Data Science Career Track
Capstone Two: Preprocessing and Training Data Development

Overview

Use the outline below as a reminder of what steps to follow while working on this part of the Data Science Method. The goal of the preprocessing work is to prepare your data for fitting models. If you identified some 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 your features numeric magnitude and creating train and test splits 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.

Project Steps

Time Estimation: 2-3 Hours

The following steps should be completed in a Jupyter Notebook.

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

Hint: you'll need to think about your old favorite pandas functions here like *get_dummies()*. Consult this guide for help.

- Standardize the magnitude of numeric features using a scaler

Hint: you might need to employ Python code like this:

```
# Making a Scaler object

scaler = preprocessing.StandardScaler()

# Fitting data to the scaler object

scaled_df = scaler.fit_transform(df)

scaled_df = pd.DataFrame(scaled_df, columns=names)
```

- Split into testing and training datasets

Hint: don't forget your sklearn functions here, like *train_test_split()*.

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?