Developing and Enhancing Predictive Models for Wine Quality and Alcohol Content

Key Steps and Performance Metrics

Module 2: Data Analysis

Sprint 3: Statistical Modeling

Part 5: Analysis of the Red Wine Quality dataset

Author : Giedrius Kaminskas

2024, Kaunas

The key stages of Model development includes:

1. Data Preprocessing: Implementation of techniques to refine dataset and ensure optimal input quality for model training.

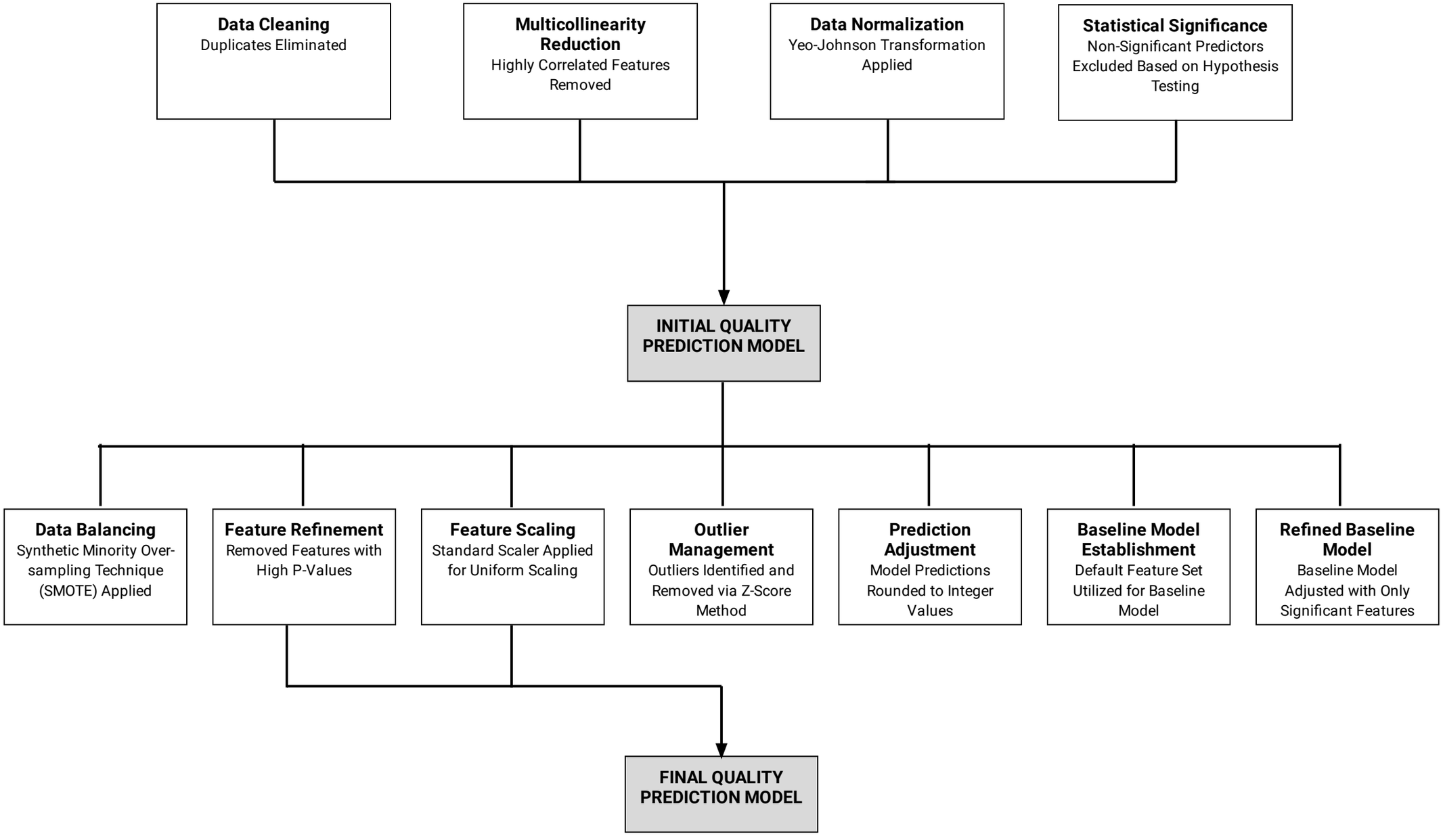
2. Model Development: Creation of initial models using the processed data to predict wine quality and alcohol content based on chemical and physical attributes.

