Team Wine

presents

THE PRICE IS RIGHT

Wine Price Analysis

Yasmine Aitouny, Daniel Carmona, Pooja Nagrecha, Terisha Kolencherry

Project Objective

Given the wine type, country, province, winery, designation, and variety of a bottle of wine, determine the estimated price range.

Data

BACKGROUND, CLEANING, AND EXPLORATION

Data Overview

- CSV titled Wine Reviews dated from 2017 that was derived from the wine review website, Wine Enthusiast.
- The dataset contains information on the countries of production, price, review points of red and white wine from different wineries around the world.

	country	description	designation	points	price	province	region_1	region_2	variety	winery
0	US	This tremendous 100% varietal wine hails from ...	Martha's Vineyard	96	235.0	California	Napa Valley	Napa	Cabernet Sauvignon	Heitz
1	Spain	Ripe aromas of fig, blackberry and cassis are ...	Carodorum Selección Especial Reserva	96	110.0	Northern Spain	Toro	Toro	Tinta de Toro	Bodega Carmen Rodríguez
2	US	Mac Watson honors the memory of a wine once ma...	Special Selected Late Harvest	96	90.0	California	Knights Valley	Sonoma	Sauvignon Blanc	Macauley
3	US	This spent 20 months in 30% new French oak, an...	Reserve	96	65.0	Oregon	Willamette Valley	Willamette Valley	Pinot Noir	Ponzi
4	France	This is the top wine from La Bégude, named aft...	La Brûlade	95	66.0	Provence	Bandol	Bandol	Provence red blend	Domaine de la Bégude

Data Cleaning

- Got rid of null values in designation and region columns by filling in with proxy information
- Manually created red vs white lookup table and dummied - carried over from Project One

	country	description	designation	points	price	province	region_1	region_2	variety	winery	Red?	wineType_encoded
0	US	This tremendous 100% varietal wine hails from ...	Martha's Vineyard	96	235.0	California	Napa Valley	Napa	Cabernet Sauvignon	Heitz	True	1
1	Spain	Ripe aromas of fig, blackberry and cassis are ...	Carodorum Selección Especial Reserva	96	110.0	Northern Spain	Toro	Toro	Tinta de Toro	Bodega Carmen Rodríguez	True	1
2	US	Mac Watson honors the memory of a wine once ma...	Special Selected Late Harvest	96	90.0	California	Knights Valley	Sonoma	Sauvignon Blanc	Macauley	False	0
3	US	This spent 20 months in 30% new French oak, an...	Reserve	96	65.0	Oregon	Willamette Valley	Willamette Valley	Pinot Noir	Ponzi	True	1
4	Spain	Deep, dense and pure from the opening bell, th...	Numanthia	95	73.0	Northern Spain	Toro	Toro	Tinta de Toro	Numanthia	True	1

Exploratory Data Analysis

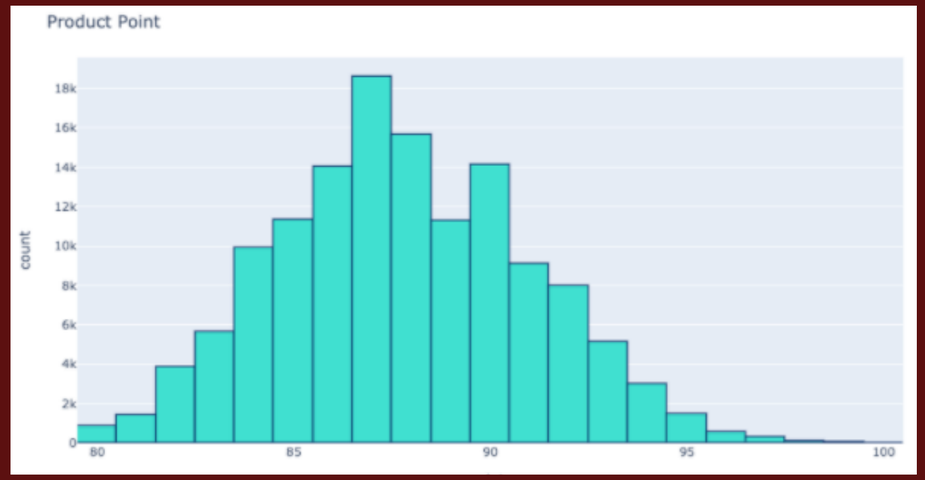
- Used Tableau to further our EDA from Project One
- Top tasted wines were Chardonnay and Pinot Noir
- The best variety of wine based on points was Nebbiolo
- In the United States the most expensive wine was in Nevada, while the best quality wine was from Washington

