Linear Regression

By Elliot Ledson

Simplifying Data Preprocessing with Custom Methods

As a data scientist, and not allowed to use Sklearn/Scipy I have had a lot of fun creating my own library of shortcuts for preprocessing data. The library is a class called "DataPreprocessor", which contains several methods that I can use to perform various data preprocessing tasks quickly and easily.

To use the library, I simply instantiate the class and call the desired method with the appropriate parameters. For example, the "convert_to_int" method converts a column in a dataframe to integer type, and also checks for any non-numeric values. Similarly, the "resolve_categorical" method maps unique values of an existing categorical column to integers and creates a new column with the mapped values.

The "round_floats" method rounds off decimal places in a column of float values, while "scale_features" scales specified columns of a dataframe using the StandardScaler from the scikit-learn library.

The "remove_outliers" method removes outliers from specified columns of a dataframe using the interquartile range (IQR) and a user-defined threshold. Finally, the "display_correlation_matrix" method displays a correlation matrix heatmap for a given list of columns to visualize the correlation between them.

With this library, I am able to preprocess data more efficiently and focus more on the actual analysis and modeling of the data.

Using Linear Regression to Compare House Prices Before and After Renovation

As a data scientist, I have used linear regression to build a model that can compare house prices before and after renovation.

Firstly, I created a function to split my data into training and testing sets. The function shuffles the indices of the data and splits them into training and testing sets using the provided percentage. I then created a MVLinearRegression class to fit my model. The class contains a fit method that trains the model and a predict method that makes predictions

using the trained model. I also defined a method for Ordinary Least Squares (OLS) regression, which I use to compute the cost function for my model.

To test my model, I split my data into training and testing sets and selected ten columns to use for training and testing. I then trained my model using the training data and plotted the training curve to visualize the model's performance. Next, I removed outliers from the input data, added a column for price after renovations, and populated the column with the new total cost of the house projected. Finally, I randomly sampled three houses and compared their sold_price before and after renovations.

Although my model functions it needs to be tweaked for accuracy as it is being fed too much information.