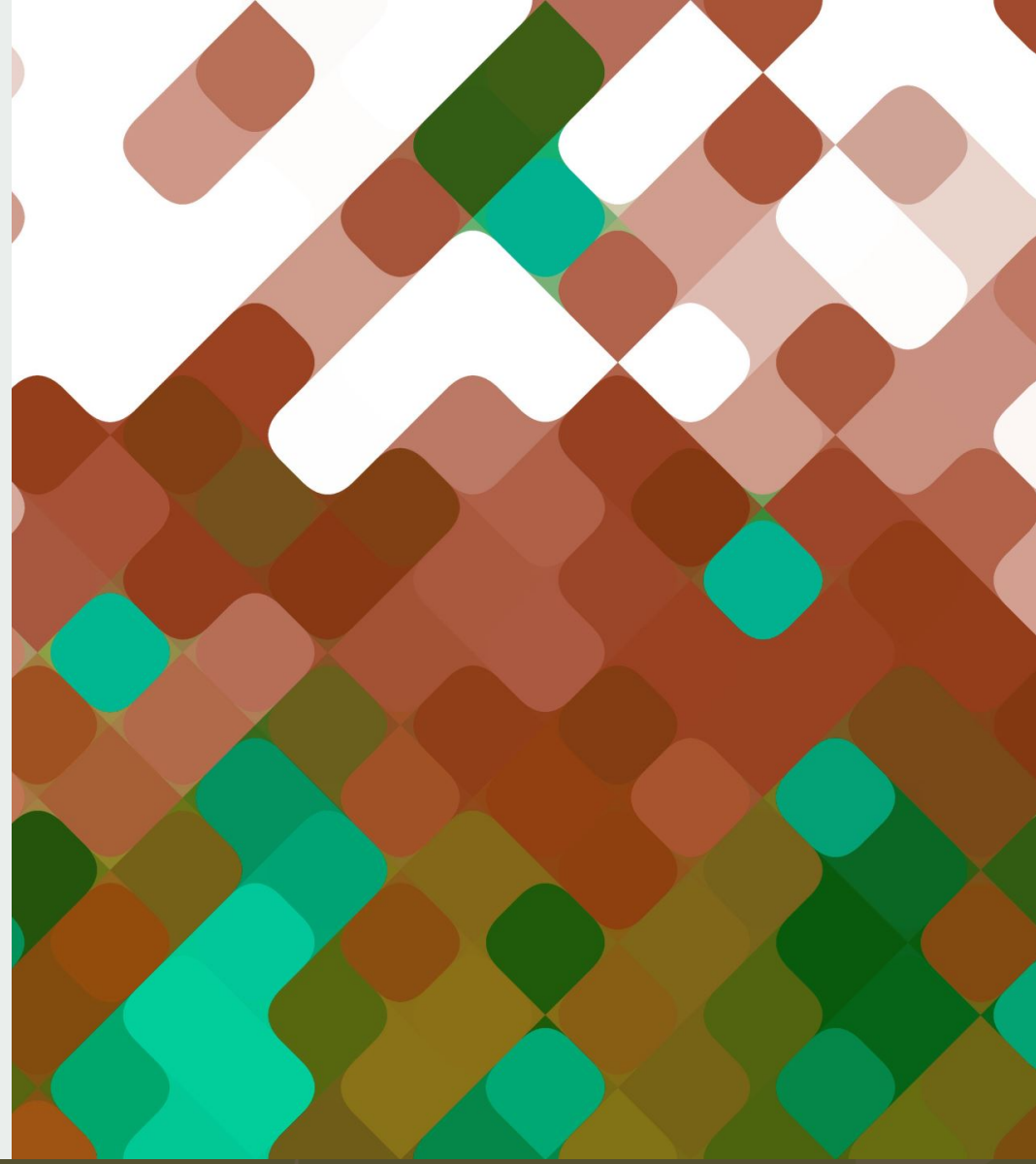
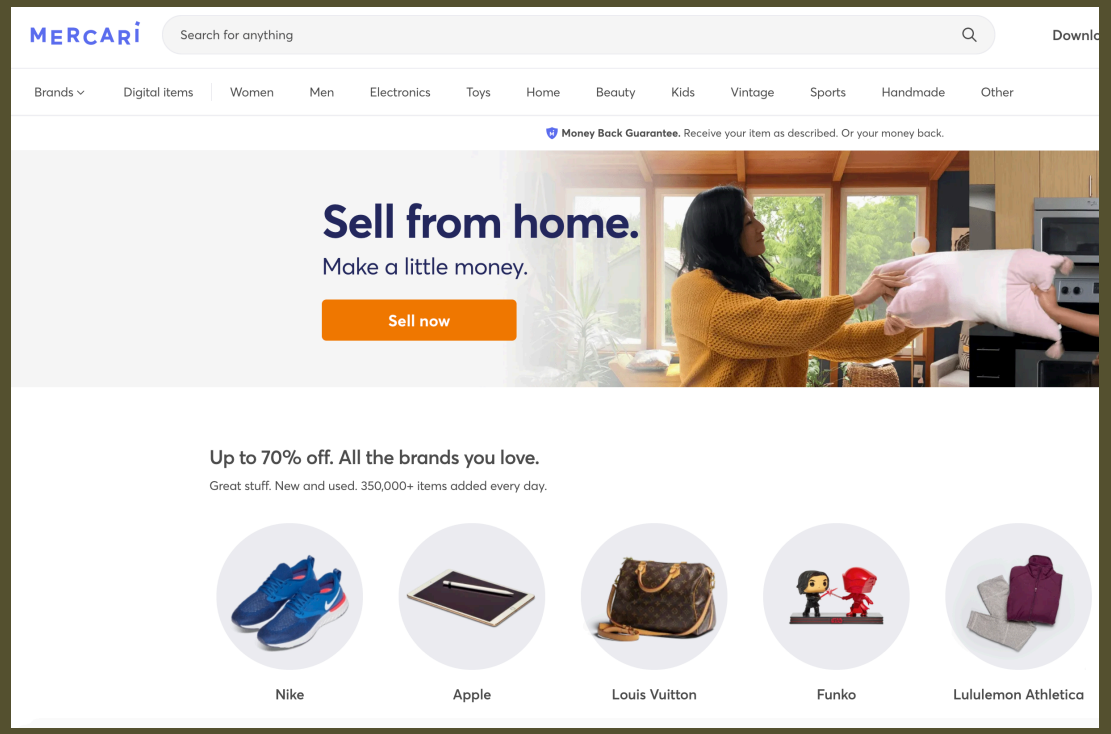
Predicting the Price of Online Shopping Listings on Mercari

