Predictive Analytics

Which iPad listing will be sold on eBay?

Class Project: MITx Analytics Edge 15.071x

By Sarah Huang

Overview

- The project challenges students to develop an analytics model that will help buyers and sellers to predict the sales success of a set of eBay listings for Apple iPads from spring 2015.
- The competition is hosted on the Kaggle platform.
- The independent variables consist of 9 pieces of product data available at the time the iPad listing is posted, and a unique identifier.
- The training set consists of 1,861 observations and the testing set consists of 798 observations.
- The goal is to predict if a listing will be sold or not.
- The code is written in R and is available on my GitHub.

Data Fields

- The provided features includes the following:
  - Text description
  - Biddable (boolean)
  - Start price of auction if biddable, or the asking sales price if non-biddable
  - Condition of the product
  - Cellular (boolean). If the product is available to be connect to a cellular network
  - Carrier
  - Color
  - Storage
  - Product line

Modeling Approach

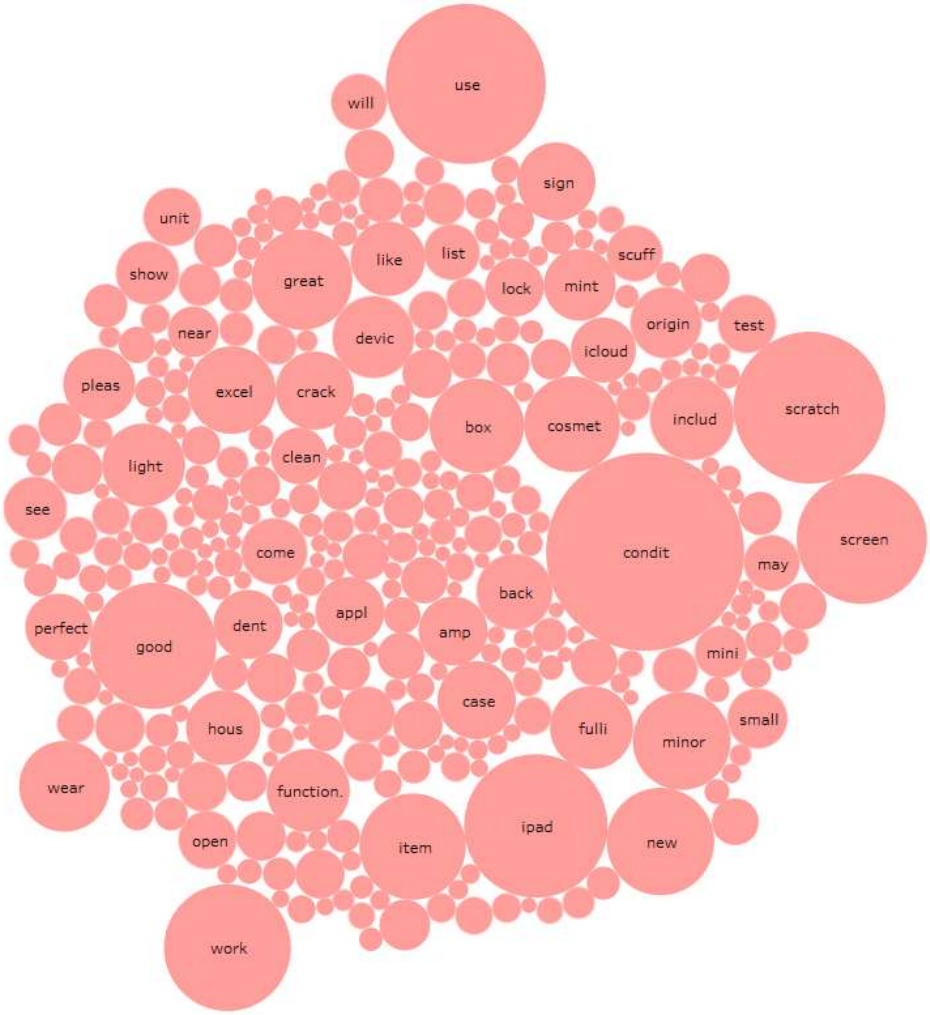
- The number one determinant of whether a product will be sold or not is price.
- The biddable items usually have start price that is far lower than the reasonable market price.
- One of the major features will be to compare the start price against a reasonable market price of the product. The assumption is that start price lower than the reasonable market price will be sold, while start price higher than reasonable market price will not be sold.
- To determine the reasonable market price, a smaller set of training set is created by filtering out observations that are non-biddable and are sold.
- An ensemble of glm (generalized linear model), random forest, gbm (generalized boosted model), and deep learning models are used to predict the reasonable market price of all observations using the above filtered set as training set. The dependent variable is the start price.
- The predicted price is used to create three features

