ORIE 4741 Midterm Report

Zhoutong Li(zl683), Yuanzheng Cao (yc2575)

November 2019

1 Abstract

This project aims to develop a deeper understanding of the wine price not only for customers, but also for the wine producer. For customers typically do not know whether those wine is overpriced or underpriced, and producers need to price their product reasonably. Our model might provide some insights about the price of the wine and help to make the market more competitive.

2 Exploring Raw Data

2.1 Data Cleaning and Pre-processing

Since our goal is to predict wine price, we removed all rows missing the *price* column. We also removed one rows missing the *points*(rating) and the *variety* column. All the others features are categorical, so we replaced all missing rows with "unknown". This avoids the bias introduced by deleting too many rows, and adds generalization to the model.

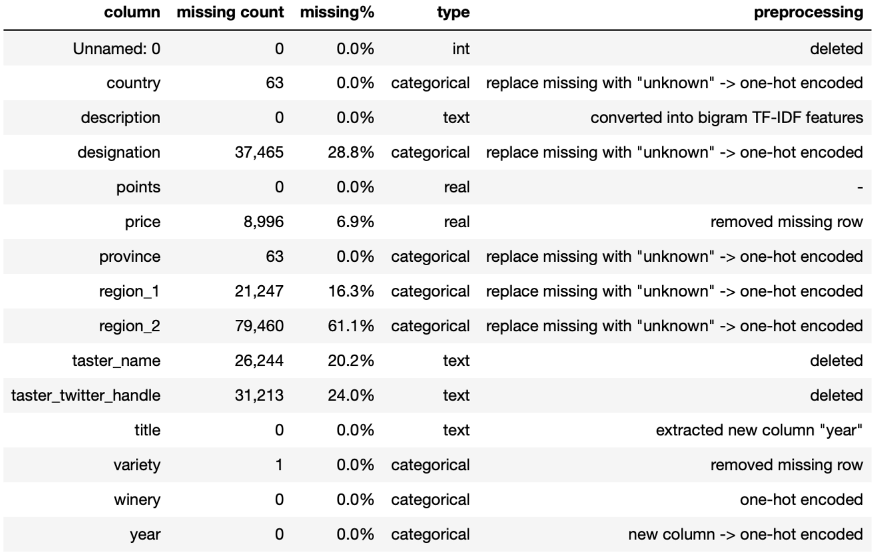
Both *taster_twitter_handle* and *taster_name*, which encodes wine reviewer information, are removed. We do not need to take into consideration who wrote the review, and it will not contribute to the prediction. All the categorical features are then processed with one-hot encoding. *Title* contains repetitive information that can be represented by other columns, so we extracted *year* from it and then removed title. For *description* with long texts,we used NLP tools to process them, which is introduced in the next subsection.

2.2 Text Pre-processing

Description is a paragraph of text describing the wine. At first, we used simple uni-gram (bag-of-words) model to represent the description. During homework 3, we have learned that this might result in overfitting, and uni-gram suffers from keeping semantics like "not good". As an improvement, we then applied bi-gram and used TF-IDF instead of simple word count as feature representation. The whole text pre-processing includes removing punctuation and stopwords, creating bi-gram tokens, calculating TF-IDF and forming feature matrix.

2.3 Exploratory Data Analysis

(See figure2 below.)



column	missing count	missing%	type	preprocessing
Unnamed: 0	0	0.0%	int	deleted
country	63	0.0%	categorical	replace missing with "unknown" -> one-hot encoded
description	0	0.0%	text	converted into bigram TF-IDF features
designation	37,465	28.8%	categorical	replace missing with "unknown" -> one-hot encoded
points	0	0.0%	real	-
price	8,996	6.9%	real	removed missing row
province	63	0.0%	categorical	replace missing with "unknown" -> one-hot encoded
region_1	21,247	16.3%	categorical	replace missing with "unknown" -> one-hot encoded
region_2	79,460	61.1%	categorical	replace missing with "unknown" -> one-hot encoded
taster_name	26,244	20.2%	text	deleted
taster_twitter_handle	31,213	24.0%	text	deleted
title	0	0.0%	text	extracted new column "year"
variety	1	0.0%	categorical	removed missing row
winery	0	0.0%	categorical	one-hot encoded
year	0	0.0%	categorical	new column -> one-hot encoded