3. Model Refinement: Evaluation and enhancement of the initial models through statistical methods and feature optimization.

4. Performance Assessment: Analysis of model metrics to identify best practices and configurations that led to the final improved models.

5. Final Models: Presentation of the key metrics of the final predictive models, underscoring the successful application of data preprocessing and model refinement techniques.

Predictive Model Development Workflow for Quality Prediction



Scheme No.1 Model development workflow.

Model Iterations and Their Impact on Quality Predictive Accuracy

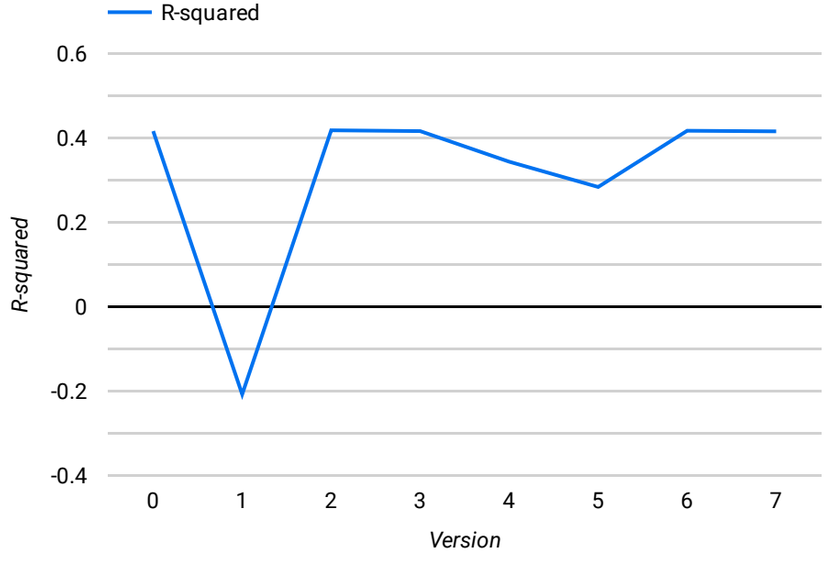
Key Insights:

1. **Resampling and Outlier Management**: Decreased the explanatory power of the model rather than enhancing it.
2. **Default Feature Baseline Model**: Showed improved results compared to methods guided solely by OLS statistical summaries.
3. **Feature Selection**: Eliminating non-significant features resulted in a slight improvement in the R-squared value, signifying a more precise model.
4. **Prediction Rounding**: Negatively affected model's predictive accuracy when rounding predictions to the nearest integer.
5. **StandardScaler Application**: Slightly reduced R-squared value, but may contribute to model stability and reduced sensitivity to noise.
6. **Combination for Final Model**: The final model integrates elements from both v2 and v3, leveraging their respective strengths.

