Improving Wine Quality Predictions with Data Analysis

# Problem Statement

The wine industry faces the challenge of balancing quality and production costs. Predicting wine quality based on chemical and sensory data is essential for producers to meet consumer expectations. This project focuses on accurately predicting wine quality using available data, supporting better decision-making throughout production and marketing. There are two distinct ways how this project can be utilized. It can be used to see which attributes in wine impact its sales, and it ca also be used to determine overall quality of a wine, for example to determine pricing. These two goals are very related but can lead to different decisions in building and selecting a model as will be illustrated below.

# Dataset and Preparation

The dataset is from a project I had to do for a different class. I thought it would be a useful way to illustrate the above dichotomy. It contains chemical and sensory variables measured from various wine samples, with a target variable for quality ratings. Data preprocessing included removing redundant columns and calculating basic statistics to understand the data. Key variables included volatile acidity, residual sugar content, alcohol percentage, and other chemical markers.

To handle missing data, visual tools like missing value matrices and heatmaps were used to assess patterns. Imputation methods, such as SimpleImputer and IterativeImputer, were applied to estimate missing values. Skewed variables were identified and transformed using logarithmic or Box-Cox transformations. All features were standardized or normalized using StandardScaler and MinMaxScaler, ensuring equal contribution from all variables.

The two categorical variables in the dataset were:

* **STARS:** A measure of how highly the wine is rated.
* **LabelAppeal:** A binary variable indicating whether the label of the wine appeals to consumers.

These categorical features were one-hot encoded to integrate them into the models appropriately.

A thorough Exploratory Data Analysis (EDA) is available in the attached .pynb file, where you can find detailed visualizations, including histograms, box plots, and correlation matrices, to help understand the data’s underlying patterns. Extensive data preparation and feature engineering code is also included in this notebook.

# Methodologies

Two approaches were used for predictive analysis:

* **Regression Modeling:** Various regression models, including ordinary least squares (OLS) and robust regression, were explored. Due to the ordinal nature of quality ratings, Poisson regression was chosen as the most appropriate model. Interaction terms, such as those between alcohol content and label appeal, were added to improve model performance. The final Poisson regression model had a Log-Likelihood of -11949, a Deviance of 4225.9, and a Pseudo R-squared (CS) of 0.2530. The model was fitted with 4 iterations using IRLS. The coefficients of the Poisson regression model are shown below:

| **Variable** | **Coefficient** |
| --- | --- |
| Intercept | 1.5527 |
| Alcohol | 0.0001 |
| Alcohol\_squared | 0.0002 |
| LabelAppeal | 0.2105 |
| STARS | 0.1778 |
| AcidIndex | -0.0484 |
| Chlorides | -0.0240 |
| CitricAcid | 0.0008 |
| Density | -0.3216 |
| FixedAcidity | 0.0003 |
| FreeSulfurDioxide | 6.902e-05 |
| ResidualSugar | -9.98e-06 |
| Sulphates | -0.0071 |
| TotalSulfurDioxide | 2.11e-05 |
| VolatileAcidity | -0.0217 |
| pH | -0.0014 |
| Alcohol\_LabelAppeal\_interaction | -0.0027 |
| STARS\_Alcohol\_interaction | 0.0005 |

* **Neural Network Modeling:** A neural network model, implemented with TensorFlow, was used to capture non-linear relationships and potentially improve predictive accuracy. The network consisted of four dense layers with decreasing neurons (256, 128, 64, 1) and LeakyReLU activation functions. Dropout and batch normalization layers were applied for regularization and training stability. Despite its strong predictive power, the model faced challenges with overfitting due to the small dataset size, reducing its interpretability. This case illustrates the trade-off between accuracy and interpretability—while the neural network may offer superior performance, the Poisson regression model provides more transparency into the relationships between variables.

**Evaluation Metrics:**

Model performance was evaluated using the following metrics:

| **Model** | **MSE** | **MAE** | **R²** | **RMSE** |
| --- | --- | --- | --- | --- |
| Poisson Regression | 0.112 | 0.267 | 0.253 | 0.335 |
| Neural Network | 0.089 | 0.232 | 0.319 | 0.298 |

* **Mean Squared Error (MSE):** MSE quantifies the average squared difference between actual and predicted values. A lower MSE indicates better predictive accuracy. The neural network performed slightly better in this metric, with a lower MSE (0.089) compared to the Poisson regression (0.112). This suggests that the neural network was able to fit the data more closely, but it’s important to note that it came at the cost of interpretability.
* **Mean Absolute Error (MAE):** MAE measures the average absolute differences between predicted and actual values. It provides an easy-to-interpret metric for model accuracy. The neural network also performed better in this regard, with an MAE of 0.232, compared to 0.267 for the Poisson regression. This further supports the neural network’s ability to make more accurate predictions.
* **R-squared (R²):** R² indicates the proportion of the variance in the target variable that is explained by the model. The neural network had a slightly higher R² value (0.319) compared to the Poisson regression (0.253), suggesting that the neural network was better at capturing the variance in wine quality ratings.
* **Root Mean Squared Error (RMSE):** RMSE is the square root of the MSE and provides an interpretation in the same unit as the target variable. The neural network again outperformed the Poisson regression, with a lower RMSE of 0.298, compared to 0.335 for Poisson regression.

