# *Classifying Chocolate*

Laura Chen

# Agenda

Objective: As aspiring chocolatiers, we want to create the most delicious bars of chocolate possible.

1. Overview of variables
2. Model selection process
3. Predictions on unseen data
4. Next Steps

# Balancing the Classes

**Total observations:** 1,852

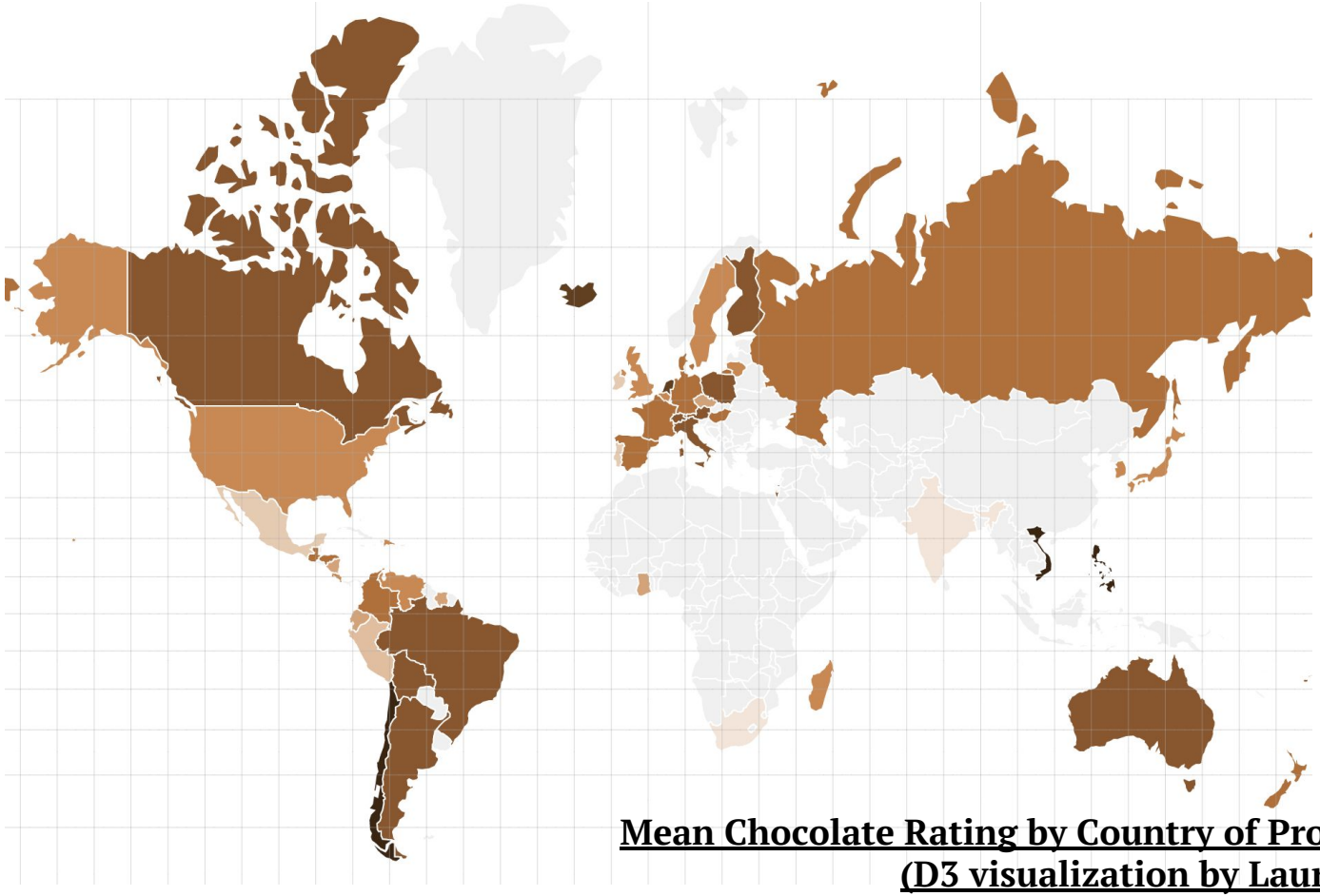
# Feature Selection

Selected Features:

- Cocoa Percentage
- Cocoa Bean Type
- Primary Region of Origin
- Country of Production
- Locally Produced
- Single Origin vs Multi

Other Considerations:

- Chocolate brands - good predictor, but not very informative
- Polynomial features - did not provide a significant difference



