Red Wine Quality

Author: Hang Yu, Ruixin Chen

1. **Who is your stakeholder**
   The stakeholders for this project are winemakers, quality control teams, and potential investors in the wine industry. These stakeholders are interested in understanding the factors affecting wine quality and want a reliable model to predict wine quality to optimize production processes.

2. **What is the problem they are trying to solve**
   The goal of this project is to predict the quality of red wine based on its physicochemical attributes, such as acidity, alcohol content, and residual sugar. This can help winemakers understand how different factors contribute to quality and ensure consistency in wine production.

3. **Where your dataset is from (link to it or include it in your submission)**
   Dataset link:
   https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009

4. **What models did you try, why did you choose those models?**
   We used two different types of models: Logistic Regression and Random Forest Regressor. The Logistic Regression model was chosen for its simplicity and interpretability, making it easy to understand the relationships between features and target variables. Random Forest Regressor was chosen for its robustness and ability to handle complex feature interactions, which is important for predicting wine quality with non-linear relationships. To further improve performance, both models underwent extensive hyperparameter tuning using grid search. Logistic Regression was tuned for its regularization strength and penalty type, while Random Forest was tuned for the number of estimators, maximum tree depth, and minimum samples for splitting nodes. We chose these two models to balance interpretability with model complexity and predictive power.

5. **What features did you select/engineer? How did you choose those?**
   We used all 11 physicochemical attributes in the dataset as features, including acidity, residual sugar, alcohol, and chlorides. These features were chosen based on their potential influence on the quality of wine as indicated in previous studies on wine production.

6. **How did you evaluate the model? What evaluation metrics did you use? Why?**
   The Logistic Regression model was evaluated using accuracy, precision, recall, and a confusion matrix. Accuracy is a basic metric to understand the percentage of correctly classified instances, while precision and recall

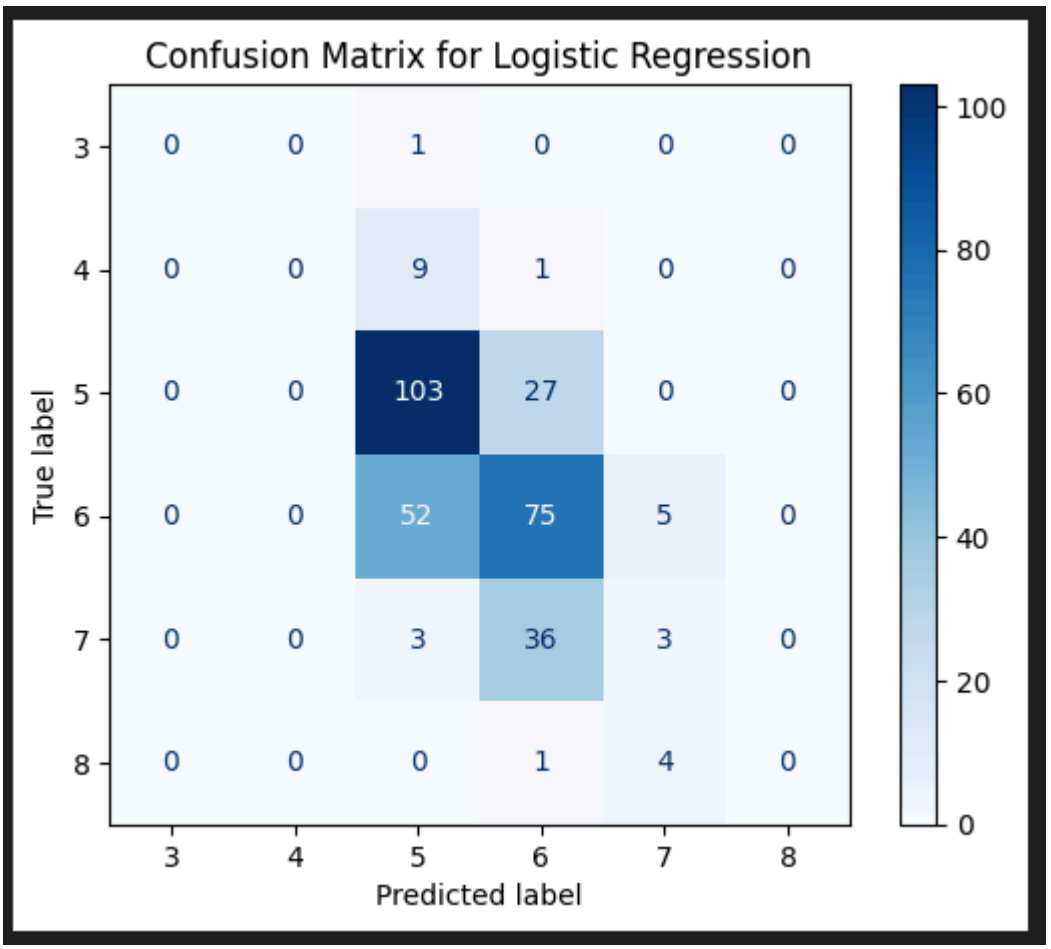
help to understand the reliability of positive predictions and the sensitivity of the model. The confusion matrix further provided detailed insights into false positives and false negatives. The Random Forest Regressor was evaluated using Mean Squared Error (MSE) and R^2 score. MSE helps understand the average squared difference between predicted and actual values, indicating the magnitude of prediction errors. R^2 was used to indicate how well the model explains the variance in the target variable, providing a comprehensive evaluation of model fit.

