MACHINE LEARNT SOMMELIER

JOON SEO

PRINCETON UNIVERSITY

PRECEPT P04

JSSEO@PRINCETON.EDU

ABSTRACT

- We analyze what factors go into a great wine
  - Use topic modeling (LDA and NMF)
  - Use classification (Logistic Regression, Bernoulli NB, and Linear SVM)
  - Use regression (Linear, ElasticNet, and Ridge)
- Combine to arrive at criteria for evaluating wines and ratings

BACKGROUND AND APPROACH

- Much of wine evaluation remains shrouded in mystery
- Look at 130,000 wine reviews from WineEnthusiast
- Much of the literature focuses upon supervised learning
  - Apply latent topic modeling
- Preprocess reviews and encode categorical variables such as origin of wine and variety

TOPIC MODELING

Table 1: LDA Topic Modeling of 5 Topics

Latent Dirichlet Allocation	
Topic 1: Mouth Feel	wine fru y it drink acid ripe s the rich
Topic 2: Wine Finish	y flavor finish palat appl c acid the fru e
Topic 3 : Wine Flavor	wine flavor it fru cherri oak y the thi rich
Topic 4 : Wine Notes	cherri palat the black aroma tannin spice red offer note
Topic 5 : Fruit Usage	flavor finish aroma fru thi palat plum berri cabernet blend

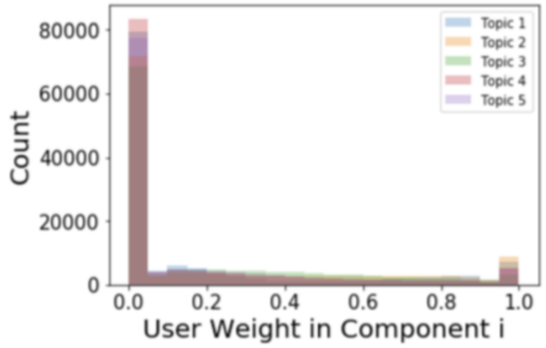

Figure 1: Topic Weighting of LDA

Table 2: NMF Topic Modeling of 5 Topics

Non-Negative Matrix Factorization	
Topic 1: Aging Process	wine fru it drink s ripe rich thi age the
Topic 2: Notes of the Wine	the palat e aroma note wh finish nose fresh appl
Topic 3: Brightness of Wine	y acid crisp fru fresh bright miner it balanc drink
Topic 4 : Scent of Wine	flavor finish thi aroma it oak feel a sweet berri
Topic 5 : Fruit Usage/ Taste	cherri black tannin palat aroma spice red the berri plum

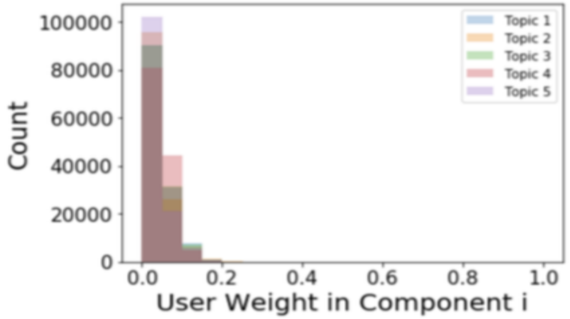

Figure 2: Topic Weighting of NMF

- More even distribution of weights with NMF
- More coherent topics with NMF
- LDA possibly limited by computation power

CLASSIFICATION MODELS

Table 3: Comparison of Classifiers

	Precision	Recall	F_1-score	Support
Bernoulli Naive Bayes				
False	0.82	0.82	0.82	13,718
True	0.80	0.79	0.80	12,276
Linear Support Vector Machine				
False	0.83	0.87	0.85	13,718
True	0.84	0.80	0.82	12,276
Logistic Regression				
False	0.83	0.86	0.85	13,718
True	0.84	0.81	0.82	12,276