**Mean Chocolate Rating by Country of Production**
**(D3 visualization by Laura Chen)**

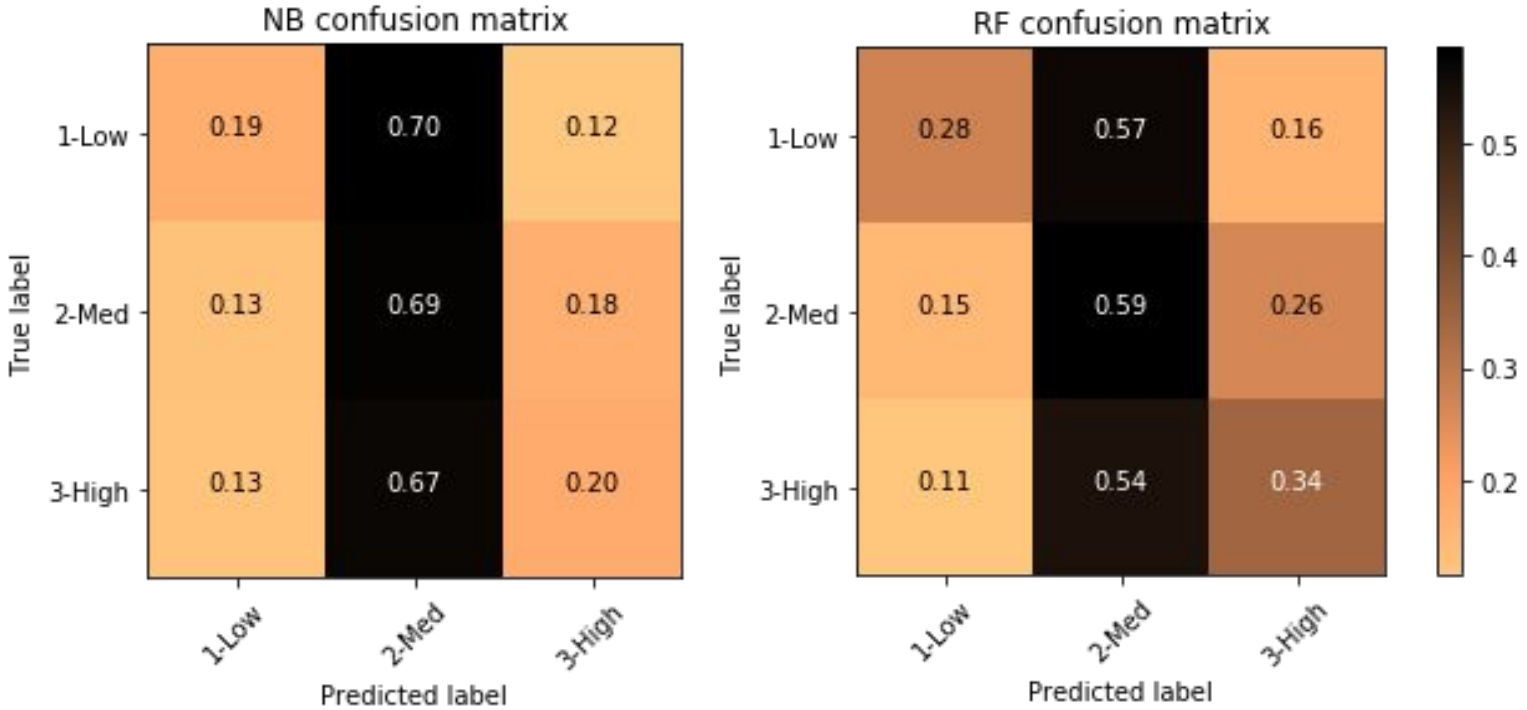
# Model Selection

Sorting by Accuracy vs F1 Score

| | CV Accuracy |
|---|---|
| SVC | 0.501623 |
| KNN | 0.499452 |
| Logistic Regression | 0.453389 |
| Random Forest | 0.436670 |
| Multinomial NB | 0.418159 |

| | CV F1 |
|---|---|
| Random Forest | 0.335353 |
| Multinomial NB | 0.271440 |
| KNN | 0.265806 |
| Logistic Regression | 0.250153 |
| SVC | 0.222703 |