Natural Language Processing

PROCESS AND TAKEAWAYS

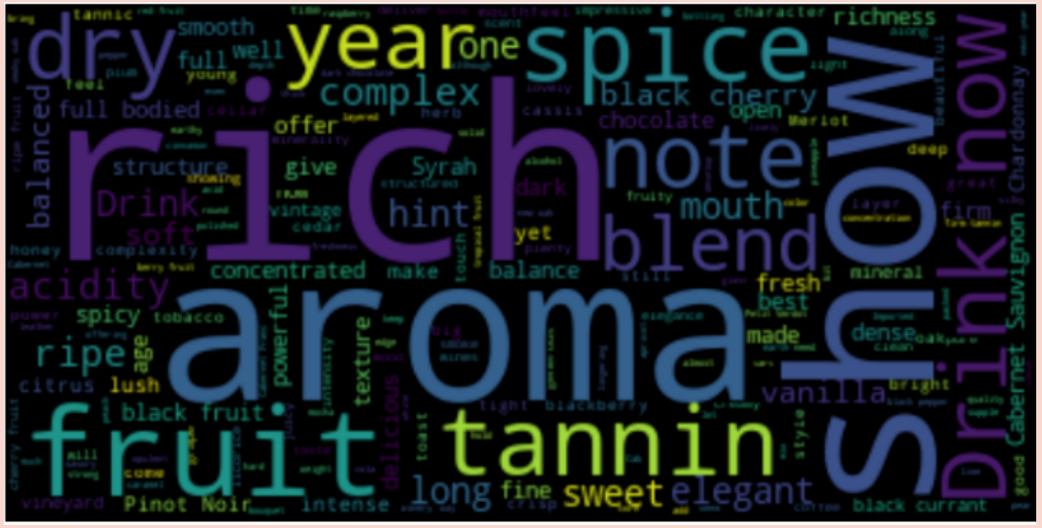
Data Preparation

- NLP model focuses on Good and Excellent Wine - first we calculated the number of wines in each category and the total number of annotations
- Reviews with points greater than or equal to 90 were classified as 1 (Excellent)
- Reviews with points less than 90 were classified as 0 (Good)



Sentiment Calculation

- We removed common words that appear in both sets of data such as "wine" and "flavours"
- Additionally, removed the word "rich", which appeared frequently in the Excellent dataset, but not in the Good dataset
- Looking across the reviews, the frequency of Good sentiments is double that of Excellent sentiments



Sentiment Model

1. REMOVED PUNCTUATION AND TRANSFORMED TO ALL LOWERCASE
2. SPLIT DATAFRAME WITH REVIEWED TEXT DATA AND TARGET VARIABLE
3. SPLIT DATA INTO TRAIN AND TEST SETS
4. UTILIZE BAG OF WORDS MODEL TO GET WORD FREQUENCY
5. THEN USED LOGISTIC REGRESSION MODEL TO PREDICT EXCELLENT OR GOOD

