HW 5

Team 2

May 15, 2019

Contents

OVERVIEW

In this homework assignment, you will explore, analyze and model a data set containing information on approximately 12,000 commercially available wines. The variables are mostly related to the chemical properties of the wine being sold. The response variable is the number of sample cases of wine that were purchased by wine distribution companies after sampling a wine.

These cases would be used to provide tasting samples to restaurants and wine stores around the United States. The more sample cases purchased, the more likely is a wine to be sold at a high end restaurant. A large wine manufacturer is studying the data in order to predict the number of wine cases ordered based upon the wine characteristics. If the wine manufacturer can predict the number of cases, then that manufacturer will be able to adjust their wine offering to maximize sales.

Objective

Your objective is to build a count regression model to predict the number of cases of wine that will be sold given certain properties of the wine. HINT: Sometimes, the fact that a variable is missing is actually predictive of the target. You can only use the variables given to you (or variables that you derive from the variables provided).

Dataset

Below is a short description of the variables of interest in the data set:

- `INDEX`: Identification Variable (do not use).
- `TARGET`: Number of Cases Purchased.
- `AcidIndex`: Proprietary method of testing total acidity of wine by using a weighted average.
- `Alcohol`: Alcohol Content.
- `Chlorides`: Chloride content of wine.
- `CitricAcid`: Citric Acid Content.
- `Density`: Density of Wine.
- `FixedAcidity`: Fixed Acidity of Wine.
- `FreeSulfurDioxide`: Sulfur Dioxide content of wine.
- `LabelAppeal`: Marketing Score indicating the appeal of label design for consumers. High numbers suggest customers like the label design. Negative numbers suggest customes don't like the design.
- `ResidualSugar`: Residual Sugar of wine.
- `STARS`: Wine rating by a team of experts. 4 Stars = Excellent, 1 Star = Poor.
- `Sulphates`: Sulfate conten of wine.
- `TotalSulfurDioxide`: Total Sulfur Dioxide of Wine.
- `VolatileAcidity`: Volatile Acid content of wine.
- `pH`: pH of wine.

There is a theoretical effect for `LabelAppeal`, which suggests many consumers purchase based on the visual appeal of the wine label design. Higher numbers suggest better sales. Additionally, a high number captured by the `STARS` variable is theorized to suggest high sales.

Dependencies

Replication of our work requires the following packages in Rstudio:

```
# descriptive analysis packages
library(psych)
library(randomForest)
library(corrplot)

# data transformations packages
library(dplyr)
library(tidyr)

# data visualization packages
library(ggplot2)
library(reshape2)
```

PART 1: DATA EXPLORATION

First, we read the data as a csv and then examined the below variable from the `training` dataset.

```
training <- as.data.frame(read.csv("wine-training-data.csv"))
test <- as.data.frame(read.csv("wine-evaluation-data.csv"))
```

Summary Statistics

We look at summary of the data below.

