Wine Project

Min-Ju Chung

2024

"Trattoria Luna," a distinguished Italian restaurant in California, is set to broaden its culinary offerings by adding a diverse selection of wines to its menu. The restaurant aims to provide patrons with a unique wine experience, underpinned by a carefully curated collection and thorough wine ratings. However, this expansion presents considerable challenges, particularly in the accurate evaluation and rating of new wine sources, including local wineries and independent sellers.

Consequently, the objective of this project is to develop a trained algorithm that accurately assesses wines based on inputs such as price, year, state, variety, region, and winery. This enhancement aims to improve the dining experience and boost customer satisfaction.

Regarding the summary of this project, the initial dataset comprises 54,503 rows and 9 wine features. In the preprocessing phase, Exploratory Data Analysis (EDA) is conducted prior to Data Cleaning to facilitate smooth preprocessing, leveraging patterns observed in both numerical and categorical data. During Data Cleaning, the "price" and "score" fields are imputed straightforwardly using the median method, whereas "regions", "variety", and "winery" are handled through a range of techniques including text mining and K-Nearest Neighbors imputation, tailored to their unique characteristics. During Feature Engineering, not only a new feature "year" is extracted from "title" exhibiting a correlation coefficient of 0.188 after cleaning, but also irrelevant features are removed, and low-frequency values are consolidated to reduce noise. In the Feature Selection and Data Encoding section, 15 features are selected using the Random Forest Feature Importance method following One-Hot Encoding. This Feature Selection approach will not assume that the main structure of the data is linear as the Principal Component Analysis (PCA) method, which may ignore the effects of noise and outliers.

In the model training section, a Decision Tree Model is implemented, followed by three ensemble algorithms: Random Forest for Bagging, Gradient Boosting for Boosting, and a

Stacking Model that combines both Random Forest and Gradient Boosting, constructed using the Linear Regression method. The models are evaluated using metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-squared (R^2), and K-Fold Cross-Validation. Among them, the Stacking Model emerges as the most effective. Subsequently, Grid Search is employed to fine-tune the Stacking Model by optimising its parameters, which increases the R-squared from 0.352 to 0.6405, demonstrating a substantial improvement in the refinement steps. However, the learning curve reveals an overfitting issue, as indicated by the significant gap between the training curve and the validation score line. To address this, adjustments are made in areas such as imputing categorical features, grouping values, removing outliers, and selecting features to mitigate overfitting and enhance machine training outcomes. Additionally, a pipeline system is implemented starting from the Data Cleaning step. While the pipeline does not explicitly show steps like text mining to create the feature "year", it facilitates a quick understanding of the project and proves beneficial in team settings. Lastly, although the Grid Search within the Pipeline may seem simplified due to the need for precisely set grid parameters to simulate a comprehensive search, it highlights the project's commitment to refining the model. This step is vital for optimising performance and ensuring the model's practicality in real-world applications.

To enhance the predictive model, it is advantageous to continuously gather customer feedback on their dining experiences. Collecting data for the machine learning models can help fine-tune the rating system, refine the wine selection, and tailor promotional efforts to better align with customer needs and preferences. Furthermore, incorporating additional characteristics such as acidity into the model can improve the rating system.

Utilising this project, several recommendations are offered to help Trattoria Luna address their business challenges and boost their profits. First, promotional campaigns leveraging the highly-rated wines identified by the rating system can be launched. These could include limited-time discounts, special tasting events, or wine pairing menus to stimulate purchases. Second, by training staff to combine wine ratings with customer preferences, upselling and cross-selling strategies can be implemented to increase average check size and overall revenue per customer. Third, the establishment of a customer loyalty program could reward regular patrons with incentives such as discounts, exclusive access to new releases, and wine tasting events, helping to cultivate long-term customer relationships. Fourth, enhancing the restaurant's social media presence by sharing engaging content such as wine pairing tips, tasting notes, and customer testimonials is likely to attract wine enthusiasts and increase restaurant visits.

Regarding the limitations of this model, there are several areas that require improvement. Firstly, models such as stacking and gradient boosting, which are capable of capturing complex patterns, are prone to overfitting. Although steps are taken to mitigate this, completely eliminating overfitting is challenging and can affect the model's generalizability. Secondly, feature selection bias could lead to a model where the choice of features is heavily influenced by Random Forest Feature Importance. This method might overlook critical predictors that are significant for human interpretation. Third, computational requirements are critical for the use of ensemble methods and grid search. They demand significant computational resources, which might not be feasible in all operational environments, especially with larger datasets. Lastly, a pipeline system which facilitates streamlined processing can also obscure the detailed steps involved in model refinement. This might make troubleshooting and iterative improvements more challenging.

In conclusion, by adopting the machine training process to refine their own rating system, "Trattoria Luna" will not only overcome the challenges associated with expanding its wine offerings but also set a new standard for excellence in wine service within the restaurant industry. This initiative promises to enrich the customer experience and reinforce Trattoria Luna's reputation as a leader in culinary innovation and customer satisfaction.