Feature Engineering based on predicted price

- feature 1:
  - Take the difference between the predicted price and the start price
- feature 2:
  - Compute average predicted price based on grouping of product line and product condition
  - Take the difference between the above average predicted price and the start price
- feature 3:
  - Compute average predicted price based on grouping of product line, product condition, and storage.
  - Take the difference between the above average predicted price and the start price

Text Analytics

- The description field is converted to word vectors
- The data is preprocessed – stemming, remove numbers, remove punctuation, remove stop words, convert to lower case, etc.
- Sparse terms are removed with 0.999 threshold
- The resulting term-matrix has about 350 words
- A new feature that counts the number of terms for each observation is added



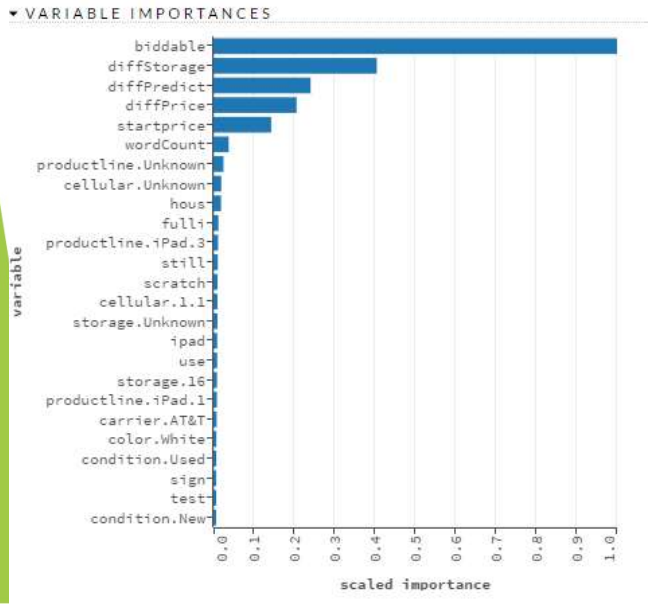
Model Performance

- An ensemble of models is used to predict if a product will be sold or not

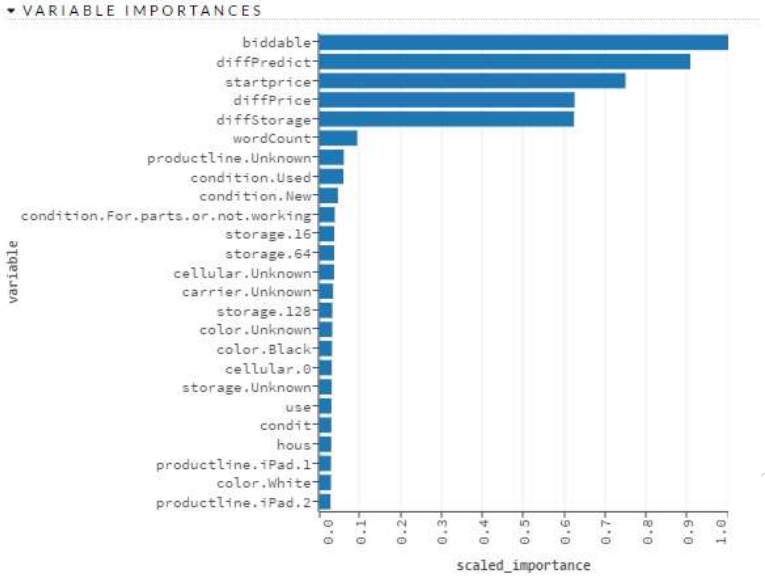
```
Base learner performance, sorted by specified metric:
                    learner       AUC
1          h2o.glm.wrapper 0.6965698
2 h2o.randomForest.wrapper 0.8851744
3          h2o.gbm.wrapper 0.9013953


H2O Ensemble Performance on <newdata>:
----------------
Family: binomial

Ensemble performance (AUC): 0.895813953488372
```

