DATA 622 Assignment 4

Wine Quality

Susanna Wong
Professor Joseph Sabeljia
March 2, 2025

Introduction

The wine quality dataset was obtained from Kaggle and it contains 1,599 observations of various red Portuguese wine and its chemical compositions. It contains 11 numerical features that includes `fixed acidity`, `volatile acidity`,` citric acidity`, `residual sugar`, `chlorides`, `free sulfur dioxide`,` total sulfur dioxide`, `density`,` pH`, `sulphates`, and `alcohol`. The target variable represents the quality of the wine based on a scale from 0 to 10, indicating poor quality to excellent quality. For this assignment, we build a binary classification model that predicts the wine quality to be 'good' or 'bad' based on its chemical compositions. Wine with score greater than 5 were classified as "good", while scores less than or equal to 5 were classified as "bad".

Exploratory Analysis

Distribution

It is important to assess skewness of the distributions of the variables because some predictive algorithms are sensitive to non-normal models. If there are skewness in the data, transformation or robust algorithm can be consider to improve accuracy of the models. The following variables show right-skewed distributions: `chlorides`, `free sulfur dioxide`, `residual sugar`, `sulphate`,` total sulfur dioxide`. This finding suggests Box Cox transformation or log transformation could benefit models that are sensitive to scaling. In contrast, the following variables show approximately normal distribution: `density` and `pH`.

Missing Data and Outliers

There are no missing data. However, there are outliers present in all variables.

Correlation between Features and Quality

Among the feature, `alcohol` has the strongest (positive) correlation with the wine quality, meaning higher alcohol wines tend to receive higher quality ratings. In contrast, `residual sugar` and `free_sulfur_dioxide` show weak correlation with

wine quality. `Sulphates` and `citric acid` each also has a positive correlation with quality but weaker than alcohol. `Volatile Acidity` and `Quality` has a strong negative correlation, meaning higher acetic acid (causes vinegary taste) tend to receive lower quality ratings. `Density` and `Chlorides` each shows moderate negative correlation with quality.

These insights highlight the potential key predictors (features showing strong or moderate correlation with wine quality) are `alcohol`, `sulphates`, `total_sulfur_dioxide`, `violatile acidity`, `density`, `citric acid` and `chlorides`.

Correlation between Features

There were some features that were highly correlated with each other, indicating potential multicollinearity. This provides an opportunity to perform feature engineering by creating new features or remove some features to improve model performance.

For example, Free Sulfur Dioxide & Total Sulfur Dioxide were highly correlated. Since `Total Sulfer Dioxide` is the sum of `Free Sulfur Dioxide` and combined sulfur dioxide, we can create a new feature `combined sulfur dioxide` = `Total Sulfer Dioxide` - `Free Sulfur Dioxide`. We can also try removing `Free Sulfur Dioxide` if `Total Sulfer Dioxide` has higher importance since `Total Sulfer Dioxide` includes `Free Sulfur Dioxide` to simplify the model.

`Density` is highly positively correlated with `fixed acidity`, `citric acid`, and `residual sugar`. `Fixed Acidity` and `citric acid` have strong positive correlations. The following pairs have strong negative correlations: `fixed acidity` and `PH`, `Density` and `Alcohol`, and `Citric Acid` and `Volatile Acidity`.

Algorithm Selection and Experimentation

Two machine learning algorithms were used for the classification taskL Random Forest, and Neural Network.

Random Forest Models

Random Forest is robust to feature scaling and outliers, which both are present in the dataset. Below are the variation of the random forest model and its result:

1. Default Random Forest: This model is the baseline that contains all features from the original dataset. The accuracy is about 0.79 and an AUC of 0.876. Feature importance consistently identify alcohol as the most significant predictor, followed by `sulphates`, `total_sulfur_dioxide`, `violatile acidity`, `density`, `citric acid` and `chlorides`. This aligns with the findings from the correlation analysis earlier.
2. Random Forest with Scaling and Box Cox Transformation: Scaling is not usually necessary for Random Forest as it's generally not sensitive to scaling as they are tree based models that split on feature values and not on magnitudes. The accuracy is identical as the default model but has a slightly higher AUC (0.8775 v. 0.8761). This suggests that a scaling and Box Cox Transformation without tuning `mtry` might offer a marginal benefit in distinguishing good and bad wine.
3. Random Tree (Tuned and No Scaling): Hyperparameter tuning for `mtry` was performed. The model has similar metrics as the default. Tuning `mtry` only did not significantly improve performance compared to default model.
4. Random Tree (Tuned + Scaled): Combining hyperpatemter tuning with scaling resulted a slight drop in performance.
5. Random Tree (Feature Removal): `residual_sugar` and `free_sulfur_dioxide`were removed as they had weak correlations with wine quality and were at the bottom of feature importance. The model is trained with the same `tune_grid`, `ntree`, and `trControl`. The accuracy, f1-score, and precision increased slightly compared to the last four variations, while maintain a comparable AOC.

