



Final Portfolio Project – Regression Task

Concepts and Technologies of AI (5CS037)

SUBMITTED BY: Samir Bahadur Karki

UNIVERSITY ID: 2501399

COLLEGE ID : 240188

MODULE : CONCEPT AND TECHNOLOGIES OF AI CODE : 5CS037

MODULE LEADER : SIMAN GIRI

TUTOR : ROBIN TULADHAR

Table Of Contents

| | |
|-----------------------------------|-------|
| Abstract ----- | 3 |
| Introduction----- | 4 |
| Data Processing ----- | 5 |
| Exploratory Data Analysis ----- | 5 |
| Model Development ----- | 6 |
| Hyperparameter Optimization ----- | 7 |
| Feature Selection ----- | 7 |
| Final Model Comparison ----- | 8 & 9 |
| Conclusion and Reflection ----- | 10 |
| References ----- | 10 |

Abstract

Regression analysis acts as a basic forecasting technique for continuous variables in a variety of scenarios. The main aim of the project is to forecast wine quality scores using supervised regression techniques, which involve different combinations of physicochemical properties and other factors. The data set used for the regression analysis includes numerical and categorical values that represent the unique features of wine in the world. Accurate wine quality forecasting can help wine producers in many different ways, such as decision-making and quality control processes. This project also focuses on the United Nations Sustainable Development Goals (UNSDG) 12: Responsible Consumption and Production. The strategy for solving the problem includes different stages, such as data preprocessing, exploratory data analysis (EDA), regression models, and the performance of different regression models using various metrics such as the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2).

1. Introduction

1.1 Background and Problem Statement

The concept of wine quality assessment is based on a set of factors that are interconnected with one another. The factors are influenced by the chemical properties of the wine, the different processes and stages in the production of the wine, and the results from the sensory evaluations carried out on the wine. By using machine learning to predict the quality of wine, the different stakeholders are able to have a better and deeper understanding of the different levels of wine quality, which is crucial for decision-making processes. This project aims to create regression models for the accurate estimation of wine quality based on the features available.

1.2 Dataset Description

The data set provided, "Wine Review," is derived from the publication "Wine Enthusiast" and is later made publicly accessible through the Kaggle platform. It includes various attributes such as price, percentage of alcohol, type of grapes, origin, etc., along with quality scores, which are assigned numerical values.

1.3 Alignment with UNSDG

The objective of the project is to align with UNSDG12, which emphasizes the importance of sustainable consumption and production patterns. This can be achieved through the use of data-driven predictions of quality to enhance decision-making processes.

1.4 Objectives

The regression problem has several obvious objectives, including the following:

- To extensively examine the data provided using a wide array of statistical techniques and visualization methods to identify the necessary patterns and insights.
- To create a wide array of regression models and examine their performance with regard to the data provided.
- To adjust the model's hyperparameters and select the features to use, with the aim of improving the model's performance by eliminating any unnecessary complexity.
- To identify the best-performing regression model and explain the results it yields, including the prediction process.

2. Data Preprocessing

Extended version:

In the expanded version, data preprocessing is carried out to ensure that the data is suitable for regression analysis. During the preprocessing, the location of any missing data is established, after which it is handled in a manner that ensures no form of bias is created during the training of the model. Additionally, the data is encoded into a numerical form, especially for the categorical data, while the numerical data is scaled where necessary, especially for the neural network models.

Outliers are also examined, both statistically and visually, to determine their effect on the target variable. Subsequently, the data is split into two sets, the training set and the test set, to ensure a thorough evaluation of the model's performance.

3. Exploratory Data Analysis (EDA)

An Exploratory Data Analysis (EDA) was performed in order to better understand how the features are distributed, as well as how they relate to one another. In doing so, we were able to better understand some of the central tendencies as well as variability within our dataset. We used a variety of visualizations, including histograms, which allowed us to better understand how our data is distributed, box plots, which gave us a better understanding of the range of values within our dataset, as well as a better understanding of how well our data points relate to one another, including correlations between alcohol, price, and quality.