By: Jason Zhou



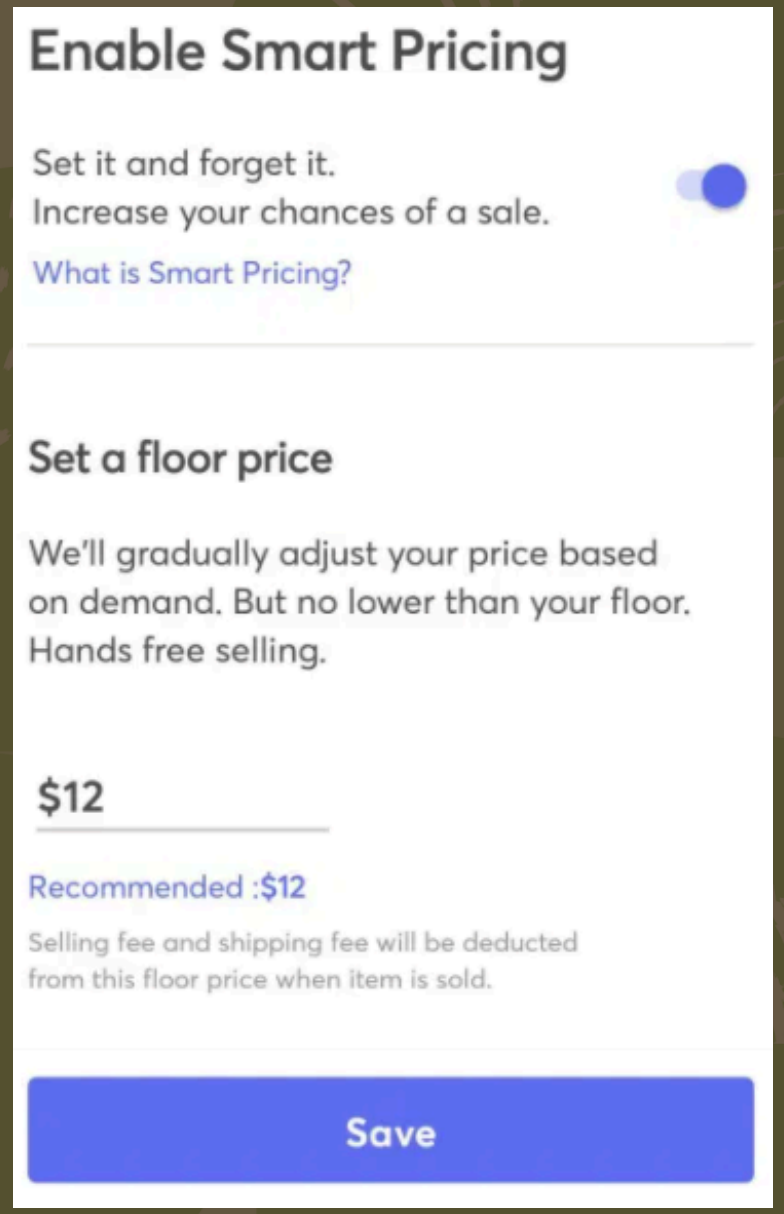
What is Mercari?

