Assessing Wine Quality

Sean Kim

Wine Industry Quality Certifications

- The wine industry is still heavily reliant on human expertise
- Wine quality certifications are essentially a requirement for any wine label/producer
  - Quality certifications help to ensure a quality product for consumers
  - Quality certifications also help to prevent illegal fraud in copying or faking certifications
- Quality certifications are carried out by human experts, or wine sommeliers
- Wine stewards/sommeliers levels:
  - Beginner sommelier
  - Certified sommelier
  - Industry experienced sommelier
  - Master sommelier
- Obtaining a master sommelier status takes years of practice and studying
  - Also a very expensive process requiring fees for courses, practice wine, examinations, etc.
- As of the beginning of 2020, there are only 269 certified master sommeliers
  - Because of this limited population of high level sommeliers (industry experienced and master level), wine certification and quality assessments for new wine labels and brands is very expensive and can be very difficult to obtain

Wine Quality

- Wine quality can be determined through a combination of physicochemical and sensory attributes
- Physicochemical = attributes relating to physics and chemistry
    - E.g. alcohol percentage, pH, sugar level
    - Physicochemical attributes can be tested and measured in a lab
- Sensory = attributes relating to sensation or the physical senses
    - E.g. appearance, odor, flavor
    - Sensory attributes require a human expert
- The relationship between physicochemical attributes and sensory analysis is still not fully understood

Wine Production Process

- Extreme emphasis on quality combined with the length of the wine production process makes the execution even more crucial
  - Very small margin for error in a complex process
- Initial wine making process:
  - Picking the grapes
  - Crushing the grapes
  - Fermenting the grapes
- Final steps of wine production process:
  - Age the wine (anywhere from a few months to multiple years)
  - Bottle the wine
- Every decision throughout the complete production process can affect both the final product's physicochemical and sensory attributes
  - When the grapes are harvested (time of the year and if it's day/night), how the grapes are crushed, how they're stored (wood or metal), how long the wine is aged, whether the wine is bottled with a cork or a screwcap, etc.
  - All of these factors can affect the wine's end quality

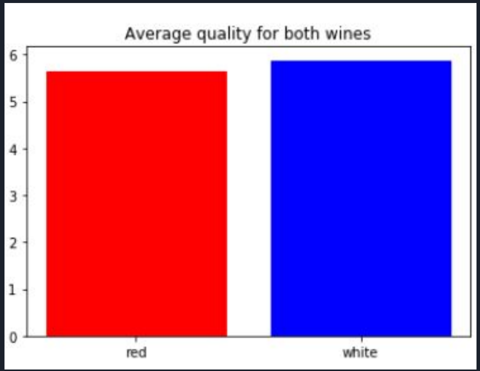
Dataset and Data Wrangling

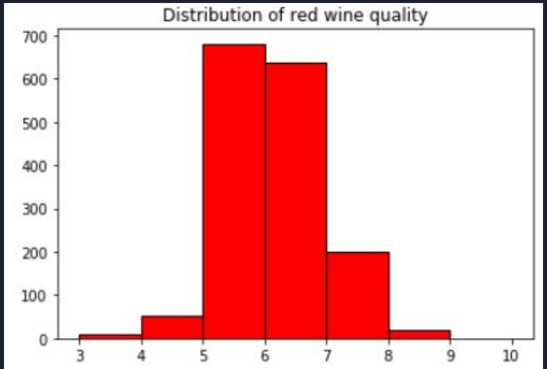
- Separate red (1,599 samples) and white (4,898) wine datasets
- 11 Physicochemical attributes:
  - Alcohol content
  - Chlorides
  - Citric acid
  - Density
  - Fixed acidity
  - Free sulfur dioxide
  - pH
  - Residual sugar
  - Sulphates
  - Total sulfur dioxide
  - Volatile acidity
- Target label: quality
  - 0-10
  - 0 being the worst possible score, 10 being the best possible score
- No records with missing data or NaN values
- Rename each feature column with whitespace and replace with an underscore
- Random split each dataset into train (80%), test (20%) sets