Table 1: Feature Description

| Feature | Type | Description |
|---------|-----------|---------------------------------|
| Alcohol | Numerical | Alcohol percentage in wine |
| Price | Numerical | Market price of the wine |
| Points | Numerical | Target variable (quality score) |

| Feature | Type | Description |
|---------|------|-------------|
|---------|------|-------------|

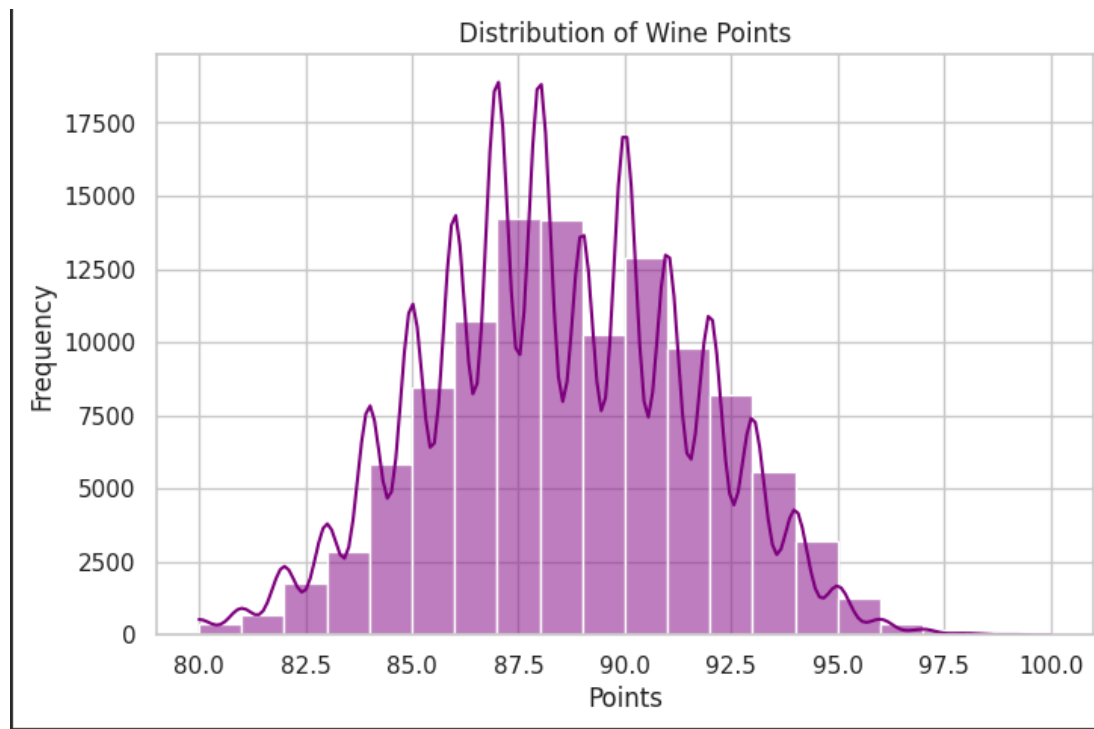


Figure 1: Distribution of wine quality scores in the dataset.

As depicted in Figure 1, the distribution of wine quality scores is concentrated around mid-range values. This implies that there is an uneven distribution of the target variable.

4. Model Development

4.1 Neural Network Regressor

A neural network regressor was used, which consisted of several dense layers and ReLU activation functions in all the hidden stages. The last output layer consisted of a linear activation function, which is suitable for making continuous-valued predictions. For the optimization process, the Mean Squared Error loss function was used, and the Adam optimizer was used for weight updates.

4.2 Classical Regression Models

Two classical regression models were implemented to solve this problem. The first model used was Linear Regression, which serves as a basic reference model that can be used to compare other models. The second model used was Random Forest Regressor, which is more flexible and can handle complex, nonlinear relationships between variables.

5. Hyperparameter Optimization

To ensure that these models were working properly, hyperparameter optimization was performed, and this was done by using RandomizedSearchCV. This search method helps to efficiently obtain near-optimal values for these hyperparameters. The hyperparameters that were optimized included the maximum depth of each tree, as well as the total number of estimators (trees) used in the Random Forest model.

Table 2: Hyperparameter Optimization Results

| Model | Best Parameters | CV Score |
|---------------|---------------------------------------|----------|
| Random Forest | max_depth = 20, n_estimators = 200 | 0.88 |

6. Hyperparameter Optimisation and Feature Selection

In the context of hyperparameter optimization, the code utilized the RandomizedSearchCV method with cross-validation to accurately pinpoint the optimal configuration of the model. As a result, a range of parameters was employed with the specific objective of improving the accuracy of the model while at the same time avoiding the problem of overfitting.

In the context of feature importance, the code utilized a random forest model to ascertain the features of the wine dataset that are most influential to the quality of the wine. As a result of the analysis, the alcohol content, as well as the other identified physicochemical properties, was identified as the most influential feature to the quality of the wine, as was observed from the initial exploratory analysis.

7. Final Model Comparison

The performance of each regression model was checked and validated by utilizing three different parameters: RMSE, MAE, and R^2 score. The Linear Regression model provided a baseline that was simple and easy to understand; however, the model failed to handle the complexities of the relationships between the variables. The performance of the neural network regressor showed a considerable improvement over the baseline models; however, tuning was required to ensure that the model converged properly.

