Phase 3 Project Description

[(https://github.com/learn-co-curriculum/dsc-phase-3-project-v3)](https://github.com/learn-co-curriculum/dsc-phase-3-project-v3) [(https://github.com/learn-co-curriculum/dsc-phase-3-project-v3/issues/new/choose)](https://github.com/learn-co-curriculum/dsc-phase-3-project-v3/issues/new/choose)

Congratulations! You've made it through another *intense* module, and now you're ready to show off your newfound Machine Learning skills

All that remains in Phase 3 is to put your new skills to use with another large project.

In this project description, we will cover:

- Project Overview
- Deliverables
- Grading
- Getting Started

Project Overview

For this project, you will engage in the full data science process from start to finish, solving a **classification** problem using a **dataset of your choice**.

Business Problem and Data

It is up to you to define a stakeholder, a business problem, and you are also responsible for choosing a dataset.

For complete details, see **Phase 3 Project - Choosing a Dataset** [(https://github.com/learn-co-curriculum/dsc-phase-3-choosing-a-dataset)](https://github.com/learn-co-curriculum/dsc-phase-3-choosing-a-dataset).

Key Points

Classification

Recall the distinction between *classification* and *regression* models:

- Classification is used when the target variable is a *category*
- Regression is used when the target variable is a *numeric value*

(Categorical data may be represented in the data as numbers, e.g. 0 and 1, but they are not truly numeric values. If you're unsure, ask yourself "is a target value of 1 *one more than* a target value of 0"; if it is one more, that is a regression target, if not, that is a classification target.)

You will have additional opportunities to work on regression problems in later phases, but **for this project, you must be modeling a classification problem**.

Findings and Recommendations

In the previous two projects, the framing was primarily *descriptive* and *inferential*, meaning that you were trying to understand the distributions of variables and the relationship between them. For this project you can still use these techniques, but make sure you are also using a ***predictive*** approach.

A predictive *finding* might include:

- How well your model is able to predict the target
- What features are most important to your model

A predictive *recommendation* might include:

- The contexts/situations where the predictions made by your model would and would not be useful for your stakeholder and business problem
- Suggestions for how the business might modify certain input variables to achieve certain target results

Iterative Approach to Modeling

You should demonstrate an iterative approach to modeling. This means that you must build multiple models. Begin with a basic model, evaluate it, and then provide justification for and proceed to a new model. After you finish refining your models, you should provide 1-3 paragraphs in the notebook discussing your final model.

With the additional techniques you have learned in Phase 3, be sure to explore:

1. Model features and preprocessing approaches
2. Different kinds of models (logistic regression, decision trees, etc.)
3. Different model hyperparameters

At minimum you must build two models:

- A simple, interpretable baseline model (logistic regression or single decision tree)
- A version of the simple model with tuned hyperparameters

Classification Metrics

You must choose appropriate classification metrics and use them to evaluate your models. Choosing the right classification metrics is a key data science skill, and should be informed by data exploration and the business problem itself. You must then use this metric to evaluate your model performance using both training and testing data.

Deliverables

There are three deliverables for this project:

- A **non-technical presentation**
- A **Jupyter Notebook**
- A **GitHub repository**

Non-Technical Presentation

Recall that the non-technical presentation is a slide deck presenting your analysis to ***business stakeholders***, and should be presented live as well as submitted in PDF form on Canvas.

We recommend that you follow this structure, although the slide titles should be specific to your project:

1. Beginning
   - Overview
   - Business and Data Understanding
2. Middle
   - Modeling
   - **Evaluation**
3. End
   - Recommendations
   - Next Steps
   - Thank you

Make sure that your discussion of classification modeling is geared towards a non-technical audience! Assume that their prior knowledge of machine learning is minimal. You don't need to explain the details of your model implementations, but you should explain why classification is useful for the problem context. Make sure you translate any metrics or feature importances into their plain language implications.

The graded elements for the non-technical presentation are the same as in **Phase 1 (https://github.com/learn-co-curriculum/dsc-phase-1-project-v3#deliverables)** and Phase 2.

