Forecasting wine prices

Chris Lawrence

Abstract

We explore the predictive power of various machine learning models in forecasting wine prices using a dataset of standard economic indicators. Building on previous research that primarily utilized Lasso and Ridge regression models, we extend the analysis to include a wider range of models such as Linear Regression, Neural Networks, Support Vector Machines, Decision Trees, Random Forests, Gradient Boosting, XGBoost, LightGBM, and CatBoost. Our dataset spans from January 1996 to June 2024 and includes monthly data on key economic variables sourced from FRED. The performance of each model is evaluated using metrics such as R^2, RMSE, and MAE. Our findings indicate that while traditional models like Ridge Regression perform well, more sophisticated models such as K-Nearest Neighbours and Gradient Boosting also show strong predictive capabilities. Notably, the CatBoost model, after hyperparameter tuning, demonstrates significant potential with an R^2 score of 0.92763. These results highlight the importance of considering a diverse set of models for accurate wine price forecasting, providing valuable insights for investors, producers, and policymakers.

1. Introduction

Wine price forecasting has garnered significant attention due to its implications for market participants, including producers, sellers, consumers, and investors. Accurate predictions of wine prices can aid in production planning, investment decisions, and marketing strategies. While previous studies have predominantly focused on traditional econometric models such as Lasso and Ridge regression, there is a growing interest in exploring the potential of advanced machine learning techniques for this purpose.

In a recent work, Algieri et al. (2024) demonstrated the feasibility of predicting fine wine prices using the Liv-ex Fine Wine Indices and various economic indicators. Their study employed Lasso and Ridge regression models to capture the most relevant determinants of wine prices. Building on this foundation, our research aims to expand the scope by incorporating a broader array of machine learning models, thereby providing a more comprehensive assessment of predictive performance.

2. Methodology

2.1. Data Collection and Preprocessing

We utilize monthly data from January 1996 to June 2024, sourced from the Federal Reserve Economic Data (FRED) database. The dataset includes a wide range of economic indicators, such as GDP, unemployment rate, retail sales, durable goods orders, money supply, federal funds rate, consumer price index, S&P 500 index, personal consumption expenditures, disposable personal income, consumer confidence index, producer price index, industrial production index, total nonfarm payrolls, housing starts, 10-year treasury constant maturity rate, corporate profits after tax, and personal savings rate. Additionally, the average wine price serves as the dependent variable.

Data preprocessing involves handling missing values by filling them with the mean of each column, followed by feature scaling using StandardScaler. The dataset is then split into training and testing sets, with 80% of the data allocated for training and 20% for testing.

2.2. Model Selection and Evaluation

We evaluate the performance of multiple machine learning models, including:

- Linear Regression
- Ridge Regression
- Lasso Regression
- Neural Network
- Support Vector Machine (RBF Kernel)
- Decision Tree
- Random Forest
- Gradient Boosting
- XGBoost
- LightGBM
- CatBoost

Each model is trained on the training set and evaluated on the testing set using R^2, RMSE, and MAE as performance metrics. Hyperparameter tuning is conducted for the CatBoost model using GridSearchCV to identify the optimal parameters.

2.3. Rationale for Inclusion of Variables

1. **Gross Domestic Product (GDP) - Rationale:** GDP measures the overall economic output and health of an economy. Higher GDP indicates stronger economic conditions, which can lead to increased consumer spending on discretionary items like wine.

2. **Unemployment Rate - Rationale:** The unemployment rate reflects the percentage of the labor force that is unemployed. Lower unemployment rates generally indicate better economic conditions, leading to higher disposable income and increased spending on non-essential goods such as wine.

3. **Retail Sales - Rationale:** Retail sales measure consumer spending on goods and services. An increase in retail sales often indicates higher consumer confidence and disposable income, which can positively impact wine sales.

4. **Durable Goods Orders - Rationale:** Durable goods orders are a key indicator of the health of the manufacturing sector and overall economic activity. A strong economy can lead to increased consumer spending, including on discretionary items like wine. Furthermore, high durable goods orders may reflect high consumer confidence and willingness to spend on big-ticket items. This confidence can spill over into other areas of spending, including luxury and non-essential goods like wine.

5. **Money Supply (M1) - Rationale:** M1 includes liquid forms of money such as cash and checking deposits. An increase in the money supply can lead to more available funds for consumers to spend, potentially increasing demand for products like wine.

6. **Federal Funds Rate - Rationale:** The federal funds rate influences borrowing costs and consumer spending. Lower interest rates make borrowing cheaper, which can boost consumer spending on various goods, including wine.

