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

Background and Cleaning

- KAGGLE DATA SET WITH BASIC WINE INFO + MANUAL LOOKUP TABLE FOR RED VS WHITE WINES
- REPLACED NULL VALUES IN GEOGRAPHIC COLUMNS WITH PROXY INFO
- REPLACED NULL VALUES IN DESIGNATION COLUMN WITH "UNKNOWN WINE"
- USED LABEL ENCODING TO DUMMIFY DATA ACROSS COLUMNS
- WOULD LATER DO MORE DATA PREP FOR THE VARIOUS MODELS

	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

EDA

- Utilized Tableau to show top trends in wine price and quality across the US and globally
- Inspired by Chelsea Argabrite's Tableau Dashboard on the same dataset
- Used the Tableau API to embed visuals on our site - will be demo'ed later.

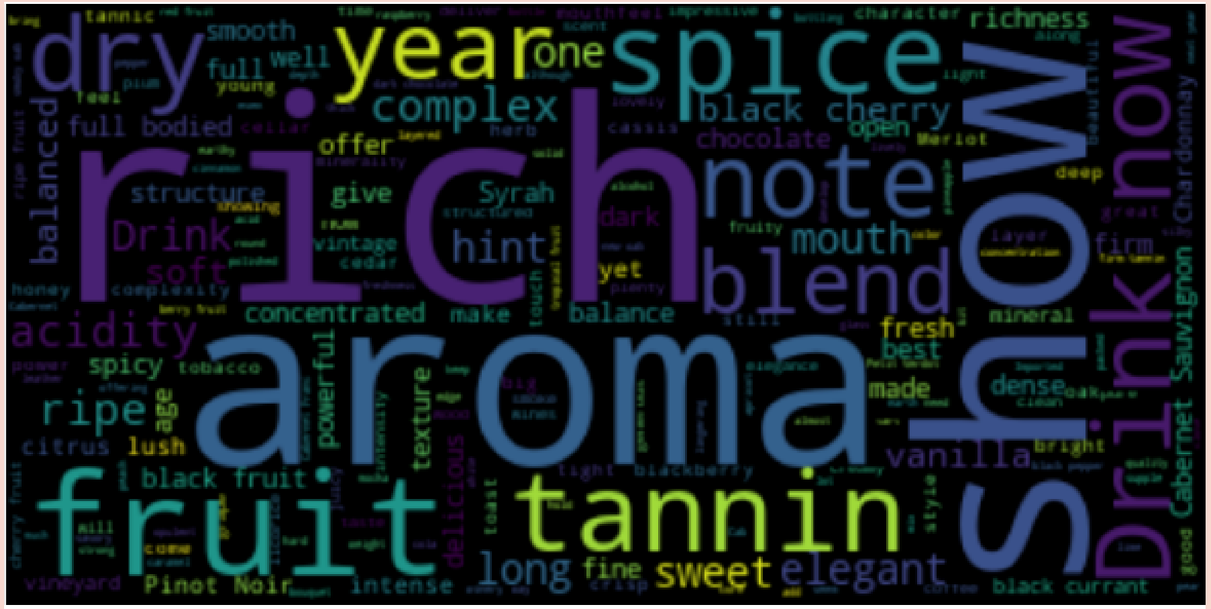
Natural Language Processing

PROCESS AND TAKEAWAYS

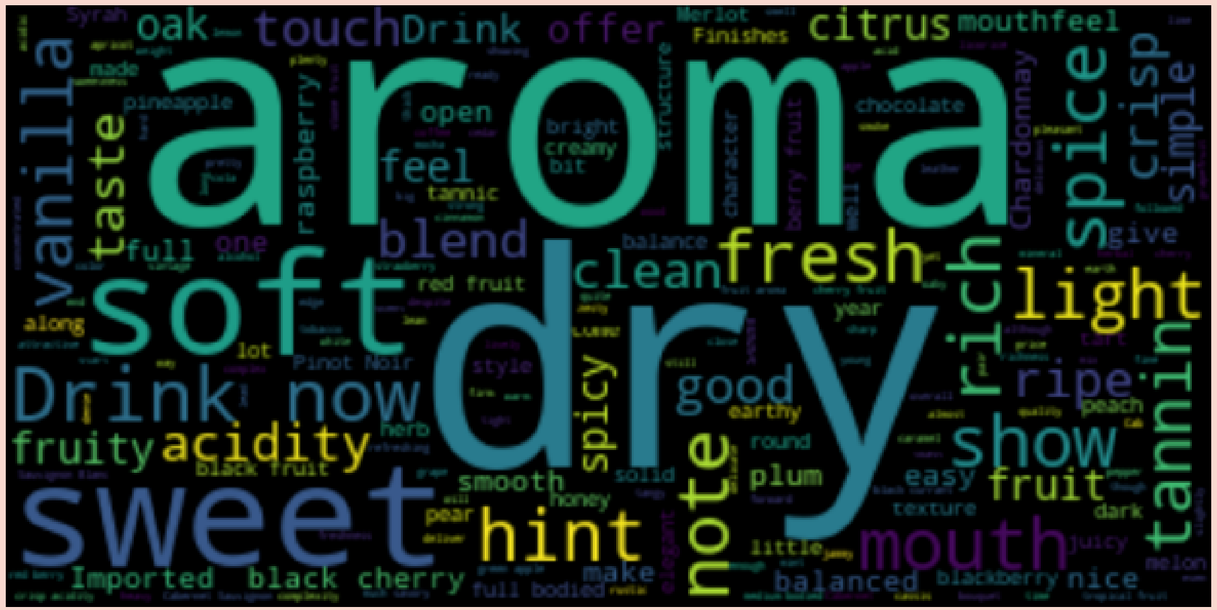
NLP - Process

- 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)
- Removed standard stop words, plus wine-specific phrases such as "wine" and "flavours"
- Used bag-of-words model to generate word clouds and eventually as part of a logistic regression model to predict whether a wine is Excellent or Good based on description