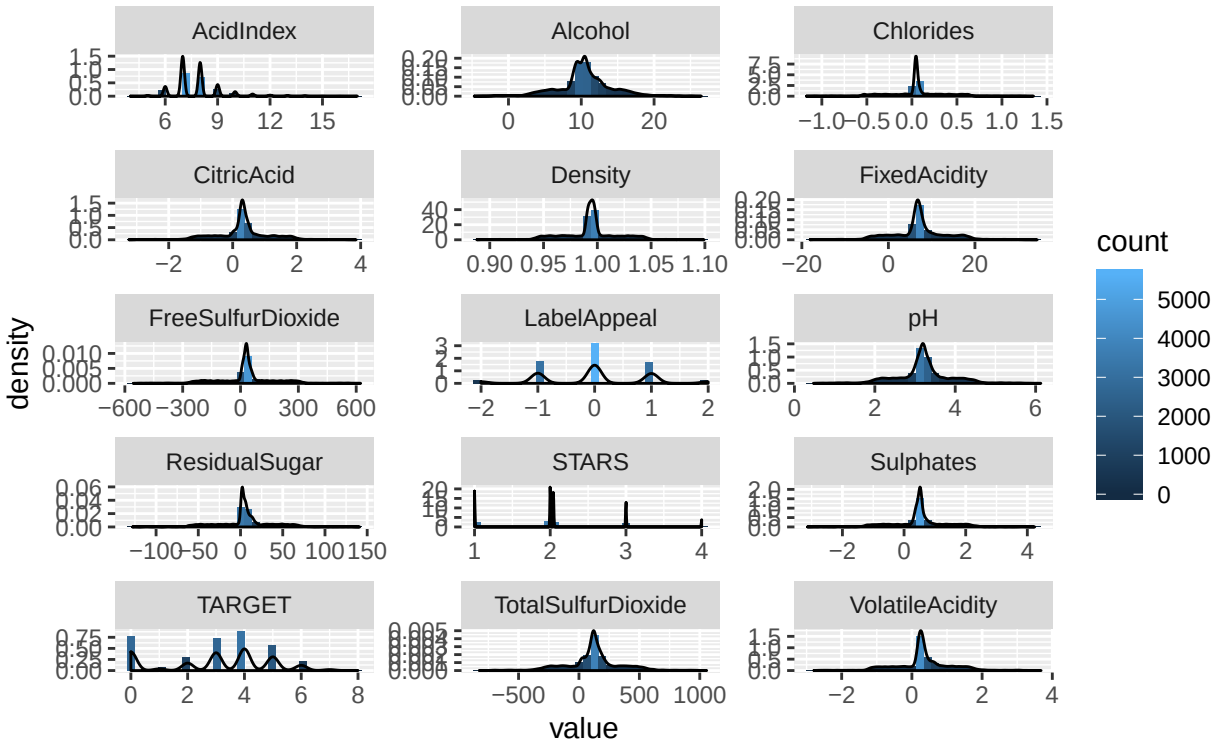
	vars	n	mean	sd	min	max	range	se
TARGET	1	12795	3.03	1.93	0.00	8.00	8.00	0.02
FixedAcidity	2	12795	7.08	6.32	-18.10	34.40	52.50	0.06
VolatileAcidity	3	12795	0.32	0.78	-2.79	3.68	6.47	0.01
CitricAcid	4	12795	0.31	0.86	-3.24	3.86	7.10	0.01
ResidualSugar	5	12179	5.42	33.75	-127.80	141.15	268.95	0.31
Chlorides	6	12157	0.05	0.32	-1.17	1.35	2.52	0.00
FreeSulfurDioxide	7	12148	30.85	148.71	-555.00	623.00	1178.00	1.35
TotalSulfurDioxide	8	12113	120.71	231.91	-823.00	1057.00	1880.00	2.11
Density	9	12795	0.99	0.03	0.89	1.10	0.21	0.00
pH	10	12400	3.21	0.68	0.48	6.13	5.65	0.01
Sulphates	11	11585	0.53	0.93	-3.13	4.24	7.37	0.01
Alcohol	12	12142	10.49	3.73	-4.70	26.50	31.20	0.03
LabelAppeal	13	12795	-0.01	0.89	-2.00	2.00	4.00	0.01
AcidIndex	14	12795	7.77	1.32	4.00	17.00	13.00	0.01
STARS	15	9436	2.04	0.90	1.00	4.00	3.00	0.01

From this, we can see that some variables have missing values. These values require imputation to conduct further analysis. We imputed the variables with missing data below using the corresponding variables' mean value.

```
NA_replace <- function(x) replace(x, is.na(x), mean(x, na.rm = TRUE))
training[] <- lapply(training, NA_replace)
```

Distribution

Below, we examine the distribution of values using histograms and density plots for each variable.