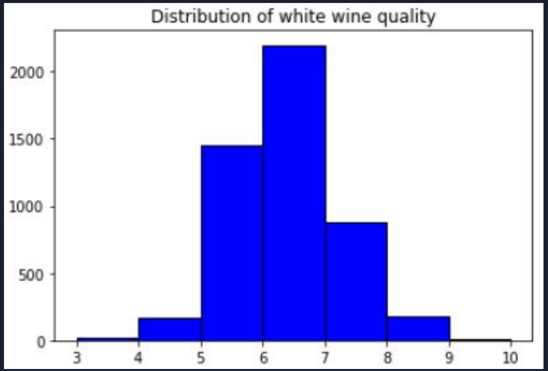
Methodology

1. Exploratory Analysis
2. Linear Regression
3. Decision Trees
4. Random Forests
5. Neural Networks
6. Support Vector Machines

Exploratory Analysis

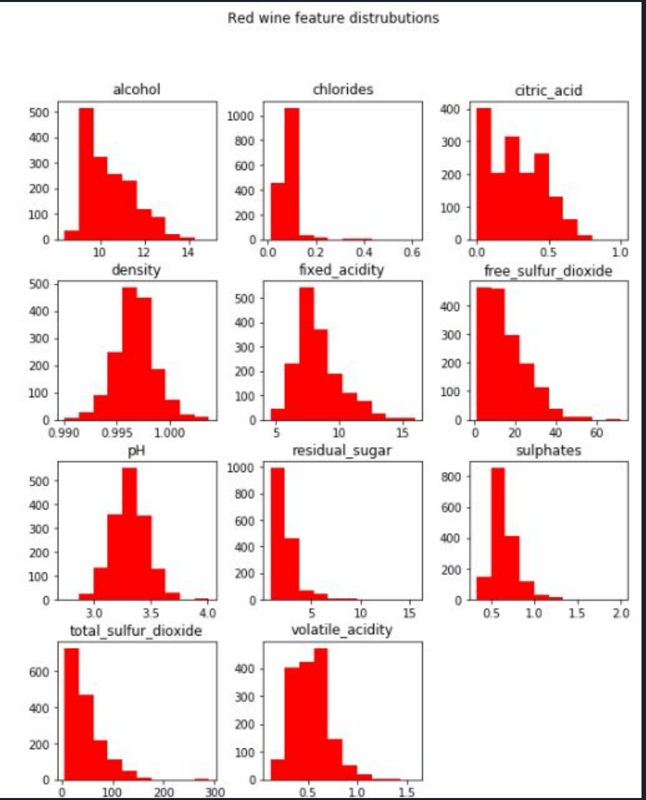






- Average wine quality
  - Red: 5.64
  - White: 5.88
- Distributions:
  - Quality is not normally distributed in either dataset
  - Red wine quality ranges from 3-8
  - White wine quality ranges from 3-9