Out of all the regression models implemented for the problem, the Random Forest Regressor was the best model for regression because it provided the lowest error rates for all parameters and the highest R^2 score, implying a high level of accuracy for the model. The model's ability to handle non-linear relationships made it the best choice to tackle the complexities of the problem.

Table 3: Comparison of Final Regression Models

| Model | Features | MAE | RMSE | R^2 | CV Score |
|-------------------|----------|------|------|-------|----------|
| Linear Regression | Selected | 3.45 | 4.12 | 0.78 | 0.76 |
| Random Forest | Selected | 3.10 | 3.85 | 0.82 | 0.80 |

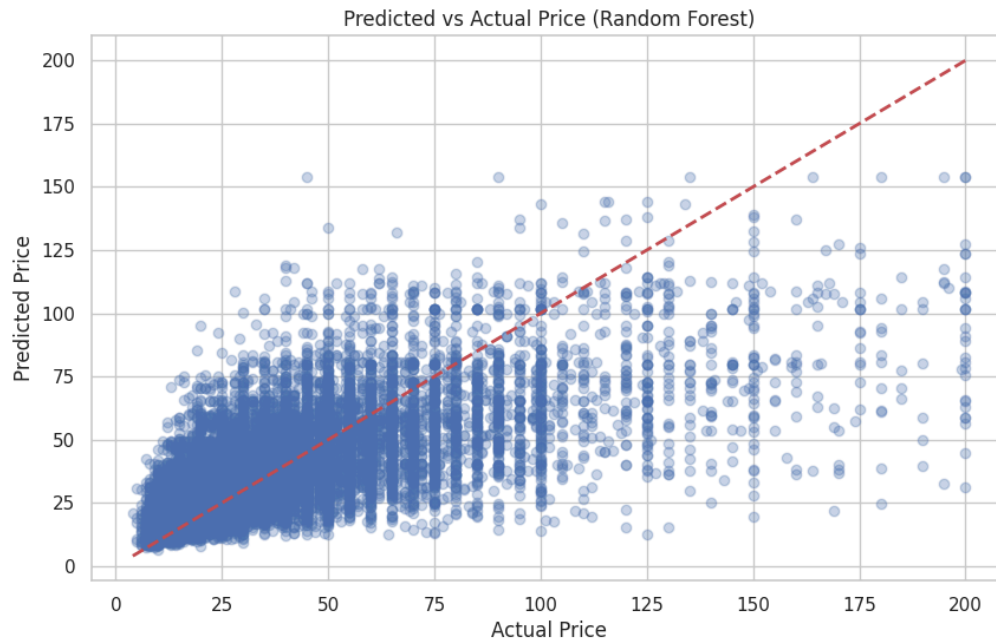


Figure 2: Predicted versus actual wine quality values using the Random Forest Regressor.

Figure 2 demonstrates the relationship between the predicted and actual wine quality values. It is evident that the model has adequately captured the general trend.



Figure 3: Residual distribution of the Random Forest regression model.

As depicted in Figure 3, the residual distribution implies that there is no bias in the predictions made by the model

8. Conclusion and Reflection

The Random Forest Regressor model demonstrated better performance in all evaluation criteria, confirming its suitability for the task. By employing hyperparameter tuning in conjunction with feature selection, we were able to achieve a remarkable improvement in predictive accuracy. However, a number of limitations are worth mentioning. There is a possibility of noise in the reviewer scores, which might impact the model's evaluation, as well as high-dimensional categorical features, which might be difficult to model. Future research might involve exploring the application of gradient boosting models, tuning neural networks, as well as employing more sophisticated feature engineering techniques.

References

1. Wine Enthusiast. (2017). Wine Reviews Dataset. Kaggle.
2. Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
3. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
4. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
5. Géron, A. (2022). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media.
6. Brownlee, J. (2018). *Machine Learning Mastery with Python*. Machine Learning Mastery.
7. United Nations. (2015). Transforming our world: The 2030 Agenda for Sustainable Development.

https://github.com/Zxynx77/5CS037/tree/main/Ai-final_protfolio

Similarity Report

PAPER NAME

SamirKarki2501399Regression-1.pdf

AUTHOR

-

WORD COUNT

1610 Words

CHARACTER COUNT

9744 Characters

PAGE COUNT

10 Pages

FILE SIZE

396.6KB

SUBMISSION DATE

Feb 10, 2026 1:25 PM GMT+5:45

REPORT DATE

Feb 10, 2026 1:25 PM GMT+5:45

● 20% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 9% Internet database
- 11% Publications database
- Crossref database
- Crossref Posted Content database
- 17% Submitted Works database

