KAVYA ANNAPAREDDY, 10/05/2020

THE WINE LAND

Data Analysis and Insights

AGENDA

Introduction

Data Exploration

5 Actionable Insights

Model Development

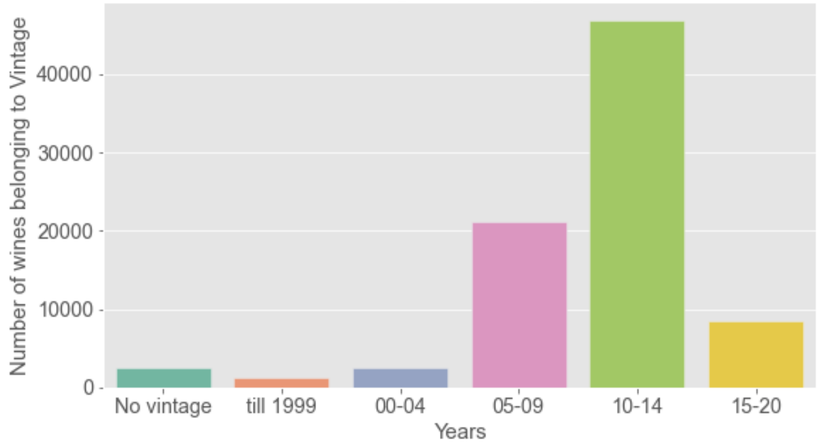
Model Validation

INTRODUCTION

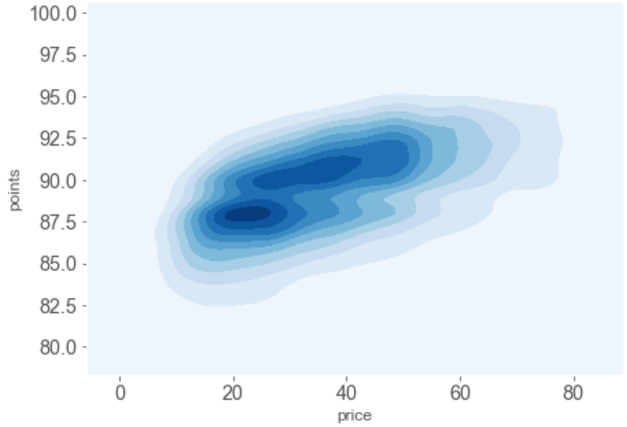
Our objective is to build a predictive model for predicting the wine ‘variety' and leverage the reviews data from dataset provided on 'The Wine Land', an online wine shop, and draw actionable insights from it.

