Stat-495 Revised Proposal

Emily Ye & Shukry Zablah

2019-11-11

Team Name: YZ Analytics

Title: Another Glass, Please: Creating a Virtual Sommelier

Purpose: Purchasing wine can be overwhelming with the sheer number of choices there are, whether they be on a restaurant's wine menu, in liquor stores, or wineries. In this project, we aim to understand the features of what makes a good wine and predict the quality of wine from its description. Our project is applicable to both newly exploring wine drinkers and experienced wine enthusiasts alike. Non-wine drinkers will be able to know what makes a good wine, and wine drinkers will be able to explore and discover new wine varieties to try.

Data: We will be working with Kaggle's wine review data, found here: https://www.kaggle.com/zynicide/wine-reviews (Thoutt, 2017). There are close to 130,000 wine reviews.

The original data set poster also included his code for scraping the data off of the WineEnthusiast website, so we will also explore to see if his code still works and scrape the data ourselves so we have more up-to-date wine reviews and more data. There are currently about 270,000 reviews on WineEnthusiast (WineEnthusiast, 2019).

Variables:

Variable descriptions from the data set:

- country - The country that the wine is from
- description - A few sentences from a sommelier describing the wine's taste, smell, look, feel, etc.
- designation - The vineyard within the winery where the grapes that made the wine are from
- points - The number of points WineEnthusiast rated the wine on a scale of 1-100 (though they say they only post reviews for wines that score >=80)
- price - The cost for a bottle of the wine
- province - The province or state that the wine is from
- region_1 - The wine growing area in a province or state (ie Napa)
- region_2 - Sometimes there are more specific regions specified within a wine growing area (i.e. Rutherford inside the Napa Valley), but this value can sometimes be blank
- variety - The type of grapes used to make the wine (ie Pinot Noir)
- winery - The winery that made the wine
- title - The title of the wine review, which often contains the vintage if you're interested in extracting that feature
- year - pulled from the vintage in the title
- taster_name - name of the taster
- taster_twitter_handle - Twitter handle of the taster

We intend to use some sort of text analysis and text mining to create additional quantitative variables from the description, in particular using Bag of Words (BOW) representation to calculate the Term Frequency-Inverse Document Frequency (TF-IDF) for the unique words of the description (Nabi, 2018). We can achieve this calculation through the `text2vec` package ("Predicting," 2018).

Model: We aim to create a model to predict the quality of wine (based on the number of points it received from WineEnthusiast) from its description. We plan to create our model using the random forest method.

End Product: We aim to create a Virtual Sommelier with dynamic visualizations that provides information on wine and its features. We envision an interactive map that will allow users to explore wine statistics and

different countries or regions and even compare them. There will also be a Create Your Own Wine tool where users can input their wine's description and see our prediction of quality.

References

Nabi, Javaid. (2018). "Machine Learning - Text Processing." Towards Data Science. https://towardsdatascience.com/machine-learning-text-processing-1d5a2d638958

Thoutt, Zack. (2017). "Wine Reviews." Kaggle. https://www.kaggle.com/zynicide/wine-reviews

2018. "Predicting Wine Ratings Using LightGBM + Text2Vec." Kaggle. https://www.kaggle.com/nnnnick/predicting-wine-ratings-using-lightgbm-text2vec

2019. WineEnthusiast. https://www.winemag.com/?s=&drink_type=wine&page=0