Feature Engineering: 1st lesson – Feature Engineering Basics

Welcome to Kaggle’s Feature Engineering course!

In this course, you'll learn about one of the most important steps on the way to building a great machine learning model: feature engineering. You'll learn how to:

* Determine which features are the most important with *mutual information*
* Invent new features in several real-world problem domains
* Encode high-cardinality categoricals with a *target encoding*
* Create determination features with *k-means clustering*
* Decompose a dataset’s variation into features with *principal component analysis*

The goal of feature engineering:

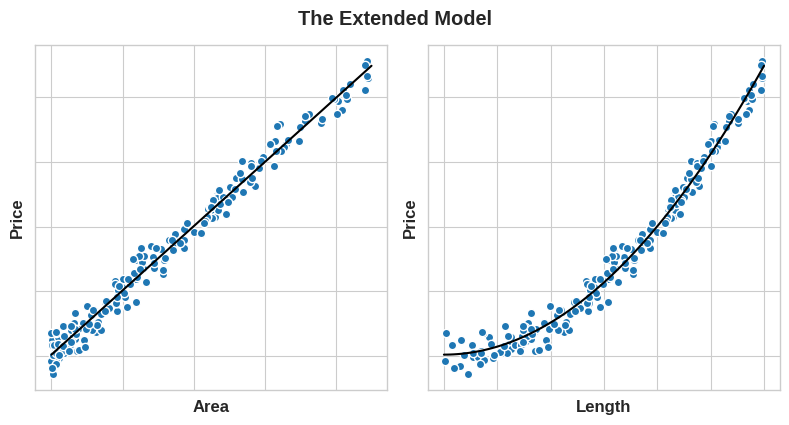
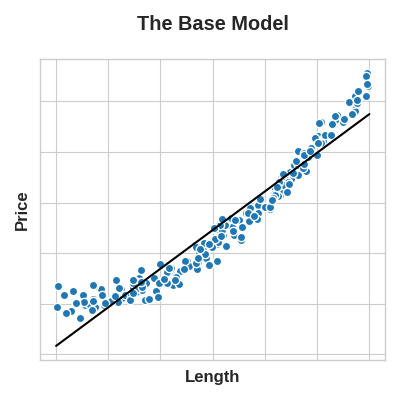
The goal of feature engineering is simply to make your data better suited to the problem at hand. Consider "apparent temperature" measures like the heat index and the wind chill. These quantities attempt to measure the perceived temperature to humans based on air temperature, humidity, and wind speed, things which we can measure directly. You could think of an apparent temperature as the result of a kind of feature engineering, an attempt to make the observed data more relevant to what we actually care about: how it actually feels outside!

You might perform feature engineering to:

* Improve a model’s predictive performance
* Reduce computational or data needs
* Improve interpretability of the results

A guiding principle of feature engineering:

* For a feature to be useful, it must have a relationship to the target that your model is able to learn. Linear models, for instance, are only able to learn linear relationships. So, when using a linear model, your goal is to transform the features to make their relationship to the target linear.
* The key idea here is that a transformation you apply to a feature becomes in essence a part of the model itself. Say you were trying to predict the Price of square plots of land from the Length of one side. Fitting a linear model directly to Length gives poor results: the relationship is not linear.



* If we square the Length feature to get 'Area', however, we create a linear relationship. Adding Area to the feature set means this linear model can now fit a parabola. Squaring a feature, in other words, gave the linear model the ability to fit squared features.
* This should show you why there can be such a high return on time invested in feature engineering. Whatever relationships your model can't learn, you can provide yourself through transformations. As you develop your feature set, think about what information your model could use to achieve its best performance.