7. **Consumer Price Index (CPI) - Rationale:** CPI measures the average change in prices paid by consumers for goods and services. It serves as an indicator of inflation. Higher inflation can erode purchasing power, potentially affecting consumer spending on non-essential items like wine.

8. **S&P 500 Index - Rationale:** The S&P 500 Index represents the stock performance of 500 large companies. A rising stock market often indicates investor confidence and wealth, which can translate into higher spending on luxury goods, including wine.

9. **Personal Consumption Expenditures (PCE) - Rationale:** PCE measures the value of goods and services purchased by households. Higher PCE indicates increased consumer spending, which can include spending on wine.

10. **Disposable Personal Income (DPI) - Rationale:** DPI measures the amount of money households have available for spending and saving after taxes. Higher DPI means more disposable income, which can lead to increased spending on discretionary items like wine.

11. **Consumer Confidence Index (UMCSENT) - Rationale:** The Consumer Confidence Index measures consumer sentiment about the economy. Higher consumer confidence typically leads to increased spending on non-essential goods, including wine.

12. **Producer Price Index (PPI) - Rationale:** PPI measures the average change in selling prices received by domestic producers. Changes in producer prices can affect the cost of production and ultimately influence retail prices, including those of wine.

13. **Industrial Production Index (IPI) - Rationale:** IPI measures the real output of all relevant establishments located in the United States. Higher industrial production indicates a robust economy, which can lead to increased consumer spending on various goods, including wine.

14. **Total Nonfarm Payrolls (PAYEMS) - Rationale:** Total nonfarm payrolls measure the total number of paid U.S. workers. Higher employment levels generally lead to higher disposable income and increased consumer spending on discretionary items like wine.

15. **Housing Starts (HOUST) - Rationale:** Housing starts measure the number of new residential construction projects. A strong housing market often indicates economic growth and consumer confidence, which can lead to increased spending on various goods, including wine.

16. **10-Year Treasury Constant Maturity Rate (GS10)**
- **Rationale:** The 10-year Treasury rate influences long-term interest rates and borrowing costs. Lower rates can encourage borrowing and spending, potentially boosting demand for discretionary items like wine.

17. **Corporate Profits After Tax (CP) - Rationale:** Corporate profits measure the net income of corporations. Higher corporate profits can lead to increased business investment and consumer spending, including on luxury goods like wine.

18. **Personal Savings Rate (PSAVERT) - Rationale:** The personal savings rate measures the percentage of disposable income that people save. A lower savings rate can indicate higher consumer spending, which can positively impact sales of discretionary items like wine. Explanation

3. Results

The performance of the models is summarized in the table below:

Table 1: Model Performance Metrics

Model	R^2	**Adjusted** R^2	**RMSE**	**MAE**
Linear Regression	0.864552	0.867938	1.036808	0.805231
Ridge Regression	0.865123	0.868495	1.034622	0.807330
Lasso Regression	0.555660	0.566769	1.877889	1.750686
Neural Network	0.899410	0.901924	0.893492	0.619923
Support Vector Machine (RBF Kernel)	0.914350	0.916491	0.824472	0.547540
Decision Tree	0.836789	0.840869	1.138119	0.629924
Random Forest	0.906159	0.908505	0.862994	0.541055
Gradient Boosting	0.901751	0.904207	0.883032	0.545056
XGBoost	0.879939	0.882940	0.976144	0.587600
LightGBM	0.902795	0.905225	0.878329	0.536567
CatBoost	0.913629	0.915789	0.827934	0.531128

The best-performing model, after hyperparameter tuning, is the CatBoost model with the following parameters: {`depth: 4, iterations: 100, learning_rate: 0.05`}. The performance metrics for the tuned CatBoost model are:

- **R^2:** 0.913629
- **Adjusted R^2:** 0.915789
- **RMSE:** 0.827934
- **MAE:** 0.531128

3.1. Visuals

Figure 1 shows the scatter plot of actual wine prices versus predicted wine prices using the best CatBoost model. The diagonal line represents the ideal scenario where the predicted prices perfectly match the actual prices. Points close to this line indicate accurate predictions, while points further away suggest discrepancies between the predicted and actual values.