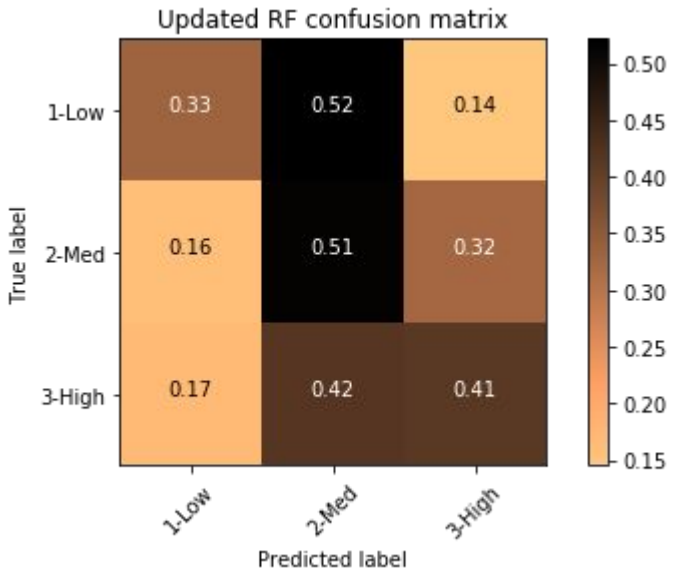
# Comparing Confusion Matrices



# Model Performance after tuning parameters

## Random Forest Classification Report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1-Low | 0.22 | 0.33 | 0.27 | 69 |
| 2-Med | 0.52 | 0.51 | 0.52 | 268 |
| 3-High | 0.48 | 0.41 | 0.44 | 218 |
| avg / total | 0.47 | 0.45 | 0.46 | 555 |



# Feature Importance

**Best predictors for a High rating:**

- **Bean Type:** Blend of Criollo and Trinitario
- **Production Location:** France
- **Primary Cocoa Bean Source:** Sub-Saharan Africa
- **Cocoa Percentage:** ~70%

**Prediction:**

80% probability of High rating

# Case Study: Lindt Chocolate

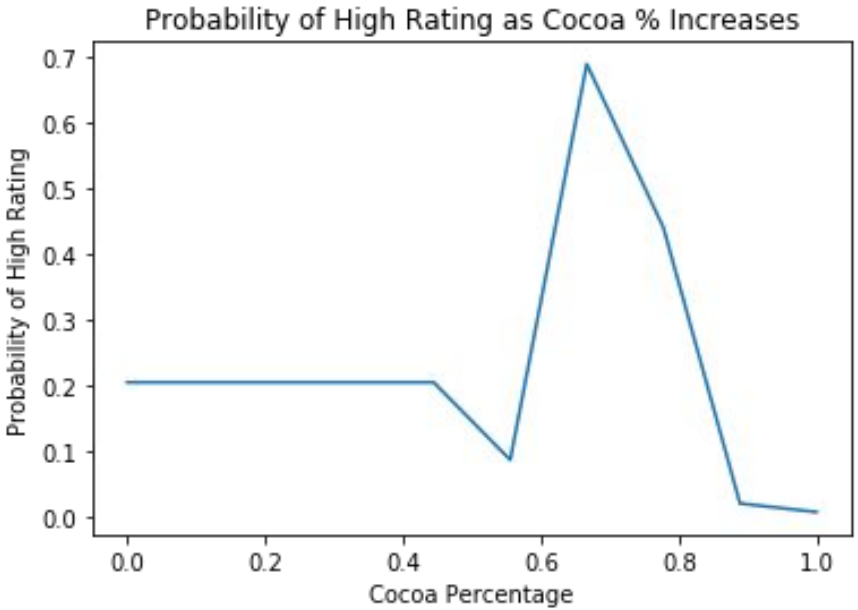
**Lindt Chocolates**

Key Features:

- Cocoa Percentage:
  - Dark: 43%
- Cocoa Bean: Blend
- Bean Origin: Latin America
- Production Facility: North America

**Prediction:** Low Rating;

Low: 55% , Medium: 32%, High: 14%

# Next Steps

- Additional data of interest:
  - Consumer preferences
  - Revenue by chocolate company
  - Ratios of other ingredients such as sugar, vanilla, emulsifiers
  - Agricultural practices: organic, fair trade, etc.
  - Processing techniques: grinding, roasting, chemically treating, etc.