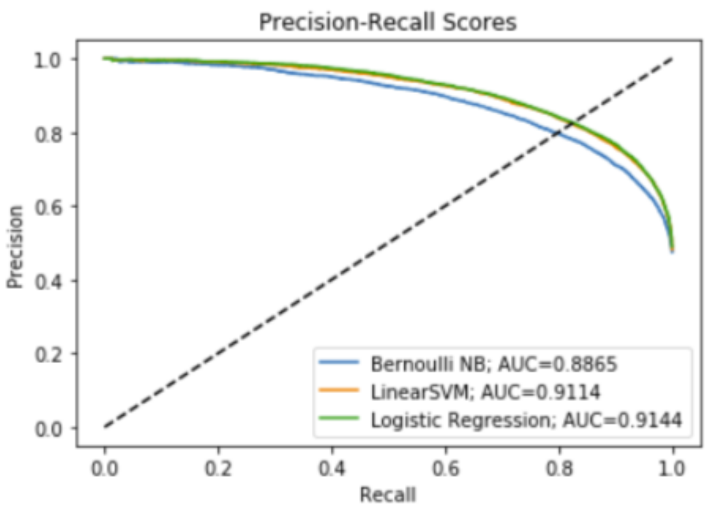

Figure 3: Precision-Recall Scores

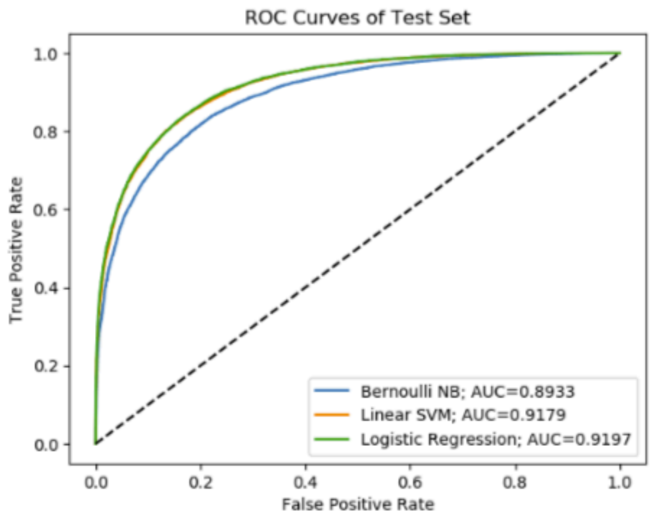

Figure 4: Receiver Operating Characteristic Curves

- Linear SVM had best performance for classification task of wines by quality

REGRESSION MODELS

Table 5: Top 20 Features of Linear Regression by Coefficient

Feature	Coefficient
country: Argentina	4.118747e+09
country: Peru	3.886923e+09
province: England	8.453873e+08
country: Ukraine	1.868244e+08
taster name: Anne Krebiehl MW	2.764144e+00
province: Wachau	2.416535e+00
province: Puente Alto	2.284224e+00
province: Kumeu	2.269647e+00
province: Kamptal	2.219636e+00
taster name: Matt Kettmann	2.178286e+00
province: Traisental	2.139456e+00
variety: Muscadelle	2.078293e+00
variety: Picolit	2.061431e+00
province: Madeira	1.936232e+00
province: Leithaberg	1.802688e+00
province: Kremstal	1.773534e+00
province: Washington	1.721713e+00
province: California	1.688816e+00
variety: Sangiovese Grosso	1.636819e+00
variety: Sagrantino	1.618324e+00

Table 6: Bottom 20 Features of Linear Regression by Coefficient

Feature	Coefficient
province: Other	-4.118747e+09
province: Mendoza Province	-4.118747e+09
province: Ica	-3.886923e+09
country: England	-8.453873e+08
province: Ukraine	-1.868244e+08
variety: Tempranillo Blanco	-4.281550e+00
variety: Brachetto	-4.115134e+00
variety: Airen	-4.078972e+00
province: Molina	-3.743532e+00
variety: Portuguese Ros	-3.453592e+00
province: Rio Claro	-3.359180e+00
province: Greece	-3.034868e+00
variety: Viura-Chardonnay	-2.956762e+00
variety: Garnacha-Syrah	-2.738010e+00
country: Brazil	-2.629350e+00
province: Bulgaria	-2.545243e+00
province: Central Valley	-2.494183e+00
variety: Pinot Noir-Gamay	-2.465141e+00
variety: Prieto Picudo	-2.421631e+00
variety: Inzolia	-2.369286e+00

Table 4: Comparison of Classifiers

Regression Model	Mean Squared Error	Residual Sum of Squares	r-squared
Ridge Regression	6.2980	152,379.097	0.3132
ElasticNet Regression	8.0851	195,617.9242	0.1183
Linear Regression	6.3456	153,532.1311	0.380

- Region of production and variety had largest impacts
- Some bias from small subsamples
- Value of recognized terroir and techniques
- Ridge and Linear had best performance

REFERENCES

- [1] Berry Bros. & Rudd.
- [2] What Makes Great Wine?
- [3] Teodoro Aguilera, Jess Lozano, Jos A. Paredes, Fernando J. Ivarez, and Jos I. Surez. Electronic Nose Based on Independent Component Analysis Combined with Partial Least Squares and Artificial Neural Networks for Wine Prediction. *Sensors*, 12(6):8055–8072, June 2012.
- [4] Paulo Cortez, Antnio Cerdeira, Fernando Almeida, Telmo Matos, and Jos Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, November 2009.
- [5] Claudia Gonzalez Viejo, Sigfredo Fuentes, Damir D. Torrico, Kate Howell, and Frank R. Dun- shea. Assessment of Beer Quality Based on a Robotic Pourer, Computer Vision, and Machine Learning Algorithms Using Commercial Beers. *Journal of Food Science*, 83(5):1381–1388, 2018.