include just a selection of important models in final notebook

Iterative Model Building: Hyperparameter Tuning

- Use cross-val to select best **hyperparameters**
- Find optimal complexity to balance **bias/variance** and **maximize validation scores**
- Use **grid-search** and **pipelines** to streamline process
- Tune to a **metric** aligned to business problem

Iterative Model Building: Final Model

- Select **final model**
- Fit on **entire training dataset**, score on test/holdout
- Inspect **feature importances** or **parameters**
- Inspect patterns of **errors**
- What did your model **do well?**
- What did your model **not do well?**
- Did your model perform **as expected** on the test/holdout set?

Project Deliverables

There are 3 Deliverables for this Project:

- **Github repository**
  - README.md
  - Clear commit history (good commit messages!)
  - Organized repository

- **Final Notebook**
  - One clean notebook

- **Non-Technical Presentation**
  - Aim for 5 minutes total

Questions?