NLP - Word Clouds



"Excellent" Wines



"Good" Wines

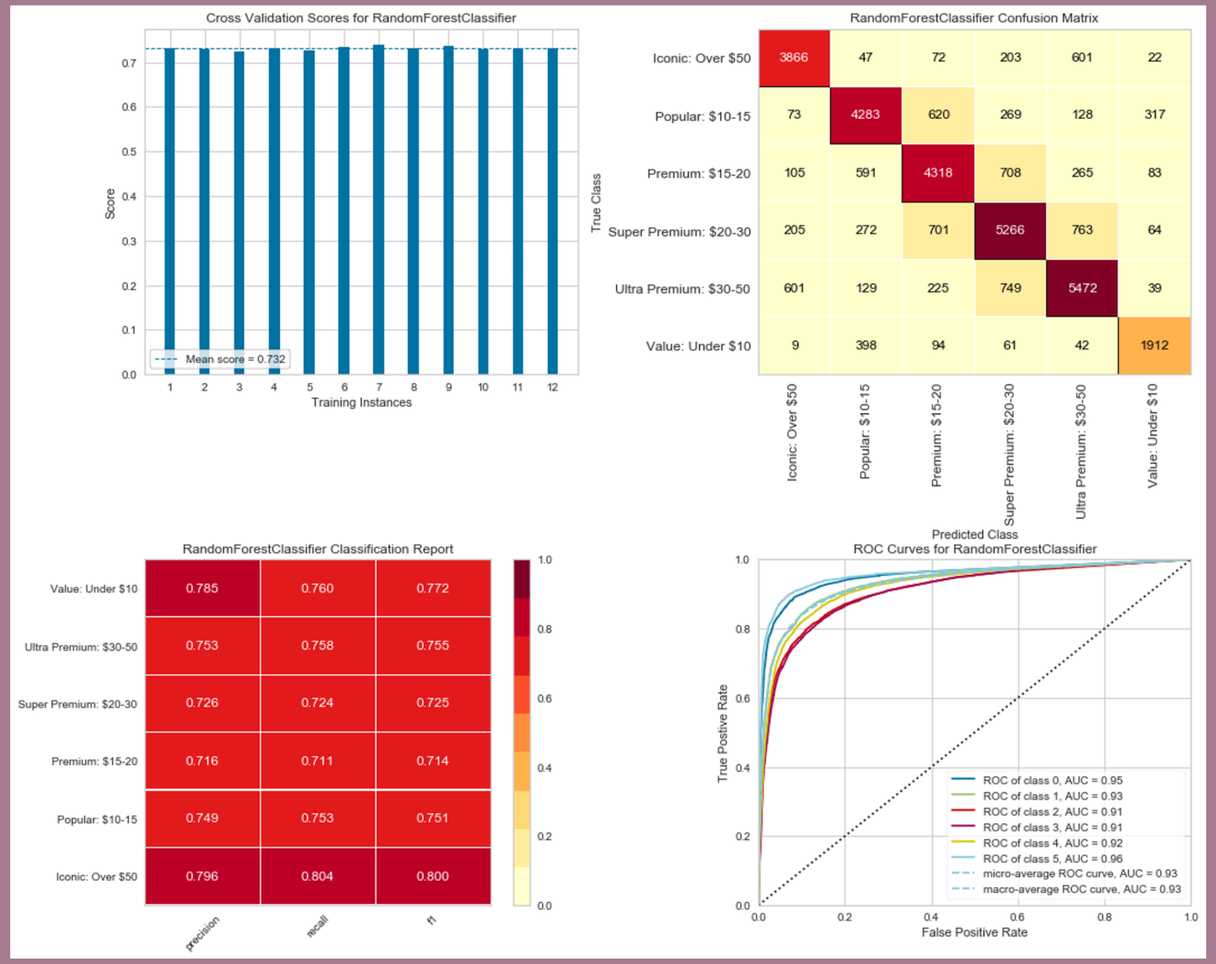
Machine Learning

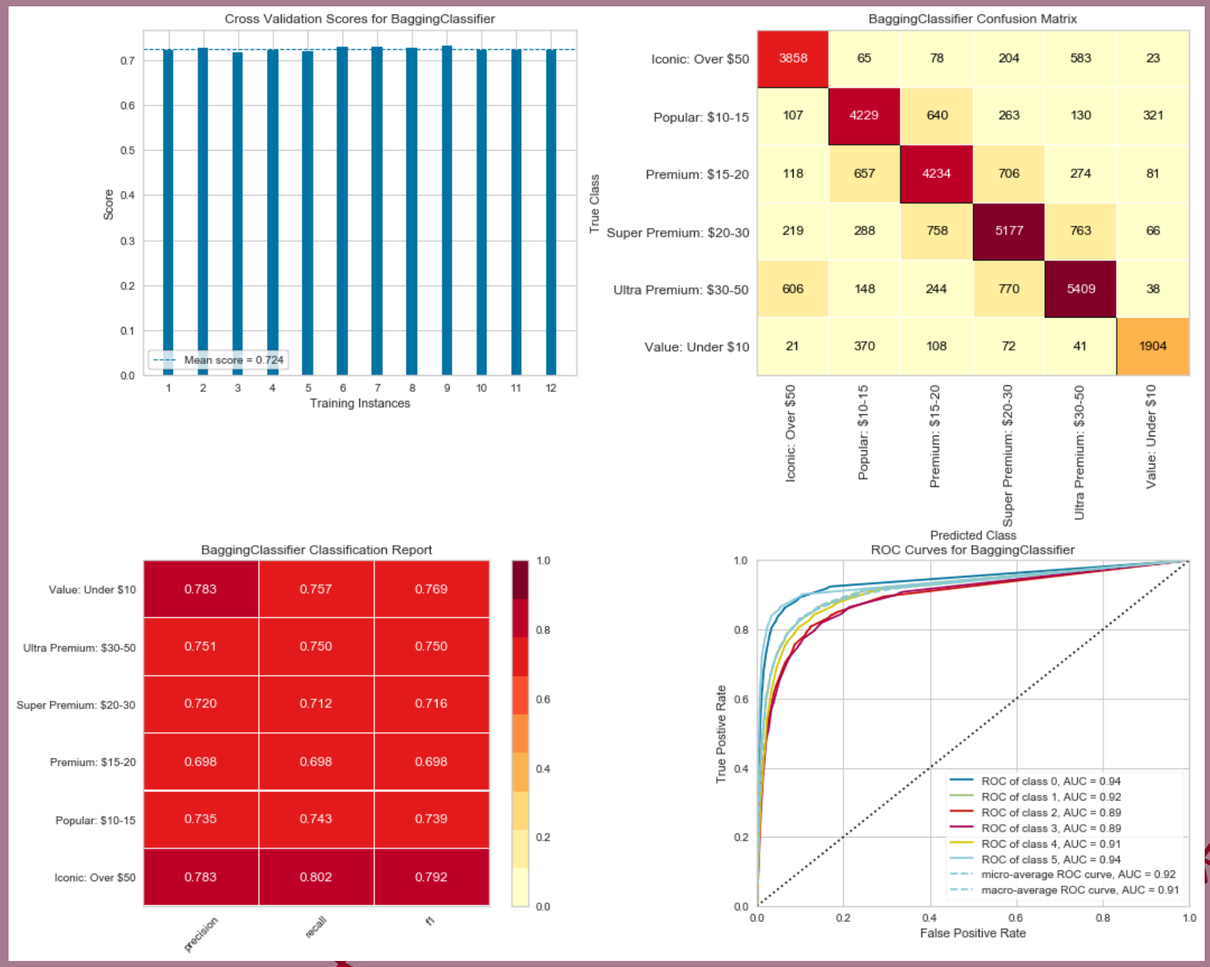
DATA PREPPING & MODEL SELECTION

