Project Luther - Predicting Desktop Prices

Luke Tibbott

Project Design

I scraped newegg.com using BeautifulSoup to retrieve data on desktop computers. My goal was to predict the price of computers based on their hardware components. This model can be used by businesses selling desktop computers to properly price their computers.

In my final product, my pipeline included feature engineering, standard scaling, and a 10-fold cross validated linear regression model. I tested Ridge regression, Lasso regression, and Elastic Net regression (all cross-validated), but found there were no real improvements to my model.

I also attempted to use Recursive Feature Elimination with Cross Validation (RFECV in scikit-learn), but did not get good results. No features were eliminated and I decided to proceed with OLS after exhausting my regularization and feature elimination methods.

Tools

BeautifulSoup4

Python

- Regular expressions (data cleaning)
- Pandas (data manipulation and storage)
- Scikit-learn (modeling)
- NumPy (data manipulation)
- Jupyter Notebook (workflow)
- Matplotlib (exploratory and explanatory data analysis)
- VSCode (learning to factor functions in to .py file)

Google Slides

Google Docs

Data

I scraped price and hardware component data from 4,600 computers listed for sale on newegg.com. In total I had 81 variables after scraping. After cleaning and dropping unusable data I had 1700 computers with 12 variables.

The data from newegg.com is the dirtiest data I've ever worked with. I was pulling multiple features from the same column using regular expressions. An example of a dirty column is something like this:

'Intel Core i7-8700K 3.70 GHz 6-core processor'

There are three pieces of information in this table entry alone: processor brand, processor speed, and number of processor cores.

My automated data cleaning procedures were not perfect -- there computers that didn't have data listed in the parts I expected, but I'm happy with how cleaning went considering how dirty the data was.

Model

I used Ordinary Least Squares linear regression with cross-validation in my final model. My pipeline included scikit-learn StandardScaling as well. I attempted to use PolynomialFeatures as well, but setting degree to anything higher than 1 resulted in overfitting issues, even when doing interaction only. This led to me using manual feature engineering rather than PolynomialFeatures to do polynomial transformations to features.

Ridge regression and Recursive Feature Elimination with Cross Validation were were fruitless efforts when it came to modeling. I did not include my attempts at using these modeling tools in my final product.

I log-transformed my response variables (price) in my final model to improve the homoscedasticity of my model, but I believe there are more things I could do in the future to achieve better homoscedasticity of the model.

Future Work

Feature engineering was the most fruitful pursuit in building my model. That being said, I think my model stands to improve the most through more feature engineering. I'm not happy with the heteroscedasticity of my residuals, and this is an area I think feature engineering could help improve.

I'd also like to collect data with more variables. Specifically, I'd like to get more graphics card data. The graphics in almost all of the computers in my data set are integrated, meaning there is no dedicated graphics card adding value to the computer. Graphics cards typically increase the price of computers by a large amount, so I think it would be a fun challenge to include this in my model.

Similarly, I'd like to get more brands in my dataset. For simplicity and consistency reasons, I only included HP, Lenovo, and Dell in my newegg.com query. Including more brands in my dataset would give my model a bigger picture of the market.

Appendix

These are the features I included in my final model:

Number of processor cores
Processor speed
Memory capacity
Storage capacity
(SSD capacity)^2
(Number of cores * speed)
(SSD or not * capacity)^2
DDR4 RAM or not
Presence of both DDR4 and SSD
Presence of both Nvidia GPU and SSD
Nvidia GPU or not
Integrated GPU or not