NOTE: DUE TO TIME CONSTRAINTS DIDN'T MAKE IT INTO THE DEPLOYED MODEL

Machine Learning

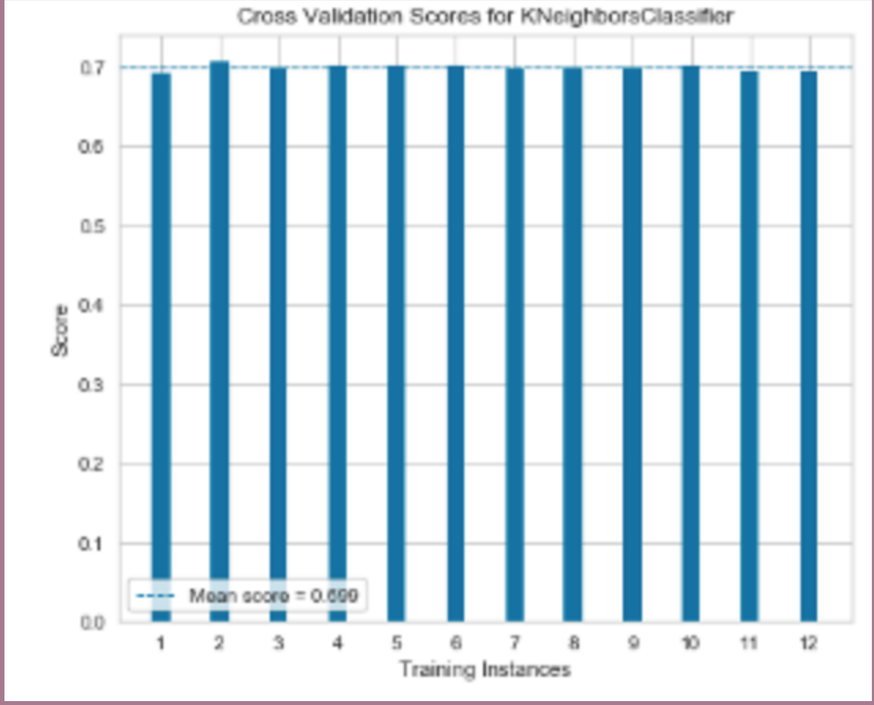
DATA PREPPING & MODEL SELECTION

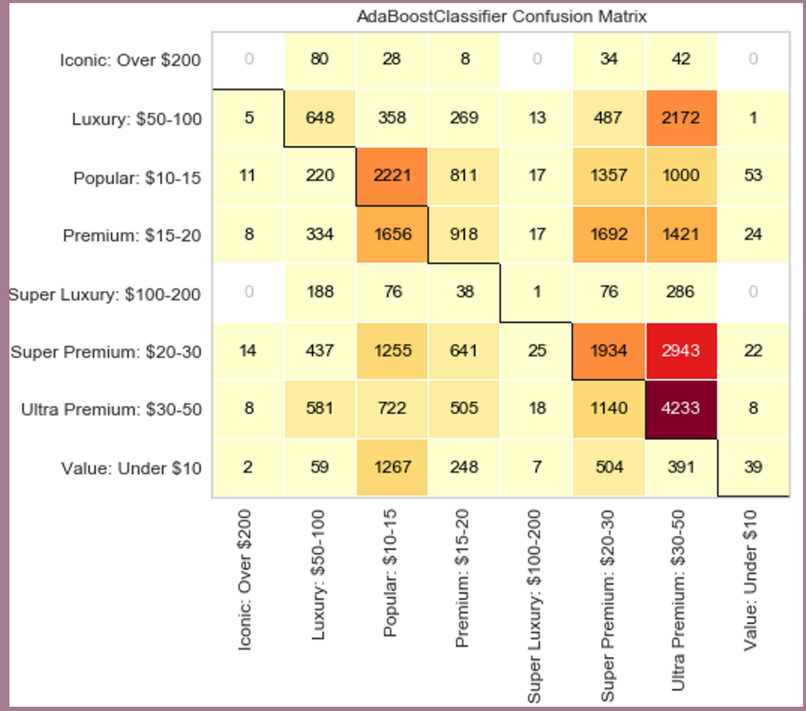
Data Preparation

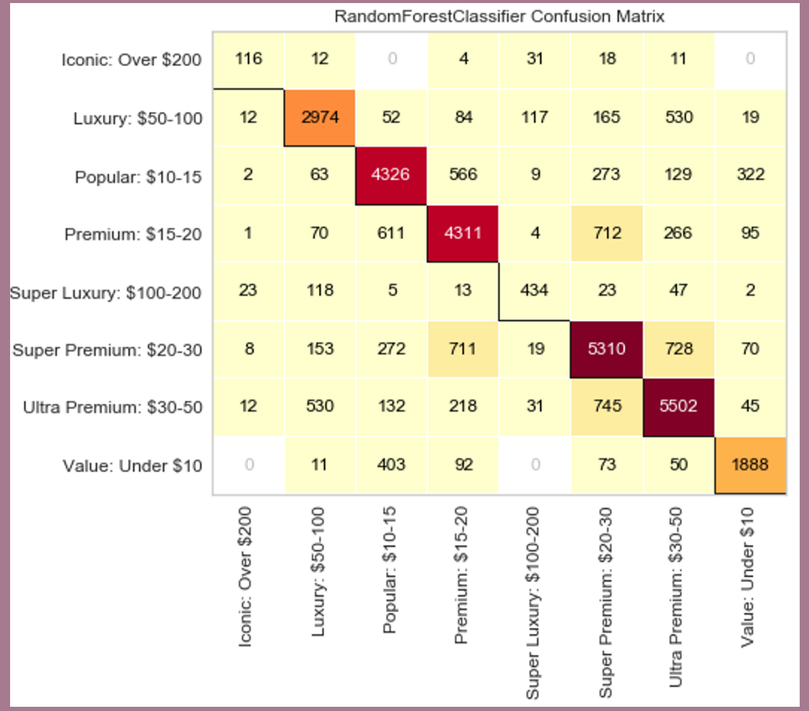
- Binned the price data to conform to a classification structure - ultimately six bins
- Label encoded different features - preference over one-hot encoding due to large number of options within each feature
- Examined correlations between price and possible features and selected different features each round - ultimately used wine type, country, variety, winery, designation, and province
- Train-test split and scaled data