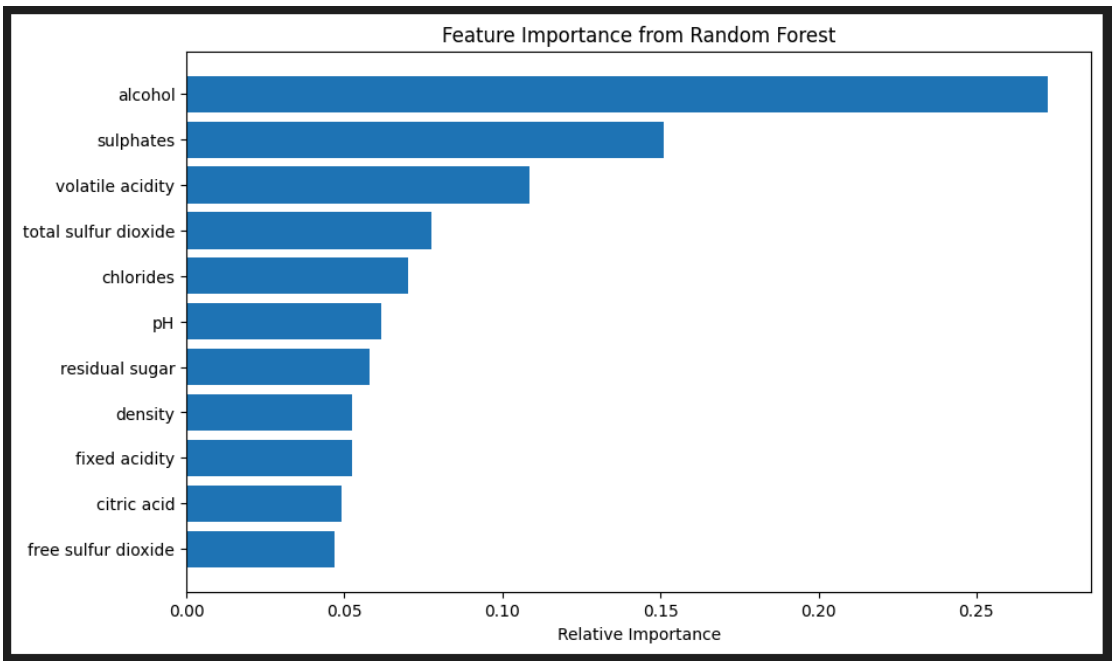
7. **What would you do different next time or given more time what would your future work be?**
   If given more time, We would experiment with more sophisticated non-linear models like Gradient Boosting Machines or XGBoost to improve prediction accuracy. Additionally, We would consider using feature selection techniques like Recursive Feature Elimination (RFE) to potentially reduce feature dimensionality.

