Wine Sales

Biguzzi, Connin, Greenlee, Moscoe, Sooklall, Telab, and Wright

11/05/2021

## Introduction

Using the information about sample wine orders by restaurants and wine stores after a tasting, how can we predict wine sales by various wine characteristics? Using the wine dataset with 12,000 entries and variables mostly related to the chemical properties of each wine, we will build a count regression model to predict the number of cases of wine that will be sold. In practice, if a wine manufacturer can predict which wines will lead to greater sales, they can choose to offer those at more tastings in restaurants and wine stores.

In this report we will:

* explore the data
* transform the data to meet conditions of count modeling
* compare models
* select an optimal model
* generate predictions for the evaluation data set

## Data Exploration

As part of our initial data exploration we want to look at the following key diagnostics:

update as sections fill in below

### Distribution

Using the skimr package we take a look at the raw dataset. There are 12,795 cases and 14 potential predictor variables. All variables are numeric.

With regards to missingness, we have 6 variables with data for every case. This leaves 8 variables that have some degree of missing data, with the worst being the STARS variable with only 73.7% of cases having a value for STARS (the wine rating by experts).

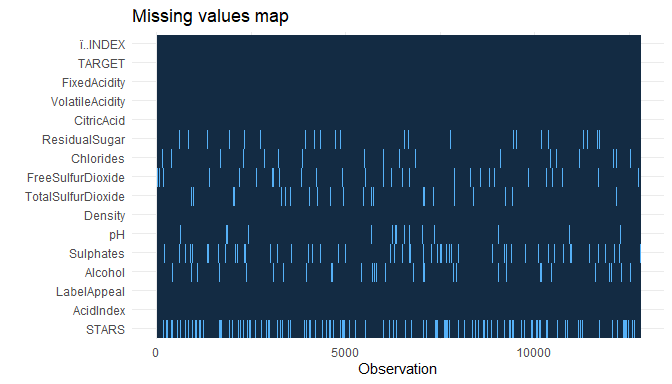
Data summary

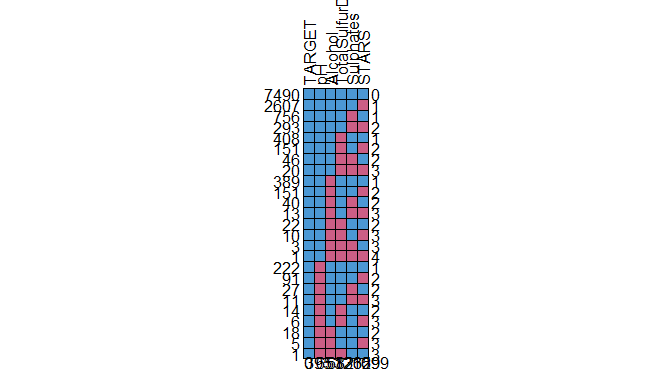
|  |  |
| --- | --- |
| Name | raw |
| Number of rows | 12795 |
| Number of columns | 16 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| numeric | 16 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ï..INDEX | 0 | 1 | 8070 | 4657 | 1 | 4038 | 8110 | 12106 | 16129 |
| TARGET | 0 | 1 | 3 | 2 | 0 | 2 | 3 | 4 | 8 |
| FixedAcidity | 0 | 1 | 7 | 6 | -18 | 5 | 7 | 10 | 34 |
| VolatileAcidity | 0 | 1 | 0 | 1 | -3 | 0 | 0 | 1 | 4 |
| CitricAcid | 0 | 1 | 0 | 1 | -3 | 0 | 0 | 1 | 4 |
| ResidualSugar | 616 | 1 | 5 | 34 | -128 | -2 | 4 | 16 | 141 |
| Chlorides | 638 | 1 | 0 | 0 | -1 | 0 | 0 | 0 | 1 |
| FreeSulfurDioxide | 647 | 1 | 31 | 149 | -555 | 0 | 30 | 70 | 623 |
| TotalSulfurDioxide | 682 | 1 | 121 | 232 | -823 | 27 | 123 | 208 | 1057 |
| Density | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| pH | 395 | 1 | 3 | 1 | 0 | 3 | 3 | 3 | 6 |
| Sulphates | 1210 | 1 | 1 | 1 | -3 | 0 | 0 | 1 | 4 |
| Alcohol | 653 | 1 | 10 | 4 | -5 | 9 | 10 | 12 | 26 |
| LabelAppeal | 0 | 1 | 0 | 1 | -2 | -1 | 0 | 1 | 2 |
| AcidIndex | 0 | 1 | 8 | 1 | 4 | 7 | 8 | 8 | 17 |
| STARS | 3359 | 1 | 2 | 1 | 1 | 1 | 2 | 3 | 4 |

Looking for patterns in missingness can be telling, we use the missing package to visualize the missing data.





## TARGET pH Alcohol TotalSulfurDioxide Sulphates STARS   
## 7490 1 1 1 1 1 1 0  
## 2607 1 1 1 1 1 0 1  
## 756 1 1 1 1 0 1 1  
## 293 1 1 1 1 0 0 2  
## 408 1 1 1 0 1 1 1  
## 151 1 1 1 0 1 0 2  
## 46 1 1 1 0 0 1 2  
## 20 1 1 1 0 0 0 3  
## 389 1 1 0 1 1 1 1  
## 151 1 1 0 1 1 0 2  
## 40 1 1 0 1 0 1 2  
## 13 1 1 0 1 0 0 3  
## 22 1 1 0 0 1 1 2  
## 10 1 1 0 0 1 0 3  
## 3 1 1 0 0 0 1 3  
## 1 1 1 0 0 0 0 4  
## 222 1 0 1 1 1 1 1  
## 91 1 0 1 1 1 0 2  
## 27 1 0 1 1 0 1 2  
## 11 1 0 1 1 0 0 3  
## 14 1 0 1 0 1 1 2  
## 6 1 0 1 0 1 0 3  
## 18 1 0 0 1 1 1 2  
## 5 1 0 0 1 1 0 3  
## 1 1 0 0 0 1 1 3  
## 0 395 653 682 1210 3359 6299