Jupyter Notebook

Recall that the Jupyter Notebook is a notebook that uses Python and Markdown to present your analysis to a ***data science audience***. You will submit the notebook in PDF format on Canvas as well as in `.ipynb` format in your GitHub repository.

The graded elements for the Jupyter Notebook are:

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- **Evaluation**
- Code Quality

GitHub Repository

Recall that the GitHub repository is the cloud-hosted directory containing all of your project files as well as their version history.

The requirements are the same as in **Phase 1** **(https://github.com/learn-co-curriculum/dsc-phase-1-project-v3#github-repository)** and **Phase 2** **(https://github.com/learn-co-curriculum/dsc-phase-2-project-v3#github-repository)** , except for the required sections in the `README.md` .

For this project, the `README.md` file should contain:

- Overview
- Business and Data Understanding
  - Explain your stakeholder audience and dataset choice here
- Modeling
- **Evaluation**
- Conclusion

Just like in Phase 1 and 2, the `README.md` file should be the bridge between your non technical presentation and the Jupyter Notebook. It should not contain the code used to develop your analysis, but should provide a more in-depth explanation of your methodology and analysis than what is described in your presentation slides.

Grading

To pass this project, you must pass each project rubric objective. The project rubric objectives for Phase 3 are:

1. ML Communication
2. Data Preparation for Machine Learning
3. Nonparametric and Ensemble Modeling

ML Communication

Recall that communication is one of the key data science "soft skills". In Phase 3, we are specifically focusing on ML Communication. We define ML Communication as:

> Communicate the **performance** of and **insights** generated by machine learning models to diverse audiences via writing, live presentation, and visualization

High-quality ML Communication includes rationale, results, limitations, and recommendations:

- **Rationale:** Explaining why you are using machine learning rather than a simpler form of data analysis
  - What about the problem or data is suitable for this form of analysis?
  - For a data science audience, this includes your reasoning for the changes you applied while iterating between models.
- **Results:** Describing the classification metrics
  - You can report multiple metrics for a single model, but make sure that indicate a reason for which metrics you are using (and don't try to use all of them at once)
  - For a business audience, make sure you connect any metrics to real-world implications. You do not need to get into the details of how the model works.

-
    - For a data science audience, you don't need to explain what a metric is, but make sure you explain why you chose that particular one.
- **Limitations:** Identifying the limitations and/or uncertainty present in your analysis
    - Are there certain kinds of records where model performance is worse? If you used this model in production, what kinds of problems might that cause?
    - In general, this should be more in-depth for a data science audience and more surface-level for a business audience.
- **Recommendations:** Interpreting the model results and limitations in the context of the business problem
    - What should stakeholders *do* with this information?

Exceeds Objective

Communicates the rationale, results, limitations, and specific recommendations generated by a classification model

> See above for an extended explanation of these terms.

Meets Objective (Passing Bar)

Successfully communicates model metrics without any major errors

> The minimum requirement is to communicate the *results*, meaning at least one overall model metric for your final model. See the Approaching Objective section for an explanation of what a "major error" means.

Approaching Objective

Communicates model metrics with at least one major error

> A major error means that some aspect of your explanation is fundamentally incorrect. For example, if you report a regression metric for a classification model, that would be a major error. Another example would be if you report the model's performance on the training data, rather than the model's performance on the test data.

Does Not Meet Objective

Does not communicate model metrics

> It is not sufficient just to display the `classification_report` or confusion matrix for a given model. You need to focus on one or more specific metrics that are important for your business case.

Data Preparation for Machine Learning

We define this objective as:

> Applying appropriate preprocessing and feature engineering steps to tabular data in preparation for predictive modeling

You still to ensure that you have a strategy for dealing with missing and non-numeric data.