Table No.1 Comparison of Model Improvement Methods and Metrics

	Version ▲	Description	R-squared	Difference
1.	0	Initial model	0.4161	0.0000
2.	1	Model with Resampled dataset	-0.2074	-0.6235
3.	2	Model without Non-Significant Features	0.4180	0.0019
4.	3	Model with Standard Scaling	0.4161	-0.0000
5.	4	Model with Potential Outliers Removed	0.3436	-0.0724
6.	5	Model with Rounded Prediction	0.2838	-0.1323
7.	6	Default Feature Baseline Model	0.4170	0.0009
8.	7	Baseline Model (v6) with Significant Features	0.4156	-0.0005

1 - 8 / 8



Graph No.1 R-Squared Values Across Different Model Version

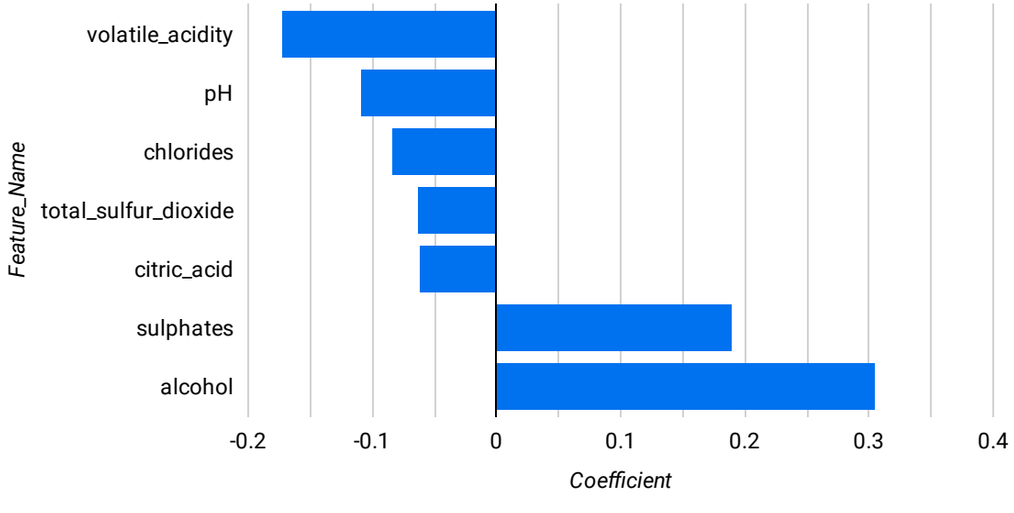
Final Quality Prediction Model Features and Key Performance Metrics

The final predictive model includes a set of key features, each with an associated coefficient that signifies the strength and direction of the relationship between the feature and the target variable.

Table No.2 Feature Importance Summary

	Feature_Name	Abs_Coefficient	Coefficient
1.	volatile_acidity	0.1731	-0.1731
2.	total_sulfur_dioxide	0.0644	-0.0644
3.	sulphates	0.1894	0.1894
4.	pH	0.1100	-0.1100
5.	citric_acid	0.0628	-0.0628
6.	chlorides	0.0852	-0.0852
7.	alcohol	0.3053	0.3053

1 - 7 / 7



Graph No.2 Feature Importance Coefficient Visualization

Final Quality Prediction Model Features and Key Performance Metrics

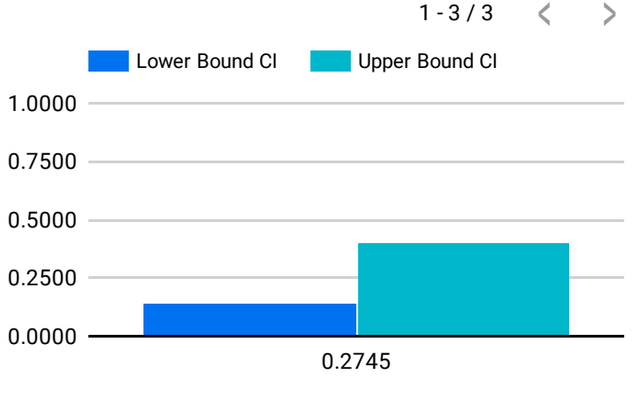
Key Points for Final Quality Prediction Model Summary:

1. R-squared Improvement: The updated model exhibits a 0.0019 increase in R-squared value, indicating a slight enhancement in explaining the target variable's variability by 0.19%.
2. MSE Reduction: There is a marginal decrease in the mean squared error, implying an improvement in prediction accuracy.
3. Consistency Across Folds: The average R-squared across 10 folds has risen by approximately 0.0029, suggesting improved model consistency across varied data subsets by 0.29%.
4. **Consideration for Model Type**: The persistently low R-squared value and prediction accuracy hint that linear regression may not be the optimal model type for this dataset, warranting the exploration and testing of alternative modeling approaches.