*Thank you!*

# Classifications of Chocolate

**According to the Manhattan Chocolate Society:**

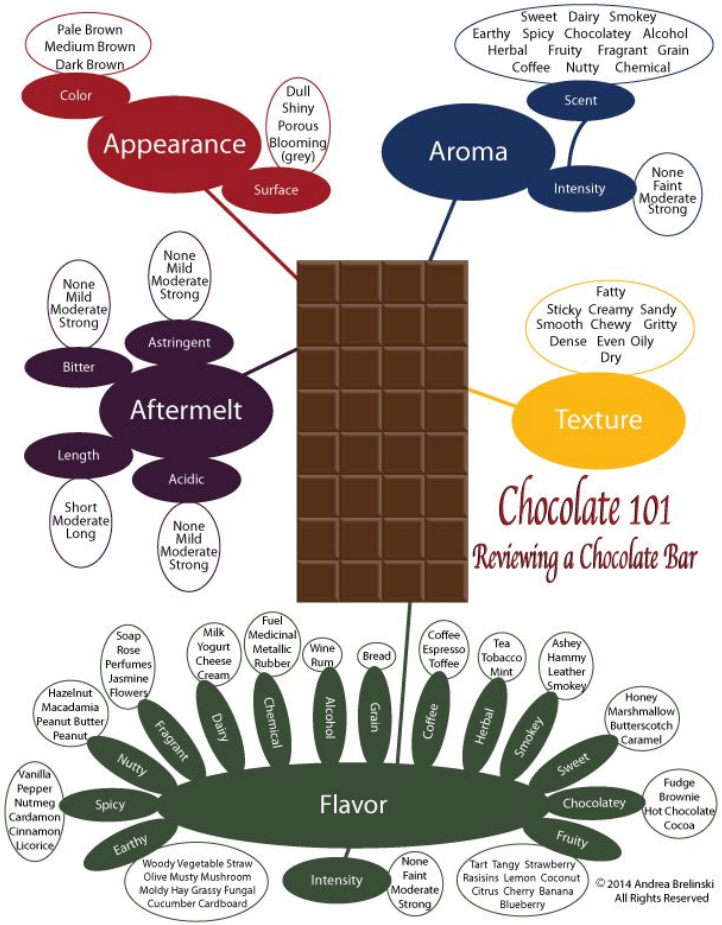
5 = Elite (Transcending beyond the ordinary limits)

4 = Premium (Superior flavor development, character and style)

3 = Satisfactory(3.0) to praiseworthy(3.75) (well made with special qualities)

2 = Disappointing (Passable but contains at least one significant flaw)

1 = Unpleasant (mostly unpalatable)




# Chocolate Review Guidelines

# Before and After Upsampling

**Before Upsampling:**

RF Accuracy: 0.467

RF CV Score 0.382096868959

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1-Low | 0.43 | 0.13 | 0.20 | 69 |
| 2-Med | 0.48 | 0.64 | 0.55 | 268 |
| 3-High | 0.44 | 0.36 | 0.40 | 218 |
| avg / total | 0.46 | 0.47 | 0.45 | 555 |

**After Upsampling:**

RF Accuracy: 0.459

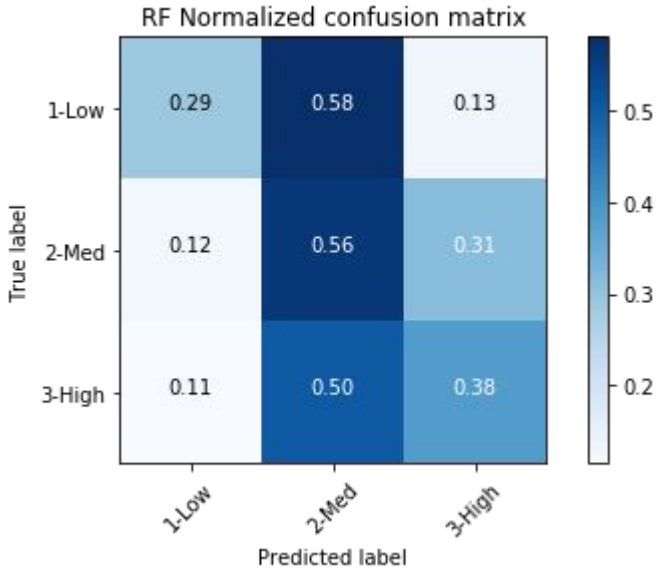
RF CV Score 0.388630043087

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1-Low | 0.25 | 0.30 | 0.27 | 69 |
| 2-Med | 0.50 | 0.58 | 0.54 | 268 |
| 3-High | 0.49 | 0.36 | 0.42 | 218 |
| avg / total | 0.47 | 0.46 | 0.46 | 555 |

# Polynomial Features

RF Accuracy: 0.467
RF CV Score 0.390222698154

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1-Low | 0.32 | 0.26 | 0.29 | 69 |
| 2-Med | 0.49 | 0.61 | 0.54 | 268 |
| 3-High | 0.48 | 0.36 | 0.41 | 218 |
| avg / total | 0.46 | 0.47 | 0.46 | 555 |



# Results from Downsampling Medium Category