- The most important features are whether a product is biddable or not, and the price related features

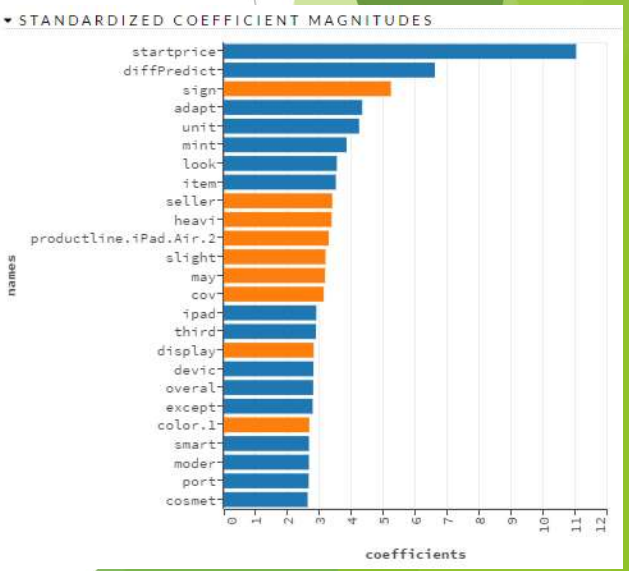
- GBM



- Random Forest

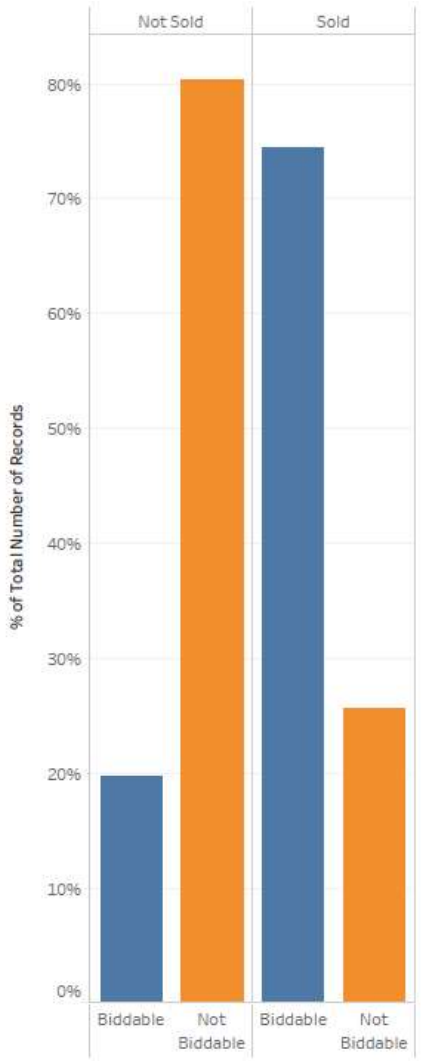


- GLM

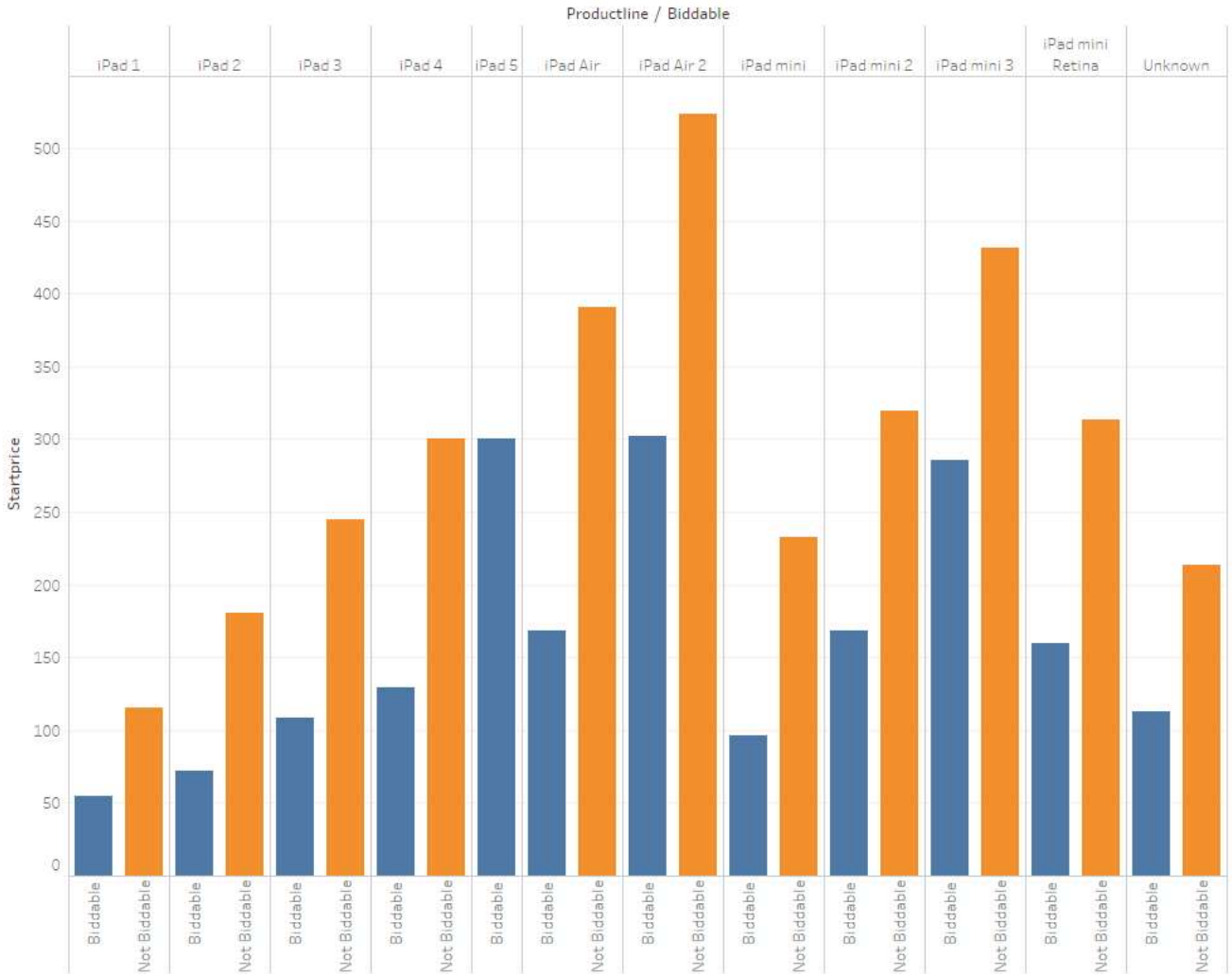


Biddable products are more sellable

