Abstract

In this project, we deployed seven different supervised machine learning methods to perform the milk grade classification task. Before fitting the models, we conducted exploratory data analysis and data preprocessing for preparation. After hyperparameter tunning with cross validation, we find the models with best performance on the test data set and evaluated the results thoroughly. In this project, we find that the model with the best performance and most suitable for this classification task is XGBoost and its accuracy score is around 0.995.

Introduction

Supervised learning methods are widely used in classification tasks across various field. These methods leverage labelled data to train predictive models that can categorize unseen data into predefined classes or categories. Supervised learning applications include spam detection, medical diagnosis, image recognition and quality control. In this project, we are tasked to classify correctly the grade of milk quality with its seven features, four of which is qualitative. Supervised machine learning methods are especially indispensable for tackling classification tasks for its providing automated solutions that enhance decision-making. If we are able to distinguish milk quality grade with the labelled data and pre-trained models, milk quality control will be made much easier. In the dairy industry, the quality and safety of dairy products are crucial for consumers' health. For businesses, these models assist in making informed decisions regarding milk processing, distribution, pricing, regulatory compliance and so on.

Some of the supervised learning methods helps identify feature importance, with which production processes for achieving desired milk grades consistently can be optimized.

Along with supervised learning, techniques like normalization, resampling, cross validation and hyperparameter tuning can be applied. By integrating these techniques into the model development process, dairy producers and processors can build more accurate, reliable, and robust supervised learning models for milk grade classification and the interpretability of models can be improved as well.

Methodology

In this project, we utilize data visualization and wrangling skills to explore the dataset and construct several supervised learning models. Subsequently, we apply cross-validation to all models to attain long-term accuracy and determine the optimal parameters. Finally, we export the results of the models for further interpretation.

Experimental Procedure:

- Data Visualization: Employing box plots and bar charts to comprehend the relationship between the response and predictors. Additionally, applying density plots to visualize the distribution of different categories of milk quality.
- Data Wrangling: Checking for NA records and utilizing the Min-Max standardization method to scale the numerical variables into similar intervals.
- Data Splitting: Dividing the dataset into training and testing sets (80% training, 20% testing).
- Selection of Machine Learning Algorithms: Including Logistic Regression, Linear Discriminant Analysis, Quadratic Discriminant Analysis, K-Nearest Neighbors (KNN), Decision Trees, Random Forest, and XGBoost.
- Model Training: Fitting the chosen algorithm to the training data.
- Model Evaluation: Assessing the performance of the trained model using accuracy and confusion matrix.
- Cross-Validation: Employing 10-fold cross-validation techniques to validate the model's generalization performance.

- Hyperparameter Tuning: Optimizing the model's hyperparameters (if applicable) using techniques like grid search and plotting the relationship between Mean Accuracy and variable values of the parameter.
- Final Model Selection: Selecting the best-performing model based on tuning data and evaluation metrics.
- Model Interpretation: Interpreting the results based on output plots such as decision trees and the importance of random forest.
- Model Comparison: Explaining the reasons why some models have worse results than others.

Experiment

Boosting algorithms outperform other machine learning models. XGBoost demonstrates the best long-term performance among all models based on our 10-fold cross-validation accuracy, followed by Random Forest. Conversely, linear models perform worse than other models, with LDA performing the worst.

Experimental Details:

- Data Wrangling:
  - We selected Taste, Odor, Fat, and Turbidity as categorical variables, excluding color. Although including color as a category and then using One-Hot encoding could yield better results, it may introduce issues when new observations with colors not included in the initial categories are added. Additionally, adding 7 dummy variables would expand the feature space, potentially destabilizing the model with only 1059 observations.
  - We utilized Min-Max standardization to scale the predictors into a similar interval. We chose Min-Max instead of normalization because it maintained excellent performance in our best boosting model while reducing accuracy and interpretability in linear models.
- Data Visualization:
  - Box plots and bar charts reveal significant differences among three types of milk quality, indicating distinct and homogenous observation clusters, which explains the strong performance of our models.
  - Density plots indicate that the categories do not adhere to a normal distribution.
- Model Selection:
  - Linear models exhibit the poorest performance, suggesting no linear relationship between the response and predictors. The logistic model violates assumptions of multicollinearity and variable independence. Additionally, LDA performs poorly due to its violation of the normal distribution assumption. Despite relaxing the assumption of homogenous variance in QDA, it still performs inadequately.
  - In the KNN model, setting the radius to 1 yields the best result, ranking it as the third-best model overall. Decision trees perform best without pruning.
  - Boosting models achieve nearly 100% accuracy with at most 2 misclassifications, benefiting from bias reduction, variance reduction, and capturing different patterns from the dataset.

Conclusion

For logistic regression, LDA and QDA, we fit models and evaluate their performances with cross validation on training data set and test data set. It is safe to say that compared with other models, these models have a relatively bad performance. The reason here is because the training data set violates the basic assumptions of these models including no multicollinearity, Gaussian distribution and observation independence. With cross validation, grid search algorithm and certain visualizations, we accomplished hyperparameter tuning. For

KNN method, the best parameter K selected is 1. For decision tree, tuned parameter ccp_alpha is 0. The max depth and number of estimators of random forest are 7 and 25 respectively. The most two importance features among all seven features are pH and temperature. We tuned five parameters of XGBoost as well and it turns out to be the most powerful and accurate model we get.

References

1. 3.2. Tuning the hyper-parameters of an estimator — scikit-learn 1.4.0 documentation
2. sklearn.ensemble.GradientBoostingClassifier — scikit-learn 1.4.0 documentation
3. sklearn.ensemble.RandomForestClassifier — scikit-learn 1.4.0 documentation