Table No.3 Model Performance Metrics Comparison

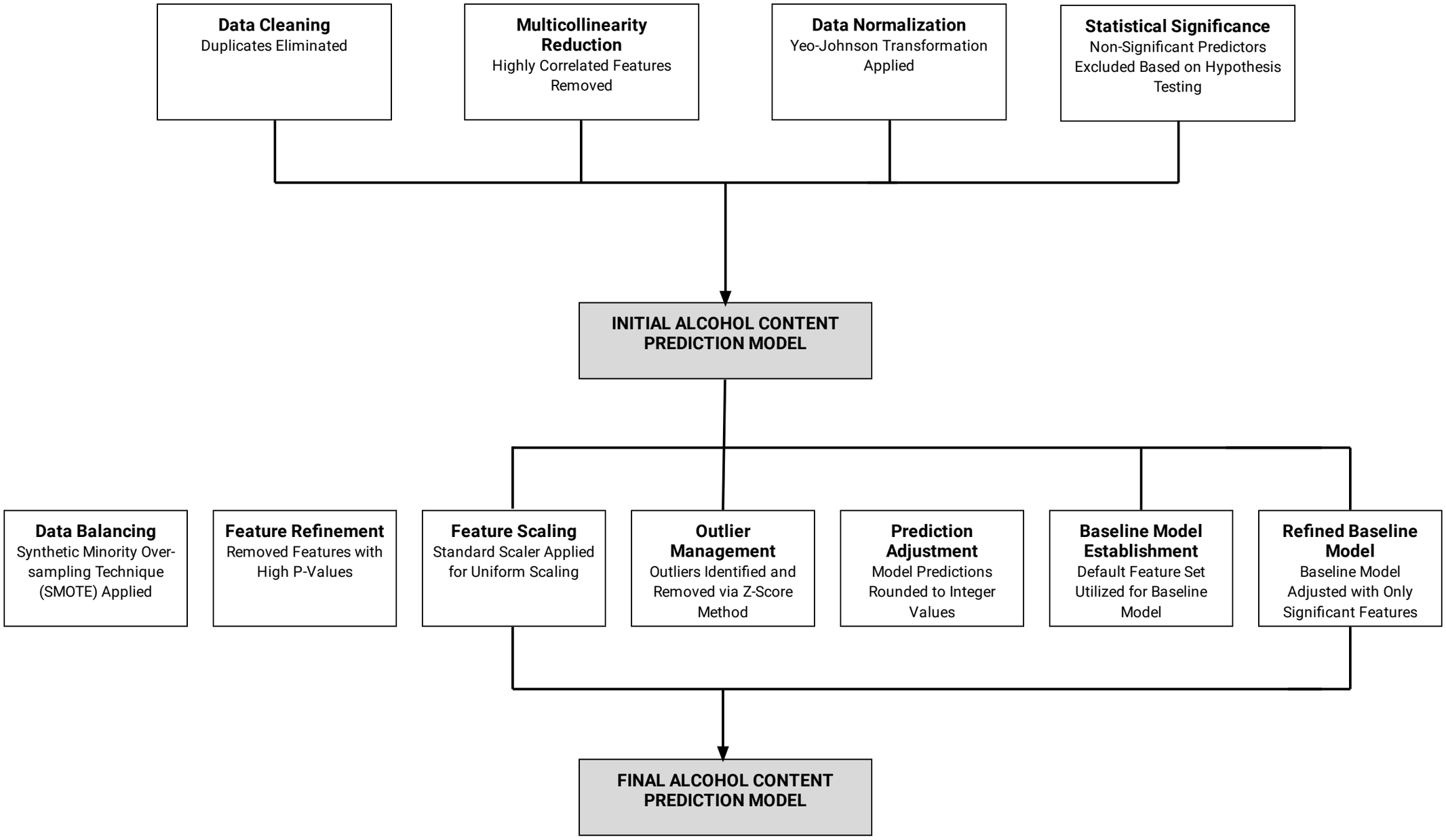
	Model ▲	R-squared	MSE	RMSE	Avarage R-squared across 10 folds
1.	Final	0.4180	0.4122	0.6421	0.2745
2.	Initial	0.4161	0.4136	0.6431	0.2716
3.	Model Differences	0.0019	-0.0014	-0.0011	0.0029