Figure 1: Features and preprocessing

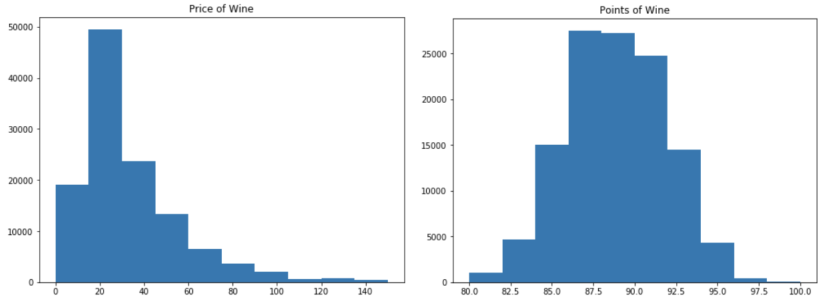


Figure 2: Visualization of numerical data

3 Preliminary Analysis

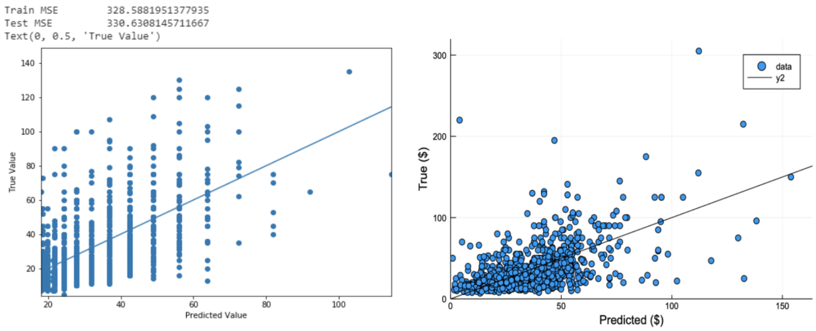
3.1 Linear Regressions

It is reasonable to assume there exist positive relation between wine rating and its price, so we first run a simple linear regression and polynomial regression with the log of the points and square of points. The square of points gave the best MSE. However, since there are only 20 unique ratings. The model is obviously underfitting.

We split our data into a training set and a test set with a common 8:2 ratio. Also, we used test MSE to evaluate our model, and detected overfitting through comparison of training and test MSEs. In this preliminary analysis we are not using cross-valitation, but we will be using it for more complex models in the future.

Then we tried combinations other features. Training MSE was smaller but the testing MSE was around 1,000, indicating overfitting. Similar results were found using text features.

Trying to avoid overfitting will be of high priority in the following works. Bottom right plot is a model using country, province, winery, and variety.



```
Train MSE        328.5881951377935
Test MSE         330.6308145711667
Text(0, 0.5, 'True Value')
```

Figure 3: Linear Regressions on different feature combinations

3.2 Overfitting

From exploratory data analysis and data pre-processing above, we can see that we have around 120k lines of data. Most of our features are categorical, and number of their unique values range from 43 (country) to more than 35k (designation). Besides, the description text bi-gram are with 736,059 columns. Taking all these into consideration, the number of features and size of data are of the same order, which may cause overfitting.

To avoid overfitting, we plan to mark low frequency features, especially ones for text data with "unknown" to reduce the number of features. Prune regression tree is another way to decrease the number of inner nodes and leaves. We will also apply cross-validation and plot errors to detect model overfitting.

4 Remaining Work

In the rest of the semester, we plan to further explore the project by:

- Fitting more models with different combinations of features to decide a better combination of features.
- Trying regression trees since there are many categorical features.
- Trying ensemble methods like Random Forest. We hope Random Forest will improve the prediction accuracy and reduce overfitting.
- Adding regularizers to linear model. With ridge and lasso regression we hope to achieve a more generalized model with lower feature weights on certain features or even reduce the number of features.
- Applying cross-validation for more complex models.