8. **Do you recommend your client use this model? Is the precision/recall good enough for the intended use case?**

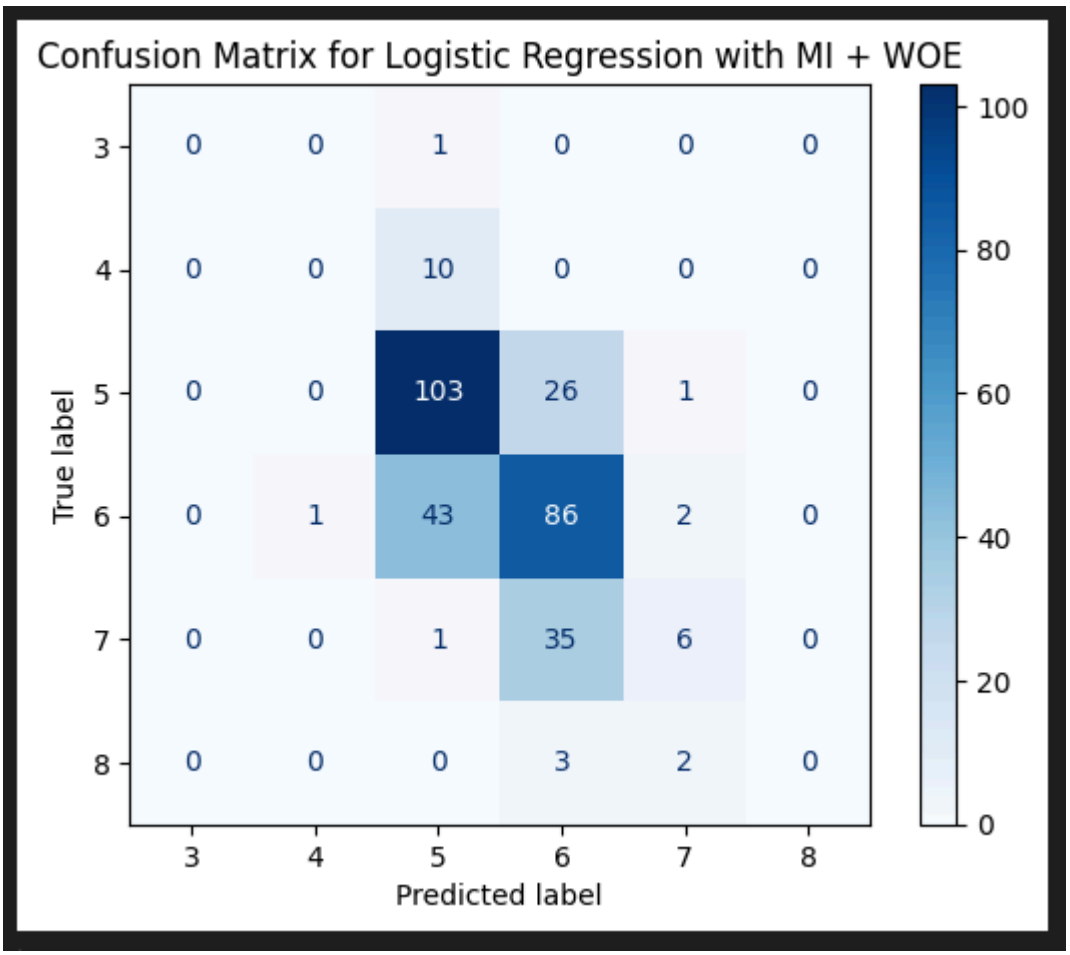
   Based on model performance, we would recommend different models depending on the feature engineering improvements. Initially, the Random Forest Regressor demonstrated better robustness and ability to capture complex interactions. However, after introducing new engineered features, the Logistic Regression model showed significant improvements in generalizability and precision/recall scores, outperforming Random Forest. This suggests that with these new features, Logistic Regression not only maintains interpretability but also offers consistent and reliable performance. Therefore, with the improved features, Logistic Regression is the preferred model for optimizing production processes.