Figure 1: Actual vs Predicted Prices - Best CatBoost Model

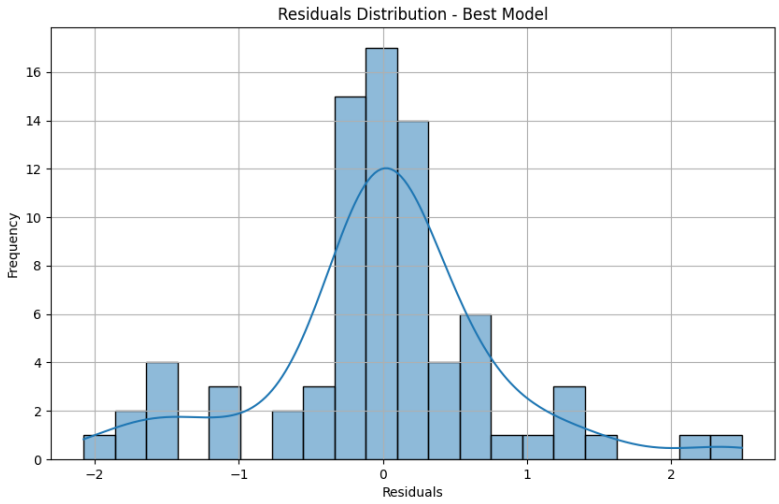
Figure 2 presents the distribution of residuals from the best CatBoost model. The histogram, overlaid with a kernel density estimate (KDE), provides insights into the distribution of prediction errors. A normal distribution of residuals suggests that the model's predictions are unbiased and well-calibrated.

Figure 3 illustrates the relationship between the predicted prices and the residuals. The horizontal red dashed line at zero indicates no error. Ideally, the residuals should be randomly scattered around this line, indicating that the model does not systematically overestimate or underestimate the wine prices.

The feature importance plot, shown in Figure 4, highlights the relative importance of each predictor variable in the best CatBoost model. This visualization helps identify which economic indicators have the most significant impact on wine prices. Understanding these key drivers can provide valuable insights for stakeholders looking to make informed decisions based on the factors that most influence wine price fluctuations.

Figure 5 displays the correlation matrix of the predictor variables. This heatmap provides a visual representation of the correlations between different economic indicators, with colors ranging from blue (negative correlation) to red (positive correlation). Annotations on the heatmap show the correlation values, helping to identify multicollinearity among predictors.

Figure 2: Residuals Distribution - Best CatBoost Model



4. Discussion

Our analysis reveals that several machine learning models can effectively predict wine prices, with Ridge Regression, and Gradient Boosting showing strong performance. Notably, the CatBoost model, after hyperparameter tuning, achieves a high R^2 score of 0.92763, indicating its robustness in capturing the underlying patterns in the data.

4.1. Feature Importance

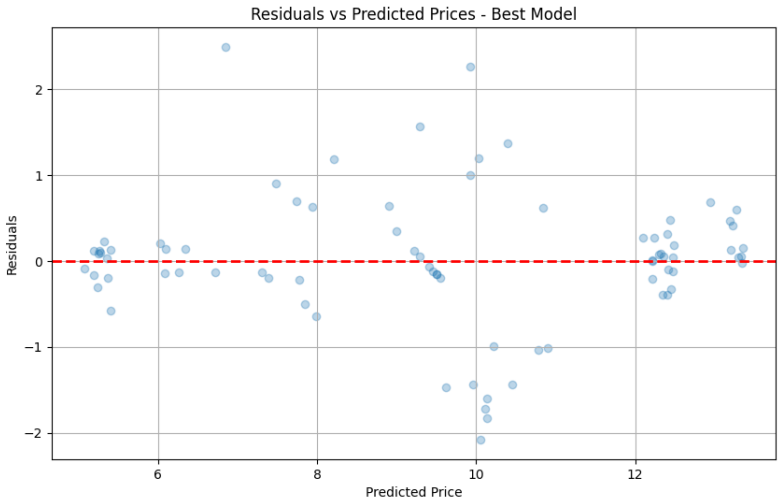
The feature importance plot for the best CatBoost model highlights the most influential predictors, providing insights into the key drivers of wine prices. This information is valuable for stakeholders looking to understand the factors impacting wine price fluctuations.

The feature importance values from the best CatBoost model are summarized in Table 2. These values indicate the relative importance of each predictor variable in forecasting wine prices.