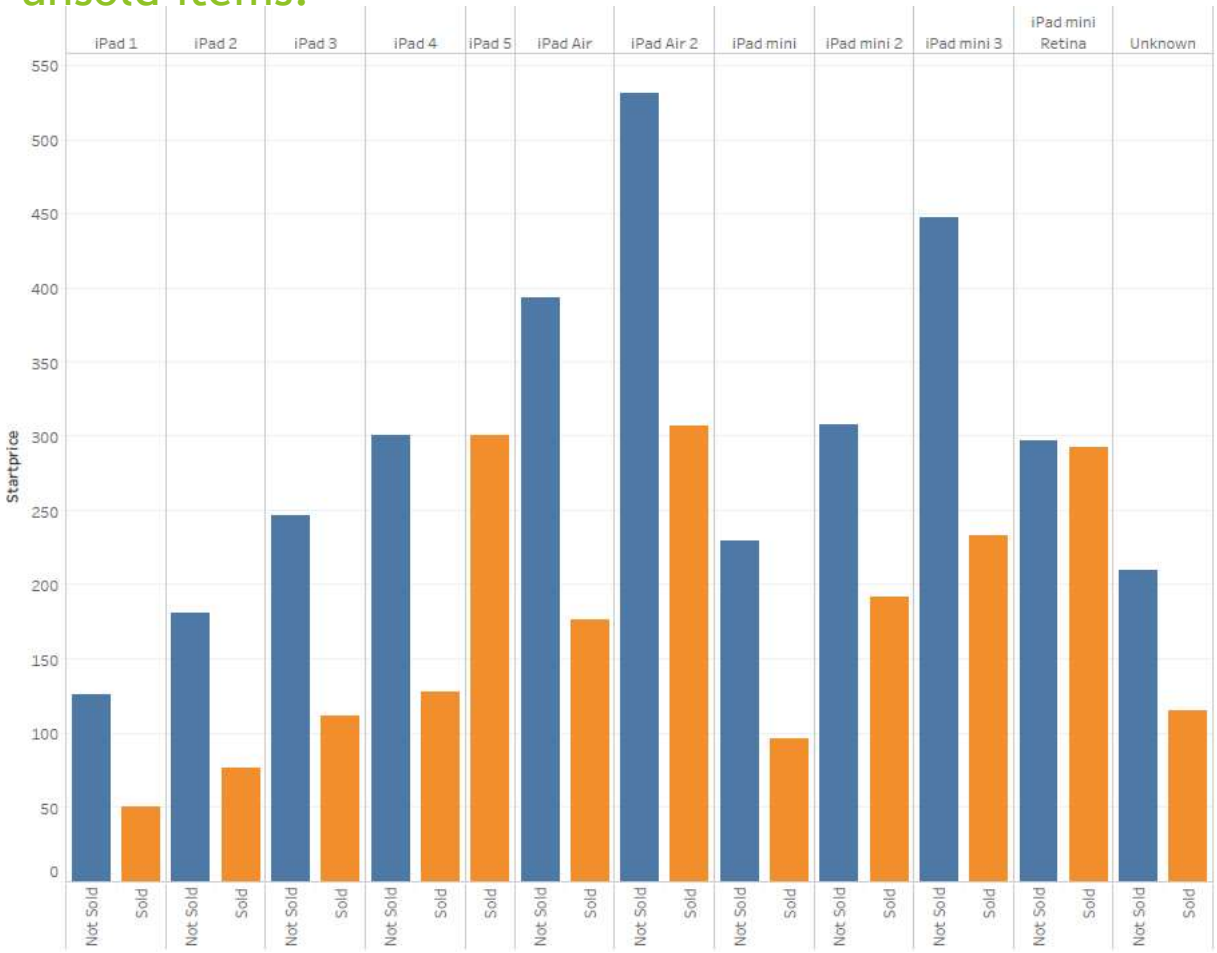
- Biddable products accounts for 74% of all sold products
- 80% of unsold products are non-biddable