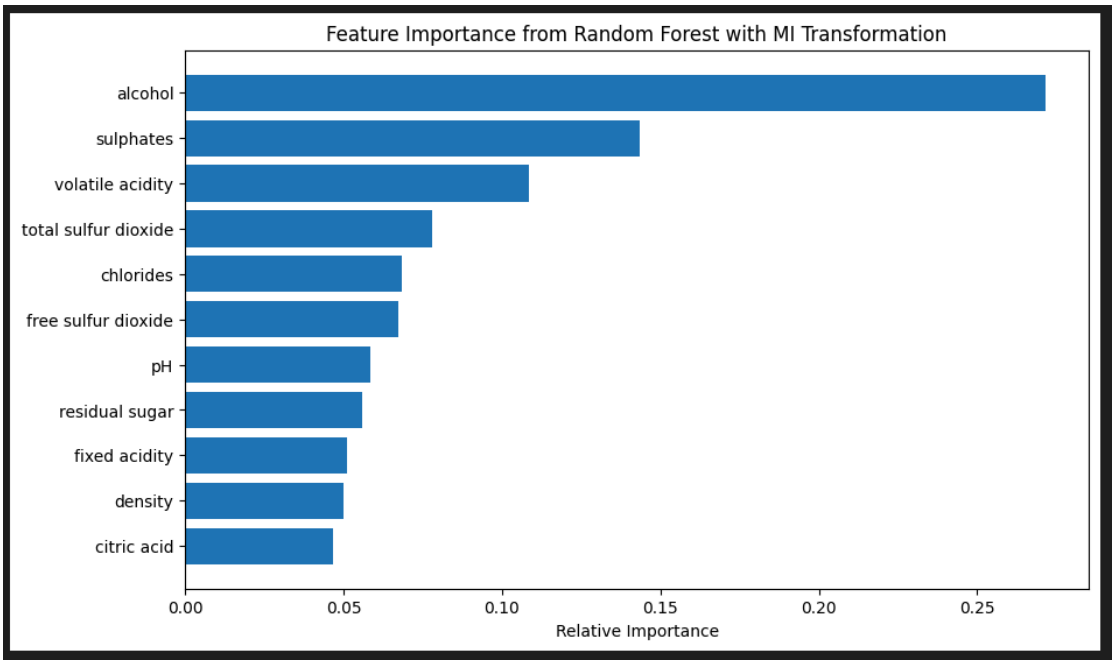
Before modification:





After modification:





9. ENGINEER AT LEAST TWO "NEW" FEATURES

Actually, we engineer at least two types of new features derived from Mutual Information (MI) and Binning + Weight of Evidence (WOE) transformations. These new features aim to capture complex relationships in the dataset and enhance model performance.

- **1. New Features from Mutual Information (MI)**

  By identifying pairs of features with high mutual information, we create new features based on their linear combinations. The formula for generating a new feature is:

$$X' = X - \alpha Y$$

  Where:

  a is the optimal coefficient calculated as:

$$\alpha^* = \frac{\text{Cov}(X, Y)}{\sigma_Y^2}$$

  It finally creates composite features based on correlations between pairs of features.

- **2. New Features from Binning and Weight of Evidence (WOE)**

  Continuous variables are binned into discrete intervals, and WOE is calculated for each bin. WOE quantifies the relationship between each bin and the target variable using the formula:

$$\text{WOE} = \ln\left(\frac{\text{Event Rate}(P(1|\text{Bin}))}{\text{Non-Event Rate}(P(0|\text{Bin}))}\right)$$

  It finally generates monotonic transformations of features to reflect their predictive power concerning the target variable.

Github link:
https://github.com/763730440/Wine-Quality.git

Slide link:
https://docs.google.com/presentation/d/1Lfi9aMHU-OowhAIYZlEreUCD_sMoNYBG/edit?usp=sharing&ouid=109997971140104995834&rtpof=true&sd=true

YouTube link:

References

1. https://blog.csdn.net/weixin_43434202/article/details/102922379
2. https://blog.csdn.net/weixin_63001635/article/details/138728867
3. https://blog.51cto.com/u_16213684/11404283
4. https://zhuanlan.zhihu.com/p/659520181
5. https://computing.hit.edu.cn/_upload/article/files/a5/9a/8f4c9c294c3bb2235626e10bf0fc/02f32893-9ebc-430e-b6eb-be86a80c5814.pdf