6. Random Tree (Feature Engineering + Removal): Since `Total Sulfer Dioxide` is the sum of `Free Sulfur Dioxide` and combined sulfur dioxide, we create a new feature `combined sulfur dioxide` = `Total Sulfer Dioxide` - `Free Sulfur Dioxide`. Both `residual_sugar` and `free_sulfur_dioxide` were still removedThis model performed the best across the board. It has the highest accuracy (0.7974948) and AUC (0.8787048). This indicates feature engineering was highly effective in capturing important information for predicting the wine quality.

Neural Network Models

Neural Network are highly sensitive to feature scaling. All neural network models are trained with Box Cox transformation and scaling. Below are the variation of the neural network model and its result:

1. Neural Network (default): With an accuracy of 0.691, this model under performs compared to all previous models.

2. Neural Network (tuned): This tuned model is evaluated with the number of neurons in the hidden layers size as 5, 10, and 15 (least to most complex) and decay parameters of 0.001, 0.01, and 0.1. The tuned Neural Network shows a clear improvement in accuracy from 0.691 to 0.7056 from the default Neural Network. This shows hyperparamenter tuning was effective in improving the model performance. However, the tuned Neural Network still under performs all Random Forest models including the default Random Forest model.

	Accuracy	Precision	F1_Score	AOC	Comment on Performance
Random Forest (default)	0.7912317	0.7860465	0.7716895	0.8761474	Baseline Performance
Random Forest (Scaling & Box Cox)	0.7912317	0.7782805	0.7747748	0.8775662	Slight improvement in Fq and AUC and scaling didn't hurt performance

	Accuracy	Precision	F1_Score	AOC	Comment on Performance
Random Forest (Tuned & No Scaling)	0.7912317	0.7860465	0.7716895	0.8770933	same as default
Random Forest (Tuned & Scaling)	0.7787056	0.7671233	0.7601810	0.8771108	Slight drop in accuracy due to simplified model
Random Forest (Feature Removal)	0.7954071	0.7906977	0.7762557	0.8756481	Precision and F1 improve slightly
Random Forest (Feature Engineering & Removal)	0.7974948	0.7863636	0.7863636	0.8787048	Best Perfomance Overall
Neural Network (default)	0.6910230	0.6543210	0.6543210	0.7755570	Baseline Performance but weaker than Random Forest
Neural Network (tuned)	0.7056367	0.6830357	0.6830357	0.7746637	Overall better performance than default neural network. However, random forest outperforms it

Conclusion

This project aimed to develop a predictive model for red wine quality, classifying the wines into good and bad based on their scores (0-10).

Through performing Exploratory Analysis, and experimenting on various models (Random Forest and Neural Network), and feature engineering features that influence the wine quality were identified to build a robust model.

Summary of Model Performance

Several Random Forest and Neural Network models were evaluated using the following metrics: F1-score, Accuracy, Precision, and AUC. All Random Forest models outperformed the neural network. Although tuning the neural network

model did improve performance of the default neural network model, the tuned neural network model still under performs the default random forest model. The Random Tree Model with feature engineering and removal performed the best with an accuracy of 0.7975 and AUC of 0.8787. In this model, `residual sugar` and `free_sulfur_dioxide` were removed and `combined sulfur dioxide` was created.

Business Conclusion

The Random Tree Model with feature engineering and removal can be used to help identify wines that likely to be bad based on the features, allowing business to maintain quality control. Ultimately, it will allow them to reduce waste and protect brand reputation. It can help business identify good quality wines for premium pricing.

Limitation and Further Improvements

In this project, we transformed the wine quality (0-10) to binary categories of `good` and `bad`, which simplify the prediction model.

For further enhancement, we can perform multi-class classification or regression to predict the wine quality score.

Another enhancement is that we can explore other feature engineering based on features that are highly correlated to each other (multi-collinearity).

For example, `Density` is highly positively correlated with `fixed acidity`, `citric acid`, and `residual sugar`.

Source

H, M. Y. (2022, January 15). *Wine quality dataset*. Kaggle. https://www.kaggle.com/datasets/yasserh/wine-quality-dataset

White, N. A. (2019, March 18). *Fixed acidity*. Waterhouse Lab. https://waterhouse.ucdavis.edu/whats-in-wine/fixed-acidity