STAT 627 Project Presentation

Ryan Heslin

April 23, 2022

Data and Problem Definition

I used a recent Tidy Tuesday dataset of chocolate bar rating, available at https://github.com/rfordatascience/tidytuesday/blob/master/data/2022/2022-01-18/readme.md.

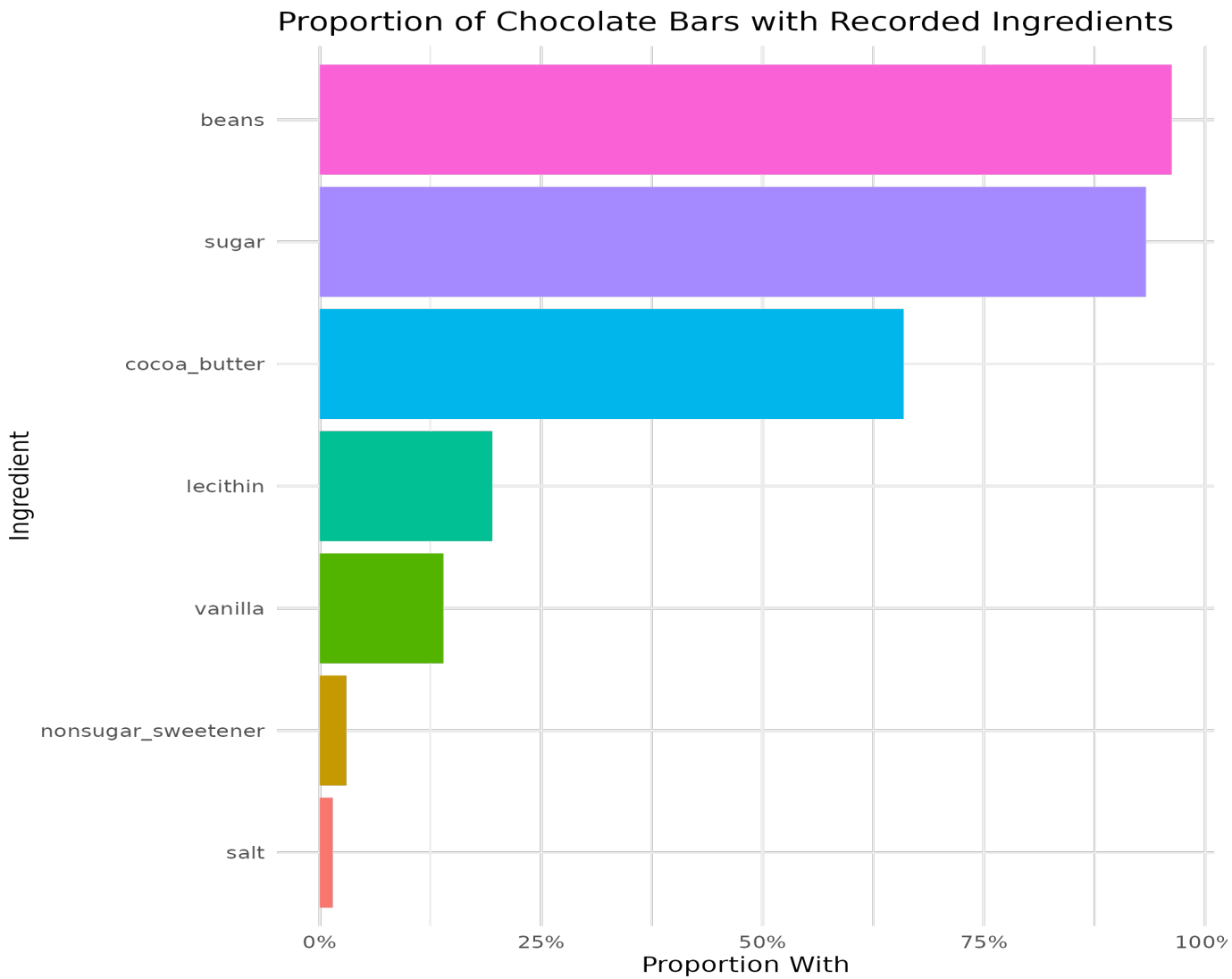
- Data about chocolate bars, with ratings from taste-testers (the response)
- 2530 observations
- Ratings on scale from 1 to 4 inclusive, in increments of 0.25

Available Features

Available features include:

- `company_manufacturer`: Chocolate manufacturer
- `company_location`: Country of chocolate manufacturer.
- `country_of_bean_origin`: Country of cacao bean used in chocolate.
- `rating`: Taste-tester's rating of the chocolate bar, on a scale of 1 to 4 in increments of 0.25. The response.
- `ingredients`: Character vectors with abbreviations for the presence of six ingredients: cacao beans, cocoa butter, lecithin, sugar, salt, nonsugar sweetener, and vanilla. I transformed these into dummy columns.
- `characteristic`: Words and phrases used by taste-testers to describe the chocolate. I split this into five columns, one for each word.
- `review_date`: Year of review.
- `cocoa_percent`: Percentage of chocolate that is cocoa