ML - Characteristics

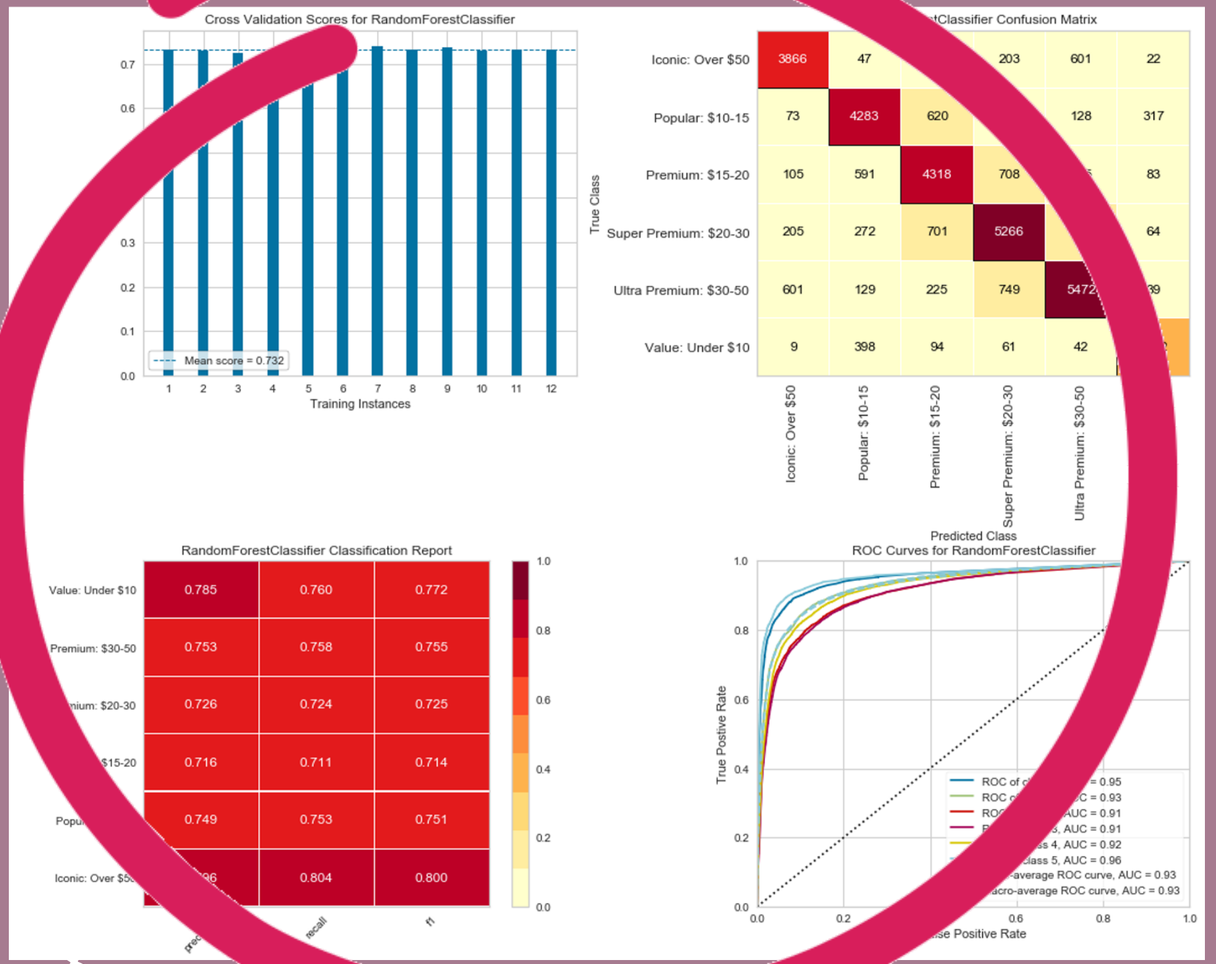
- X: WINE TYPE, COUNTRY, PROVINCE, DESIGNATION, WINERY, VARIETY
- Y: BINNED PRICE DATA (SIX CATEGORIES)
- 75% - 25% TRAIN- TEST SPLIT AND SCALED DATA
- SELECTION BASED ON CLASSIFICATION REPORTS, CONFUSION MATRICES, CROSS VALIDATION SCORES, AND ROC CURVES
- MODEL TUNING USING FEATURE IMPORTANCE AND PEARSON COEFFICIENTS

