STAT 542 Individual Project: Regression Analysis on Wine Review & Recommendations

By Yuan Shi

Section1: Project description and summary

For this project, we are interested in creating predictive models that predicts points (ratings) rated by WineEnthusiast on a scale of 1 to 100. Such points can be meaningful to determine how good a out-of-book wine is. The Wine Review dataset contains 129971 observations and 14 variables, among the 14 variables, there are only 1 variable is numeric while everything else are categorical. Also, missing data is involved and there are a lot of levels in those categorical variables. I decided to build a Catboost model, which a gradient boosting model with categorical support out of the box, and a random forest model, which is also a tree-based model, but it takes a different approach.

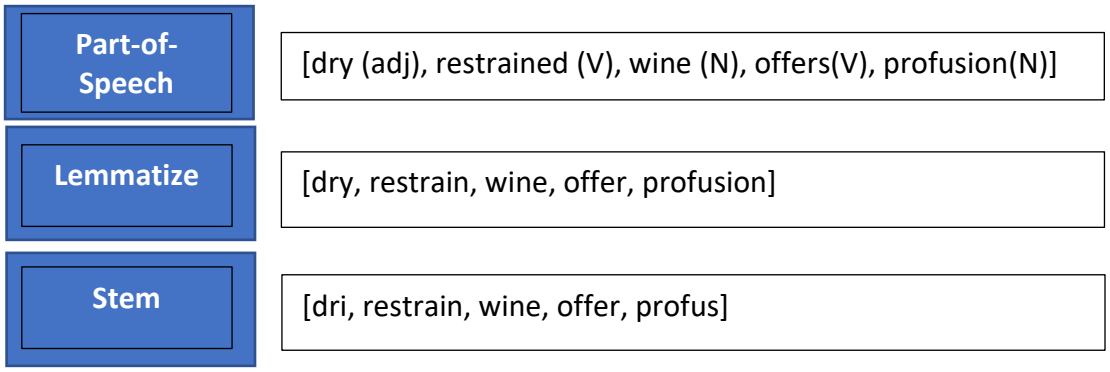
In summary, the Catboost model performs better than Random Forest model with a RMSE of 1.71 on Test set compared to a RMSE of 1.85 from a random forest model. Parameter-tuning and additional processing taken for each method will be further discussed in this report in the next few sections.

Section 2: Data processing

- Pre-process Text Data

This dataset contains 13 categorical variables and 1 numeric variable. The only numeric variable besides response variable, price, has no missing and good to go. Among the categorical variable, the 'description' variable contains detailed description of wine written by actual humans with a good amount of words. Such data can be considered as text data, and every observation of from this variable is very likely unique, a traditional encoding method like dummy coding will not work well. However, like dummy coding, our goal is to process this variable from text variable to numeric variable, just through a different approach. In short, I took 5 steps to process text data: Let's see an example: "This dry and restrained wine offers spice in profusion." to demonstrate what each step does:

Tokenize	[this, dry, and, restrained, wine, offers, spice, in, profusion]
Remove Stopwords	[~~this~~, dry, ~~and~~, restrained, wine, offers, spice, ~~in~~, profusion]



Stem is an extra step that reduce vocabulary to its root form, sometimes it is not readable by people, but it helps the machine 'read'. After all pre-processing, TF-IDF (term-frequency-inverse-document-frequency) is applied to cleaned description so 200 columns (one per important term) with weights replaced the 'description' variable.

Processing other categorical variables

Dummy coding is applied to other categorical variables for random forest model (Catboost handle categorical automatically), now, all variables are converted to numeric.

Section 3: Descriptive statistics

Missingness

The table below (table 1) shows the missing information for this dataset, as we can see from the table, 9 variables involve missing values and 4 variables have more than 20% missing. Tree-based methods have advantages dealing with missing value, although most missing here are categorical, I still decided to pick two tree-based methods from different approaches.

	total_na	percent_missing
country	63	0.048472
description	0	0.000000
designation	37465	28.825661
points	0	0.000000
price	8996	6.921544
province	63	0.048472
region_1	21247	16.347493
region_2	79460	61.136715
taster_name	26244	20.192197
taster_twitter_handle	31213	24.015357
title	0	0.000000
variety	1	0.000769
winery	0	0.000000

Table 1: Missing percentage per variable

Points Distribution

The graph below (figure 1) provided a distribution plot of 'points' variable, which is our response variable. As we can see from the figure, the data follow normal distribution.