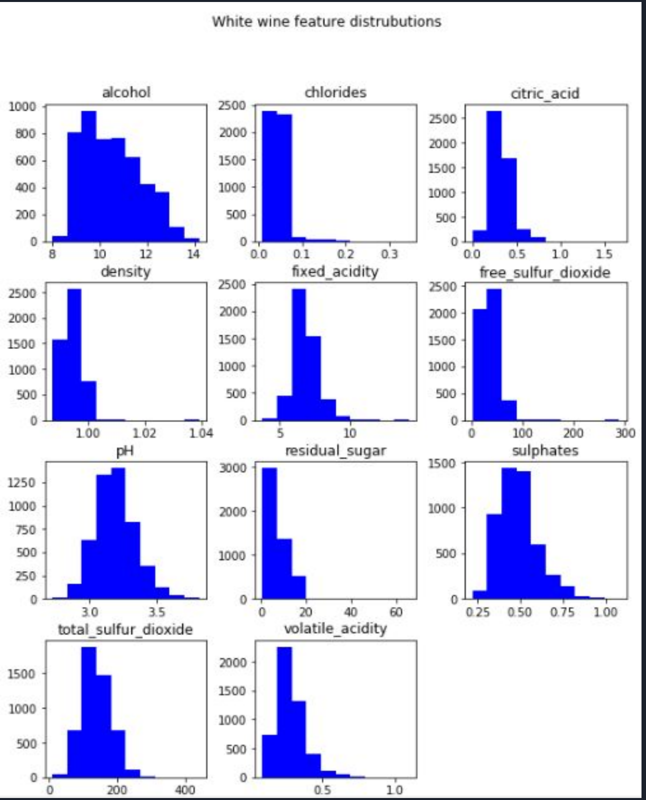
Exploratory Analysis



	feature	p-value	is_normal
1	fixed_acidity	1.7528277735470436e-49	false
2	volatile_acidity	7.192589039756692e-32	false
3	citric_acid	9.662822259281018e-34	false
4	residual_sugar	0	false
5	chlorides	0	false
6	free_sulfur_dioxide	4.779365332171477e-75	false
7	total_sulfur_dioxide	1.433890834343538e-106	false
8	density	2.1473202738102222e-7	false
9	pH	4.846864534772716e-8	false
10	sulphates	1.1759065222978855e-197	false
11	alcohol	3.3163288473185496e-34	false

- None of the physicochemical attributes are normally distributed in the red wine dataset