With 95% confidence, the true prediction accuracy of the linear regression model for predicting wine Quality lies between **14.53%** and **40.37%**, based on current red wine dataset.



Graph No.3 Confidence Interval for Model Predictions

Predictive Model Development Workflow for Alcohol Content Prediction



Scheme No.2 Model development workflow.

Model Iterations and Their Impact on Alcohol c Content Predictive Accuracy

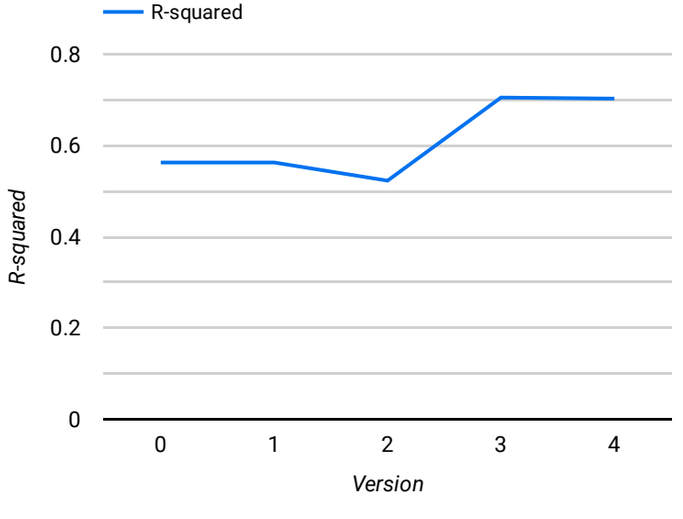
Key Insights:

1. **Model Version v3 Performance**: Showed the most notable improvement in R-squared value among the iterations.
2. **Final Model Selection**: Version v4 was chosen for its simplified complexity through the exclusion of non-significant features.
3. **Condition Number Concern**: v4 alone exhibits a high condition number, indicating potential multicollinearity or numerical issues.
4. **StandardScaler Integration**: The application of StandardScaler aims to mitigate multicollinearity and enhance the robustness of the final model.
5. **Combination for Final Model**: The final model integrates elements from both v1 and v4, leveraging their respective strengths.

Table No.4 Comparison of Model Improvement Methods and Metrics

	Version...	Description	R-squared	Difference
1.	0	Initial model	0.5626	0.0000
2.	1	Model with Standard Scaling	0.5626	0.0000
3.	2	Model with Potential Outliers Removed	0.5229	-0.0397
4.	3	Default Feature Baseline Model	0.7049	0.1423
5.	4	Baseline Model (v4) with Significant Features	0.7026	0.1400