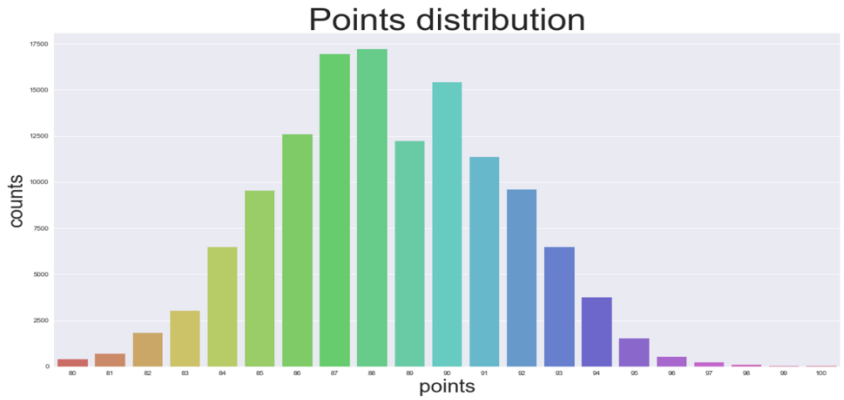


Figure1: Points Distribution

In short, the variables from the wine dataset are mostly categorical, and it does involve a lot of missing values. Tree-based methods require less cleaning and is generally performing well on complex data. The first regression model comes to mind is Catboost, because Catboost does not require coding categorical features into numeric, it handle categorical features very well automatically. Random forest is also a good option, such tree model is not affected by missing values significantly.

Section 4: Regression Model Analysis

Catboost Regression Model

1. Model benefits

Catboost model name comes from two words: "categorical" and "Boosting". Catboost support both categorical and numeric features, one huge benefit of Catboost is that it handles categorical features **automatically**, using various statistics on combinations categorical features even some interaction terms, which is better than just dummy coding.

2. Model Validation

In order to prevent overfitting, I split the data into three subsets: training (70%), validation (15%) and test (15%). Usually we just do training and validation, but this time I decide to have a validation set using Catboost feature to tune towards validation set. And have a holdout test set that is not involving in model-building stage for final result (RMSE).

3. Model Tuning

Just like other boosting method, Catboost creates a number of weak learner(trees) and combined into a strong model. One critical parameter to tune is number of iterations. I used the validation set to calculate the RMSE at each iteration, as we can see in Figure 2 below, the RMSE on validation set appears to turn flat after 500 iterations, so 500 iterations will be sufficient for a good model.

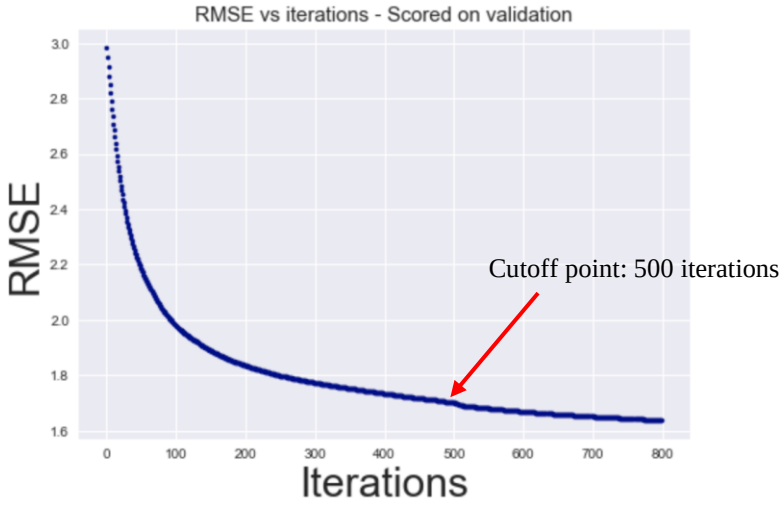


Figure 2: Catboost Validation RMSE per iterations

4. Variable Importance

Catboost can general a variable importance table, among all variable entered, 'description' is the most important one, it does not show here because the TF-IDF applied breaks it into 200 variables. Other than that, all variable importance can be viewed in Figure 3 below. Clearly, besides 'description', price is the most important variable.

