What Makes Specific Wines Better than Others?

AJ Rallo, arallo@bellarmine.edu & Shay Korhorn, skorhorn@bellarmine.edu & Anthony Streib, astreib@bellarmine.edu

ABSTRACT

Up to 150 word summary of your project.

I. INTRODUCTION

For our final Project, We decided to use a wine quality data set because it was simple and easy to create a training module. In past assignments, we have had trouble using more complicated datasets, so we chose to use something simple that would be easy to understand and create linear regression models. We wanted to predict simple correlations between significant variables in the dataset. We were curious whether sugar affected wine quality because it always makes food taste better. So, we decided to make that our hypothesis one to test. Using the trained regression model to draw a correlation between the two variables. So, for our second test, we decided to go with something a little different. We decided to find the correlation between alcohol content and wine quality. Using the same method we implemented for hypothesis #1.

II. BACKGROUND

In this section, provide some background for the problem for which the data were collected. For example, if you were using the mushroom data set, you write up some background on what a mushroom is, why the data were originally collected, what question(s) the authors were trying to answer, etc.

Wine is an alcoholic beverage made from fruits (mainly fermented grapes), crushing and pressing, fermentation, and aging are the steps to creating wine. This data was initially collected to correlate different variables and use these to figure out why certain wines taste better than other kinds.

III. EXPLORATORY ANALYSIS

This section will be similar to your exploratory analysis project. First, provide a summary of the data set similar to your first exploratory analysis: *e.g. this data set contains 398 samples with 7 columns with various data types*. In

this summary, provide the data types of your columns (in a table) and then rather than providing tabular statistics and plots for each variable, provide only statistics and plots that seem unusual. For example, if one or two variables have significant missing values or the distribution of the variable is skewed or looks unusual note that. Provide the unusual statistics or plots in this section. Provide any other appropriate plots (e.g. correlation matrix, heatmaps, bar charts, etc.) that you deem necessary.

Table 1: Data Types

Variable Type	Data Type
Fixed Acidity	**Float**
Volatile Acidity	**Float**
Citric Acid	**Float**
Residual Sugar	**Float**
Chloride	**Float**
Free sulfur dioxide	**Float**
Total sulfur dioxide	**Float**
density	**Float**
pH	**Float**
sulphates	**Float**
Alcohol	**Float**

	coef	std err
const	21.9652	21.195
fixed acidity	0.0250	0.026
volatile acidity	-1.0836	0.121
citric acid	-0.1826	0.147
residual sugar	0.0163	0.015
chlorides	-1.8742	0.419
free sulfur dioxide	0.0044	0.002
total sulfur dioxide	-0.0033	0.001
density	-17.8812	21.633
pH	-0.4137	0.192
sulphates	0.9163	0.114
alcohol	0.2762	0.026

IV. METHODS

A. Data Preparation

For our preparation of the data, We didn't necessarily need to do any data preparation. It was a perfect dataset that didn't have any missing data. There was a ton of independent variables and one major dependant variable which made it perfect to do linear regression on. The biggest problem when preparing the data was deciding what factors were unimportant, versus which ones we should have tried to pay more attention to, in order to correlate the variables (based on our hypothesis) correctly. There was simply no need to prepare the data in any way. We didn't need to normalize any of the variables because the variables all had a similar range that wasn't too significant.

B. Experimental Design

You will run your model several times with different parameters to see what different results you get. In a table, describe your experimental parameters. Three or four experiments are sufficient. This is where you will describe how you divided your data into train, validate and test data sets. For example:

Table X: Experiment Parameters

Experiment Number	**Parameters**
1	Quality vs Sugar Content with 80/20 split for train and test

