Craigslist Bike Price Prediction

Dan Weiss

Question

Can I predict the prices of bikes for sale on Craigslist based on features present in the listing?

Motivations:

- If I'm selling a bike on Craigslist, what price should I set?
- If I see a bike for sale, is the price fair?

Data

- 3000 most recent bikes listed on Craigslist Bay Area
  - Taken on 4/17/2018
- Scrape all aspects of the listing
  - Price
  - Listing text
  - Listing attributes
  - Number of Images
  - Area

reply ✉ ☎ prohibited [?] Posted about 12 hours ago print

2013 Cervelo S5 Cherry Condition/Custom Paint - $1000 (willow glen / cambrian)

image 1 of 6

(google map)

condition: **excellent**

make / manufacturer: **Cervelo**

model name / number: **S5**

size / dimensions: **56cm**

2013 Cervelo S5 56cm aero road bike. Custom metallic blue sparkle paint.

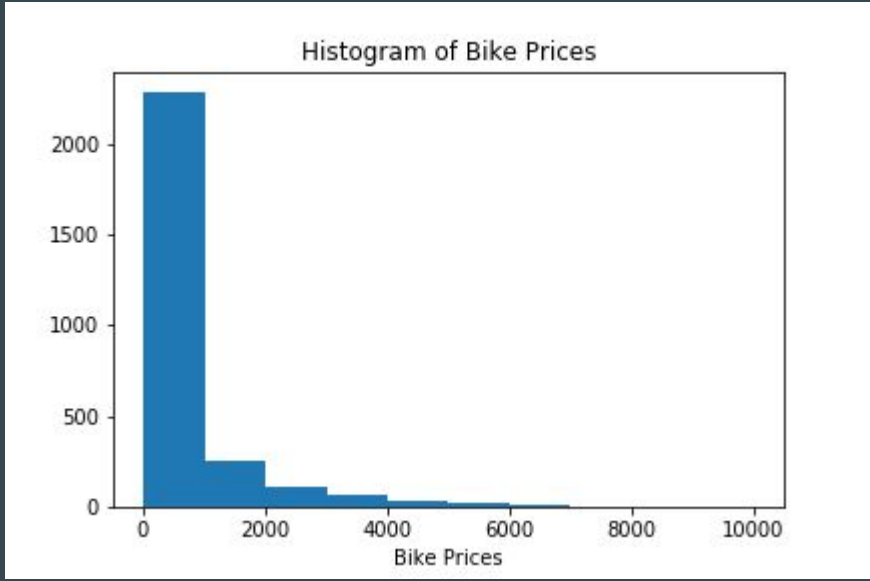
Full Ultegra 6800
3T Ergonova bar
Look Keo Blade Titanium pedals
Fulcrum Racing 5.5 wheels
Ultegra 11-25 cassette
Brand new Fizik Antares saddle
New Elite bottle cages

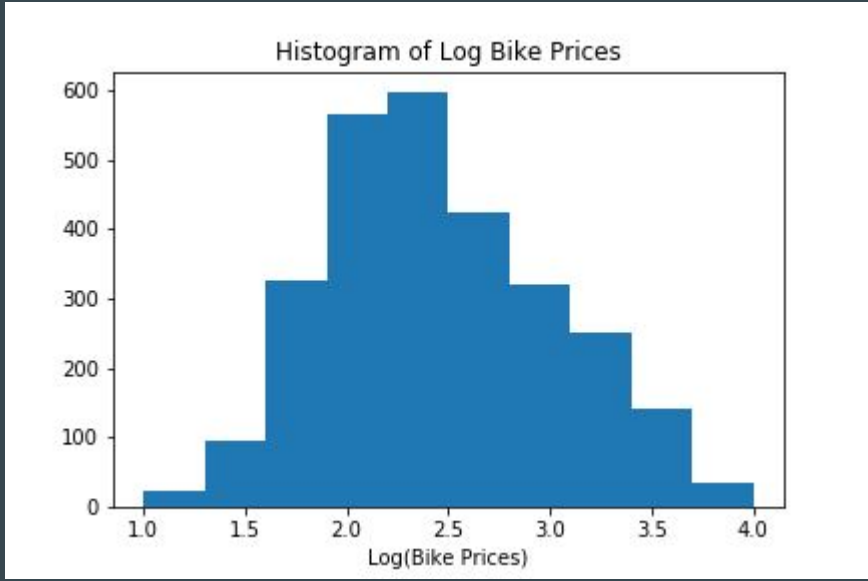
This bike rips. Stiff, super aero and uncompromisingly fast. Gorgeous custom paint by Joe's Carbon Solutions. You'll be the only person in the world with this Cervelo.

Data Cleaning

- Filter out non- bikes (e.g. Bike racks)
- Calculate the length of the post, number of images
- Search title and post to classify the bike as:
  - Kid’s
  - Electric
  - Having professional level, medium level, or entry level components

Transform Price Variable





Results

- Selected linear model using cross-validation
- R^2 on Holdout Data: 0.41
- RMSE on Holdout Data: ~ $967

Future Work

- Add Features: Classify Bikes by frame materials (e.g. Carbon)
- **Algorithms: Try** other model types (e.g. Random Forest)
- CV: Use computer vision on images to detect quality
- NLP: Natural Language Processing techniques
- Attempt to determine whether bikes sold or not

Thanks!

Link to slides: https://github.com/DanWeiss1/Luther/blob/master/Slides.pdf

Contact me: Dan Weiss

Email: dweiss89@gmail.com

Phone: 515-988-2603

Appendix

1. More details on model selection strategy
2. More info on features included
3. Details on Model Results

Model Selection

- Linear Models
  - Log Price = B0 + B*X
- Test various combinations of features
- Cross validate models on 75% of data (Training Set)
- Chose model with best average R^2 over 5- Fold CV
- Overfitting not an issue as R^2 on training and test data similar

Results Detailed

- Final Model Includes:
  - Categorical Variables:
    - Condition of the bike (New, Like New, Excellent, Good, Fair, N/A)
    - Location of the bike (North Bay, South Bay, East Bay, Peninsula, San Francisco)
    - Quality of Derailleur (Pro, Medium, Entry, N/A)
    - Kids Bike
    - Electric Bike
    - Size, Brand, and Model listed
  - Length of the Post (in words)
  - Number of Pictures

Model Results (Detailed)

Feature	coef	std err	t	P>|t|
Constant	2.1887	0.03	74.135	0
E- bike	0.3428	0.063	5.481	0
Kids bike	-0.5435	0.038	-14.378	0
Pro components	0.2834	0.025	11.417	0
Medium components	0.375	0.042	8.974	0
Entry components	0.0674	0.069	0.975	0.33
Number of Pictures	0.019	0.002	8.674	0
Length of Post	0.0001	2.09E-05	5.494	0
East Bay	-0.0971	0.028	-3.427	0.001
North Bay	-0.002	0.034	-0.057	0.954
Peninsula	0.0422	0.037	1.15	0.25
South Bay	-0.0594	0.032	-1.86	0.063
Santa Cruz	0.0442	0.043	1.033	0.302
Model Listed	0.1054	0.029	3.587	0
Size Listed	0.1701	0.027	6.289	0
Brand Listed	0.0719	0.03	2.374	0.018
Condition: Fair	-0.4765	0.073	-6.566	0
Condition: Good	-0.2495	0.031	-8.158	0
Condition: Excellent	-0.0297	0.026	-1.136	0.256
Condition: Like new	0.0846	0.035	2.448	0.014
Condition: New	0.008	0.047	0.172	0.863