Project Proposal: Wine Reviews

INLS 613.001.S18 Text Mining

Patric Carman Jordan Chasteen Andrew Garcia

19 February 2018

1 Problem description

A description of the problem area you wish to investigate

Representing an enigmatic global economic structure, the world wine market is driven by a large number of feature variables, each with varying weight. A certain wine's price is determined by its feature variables which can range from obvious things like region, vintage (production year), grape quality to not so obvious things like soil nutrients, vine height and even varietal of barrel used for fermenting. Due to the high number and specific variety of interesting features, the problem of predicting a wine's price given its production variables is a problem that is almost a perfect fit for machine learning. Using machine learning we hope to gain insight on how the price of any given wine is determined, trained and tested on a database of 130k wines, listed with all sorts of variables such as origin, description, designation, points, provenance, regions, title, variety and winery.

2 Literature review

A list of 6-8 papers you plan to survey as part of your literature review. For each paper, provide a brief justification for why you selected it (3-5 sentences).

1. *Predicting the Quality and Prices of Bordeaux Wine*
   Author: Orley Ashenfelter
   JSTOR permalink
   This paper provides a more economics driven analysis of wine prices, and pertains to predicting prices of a specific type of wine – namely the Bordeaux varietal. This paper provides some much needed information on the economics of wine, and give us a bit of background in the area of study we chose – an area in which we are not particularly experts.

2. *Statistical analysis of the price and subjective quality ratings on Australian wines*
   Author: Peter M. Visscher
   ArXiv permalink
   This paper from a researcher from Queensland focuses on analyzing price and subjective quality ratings on Australian wines. This is a similar research area to that of our project, and will be useful moving forward through our project. It provides detailed statistics on relationships between price and rating, variations in price and variations in ratings. If we determine at any point in our research that we need to adjust for these variables, we will have adjustment coefficients readily available for a sample subset.

3. *Region of origin and its importance among choice factors in the wine-buying decision making of consumers*
   Authors: Emily McCutcheon, Johan Bruwer, Elton Li
   ResearchGate permalink
   This paper examines the effect that region of origin has on a consumer's desire to purchase a wine. It also tests the effects that grape variety, price, quality, and wine style have on the sells of a wine. This article can be helpful in our analysis because it examines many of the same variables that are present in our dataset. This means that this article shows us many of the most important variables that affect a consumer and allows us to dive deeper into the subject of which variables are the best predictors.

4. *What Determines Wine Prices: Objective vs. Sensory Characteristics*
   Authors: Sébastien Lecocq and Michael Visser
   Pittsburg State University permalink
   This article attempts to relate wine prices and wine characteristics to see how wine characteristics influence the prices of a wine. It also attempts to tackle the issue of how price influences the perceived quality of a wine. This article is helpful for us because it could be used to help explain why higher priced wines also have better ratings. This could help show us that wine price is a good predictor of the rating of a wine, in contrast to the belief that the rating a person gives a wine can be used to predict the price. This article will also show us how certain wine characteristics affect a price, which is the goal of our analysis.

5. *What Drives Consumer Choices? Mining Aspects and Opinions on Large Scale Review Data using Distributed Representation of Words*
   Authors: Kasturi Bhattacharjee, Linda Petzold
   Sentic permalink
   Bhattacharjee and Petzold explore opinion mining and sentiment analysis on restaurant reviews from Yelp. The authors detail the challenges that they experienced throughout their research: fixing noisy data, accounting for differences in how people express themselves, and finding good datasets. The article can be especially helpful because the authors show how they aggregate similar words to reduce feature set. We might have to do this if our feature set is too wide.

6. *Featurizing Text: Converting Text into Predictors for Regression Analysis*
   Authors: Dean P. Foster, Mark Liberman, Robert A. Stine
   Wharton permalink
   Dean Foster tries to establish a relationship between home list prices and their text-based listing descriptions. Though the paper isn't about the domain we're researching (wine), it will be helpful because it is one of the few papers that creates a model for guessing price given a text description. Foster spends much of the paper discussing feature selection and how he can derive quantitative features from unstructured text.

3 Experiment description

A description of the purpose of your experiments. What are you testing?

According to the Wine Institute, 62% of Americans aged 21+ regularly make wine purchases. Additionally, the Wine Genome study performed by Constellation Brands found that one-in-five wine purchasers makes their wine decision purely based on price, and on no other factors. These facts together imply that close to 30 million (27.5 million) Americans are blindly picking up a wine off the shelf based on price, with no regard for quality or popularity. What's even more surprising is that the entities who set the price of wine are those who produce and profit from it. This creates a positive feedback loop, where by setting higher prices, yet keeping quality consistent, you will maintain a large swath of customers, only encouraging higher prices on lower quality wines. It makes the wine pricing process seem to have the same randomness as a blind dart throw. We selected this problem in order to give method to this current madness in wine pricing.

We are testing to see whether there is a correlation between the language used in wine reviews and the price of wine. To do this, we will derive features from the text corpus using Lightside and then move on to more advanced modeling with Weka.

4 Dataset

A description of the dataset you will use to conduct your experiments. If you are using an existing dataset (highly recommended), provide a short description and a hyperlink so that I can review it. If you plan to collect our own dataset, provide a short description of how you will collect it (e.g., By scraping text from the web? Will you annotate it manually?).

Dataset: *https://www.kaggle.com/zynicide/wine-reviews/data*

This dataset gathers information from people that have reviewed many different types of wine. It has over 100,000 entries describing variety, location, winery, price, and description,etc., of the wine. These 130,000 entries were scraped from the website WineEnthusiast, which is a website that specializes in recommending wines. The original creator of the dataset scraped this data in order to be able to predict the identity of the wine based on many different factors. Since it has all of this data available, it can also be used to predict wine prices based on other variables.

5 Risks

A description of any risks associated with your proposed plan.

As with any machine learning experiment, the hypothesis that there is a significant correlation between features and class could be wrong. In this instance the risk is that wine price is not dependent on the feature variables we possess in our dataset. However, even if we encounter this problem, we will still have garnered a valuable piece of information about wine pricing models – that it is more arbitrary than we initially thought.

Another risk that we could have is how the data was gathered. Since wine prices often influence the perceived quality of a wine, if the reviewers knew the price of the wine before reviewing, it could lead to bias in the data. This could lead to inaccuracies in the predictions or false predictors because the data would be influenced by how a reviewer believed the wine should have tasted compared to how it actually tasted.