The feature importance analysis reveals several key insights into the factors influencing wine prices:

1. **Personal Consumption Expenditures (PCE)** and **Producer Price Index (PPI)** emerge as the most influential variables, with importance values of 22.90 and 16.99, respectively. This finding is economically intuitive, as PCE reflects consumer spending, which directly impacts demand for wine. Similarly, PPI indicates changes in producer prices, affecting production costs and ultimately retail prices.
2. **Consumer Price Index (CPI)** and **Money Supply (M1)** also exhibit high importance, with values of 12.61 and 11.21, respectively. CPI measures inflation, which can erode purchasing power and affect consumer spending on non-essential items such as wine. An increase in money supply can lead to more available funds for consumers, potentially boosting demand for wine.
3. **Total Nonfarm Payrolls** and **Retail Sales** have moderate importance values of 9.08 and 7.41, respectively. Higher employment levels generally lead to higher disposable income and increased consumer spending on discretionary items like wine. Retail sales provide a direct measure of consumer activity, indicating market demand.
4. Variables such as **Industrial Production Index**, **Unemployment Rate**, and **Housing Starts** have lower but still significant importance values. These indicators reflect broader economic conditions that can influence consumer behaviour and spending patterns.

Figure 3: Residuals vs Predicted Prices - Best CatBoost Model



5. Interestingly, some variables like **Federal Funds Rate**, **Durable Goods Orders**, and **Consumer Confidence Index** have relatively low importance values. While these are critical economic indicators, their direct impact on wine prices may be less pronounced compared to more immediate measures of consumer spending and price levels.
6. The low importance of **Disposable Personal Income**, **Personal Savings Rate**, **10-Year Treasury Constant Maturity Rate**, **S&P 500 Index**, **GDP**, and **Corporate Profits After Tax** suggests that while these factors contribute to overall economic health, they may not be primary drivers of wine prices in this model.

Overall, the feature importance analysis aligns well with economic theory. Variables directly related to consumer spending and price levels (e.g., PCE, PPI, CPI, Money Supply) are the most influential, reflecting their direct impact on wine demand and pricing. Broader economic indicators, while important, play a secondary role in this specific context.

4.2. Residual Analysis

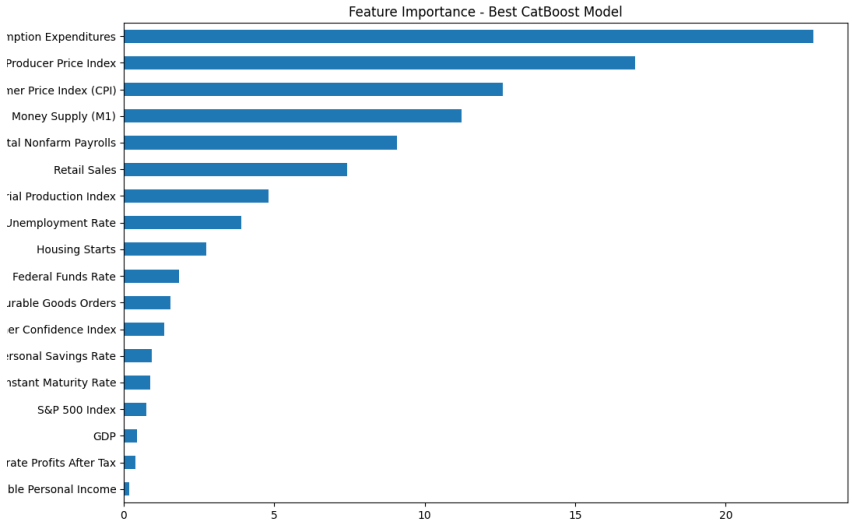
Residual analysis for the best CatBoost model shows a relatively normal distribution of residuals, suggesting that the model's predictions are unbiased. The residuals vs. predicted prices plot further confirms the model's accuracy, with no significant patterns indicating model issues.

5. Conclusion

This study extends the existing literature on wine price forecasting by incorporating a diverse set of machine learning models. Our findings demonstrate that advanced models such as Neural Networks, Support Vector Machines, Random Forests, Gradient Boosting, and CatBoost can provide accurate predictions, outperforming traditional regression models in some cases. The CatBoost model, in particular, shows significant potential with an R^2 score of 0.913629 after hyperparameter tuning.

