Wine Price Predictions

Author: Paul Foy

Project Overview

Customer Description:

Winerate.com. Winerate.com is a website that provides wine enthusiast information about popular wines from different regions of the world.

Objective:

Their analytics department wants to away to easily predict the cost of a trending bottle of wine over time. This feature allows their customers to know if a bottle is trending and is expected to increase its value over time.

Using existing data from previous wine sales, create a model that accurately predicts the price of a bottle of wine.

Source Data Description

Overview:

The dataset contains 7,500 different types of red wines from Spain with 11 features that describe their price, rating, and even some flavor description. The was collected by me using web scraping from different sources (from wine specialized pages to supermarkets).

Data Source: https://www.kaggle.com/datasets/fedesoriano/spanish-wine-quality-dataset

Data Dictionary:

- **winery**: Winery name
- **wine**: Name of the wine
- **year**: Year in which the grapes were harvested
- **rating**: Average rating given to the wine by the users [from 1-5]
- **num_reviews**: Number of users that reviewed the wine
- **country**: Country of origin [Spain]
- **region**: Region of the wine
- **price**: Price in euros [€]
- **type**: Wine variety
- **body**: Body score, defined as the richness and weight of the wine in your mouth [from 1-5]
- **acidity**: Acidity score, defined as wine's "pucker" or tartness; it's what makes a wine refreshing and your tongue salivate and want another sip [from 1-5]

Note: For our modeling, the following features were removed:

- winery, wine, region: These features have high cardinality
- country: All values were "Espana"

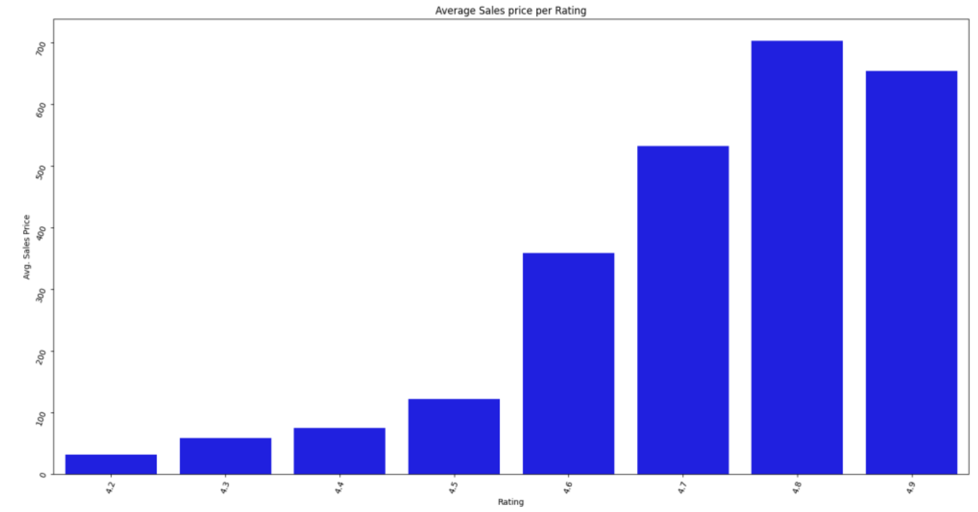
Key Insights

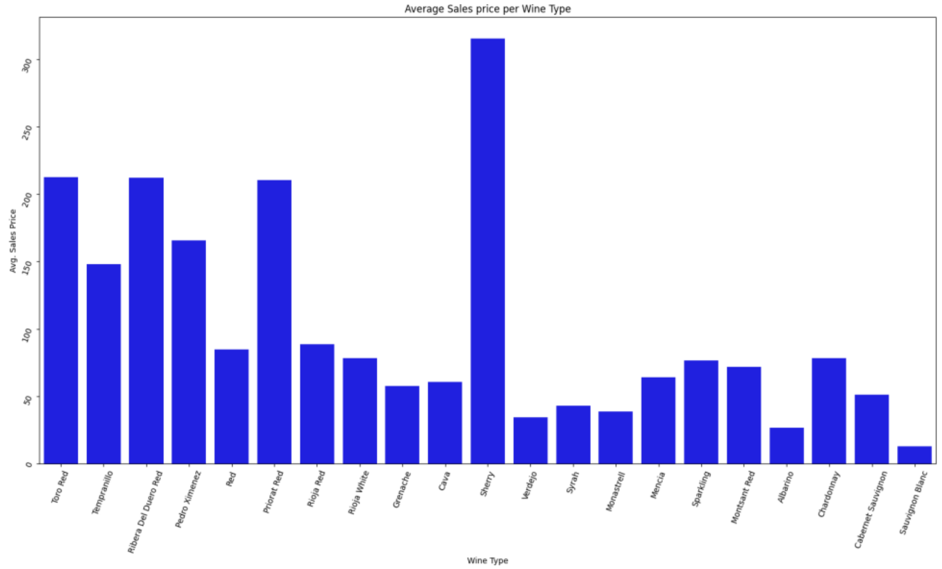
Higher ratings produce higher prices

There is a moderate correlation between the wine's rating and Price. Higher-rated wines can demand higher prices. In Figure 1, you can see that higher-rated wines (4.8-4.7) are significantly more expensive than lower rated wines.

Sherry commands the highest price

Sherry has the highest average price of all wine types. See figure 2. The second most expensive wine is almost half the price.





Modeling Approach and Initial Findings

Approach: Create a model that predicts if a wine is expensive or not

- The Tuned LR model produced the best results. For this model, it's important that we have accurate True Positive and True Negative rates. The tuned LR model produced True Positives 85% of the time and True Negatives 83% of the time.
- While the Default LR model produced more accurate True Negative predictions, its True Positive performance was significantly below the Tuned LR model's performance with only 57%.

Final Results (Binary Classification Model)

	LR Default	LR Tuned
True Pos. rate	57%	86%
True Neg. rate	96%	82%

Recommendation

As a result of these evaluations, here are my recommendations:

1. To create a more accurate regression model, we need to pull in more data that might help increase the accuracy and reliability of the models. The results are not good enough for production at this point.
2. Start by deploying the tuned LR model that classifies the wine as above or below the average cost. This first model could help classify where specific wines would be sold. For example, more expensive wines would be sold in more high-end establishments like restaurants and high-end grocers.