Feature Selection: Takeaways

Syntax

- Using Seaborn to generate a correlation matrix heatmap:

```
sns.heatmap(DataFrame)
```

- Rescaling the features for a model:

```
data = pd.read_csv('AmesHousing.txt',  delimiter="\t")

train = data[0:1460]

unit_train = (train[['Gr Liv Area']] - train['Gr Liv Area'].min())/(train['Gr  Liv

Area'].max()  - train['Gr Liv Area'].min())
```

Concepts

- Once we select the model we want to use, selecting the appropriate features for that model is the next important step. When selecting features, you'll want to consider: correlations between features and the target column, correlation with other features, and the variance of features.
- Along with correlation with other features, we need to also look for potential collinearity between some of the feature columns. Collinearity is when two feature columns are highly correlated and have the risk of duplicating information.
- We can generate a correlation matrix heatmap using Seaborn to visually compare the correlations and look for problematic pairwise feature correlations.
- Feature scaling helps ensure that some columns aren't weighted more than others when helping the model make predictions. We can rescale all of the columns to vary between 0 and 1. This is known as min-max scaling or rescaling. The formula for rescaling is as follows:

$$\frac{x - \min(x)}{\max(x) - \min(x)}$$

where x is the individual value, $\min(x)$ is the minimum value for the column x belongs to, and $\max(x)$ is the maximum value for the column x belongs to.

Resources

- seaborn.heatmap() documentation
- Feature scaling