DATA EXPLORATION

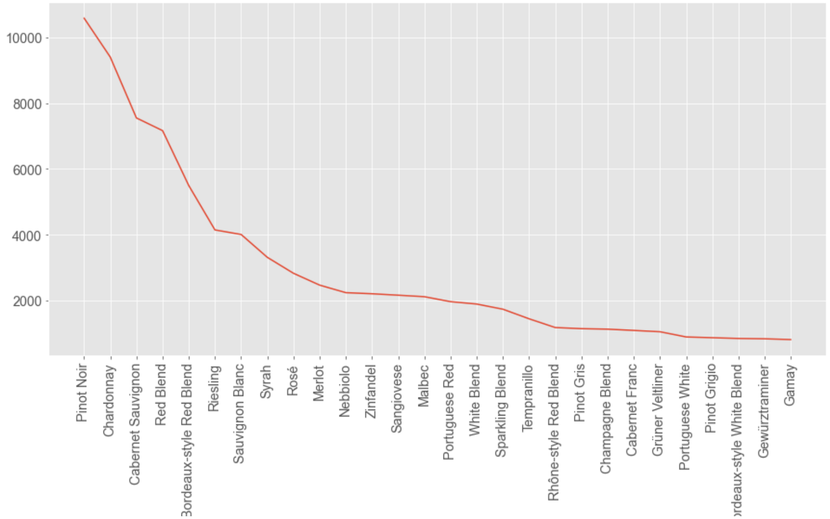
Range of Vintages

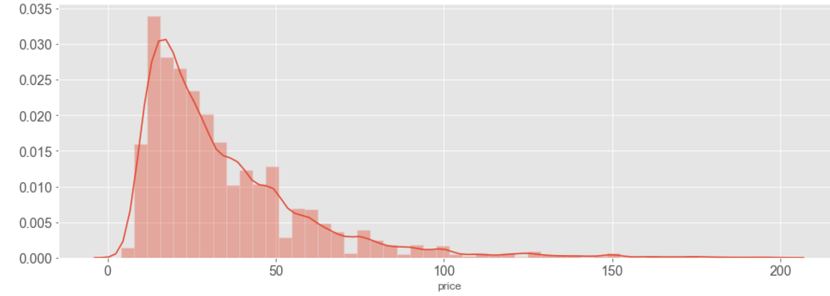


Contour plot of price and points

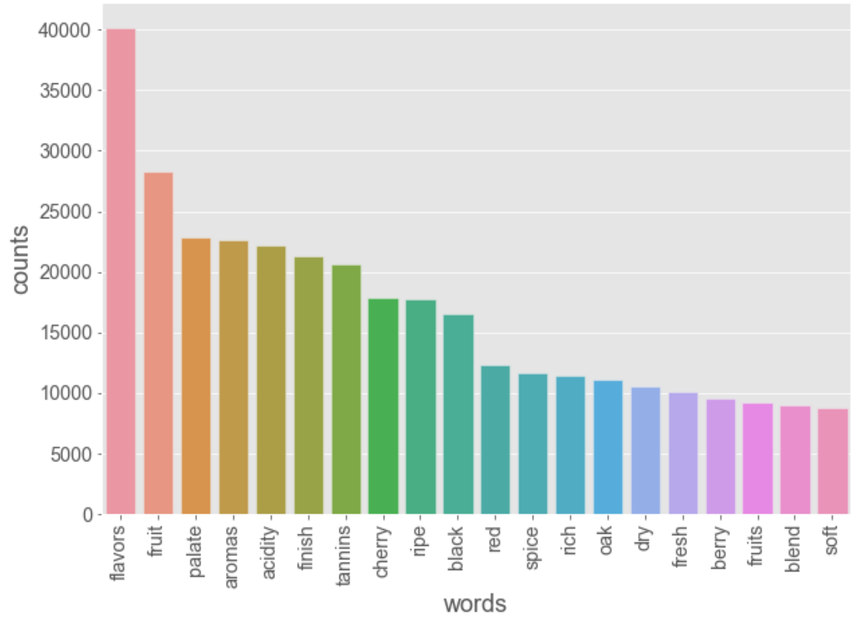


Wine Varieties and their counts

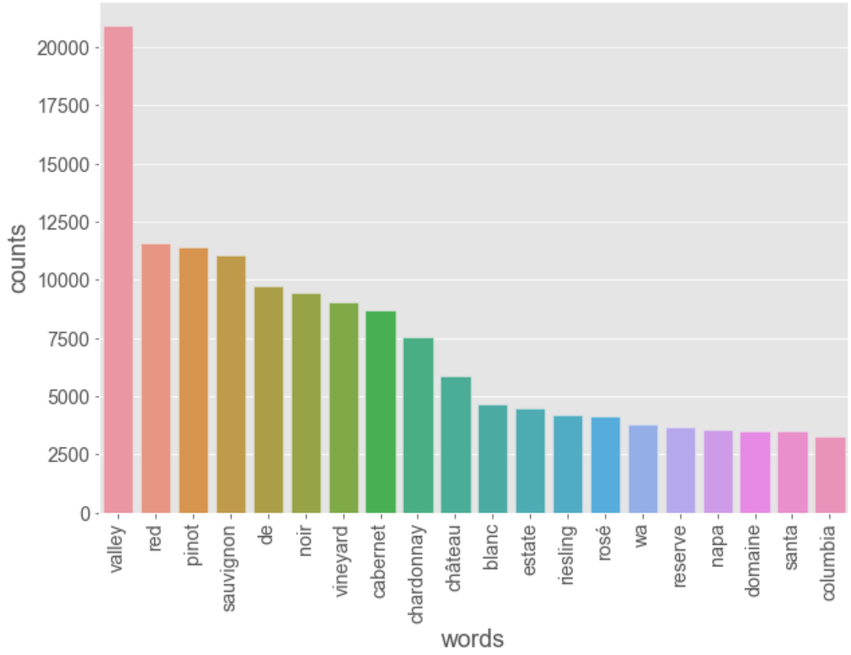




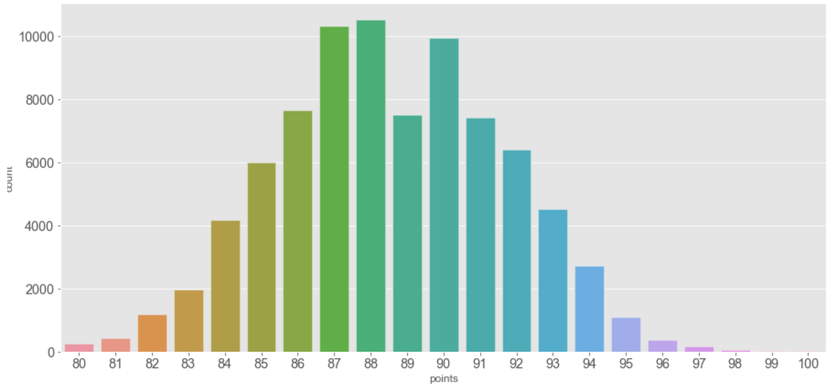
20 most common words in review description



20 most common words in review title



Distribution of points



ACTIONABLE INSIGHTS

Wine Variety and Location

- Insight: Each region is famous for a specific variety of wine. For example, Napa Valley in California got lot of reviews on Cabernet Sauvignon. Likewise, Tuscany is famous for its Red Blend and Sangiovese varieties. We also obtain a list of wineries with most reviews for a variety of wine.
- Action: The location of the wineries can be geocoded and shown on a map for the browsers online to see from where their wine is coming from. And the distance to the nearest winery, if they wish to visit one.

User name and wine variety reviewed

- Insight: We observe that each user reviews certain varieties of wine. The most popular user_name @vossroger reviewed mostly Bordeaux-style Red Blend followed by Chardonnay, Portuguese Red and others. Whereas users such as @vboone reviewed about specific varieties of wine such as Zinfandel, Sauvignon Blanc, Cabernet Sauvignon, etc
- Action: Since each of the wine opinion leaders cater to the interests of that specific customer demographic, this data could be used to suggest influencers for social media following to customers who visit The Wine Land.
- Action: These wine experts can be targeted by wineries and vineyards for marketing purposes.