1 - 5 / 5



Graph No.4 R-Squared Values Across Different Model Version

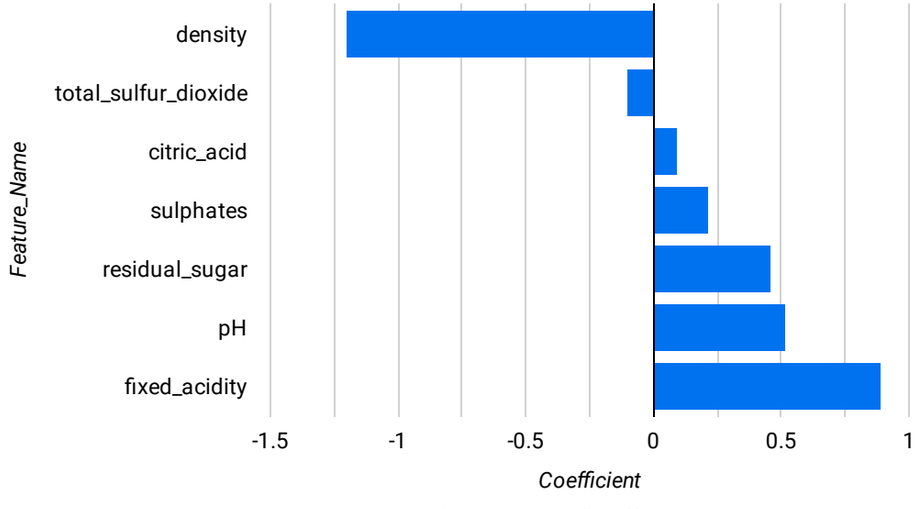
Final Alcohol Content Prediction Model Features and Key Performance Metrics

The final predictive model includes a set of key features, each with an associated coefficient that signifies the strength and direction of the relationship between the feature and the target variable.

Table No.5 Feature Importance Summary

	Feature_Name	Abs_Coefficient	Coefficient
1.	total_sulfur_dioxide	0.1063	-0.1063
2.	sulphates	0.2176	0.2176
3.	residual_sugar	0.4592	0.4592
4.	pH	0.5176	0.5176
5.	fixed_acidity	0.8944	0.8944
6.	density	1.2088	-1.2088
7.	citric_acid	0.0958	0.0958

1 - 7 / 7



Graph No.5 Feature Importance Coefficient Visualization

Final Quality Prediction Model Features and Key Performance Metrics

Key Points for Final Quality Prediction Model Summary:

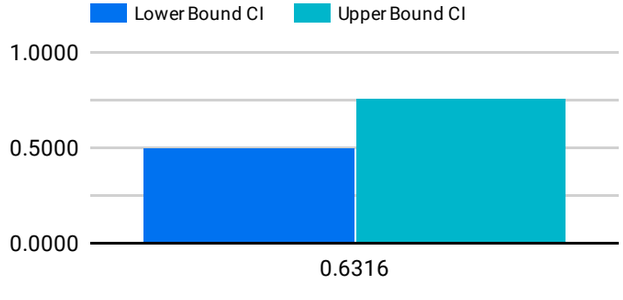
1. **Explanatory Power**: The final model shows a significant R-squared increase of 0.14, marking a substantial gain in its ability to explain the variability of the target variable.
2. **Prediction Accuracy**: A reduction of 0.15 in the Mean Squared Error (MSE) reflects a noteworthy improvement in the accuracy of the model's predictions.
3. **Consistency and Reliability**: The average R-squared value's increase from 0.44 to 0.63 across 10-fold cross-validation underscores enhanced consistency and hints at the model's increased reliability by 19%.

Table No.6 Model Performance Metrics Comparison

	Model	R-squared	MSE	RMSE	Avarage R-squared across 10 folds
1.	Final	0.7026	0.3242	0.5694	0.6316
2.	Initial	0.5626	0.4768	0.6905	0.4411
3.	Model Differences	0.1400	-0.1526	-0.1211	0.1904

1 - 3 / 3

With 95% confidence, the true prediction accuracy of the linear regression model for predicting Alcohol Content in wine lies between **50.10%** and **76.21%**, based on current red wine dataset.



Graph No.6 Confidence Interval for Model Predictions