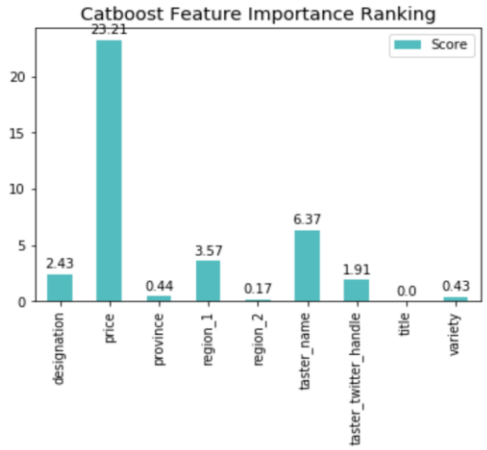


Figure 3: Variable Importance from Catboost

5. Model Results

The Catboost model performs well on both Train, Validation and Test set. Table 2 is a summary of model performance and top 3 variables that affect wines' rating points.

RMSE	Top 3 Important Predictors
Train:1.609	Description (1)
Validation: 1.702	Price (2)
Test(holdout): 1.710	Taster_name (3)

Table 2: Catboost Model Result

The RMSE is low given a sample size over 120,000 observations, also, the top 3 variables that are selected as most important hold an intuitive relationship with wine's scores.

Random Forest Model

- Comparison with Catboost

Unlike Catboost model, the random forest model requires us to dummy code all categorical variables, the runtime took longer, the RMSE from RF model is similar with what we have with Catboost model. See Comparison summary below in Table 3:

Model	RMSE on Test set	Top 3 Variables	Runtime	Dummy Coding?
Random Forest	1.852	Description, Price, Taster_name	142s	Yes
Catboost	1.710	Description, Price, Taster_name	328s	No(automatically)

Table 3: Model Comparison and Regression Results

In summary, both models work well by having a low RMSE and identify the top 3 variables that affect wine's rating(points). However, in term of RMSE, runtime, and processing required, **Catboost is a better choice** than random forest.

Section 5: Recommendations of five wineries

Top 10 within the selected category

The customer is interested in purchasing a pinot noir that has a fruity taste, as well as with a price less than 20 dollars. There are 39401 wineries sell wine that is under 20 dollars, among those wineries, there are 20357 wineries that also contains the keyword 'fruit' in its descriptions. If we select the variety to be 'Pinot Noir' among the 20357 wineries, we will end up with 550 wineries that produce fruity Pinot Noir that is under 20 dollars.

We can calculate the average points for the 550 wineries here and produce a table that contains the top 10 wineries that have the highest average scores, they are: Stadlmann, Tangley Oaks, Starmont, Schug, Meinklang, Markowitsch, Garnet, J. Lohr, Louis Max.

Top 5 in general among the Top 10 within category

We have the top 10 wineries that has the highest average scores in the category, however, this is not good enough, we need to make sure that those wineries are good in general. So, we calculate the average points for all wine that are made in each of the 10 wineries selected from previous steps, and we select top 5. This is the five wineries that are not only good in making fruity pinot noir but also have good points for all the wine they made. Let's see the table 4 below

avg_points_general	winery	fruity_pinot_points
91.000000	Stadlmann	93
90.619048	Markowitsch	92
90.571429	Meinklang	92
88.807692	Starmont	92
88.397849	J. Lohr	92
88.300000	Garnet	92
87.958333	Schug	92
87.500000	Louis Max	91

Selected for recommendations

Table 4: fruity wine points vs general wine points per winery

In summary, I will recommend: **Stadlmann, Markowitsch, Meinklang, Starmont and J. Lohr** to the customer, because they are good at making fruity pinot noir as well as all general wines.