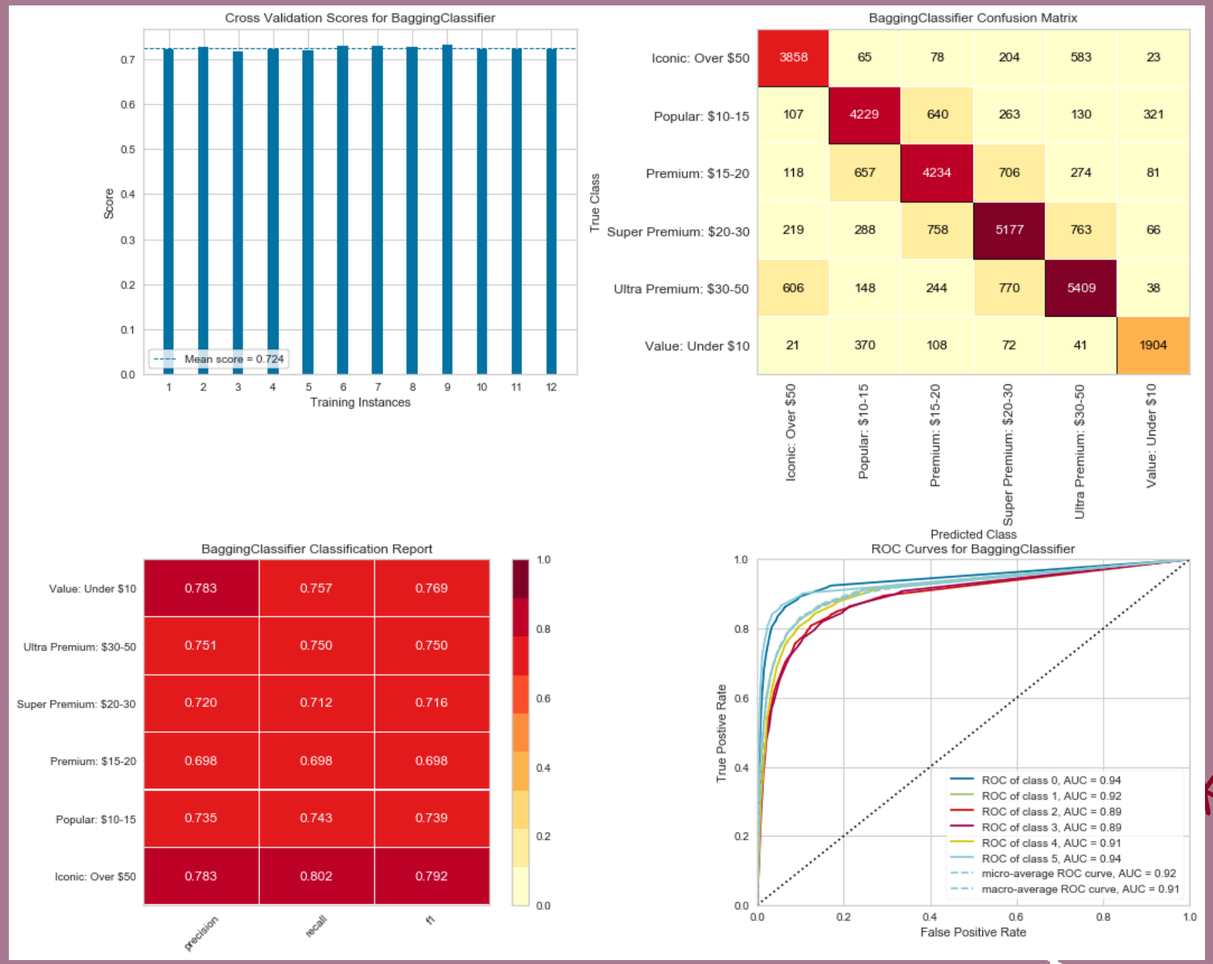
Model Selection





Model Selection





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

APP DEMO

Questions & comments?