- Online marketplace platform
- Sellers are individuals, not corporations
- Requires that sellers ship out items to the buyers



The Problem

- Sellers often don’t know how much they should charge for their item
- Mercari has an algorithm in place to suggest a price based on demand
- Our job is to see if we can improve upon this



Stakeholders

- Mercari
- Sellers
- Buyers

What determines the price/value of an item?

- Item category
- Condition
- Brand
- Availability

Working Data Set

- 1,000,000+ Item listings
- Number of columns: 7
- Over 4000 unique Brands, 1000+ unique categories
- Found on Kaggle

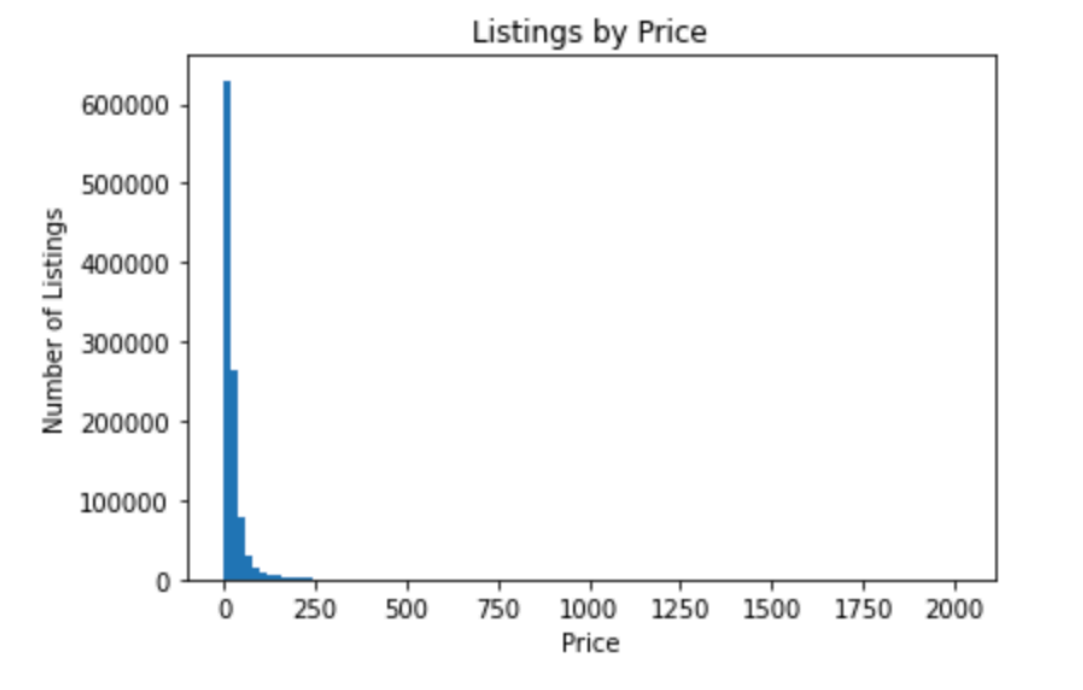
Data Fields

- Name
- Condition
- Category
- Brand
- Price
- Shipping
- Item Description

Data Cleaning Decisions

- Dropped all rows with missing values
- Dropped all rows with ‘Brand’s and ‘Category’s that occur in less than 50 instances in the data set
- Dropped every row that has a value of 0 in the ‘Price’ column
- Took the log of the ‘Price’ column to account for skewing
- Dropped the ‘Name’ and ‘Item Description’ columns