Review points of various wines

- Insight: We spot that the points range from 80-100 with most wines lying in the range 85-89 followed by 90-94.
- Action: A recommender system could be built based on the points by suggesting wines within the same point range and enhance customer buying choice. Apart from price, points can act as information for customer decision making.

Wine variety, vintage, price and points

- Insight: The top 10 popular vintages are years 2006-2014. And the average price range for these bottles is from $ 35-40. A wine bottle with no vintage is cheaper than the popular vintages selling at average of $ 33.7.
- Action: The online listings could be curated in the price range of $ 30-40 and whose points range from 85-93 points as they seem to be most popular. From our dataset, we observe that wine varieties Pinor Noir and Chardonnay are extremely popular amongst reviewers. This would lead to more turnover and less inventory costs, thereby improving the margins of The Wine Land.

Review title and description

- Insight: We identify a certain set of words that occur with great frequency overall and specific to each wine variety in review description. These combination of terms such as finish, apple, citrus, acidity, berry, dry , spice, etc describe a specific kind of wine.
- Action: Given the taste preferences, the online store interface would be able to suggest wine varieties to the consumer.
- Action: This corpus of data allows us to find a fake reviewer from the genuine and help business post trustworthy reviews.
- Action: We can perform sentiment analysis and see how each user felt about the different varieties of wine.