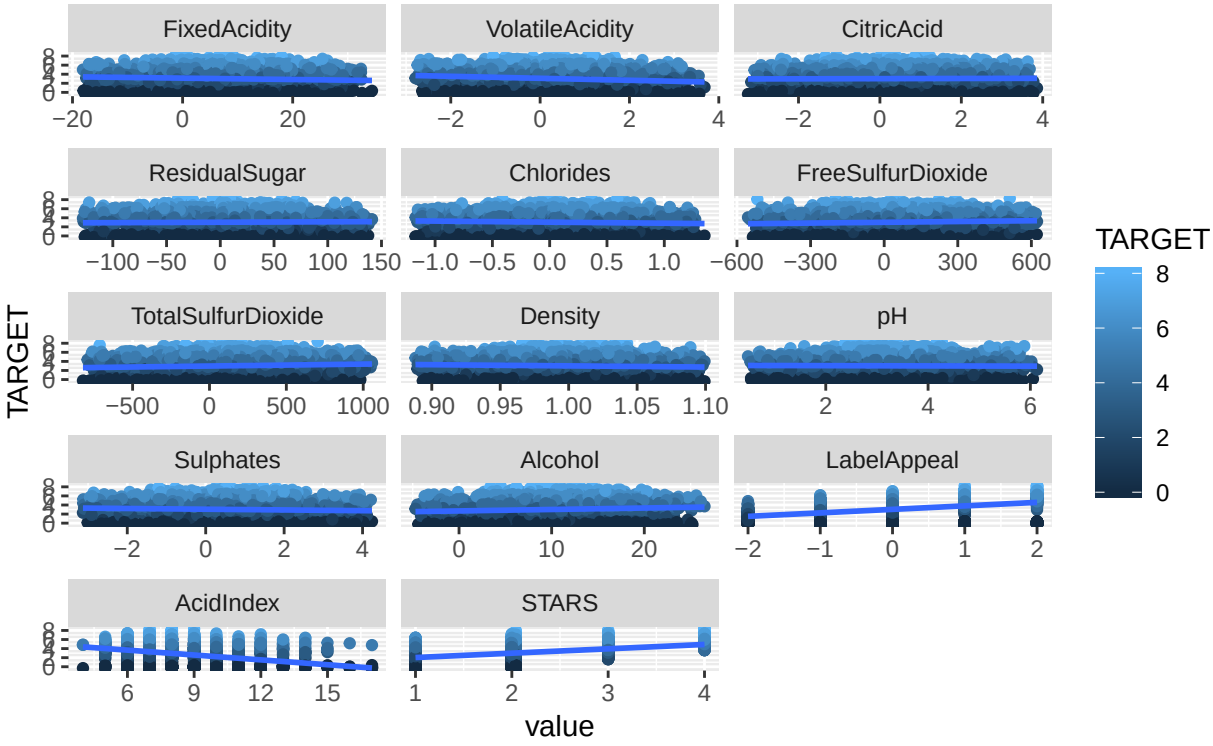


Most variables appear to be fairly normally distributed. However, we can tell that `LabelAppeal` and `STARS` are ordinal variables that measure specific rankings. Additionally, we can tell that the `AcidIndex` has a multimodal distribution that is skewed to the left.

Correlation Plot Matrix

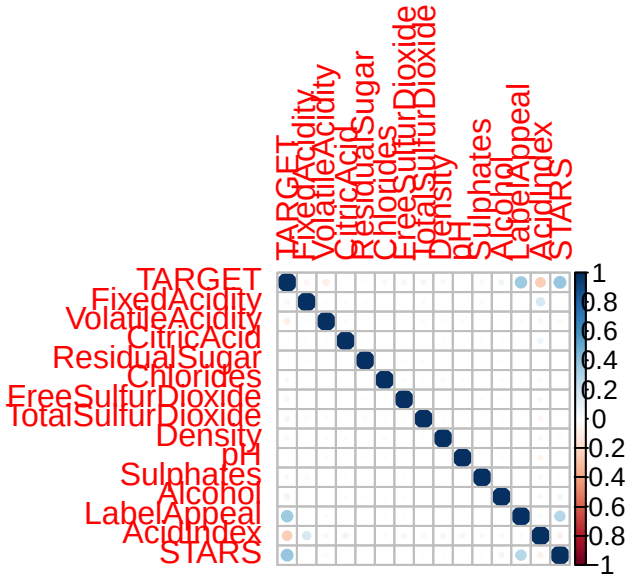
Scatter plot matrix with the `pairs` function does not seem useful given the amount of variables. Recommend using the ggplot correlation plot matrix, which shows linear relationship between the predictor and response variables.

The correlation plot matrix below shows linear relationship between the predictor and response variables.