Price Distribution Before and After



Exploratory Data Analysis Findings

- Target: Price
- Features: Condition, Brand, Category, Shipping

- Correlations with Price: Brand, Category, Shipping
- No Correlations with Price: Condition

Preprocessing

- Had to take a 20% sample of our working data set due to computational constraints
- Target: Price
- Features: Brand, Category, Shipping
- 70/30 train/test split

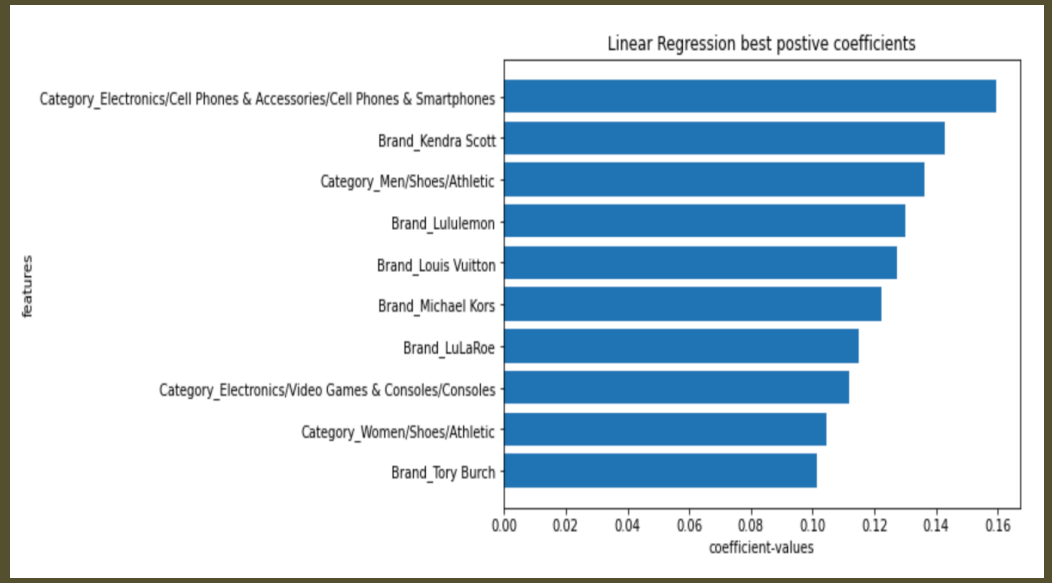
Model Selection

- Linear Regression
- Tuning:
  - Selecting K best neighbors

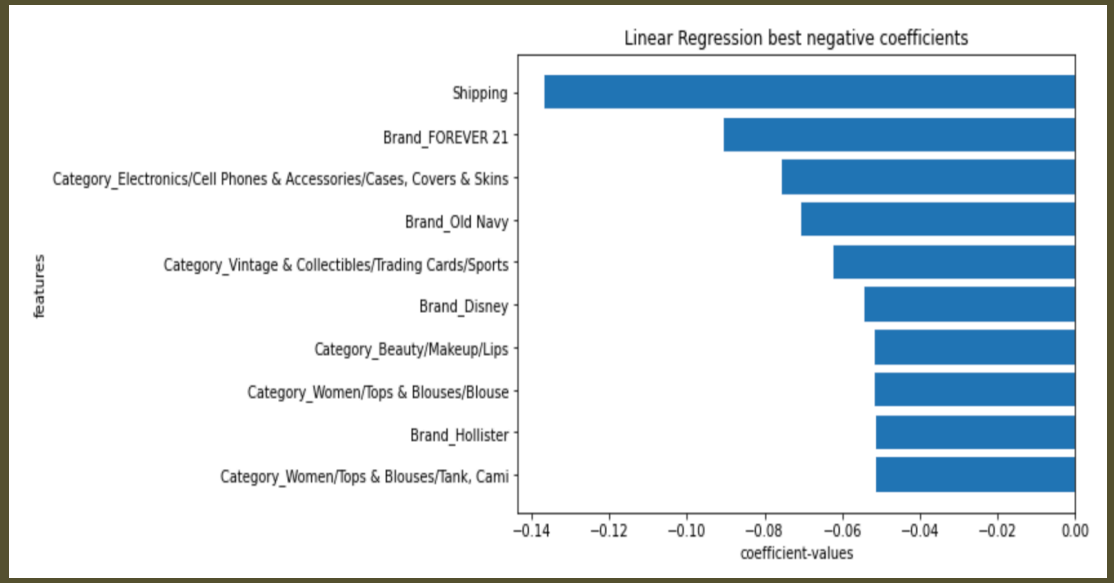
- Random Forest
- Tuning:
  - Increasing max_depth

