Executive Summary

Wine Price Prediction Project

Yasmine Aitouny, Daniel Carmona, Terisha Kolencherry, Pooja Nagrecha

Objective and Methodology

This project aims to develop a model to predict wine price ranges based on selected features. The group used publicly available data from Wine Enthusiast magazine, cleaned the data to make sure each entry utilized in the model had a complete profile, explored the data, and then ran multiple machine learning models to build a model to achieve the project objective. The group then hosted a live application that takes in user input and delivers an estimate of wine price using the finalized model.

The Model

After four rounds of machine learning testing, with 6-7 different types of models in each round, the group selected a Random Forest Classification Model. The group focused on the accuracy of the model and determined that this model had the highest accuracy rate, while maintaining an acceptable level of precision. The model takes in six features: wine type (red vs white wine), country or origin, province of origin, winery, designation of wine, and grape variety. Model outputs are also divided into six categories: "Value: Under $10", "Popular: $10-$15", "Premium: $15-$20", "Super Premium: $20-$30", "Ultra Premium: $30-$50", and "Iconic: Over $50". Output bins were structured to allow for more even spread of wines across categories. The model displays moderate accuracy across all categories, with higher accuracy across mid-priced wines.

Next Steps

Alongside the selected machine learning model, the group also built a natural language processing model to take in the wine descriptions from the magazine and transform them into a set of numerical values that could be fed into a model to distinguish "good" wines from "excellent" wines. This distinguishment is important because the current model has some difficulty sorting wines on either end of the price scale. Due to time constraints this separate track was not integrated into the larger price prediction model, but by sorting wines into these two distinct buckets first and then running the current model, the group might be able to better the application's predictive ability.