Chocolate by Included Ingredient



Most Common Chocolate Descriptors

The words most commonly used to describe chocolate bars described flavors (e.g., “creamy”, “cocoa”).

sweet	nutty	cocoa	roasty	earthy	creamy	sandy	fatty
273	261	252	213	190	188	170	166

Preprocessing

- I observed that the most common words in the text descriptions referred to flavors or ingredients
- I created dummy columns indicating the presence of the 10 most common of these
- I also created dummies for country of bean origin and company, year of rating, and the ingredients columns
- Also computed text sentiment for words recorded for each bar and used as a feature
- Defined custom metric measuring prediction error as rounded number of increments on the rating scale from the true value

Modeling Strategy

- Fit initial linear model with the following features:
  - Country of origin
  - Company of origin
  - Ingredients present
  - Review date
  - Presence of most common flavor words
  - Summed sentiment of phrases used to describe chocolate
- Then use this recipe to fit random forest and xgboost models and compare results

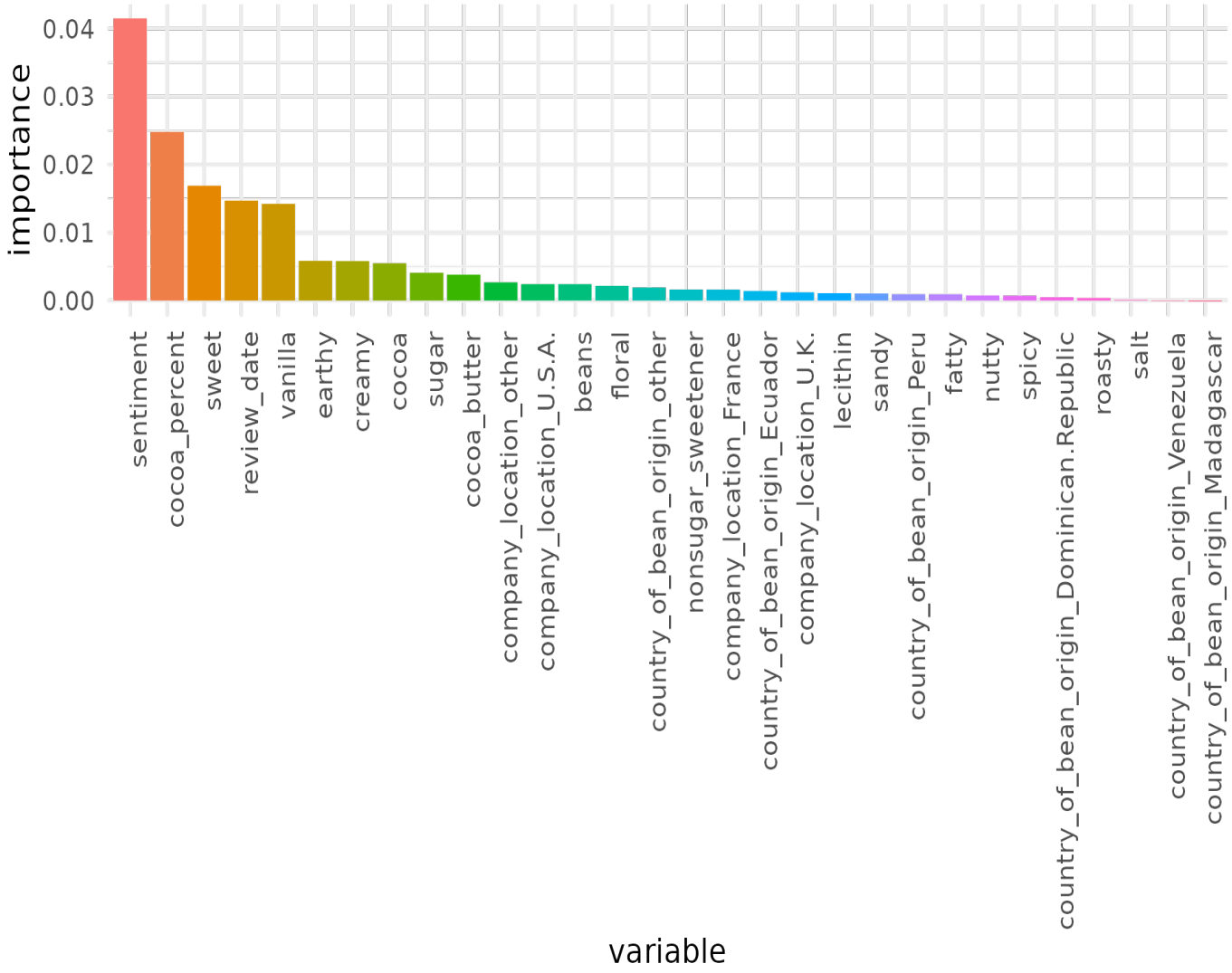
Results

- Linear regression and random forest performed about the same, with an interval error of about 1.25
- xgboost performed badly
- Sentiment feature, some ingredient words, and some dummies for country of bean origin proved most important
- Many dummies contributed little to models
- Biggest errors on low-rated bars