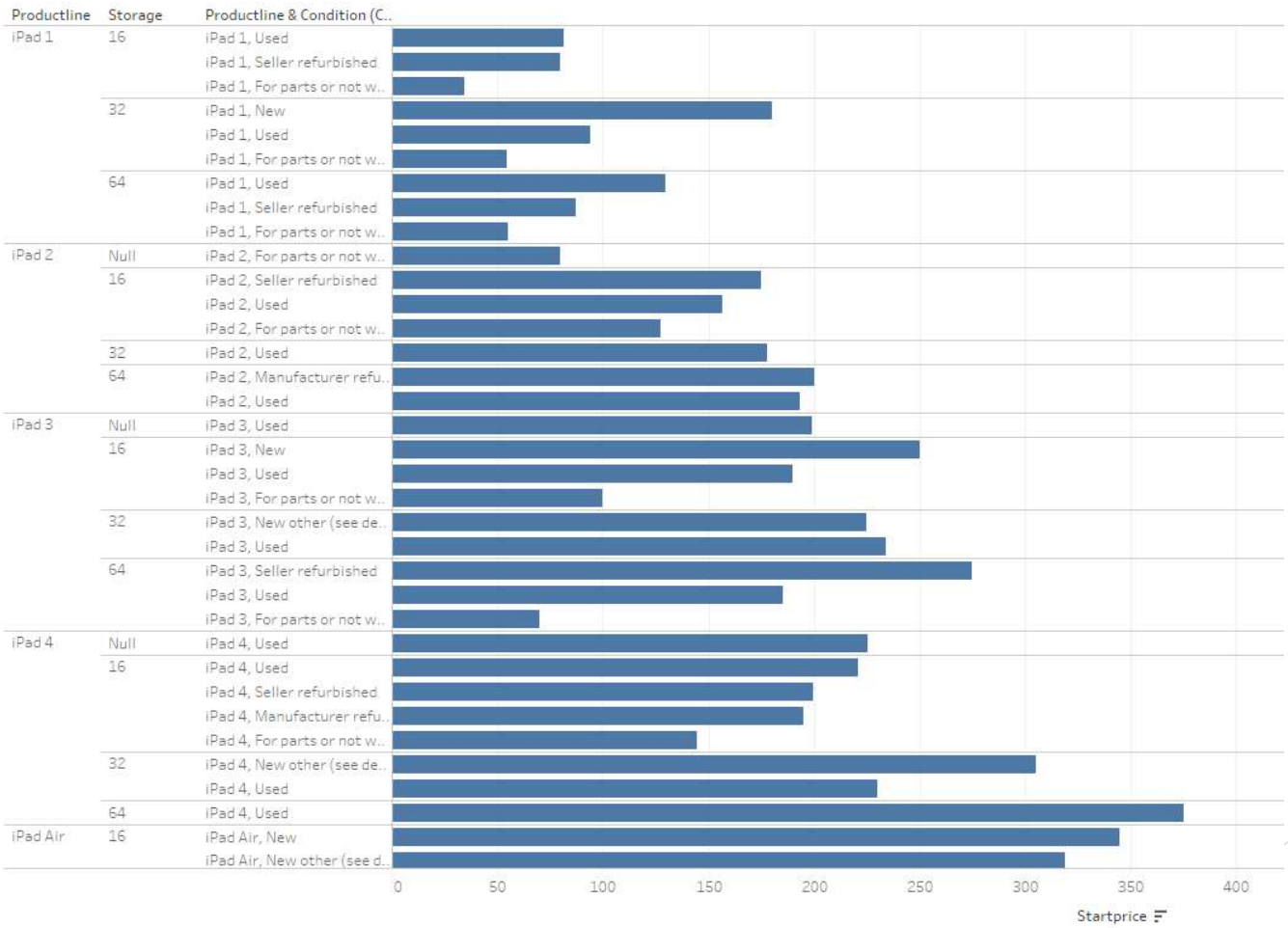
Biddable start prices are in general lower than non-biddable prices



Price is the primary determinant of whether a listing will be sold or not. The average start price of sold items are lower than unsold items.

