Wine Quality Prediction

Yucheng Zhu
May 13, 2019

Highlights

- Conducted exploratory data analysis and data visualization
- Prepared new column (recommend or not) for binary classification
- Developed a model based on Random Forest that provides an accuracy of ~93% on the test data set (20% of total data).

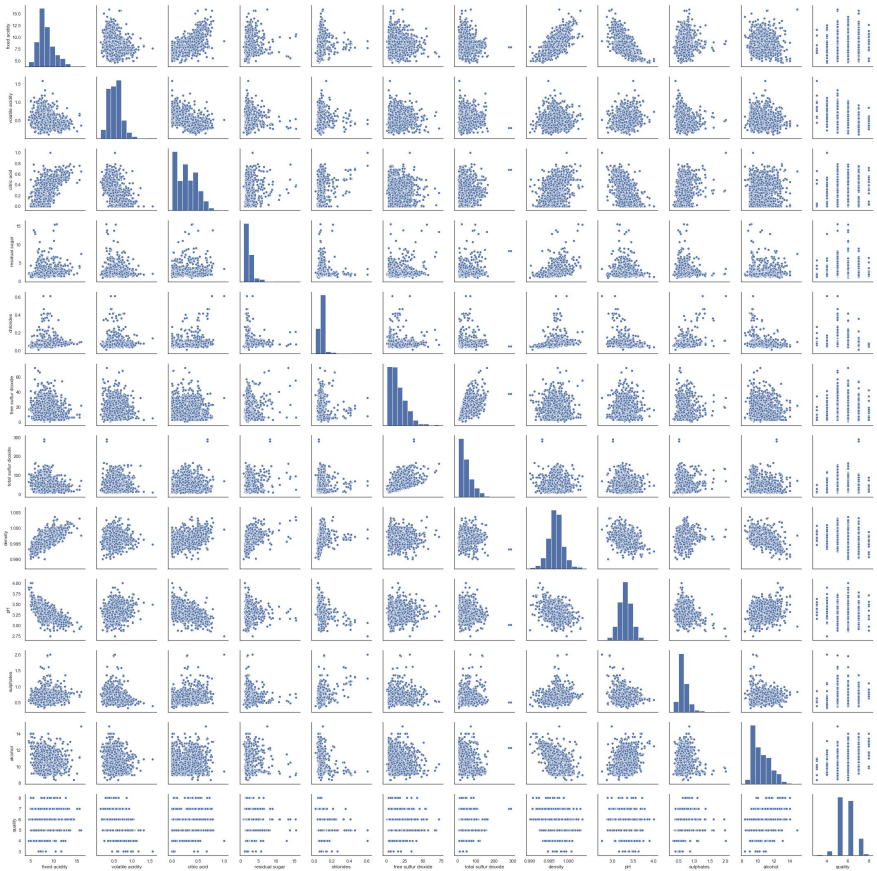
Review Process

- Stories completed during this sprint:
  - 2 Stories in Epic 1(Data Preparation)
    - Downloaded the data file
    - Conducted EDA and Data Cleaning
  - 1 Story in Epic 2 (Modeling and Model Selection)
    - Builded a classification model on randomly selected training data using Random Forest

Demo/analysis

- Features generally are not highly correlated with each other, so multicollinearity is not an issue
- Some features are right-skewed

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1.000000	-0.256131	0.671703	0.114777	0.093705	-0.153794	-0.113181	0.668047	-0.682978	0.183006	-0.061668	0.124052
volatile acidity	-0.256131	1.000000	-0.552496	0.001918	0.061298	-0.010504	0.076470	0.022026	0.234937	-0.260987	-0.202288	-0.390558
citric acid	0.671703	-0.552496	1.000000	0.143577	0.203823	-0.060978	0.035533	0.364947	-0.541904	0.312770	0.109903	0.226373
residual sugar	0.114777	0.001918	0.143577	1.000000	0.055610	0.187049	0.203028	0.355283	-0.085652	0.005527	0.042075	0.013732
chlorides	0.093705	0.061298	0.203823	0.055610	1.000000	0.005562	0.047400	0.200632	-0.265026	0.371260	-0.221141	-0.128907
free sulfur dioxide	-0.153794	-0.010504	-0.060978	0.187049	0.005562	1.000000	0.667666	-0.021946	0.070377	0.051658	-0.069408	-0.050656
total sulfur dioxide	-0.113181	0.076470	0.035533	0.203028	0.047400	0.667666	1.000000	0.071269	-0.066495	0.042947	-0.205654	-0.185100
density	0.668047	0.022026	0.364947	0.355283	0.200632	-0.021946	0.071269	1.000000	-0.341699	0.148506	-0.496180	-0.174919
pH	-0.682978	0.234937	-0.541904	-0.085652	-0.265026	0.070377	-0.066495	-0.341699	1.000000	-0.196648	0.205633	-0.057731
sulphates	0.183006	-0.260987	0.312770	0.005527	0.371260	0.051658	0.042947	0.148506	-0.196648	1.000000	0.093595	0.251397
alcohol	-0.061668	-0.202288	0.109903	0.042075	-0.221141	-0.069408	-0.205654	-0.496180	0.205633	0.093595	1.000000	0.476166
quality	0.124052	-0.390558	0.226373	0.013732	-0.128907	-0.050656	-0.185100	-0.174919	-0.057731	0.251397	0.476166	1.000000



Lessons Learned

- Chemical-wise, different types of wine generally have very similar levels of “Density” and “pH”.
- Compared with low quality wine (quality <=4), high quality wine (quality >=6) generally:
  - Higher in “Fixed Acidity,” “Citric Acid,” “Free Sulfur Dioxide”
  - Lower in “Volatile Acidity” and “Chlorides”
- Measurable physicochemical properties of wines can help people differentiate good and bad wines.

Recommendations

- Stories should be completed in the sprint:
  - Epic 2 (Modeling and Model Selection)
    - Attempting more models (XGBoost and Neural Net)
    - Conducting final model selection based on performance
  - Epic 3
    - Designing Web app UI
    - Deploying web app (Flask) on AWS
    - Testing RDS instance