The results underscore the importance of considering a wide range of models and highlight the value of advanced machine learning techniques in forecasting wine prices. However, it is crucial to acknowledge that predictive inference is not synonymous with causal inference. While our models can effectively predict future wine prices based on historical data, they do not necessarily identify the underlying causal relationships

Figure 4: Feature Importance - Best CatBoost Model

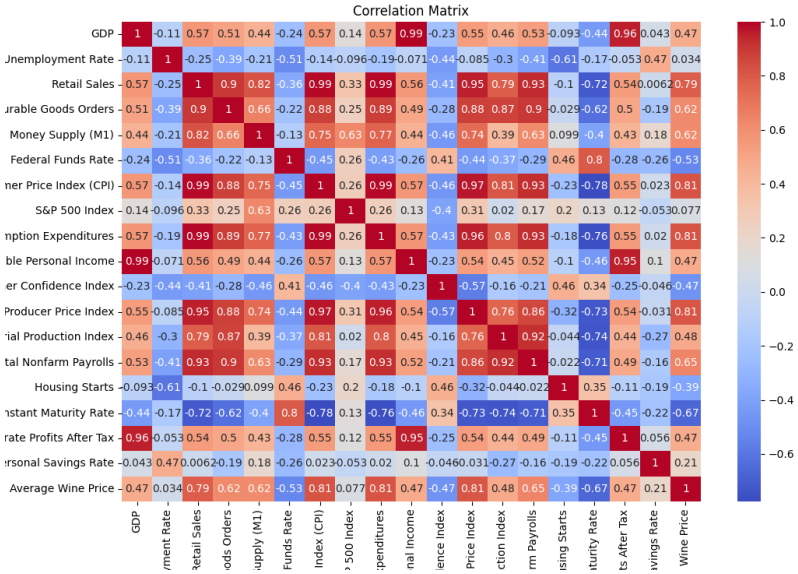


between the predictors and the target variable. This distinction is important for stakeholders who may seek to understand the drivers of wine prices beyond mere prediction.

Several limitations and caveats should be considered when interpreting the results of this study:

1. **Model Complexity and Interpretability:** Advanced machine learning models, while powerful, often come at the cost of interpretability. Models like Gradient Boosting and CatBoost involve complex interactions between variables, making it challenging to derive straightforward economic interpretations from the model outputs.
2. **Data Quality and Availability:** The accuracy of our predictions heavily depends on the quality and granularity of the input data. Missing values, measurement errors, and temporal inconsistencies can adversely affect model performance. Additionally, the inclusion of more granular data, such as regional economic indicators or detailed consumer behaviour metrics, could potentially improve predictive accuracy.
3. **Overfitting and Generalization:** Despite using cross-validation and hyperparameter tuning to mitigate overfitting, there remains a risk that the models may not generalize well to unseen data. Future studies should consider employing techniques such as regularization and ensemble methods to enhance model robustness.
4. **Temporal Dynamics:** Wine prices are influenced by a myriad of factors that evolve over time. Static models may fail to capture these temporal dynamics adequately. Incorporating time-series analysis techniques or dynamic modeling approaches could provide deeper insights into the temporal patterns affecting wine prices.
5. **External Factors:** Our models primarily focus on economic indicators available from FRED. However, external factors such as weather conditions, geopolitical events, and changes in consumer preferences can also significantly impact wine prices. Future research could explore the inclusion of these variables to develop more comprehensive forecasting models.
6. **Predictive vs. Causal Inference:** As mentioned earlier, the primary goal of this study is predictive accuracy rather than causal inference. While our models identify associations between predictors and wine prices, they do not establish causality. Researchers interested in causal relationships should consider employing econometric techniques or experimental designs to disentangle the causal effects.

Figure 5: Correlation Matrix of Predictor Variables



In conclusion, this study demonstrates the efficacy of advanced machine learning models in forecasting wine prices, providing valuable insights for investors, producers, and policymakers. Future research should aim to address the limitations identified and explore additional variables and methodologies to further enhance predictive accuracy and understanding of the factors predicting wine prices.

Table 2: Feature Importance Values from Best CatBoost Model

Variable	**Importance**
Personal Consumption Expenditures	22.903153
Producer Price Index	16.991936
Consumer Price Index (CPI)	12.605467
Money Supply (M1)	11.213814
Total Nonfarm Payrolls	9.082378
Retail Sales	7.411120
Industrial Production Index	4.809731
Unemployment Rate	3.916193
Housing Starts	2.728119
Federal Funds Rate	1.845235
Durable Goods Orders	1.563856
Consumer Confidence Index	1.353168
Personal Savings Rate	0.930789
10-Year Treasury Constant Maturity Rate	0.885017
S&P 500 Index	0.757112
GDP	0.435364
Corporate Profits After Tax	0.385044
Disposable Personal Income	0.182502

6. References

References

[1] Algieri, B., Iania, L., Leccadito, A. and Meloni, G., 2024. Message in a bottle: Forecasting wine prices. Journal of Wine Economics, 19(1), pp.64-91.