Selection Criteria pt 1

Cross Validation and Confusion Matrices





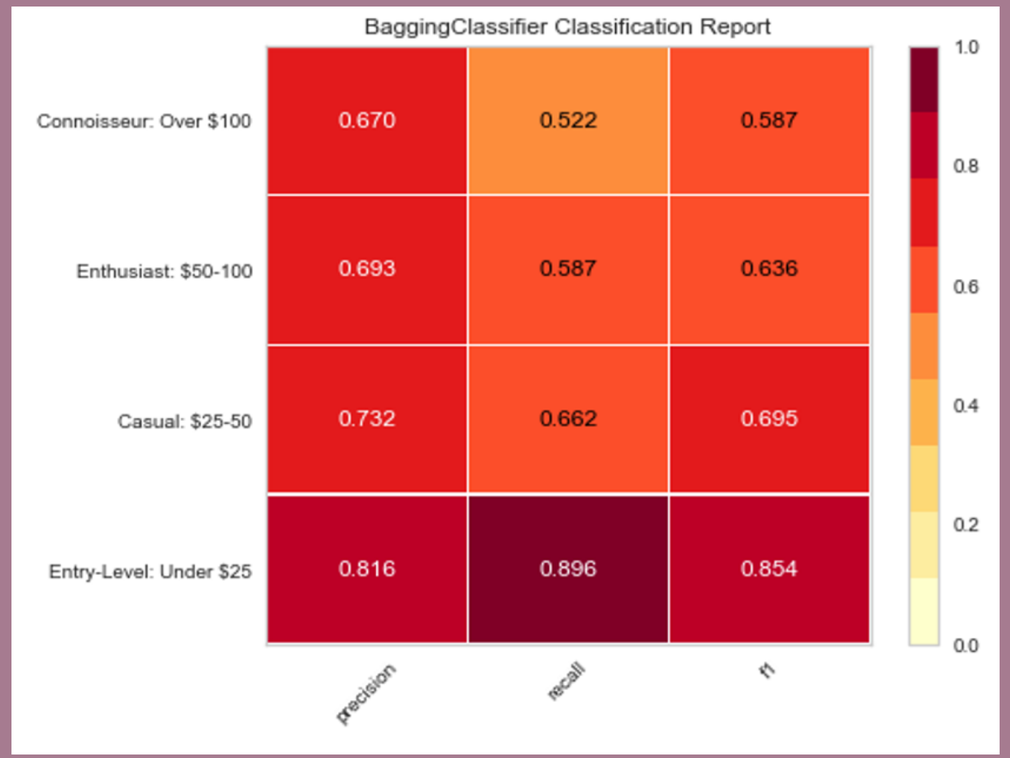


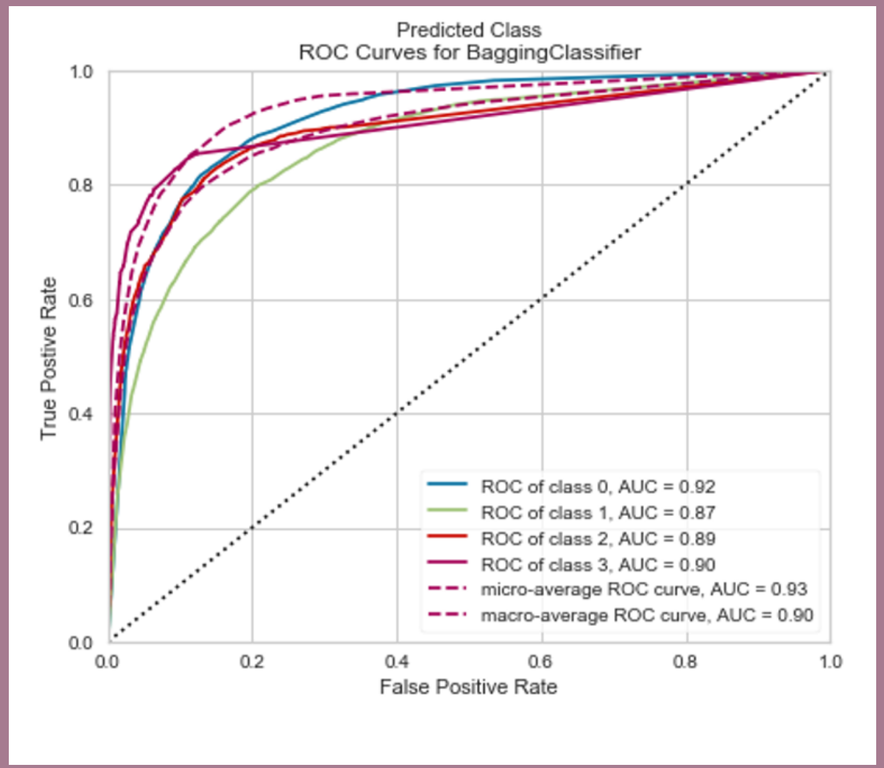
Selection Criteria pt 2

Classification Reports

and

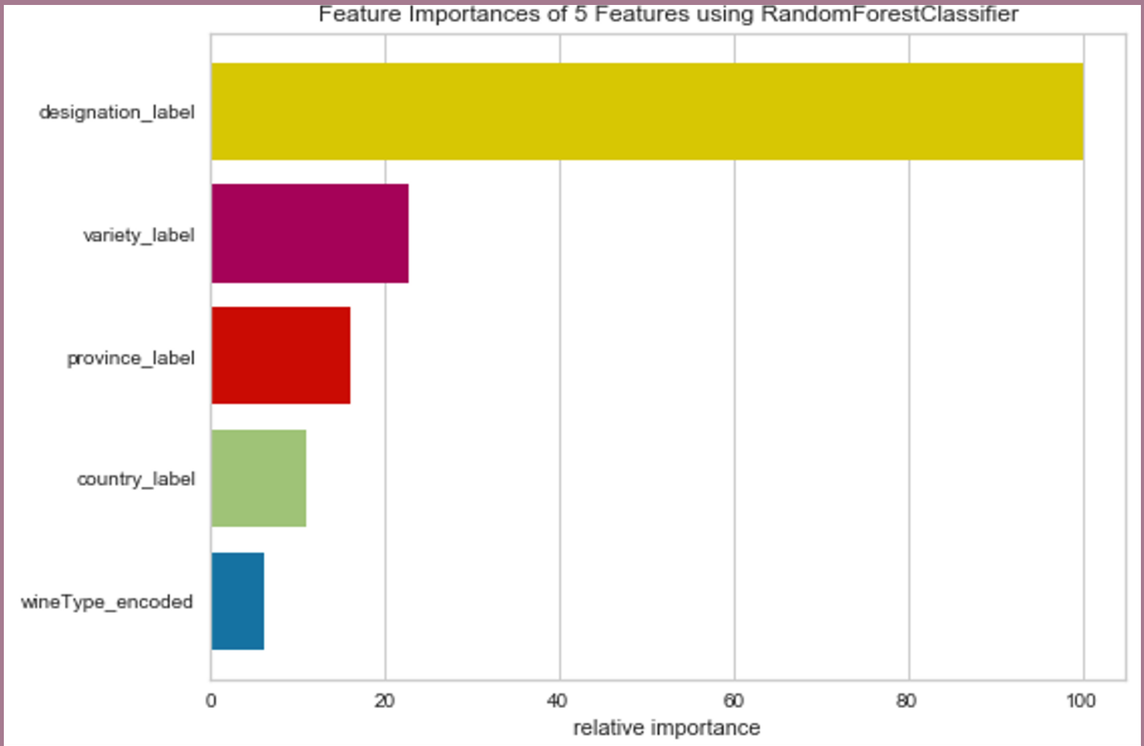