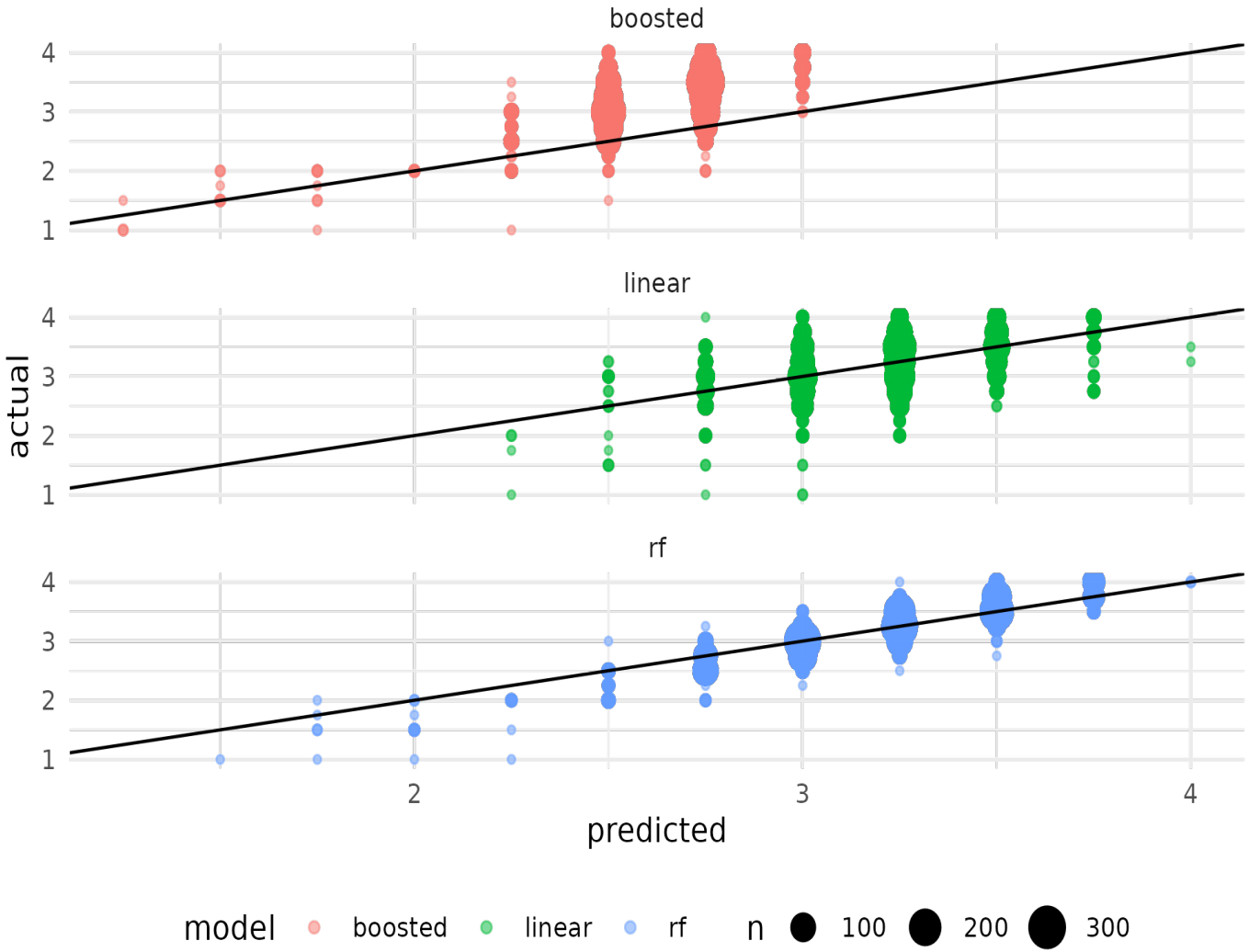
Overall Test Metrics

model	.metric	.estimator	.estimate
linear	mae	standard	0.3079662
linear	rmse	standard	0.3970624
linear	interval_error	standard	1.2211690
linear	rsq	standard	0.2176920
rf	mae	standard	0.3134230
rf	rmse	standard	0.4021690
rf	interval_error	standard	1.2527646
rf	rsq	standard	0.2092209
boost	mae	standard	0.5831447
boost	rmse	standard	0.6752469
boost	interval_error	standard	2.3270142
boost	rsq	standard	0.2039640