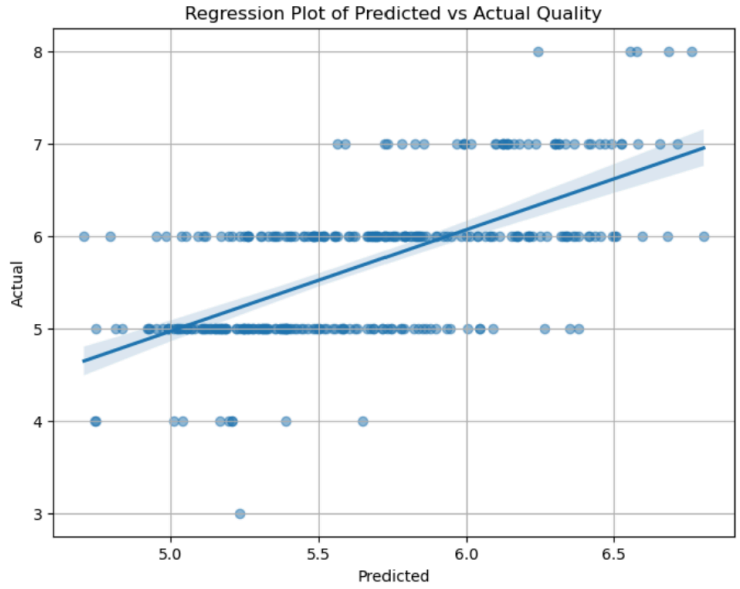
2	Quality vs Alcohol Content for 80/20 split for train and test
3	Quality vs Sugar Content with 70/30 split for train and test
4	Quality vs Alcohol Content with 70/30 split for train and test

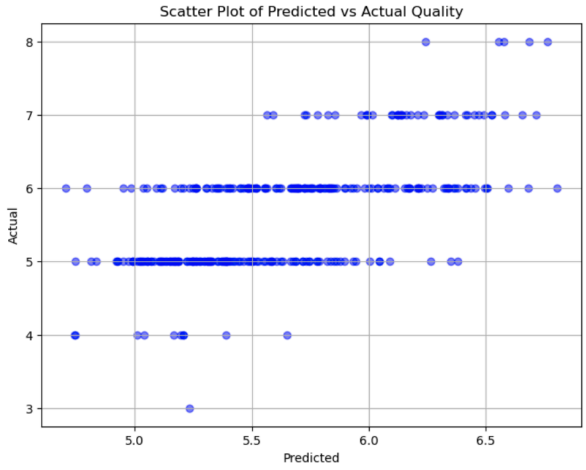
C. Tools Used

The following tools were used for this analysis: Python v3.5.2 running the Anaconda 4.3.22 environment for Windows Surface Pro v5. This Surface Pro computer was used for all analysis and implementation. In addition to base Python, the following libraries were also used: Pandas 0.18.1, SKLearn 0.18.1, and Statsmodels.api. Pandas is an essential toolkit for analyzing excel documents of data, it was necessary. We also used SKLearn to implement machine learning elements and to train the linear regression models that we came up with.

V. RESULTS

A. Classification Measures/ Accuracy measure





This scatterplot shows that, although it is sort of accurate when predicting the wine quality, it is not something that should be relied on all the time and trusted.

Model Evaluation:

In [22]:
```
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
```

In [23]:
```
print("Mean Squared Error:", mse)
print("R-squared:", r2)
```

```
Mean Squared Error: 0.39002514396431653
R-squared: 0.40318034127906854
```

B. Discussion of Results

I think that the first model that compared the relationship between sugar and wine quality was my worst. My second model that had a much stronger relationship (Alcohol-Quality) and gave me a much better idea of what gives better quality wine its edge was the best model.

C. Problems Encountered

One of the main problems we faced was understanding how this dataset was so perfect. We spend over an hour looking for things normalize and errors in the dataset/ incorrect data types, but couldn't find anything.We were very confused but then attempted to train the model and everything went off without a hitch,

D. Limitations of Implementation

I think that their are a few limitations to the model because it is not as good at predicting what it is supposed to predict. I think that one of the main factors for this is there is not category for what type of fruit/unique fruit strain is being used in that specific wine. That is a factor that likely makjes a big different in quality because certain types will create a better wine than others. Another reason could be that people that judge the quality may have a different 'grading scale' than other wine quality judgers. I think if these factors were in the dataset (I don't know how they would be able to be put in there) than it would make the model a better predictor.

E. Improvements/Future Work

I would like to try taking a few of the variables out that wouldnt make much of a difference to the quality and mess around with that and see whether it creates a more accurate prediction model. I'd also try to maybe add a variable or two to see if that does the same thing.

VI. CONCLUSION

We think that this was an easy and simple dataset to use for such a complicated task. The model we created was alright, it wasn't super accurate but it also wasn't inaccurate. We think that the dataset itself was not specific enough with its variables to create a super accurate prediction model. We would try to experiment with adding and taking away variables to make it as accurate as possible. Overall, this has been a very interesting project to work and has made us all understand linear regression and correlation better.