ROC Curves + AUC Scores

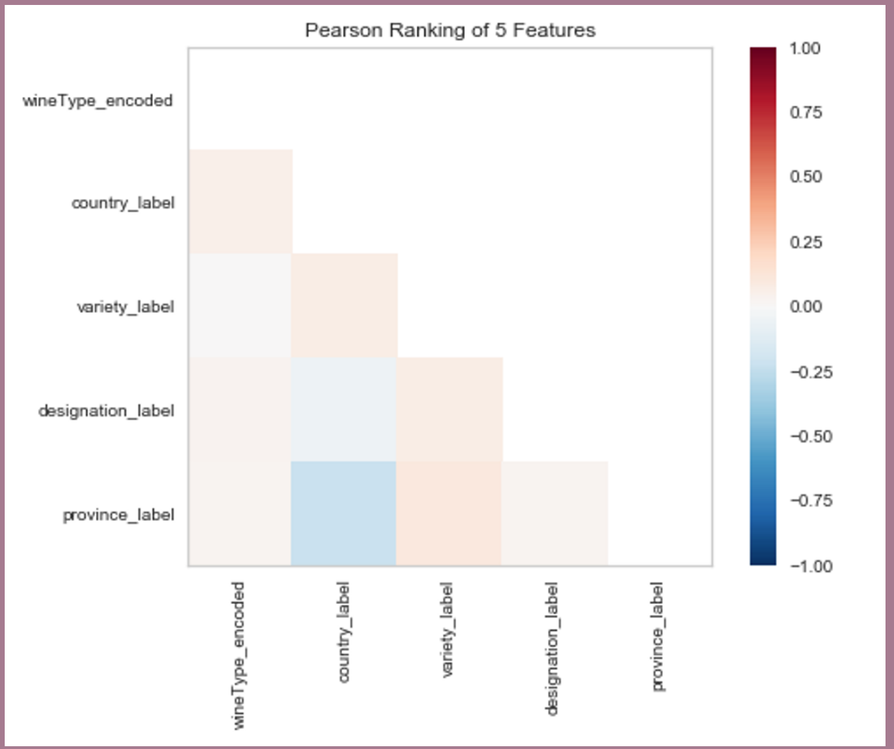




Tuning Criteria

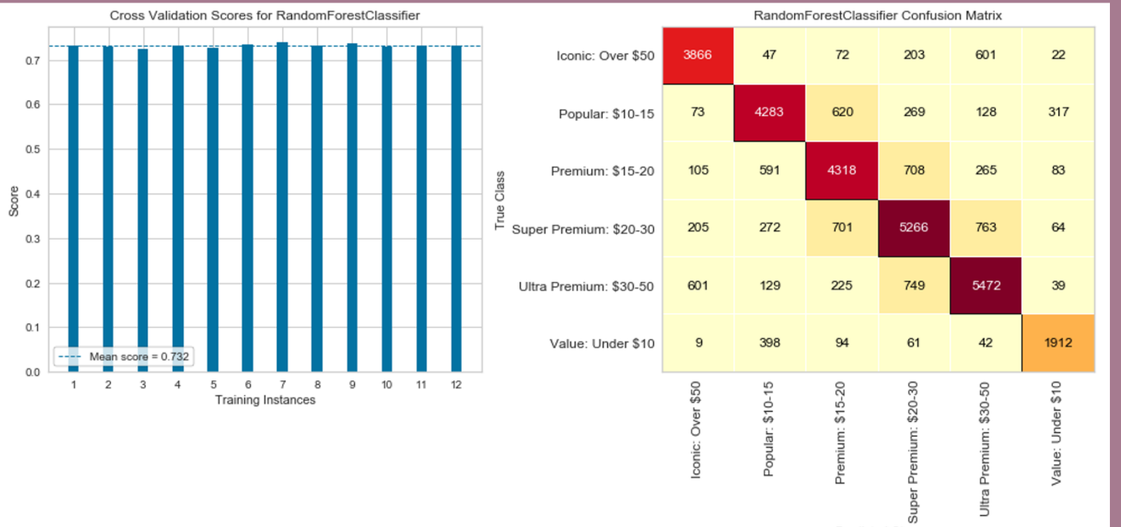
Feature Importance and Collinearity

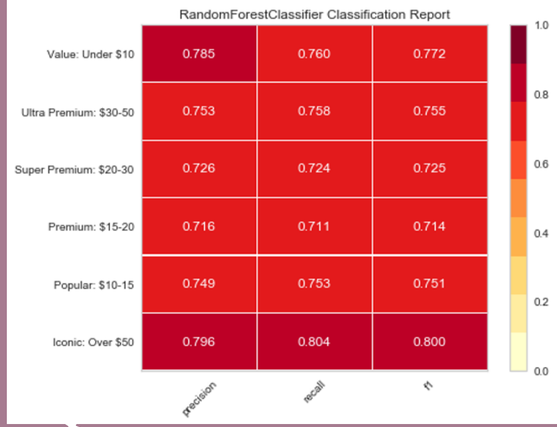


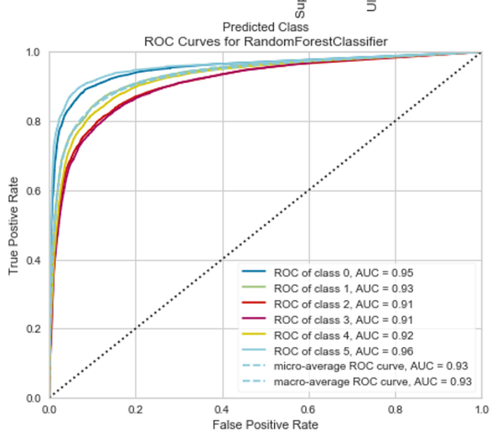