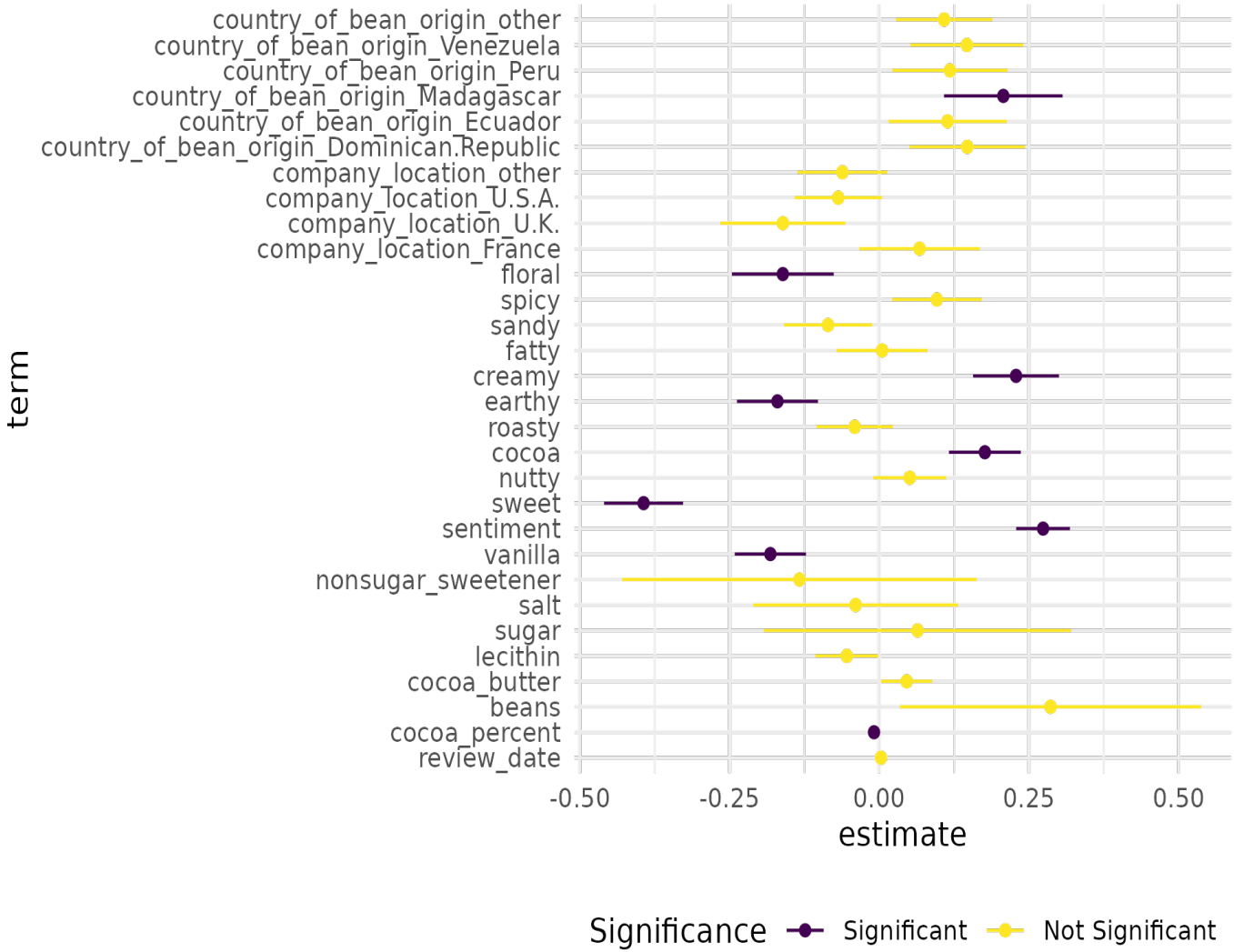
Random Forest Variable Importances



Test Fitted-Actual Plot



Comparison of Regression Coefficients



Self-Assessment

- Did well:
  - Made good use of text features
  - Defined interesting custom metric
  - Achieved reasonably good performance
- Should have done better:
  - Not using a list column to store text feature
  - Not considering feature selection (PCA, lasso, etc.)

For comparison: Julia Silge's model on the same data (https://juliasilge.com/blog/chocolate-ratings/) achieved an R^2 of 0.348 and an RMSE of 0.38