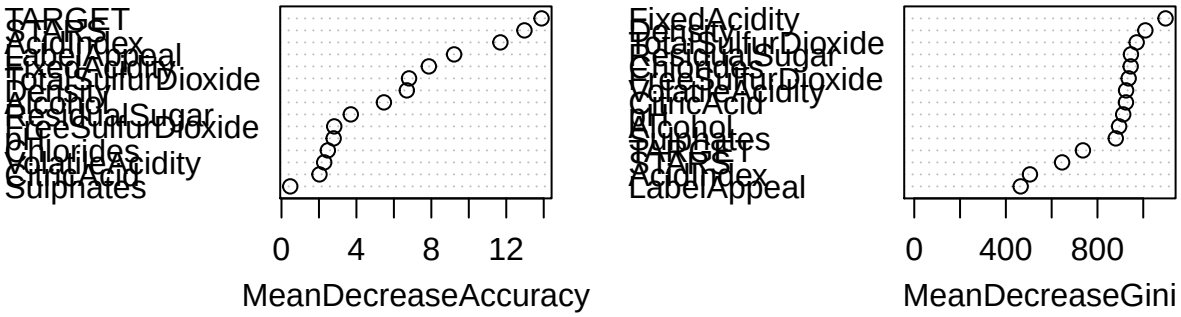
Correlation

We can see our correlation matrix below. A dark blue circle represents a strong positive relationship and a dark red circle represents a strong negative relationship between two variables.

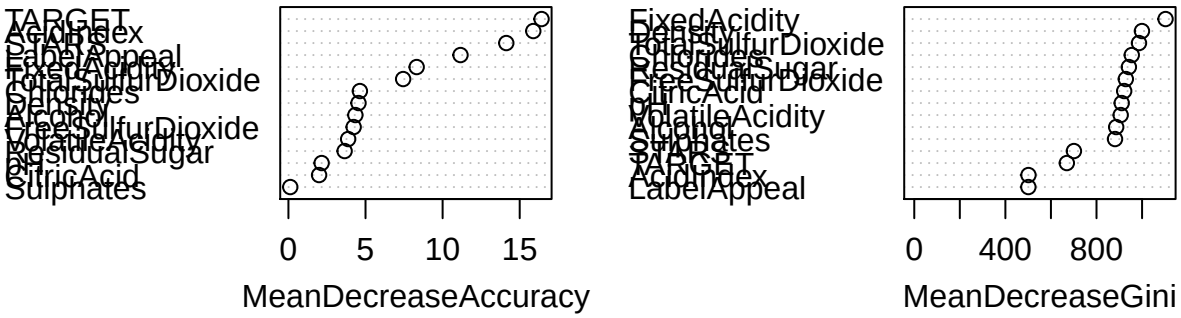


Finally, we can use the `randomforest` package to verify our assumptions from the correlation plot.

fit1



fit2



PART 2: DATA EXPLORATION

Describe how you have transformed the data by changing the original variables or creating new variables. If you did transform the data or create new variables, discuss why you did this. Here are some possible transformations.

Transformations

- Transform data by putting it into buckets.
- Mathematical transforms such as log or square root (or use Box-Cox).

New Variables

- Combine variables (such as ratios or adding or multiplying) to create new variables

PART 3: BUILD MODELS

Using the training data set, build:

- 2 different negative binomial regression models.
- 2 different poisson regression models.
- 2 different multiple linear regression model

Sometimes poisson and negative binomial regression models give the same results. If that is the case, comment on that. Consider changing the input variables if that occurs so that you get different models.

Although not covered in class, you may also want to consider building zero-inflated poisson and negative binomial regression models. You may select the variables manually, use an approach such as Forward or Stepwise, use a different approach such as trees, or use a combination of techniques. Describe the techniques you used. If you manually selected a variable for inclusion into the model or exclusion into the model, indicate why this was done.

Discuss the coefficients in the models, do they make sense? In this case, about the only thing you can comment on is the number of stars and the wine label appeal. However, you might comment on the coefficient and magnitude of variables and how they are similar or different from model to model. For example, you might say "pH seems to have a major positive impact in my poisson regression model, but a negative effect in my multiple linear regression model". Are you keeping the model even though it is counter intuitive? Why? The boss needs to know.

Negative Binomial Regression

Model 1

Model 2

Poisson Regression

Model 3

Model 4

Multiple Linear Regression

Model 5

Model 6

PART 4: SELECT MODELS

Decide on the criteria for selecting the best count regression model. Will you select models with slightly worse performance if it makes more sense or is more parsimonious? Discuss why you selected your models.

Model Evaluation

For the count regression model, will you use a metric such as AIC, average squared error, etc.? Be sure to explain how you can make inferences from the model, and discuss other relevant model output. If you like the multiple linear regression model the best, please say why. However, you must select a count regression model for model deployment.

Using the training data set, evaluate the performance of the count regression model.

Forecasting

Make predictions using the evaluation data set