# Discussion:

The results from both the Poisson regression and the neural network models reveal a fundamental trade-off between \*\*predictive accuracy\*\* and \*\*interpretability\*\*. Both models achieved reasonable performance metrics, but their distinct characteristics emphasize different priorities in the analysis of wine quality prediction.

## Interpretability vs. Accuracy

The neural network model, while outperforming the Poisson regression in terms of Mean Squared Error (MSE), Mean Absolute Error (MAE), R-squared (R²), and Root Mean Squared Error (RMSE), is less interpretable. Neural networks, especially deep ones, are often referred to as "black boxes" because, despite their high accuracy, they do not provide clear insights into how individual features influence predictions. In this case, the network’s ability to more accurately predict wine quality comes at the expense of our ability to directly explain the model’s behavior.

For example, the neural network does not provide easily interpretable coefficients or feature importance metrics. As a result, stakeholders in the wine industry who may need to understand why a certain wine received a higher or lower quality rating would find the Poisson regression model much more useful. With Poisson regression, each variable’s effect on wine quality is explicitly modeled, and the estimated coefficients allow for direct interpretation of how factors like alcohol content, label appeal, and volatility acidity contribute to predicting wine quality. The transparency provided by the Poisson regression helps to clarify the relationship between these factors and the outcome, providing actionable insights for decision-making in wine production and marketing.

## Feature Engineering and Model Flexibility

One key difference between the two models is the feature engineering process. In the Poisson regression model, I did perform some\*feature engineering to capture potential interactions between features. For example, I manually created interaction terms for predictors such as alcohol content and label appeal, assuming that these factors might combine to influence wine quality in ways that their individual contributions alone would not capture. However, this feature engineering was limited to the interactions that I thought of during the model development process.

In contrast, the neural network model has a significant advantage in terms of automatically detecting complex, non-linear relationships and interactions between features. The network can learn these intricate relationships on its own, without the need for explicit feature engineering. For instance, interactions between alcohol content and label appeal could influence wine ratings in ways that I might not have considered when manually creating interaction terms. The neural network’s flexibility allows it to account for these interactions automatically, offering a more comprehensive model of the data without requiring a pre-defined set of interactions.

However, this flexibility comes with challenges. While the neural network model can capture more complex relationships, its results are harder to interpret, and it may overfit the data, particularly when there is insufficient data or when the model is too complex. The risk of overfitting can lead to poorer generalization on unseen data, making the neural network less robust in certain contexts. This is something that the Poisson regression, despite its limitations in modeling interactions, handles more effectively with its simpler, more transparent structure.

## Categorical Variables: The Role of STARS and LabelAppeal

The presence of categorical variables like STARS (wine quality rating) and LabelAppeal (whether the wine label appeals to consumers) adds another layer of complexity to the models. These variables were handled differently in the two approaches. In the Poisson regression, STARS was treated as an ordinal variable and integrated as a continuous predictor, while LabelAppeal was binary and one-hot encoded for inclusion in the model.

The model showed that LabelAppeal had a significant positive effect on wine quality, while STARS had a more nuanced, interactive relationship with other predictors like alcohol content.

In contrast, the neural network’s ability to model non-linear relationships automatically allowed it to identify and exploit these interactions more effectively, but without offering a straightforward explanation of how these relationships affected the model's predictions. The neural network handled the categorical features by transforming them into numerical representations during training, but this transformation is opaque in comparison to the clear coefficient estimates provided by Poisson regression.

# Model Choice Based on Use Case

Ultimately, the choice between a Poisson regression model and a neural network hinges on the specific goals and requirements of the project:

1. Accuracy: If the primary objective is maximizing prediction accuracy, especially when the focus is purely on model performance (e.g., in a recommendation system for wine or to determine a way to price the bottle), the neural network may be the preferred choice. Its ability to model complex interactions and non-linear relationships often results in superior performance metrics, as evidenced by its lower MSE, MAE, and RMSE.

2. Interpretability: If the goal is to provide explanations and insights into the relationships between wine characteristics and quality (such as in a report for winemakers or marketing teams), the Poisson regression is more valuable. Its clear coefficients and the ability to interpret the impact of individual variables make it easier for stakeholders to understand what drives the predicted quality scores. This is particularly useful in scenarios where regulatory bodies or consumers may need explanations for product assessments.

3. Transparency and Communication: In industries like wine production, where decisions around quality control, marketing, and production are often data-driven, having a model that can be clearly understood and communicated is crucial. The Poisson regression, being simpler and more interpretable, may be the better choice in these contexts, where it is important to not only predict wine quality but also justify decisions and build trust in the predictions.

# Final Thoughts

In conclusion, while the neural network model provided higher predictive accuracy, it does so at the cost of interpretability, which may limit its practical use in certain contexts where understanding the model’s decision-making process is essential. The Poisson regression, with its clarity and ease of explanation, provides valuable insights that can be used to guide business decisions. The key takeaway from this analysis is that the right model choice depends on the specific needs of the application—whether it's accuracy, transparency, or a balance of both. For projects where interpretability is paramount and the underlying relationships are relatively straightforward, the Poisson regression model remains a powerful tool. However, for applications where prediction accuracy is the main objective and interpretability is secondary, the neural network may offer a significant edge.

As with any model selection, it’s important to consider the trade-offs involved and to select the approach that aligns best with the project’s goals and the stakeholders’ needs.