FEATURES EXTRACTED

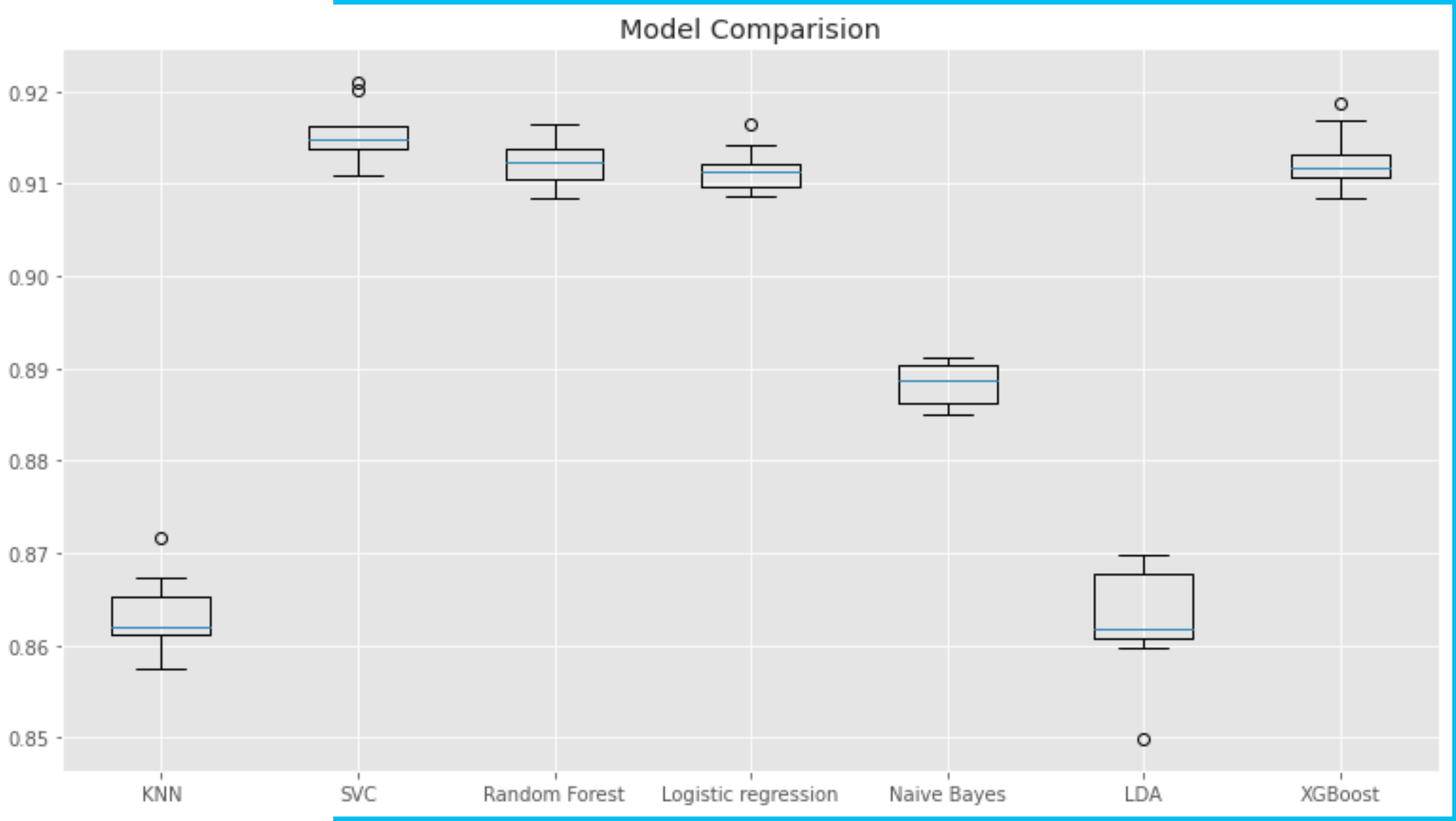
VINTAGE BAG OF WORDS

- Although location plays an important role in defining the taste and quality of wine, I select only bag of words as my explanatory variable
- I observe the performance and add other features as necessary. However, there is missing data in the other variables and it needs to be dealt with before using it for training.
- PCA could be used to reduce the dimensionality of data but at cost of reducing accuracy.

MODEL DEVELOPMENT

MODELS AND ACCURACY SCORES

- K Nearest Neighbors
- Support Vector Machine
- Decision Tree/ Random Forest
- Logistic Regression
- Naive Bayes
- Linear Discriminant Analysis
- XGBoost



MODEL VALIDATION

Apply XGBoost on test data to obtain wine variety prediction output
Git hub link: https://github.com/Artemis601/ML/tree/master/Knight_ML_Assignment