Data Science Career Track

Capstone Three: Preprocessing and Training Data Development, Modeling, and Documentation

---

Overview

Time Estimation: 2-3 Hours

The goal of the preprocessing work is to prepare your data for fitting models. If you identified any categorical features in your dataset in the EDA step, now is the time to create dummy features to allow for the inclusion of those features in your model development. Additionally, standardizing the features numeric magnitude and creating train and test data subsets happen in this step. You may want to save a version of your clean, preprocessed data frame as a CSV to access later.

If you need a refresher about how to complete this work, review the work you did during the guided capstone and revisit the DSM Medium article.

Pre-processing and training data development

The following steps should be completed in a Jupyter Notebook, python scripts, or in Paperspace.

Preprocessing and Training Data Development

Goal: Create a cleaned development dataset you can use to complete the modeling step of your project.

Steps:

- Create dummy or indicator features for categorical variables
- Standardize the magnitude of numeric features using a scaler
- Split into testing and training datasets

Review the following questions and apply them to your dataset:

- Does my data set have any categorical data, such as Gender or day of the week?
- Do my features have data values that range from 0 - 100 or 0-1 or both and more?

Modeling

Time Estimation: 10-15 Hours

The goal of the modeling step is to develop a final model that effectively predicts the stated goal in the problem identification section. Review the types of models that would be appropriate given your modeling response variable and the features in your dataset and build two to three models. In addition to considering different algorithm types in your model selection, also consider applying model hyperparameter tuning operations. Be sure to define the metrics you use to choose your final model.

If you need a refresher about how to go about building a model, review the work you did during the guided capstone and the unit about modeling.

Goal: Build two to three different models and identify the best one.

- Fit your models with a training dataset
- Review model outcomes — Iterate over additional models as needed
- Identify the final model that you think is the best model for this project

Review the following questions and apply them to your analysis:

- Does my data involve a time series or forecasting? If so, am I splitting the train and test data appropriately?
- Is my response variable continuous or categorical?