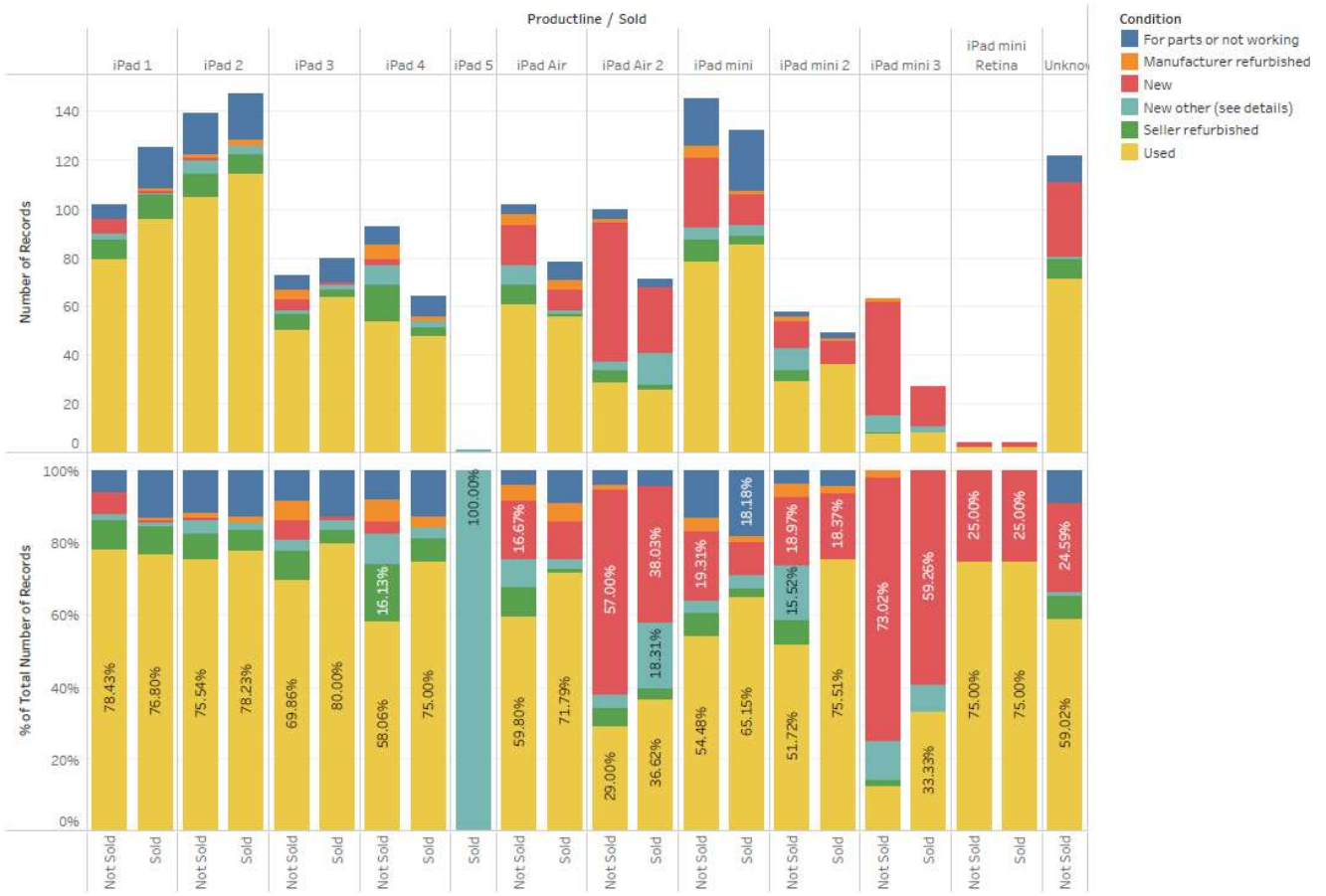


Average price varies depending on product line and condition of products



However, product condition itself does not determine if a product will be sold or not.

- % of sold and unsold products are similar regardless of product conditions.



Price is the key

Products that are deemed expensive based on predicted value are mostly not sold (in blue), and products deemed cheap by prediction are mostly sold (in orange).