MODULE #1

Final Project

Noah X. Deutsch
Self-Paced Data Science Program

Objective

Clean, explore, and model the King County House Sales dataset with a multivariate linear regression to predict the sale price of houses as accurately as possible.

Methodology (OSEMiN)

Obtain

* Import the data
* Familiarize myself with the data

Scrub

* Ensure Right Column Data Types
* Deal With NaNs
* Check For Multicollinearity
* Normalize Numerical Data
* One-Hot Encode Categorical Data

Explore

* Understand the Distribution (Hist + KDE, Joint Plots)
* Check the linearity assumption between predictors and target variable

Model

* Fit single reg models for each cont. Variable
* Fit reg models for each group of cat variables
* Fit multi-reg model and check Multicollinearity Normality, and Homoscedasticity
* Use recursive feat. elimination and cross val to protect against overfitting

iNterpret

* Understand the implications of the model and reflect

Key Challenges

* Some of our predictors did not have clear linear relationships with price.

* Many predictors violated the normality assumption.

* Outliers were common for many predictors. This was especially true with values where house ‘price’ was greater than ~$1M.

Necessary Adjustments

* The final model used bedrooms, bathrooms, sqft_living, view, sqft_living15 and 67 different zip-codes to predict the price value. Other predictors including waterfront, condition, grade,yr_built, yr_renovated, floors, lat and long were not included in the model for various reasons.

* The biggest adjustment by far was removing all data for houses with a price of over $1.25M. This was necessary to remove troublesome outliers and ultimately help our model more closely align with normality and homoscedasticity assumptions.

Result

Ultimately the final model was able to explain roughly 80% of the variation in the response variable around its mean, with high confidence, but only for houses under $1.25M in price.

Zip Codes and Square Foot Living were the two features that contributed most to housing prices.

Thank you!

Noah X. Deutsch