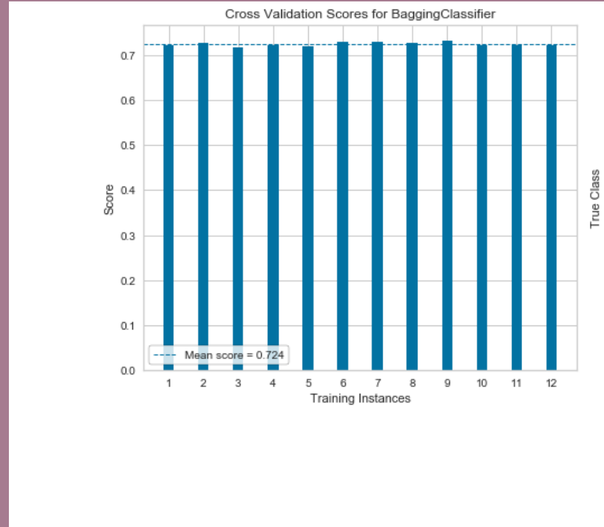
Model Selection

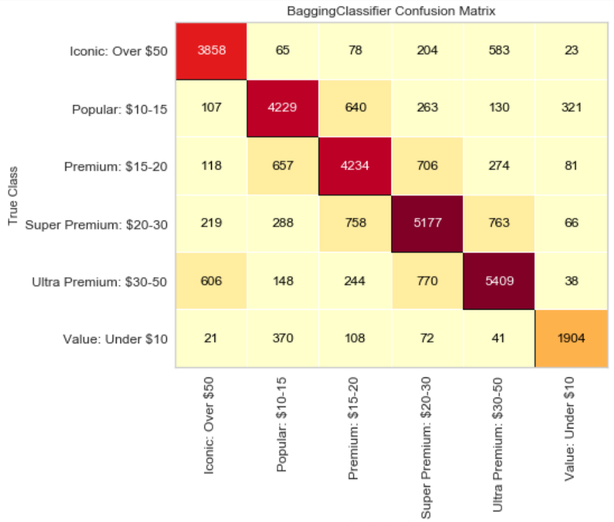
- Four rounds of machine learning models - features changed from round to round and binning changed from round to round
- Two finalists in Round Four:

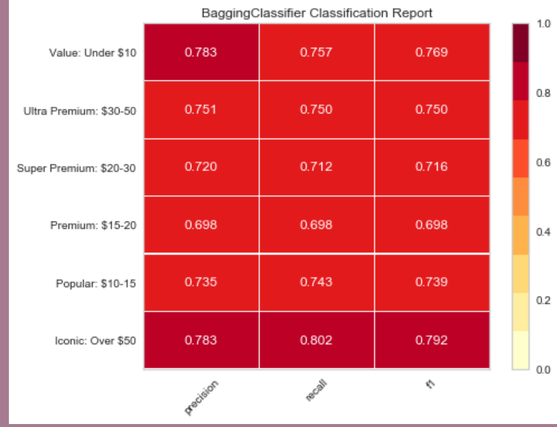


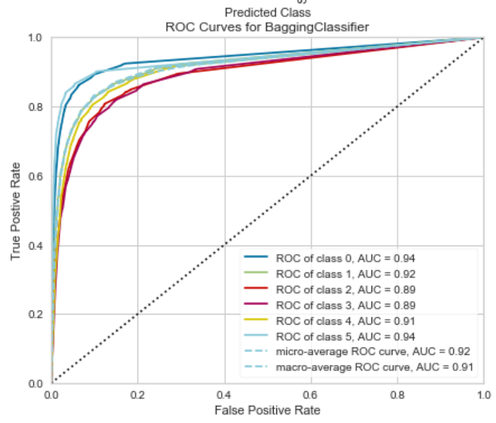






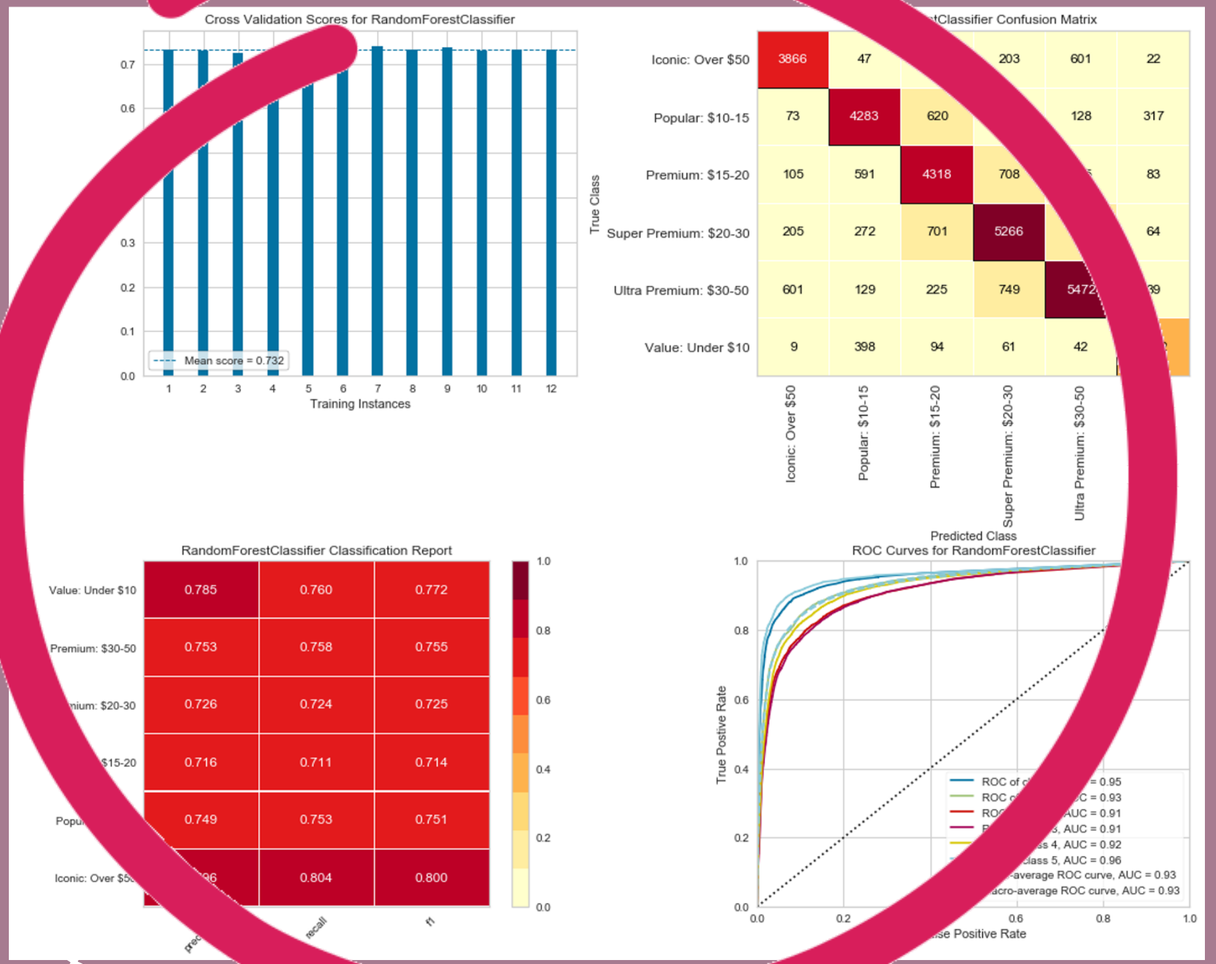


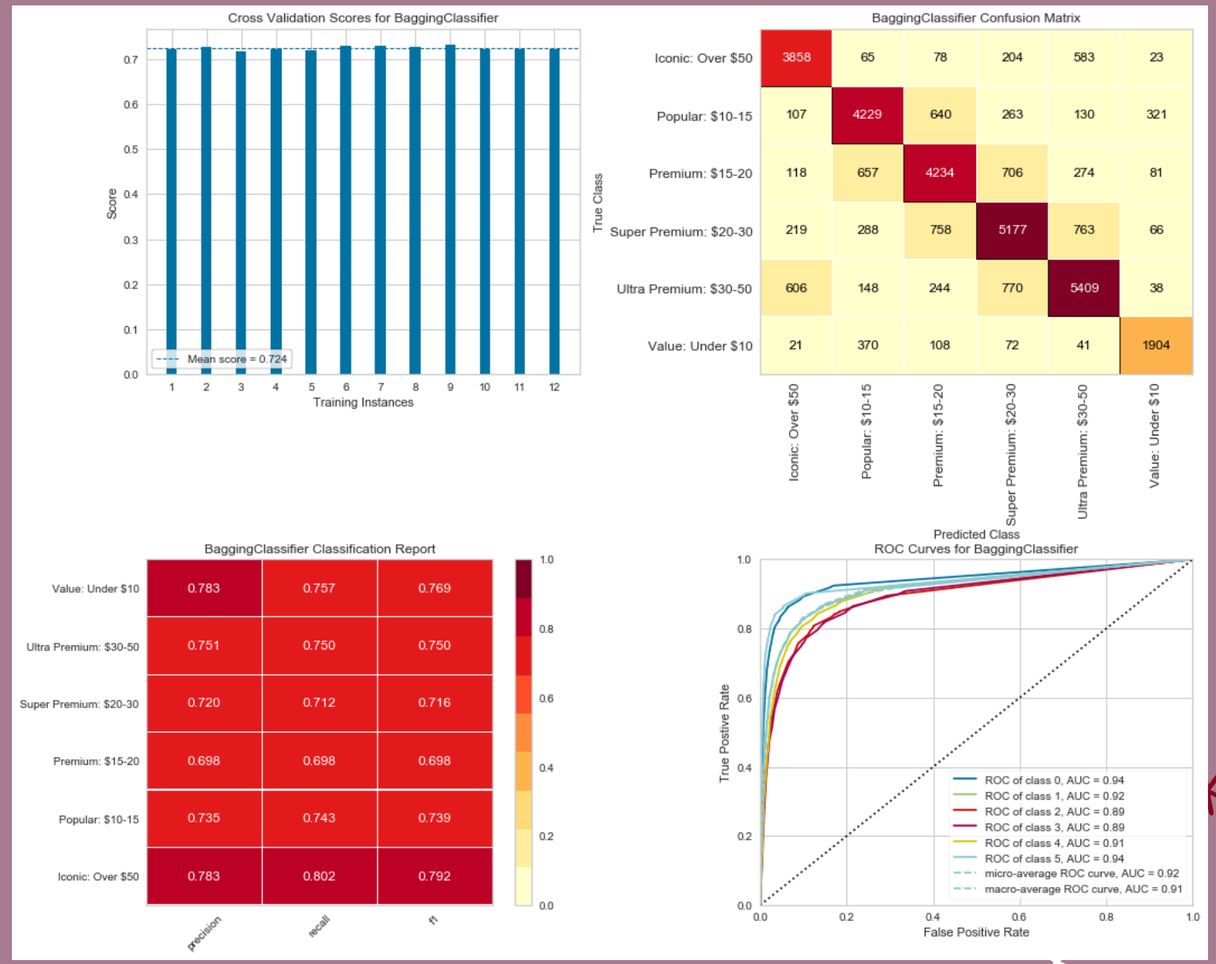




Model Selection

- Four rounds of machine learning models - features changed from round to round and binning changed from round to round
- Two finalists in Round Four:





APP DEMO

Conclusions

CHALLENGES & FUTURE CONSIDERATIONS

Challenges

1. TRANSITION FROM LIVE SERVER TO HEROKU
2. NATURAL LANGUAGE PROCESSING
3. MACHINE LEARNING PREDICTIONS

Future Considerations

1. GENERALIZE MODEL TO OUTSIDE WINE ENTHUSIAST BLENDS
2. INCORPORATE IN NLP DATA
3. UTILIZE A STACKED MACHINE LEARNING MODEL

Questions & comments?