Exploratory Analysis

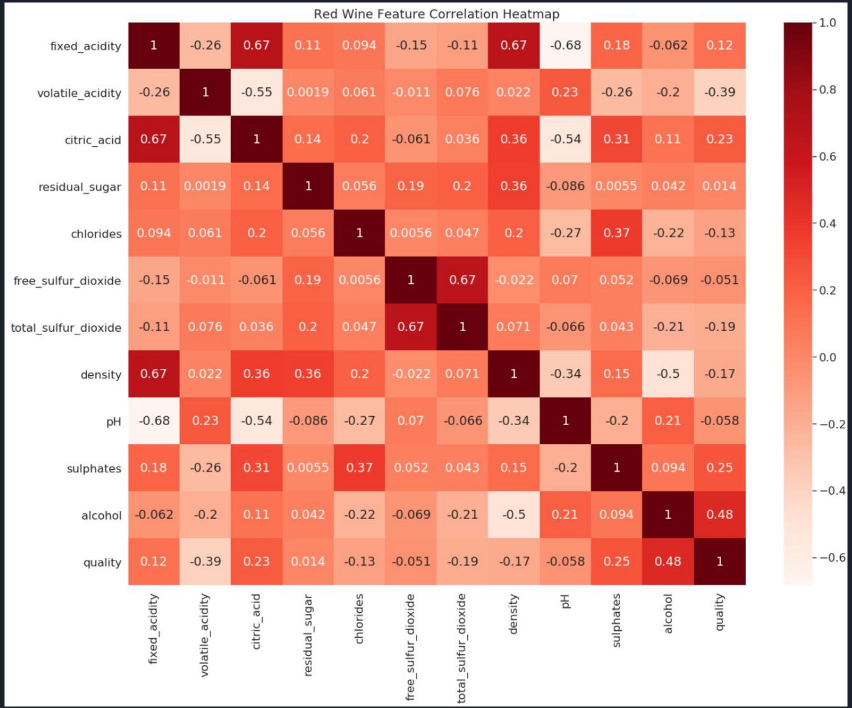


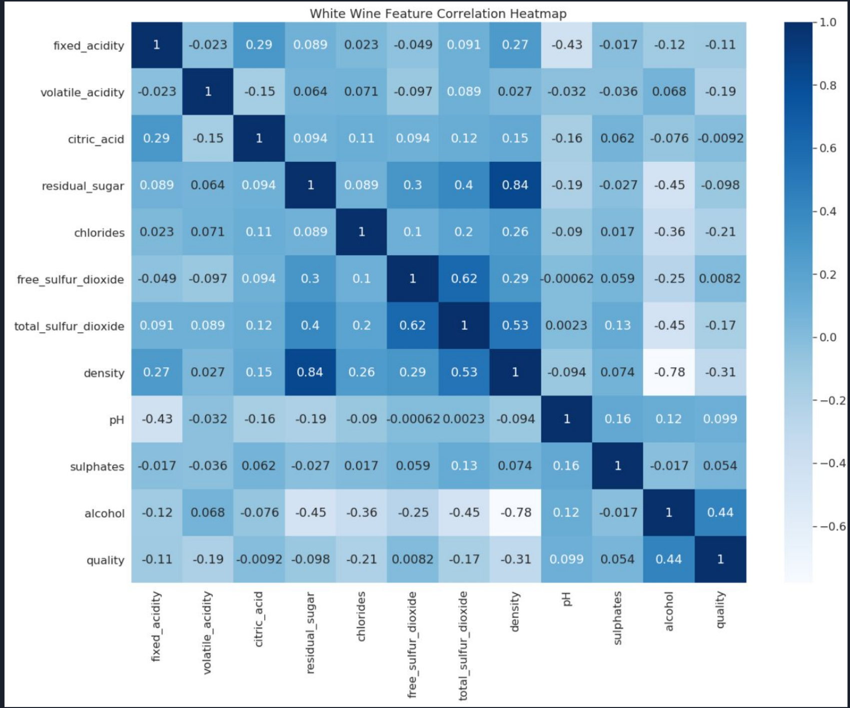
	feature	p-value	is_normal
1	fixed_acidity	5.192296946299217e-118	false
2	volatile_acidity	0	false
3	citric_acid	0	false
4	residual_sugar	1.3781731594418583e-228	false
5	chlorides	0	false
6	free_sulfur_dioxide	0	false
7	total_sulfur_dioxide	4.330752711423771e-35	false
8	density	1.0995470514104443e-307	false
9	pH	1.0892373353367272e-42	false
10	sulphates	8.19207367045622e-161	false
11	alcohol	3.609332261783645e-94	false

- None of the physicochemical attributes are normally distributed in the white wine dataset

Pearson Correlation Coefficient Heatmaps

- In both datasets, alcohol content shows the highest correlation coefficient to quality





Linear Regression Results

Input Feature(s)	Red Wine Model RMSE	White Wine Model RMSE
Alcohol	0.7039	0.8106
Sulphates	0.7888	0.8853
pH	0.8077	0.8819
All	0.6294	0.7755

Decision Trees

- Evaluated using 3 fold cross validation
  - Testing max tree depths of 2, 5, 10 and max bins of 10, 20, 40

	Best Parameters	**RMSE**
Red Wine Model	Max Depth: 5 Max Bins: 20	0.6768
White Wine Model	Max Depth: 5 Max Bins: 40	0.7468

Random Forests

- Hyperparameter tuning done with hyperopt
  - Tuning done on max bins, max depth, and num trees

	Best Parameters	**RMSE**
Red Wine Model	Max Bins: 14 Max Depth: 26 Num Trees: 54	0.5859
White Wine Model	Max Bins: 21 Max Depth: 29 Num Trees: 94	0.6356

Neural Networks (Keras)

- Hyperparameter tuning done with hyperopt
  - Tuning done on # of units in two dense layers, number of epochs, and learning rate
- Model trained and tested on scaled data
  - Data scaled with sklearn's StandardScaler

	Best Parameters	**RMSE**
Red Wine Model	Dense Layer 1: 86 Dense Layer 2: 188 Epochs: 46 Learning Rate: 0.1679	0.6369
White Wine Model	Dense Layer 1: 186 Dense Layer 2: 204 Epochs: 43 Learning Rate: 0.0003	0.6855

Support Vector Machines

- Hyperparameter tuning done with hyperopt
  - Tuning done on C, epsilon, and kernel
- Model trained and tested on scaled data
  - Data scaled with sklearn's StandardScaler

	Best Parameters	**RMSE**
Red Wine Model	C: 20.9674 Epsilon: 0.2483 Kernel: linear	0.6207
White Wine Model	C: 7.6954 Epsilon: 0.4776 Kernel: rbf	0.6784

Conclusions and Future Improvements

- Random forest algorithms produced the most accurate models for both datasets
- Support vector machines also showed promise
- Dataset contains some variables with high multicollinearity (e.g. acidity/pH)
  - Models could benefit from some more exploratory analysis, specifically more advanced feature selection
- Models could also benefit from more hyperparameter tuning done with hyperopt
  - Hyperopt training took very long for some models and the databricks community edition cluster timed out in some cases
  - Could benefit from further studying of the hyperparameter models
    - Create a more efficient hyperopt search space
- Could consider the two datasets together
  - Adding one boolean feature columns for is_red