Feature Analysis – Linear Regression

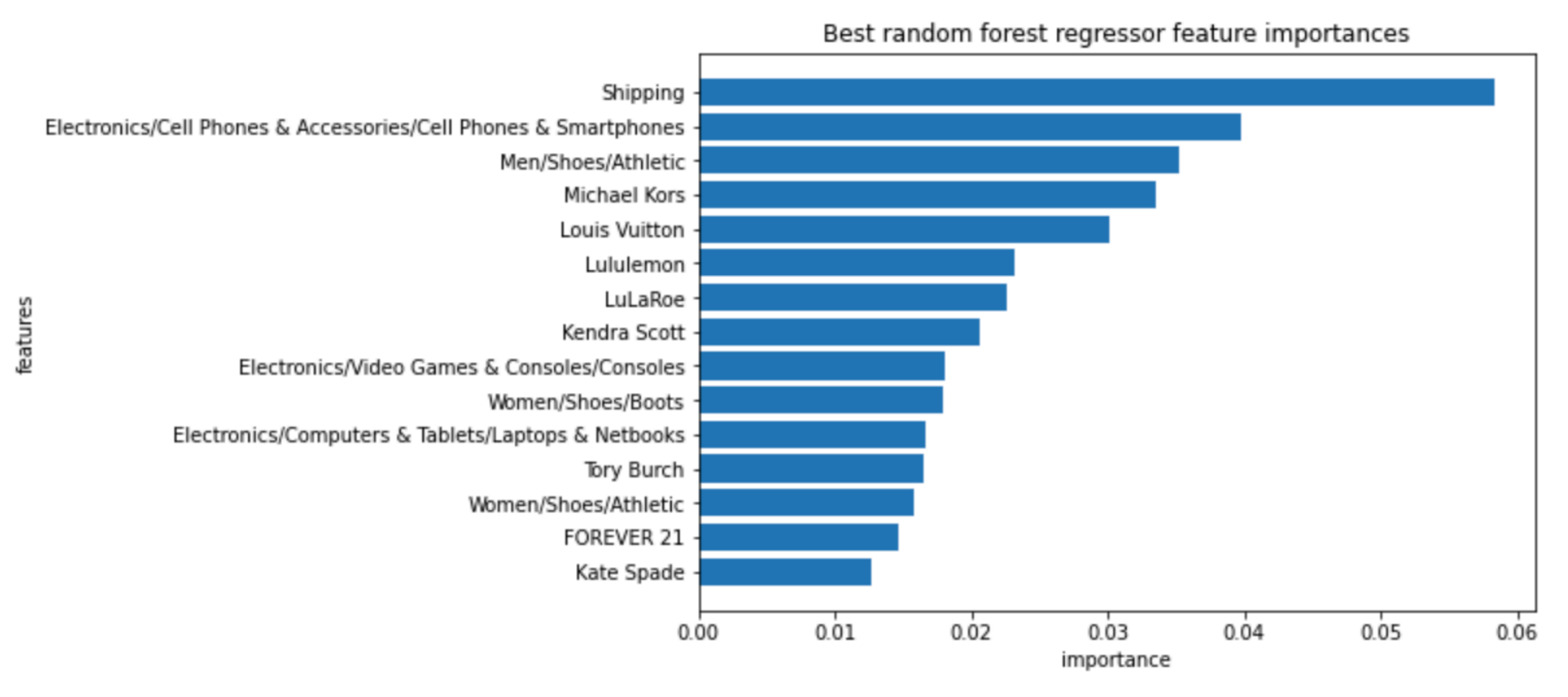
- Most positive coefficients



- Most negative coefficients



Feature Analysis – Random Forest



Final Model Evaluation

Model	Linear Regression	Random Forest
R score	0.4109	0.4584
Mean Absolute Error	0.6653	0.6333
Mean Absolute Percentage Error	16.3564	15.4824