For the Phase 3 project, make sure you also consider:

- **Preventing Data Leakage:** As you prepare data for modeling, make sure that you are correctly applying data preparation techniques so that your model's performance on test data realistically represents how it would perform on unseen data. For scikit-learn transformers specifically, ***make sure that you do not fit the transformer on the test data***. Instead, fit the transformer on the training data and use it to transform both the train and test data.
- **Scaling:** If you are using a distance-based model algorithm (e.g. kNN or logistic regression with regularization), make sure you scale your data prior to fitting the model.

Feature engineering is encouraged but not required for this project.

Exceeds Objective

Goes above and beyond with data preparation, such as feature engineering or using pipelines

> Relevant examples of feature engineering will depend on your choice of dataset and business problem.
>
> Pipelines are the best-practice approach to data preparation that avoids leakage, but they can get complicated very quickly. We therefore do not recommend that you use pipelines in your initial modeling approach, but rather that you refactor to use pipelines if you have time.

Meets Objective (Passing Bar)

Successfully prepares data for modeling, using a final holdout dataset that is transformed by (but not fitted on) transformers used to prepare training data AND scaling data when appropriate

> See the descriptions above for explanations of how to use transformers and scaling.

Approaching Objective

Prepares some data successfully, but has at least one major error

> A major error means that some aspect of your data preparation is fundamentally incorrect. Some examples of major errors include: (1) fitting transformers on test data, (2) not performing a train-test split, (3) not scaling data that is used in a distance-based model.

Does Not Meet Objective

Does not prepare data for modeling

> This includes projects where data is partially prepared, but the model is unable to run.

Nonparametric and Ensemble Modeling

Your project should consider the different types of models that have been covered in the course so far and whether they are appropriate or inappropriate for the dataset and business case you are working with.

Your final model can still be a linear model (e.g. logistic regression) but you should explore at least one nonparametric model (e.g. decision tree) as well and articulate why one or the other is a better approach.

Exceeds Objective

Goes above and beyond in the modeling process, such as articulating why a given model type is best suited to the problem or correctly using scikit-learn models not covered in the curriculum

> Another way you might go above and beyond would be to create custom Python classes, possibly inheriting from scikit-learn classes.

Meets Objective (Passing Bar)

Uses at least two types of scikit-learn model and tunes at least one hyperparameter in a justifiable way without any major errors

> See the "Iterative Approach to Modeling" section above for a more-lengthy explanation.
>
> Once again, ideally you would include written justifications for each model iteration, but at minimum the iterations must be *justifiable*.
>
> For an explanation of "major errors", see the description under "Approaching Objective".

Approaching Objective

Builds multiple classification models with at least one major error

> A major error means that some aspect of your modeling approach is fundamentally incorrect.
>
> Once again, the number one major error to avoid is including the target as one of your features. If you are getting metrics that are "too good to be true", make sure that you removed the target (`y`) from your data before fitting the model.
>
> Other examples of major errors include: using a numeric target value (since this is a classification project), not starting with a baseline model (e.g. proceeding directly to a Random Forest model), or not tuning hyperparameters in a justifiable way (e.g. reducing regularization on a model that is overfitting)

Does Not Meet Objective

Does not build multiple classification models

Getting Started

Please start by reviewing the contents of this project description. If you have any questions, please ask your instructor ASAP.

Once you are ready to begin the project, you will need to complete the Project Proposal.

Recall that more information is available in **Phase 3 Project - Choosing a Dataset** **(https://github.com/learn-co-curriculum/dsc-phase-3-choosing-a-dataset)** .

To get started with project development, create a new repository on GitHub. For this project, we recommend that you do not fork the template repository, but rather that you make a new repository from scratch, starting by going to **github.com/new** **(https://github.com/new)** .

Summary

This project is an opportunity to expand your data science toolkit by evaluating, choosing, and working with new datasets. Spending time up front making sure you have a good dataset for a solvable problem will help avoid the major problems that can sometimes derail data science projects. You've got this!

How do you feel about this lesson?

Have specific feedback?

Tell us here! (https://github.com/learn-co-curriculum/dsc-phase-3-project-v3/issues/new/choose)