cellister / Brewing_Chemistry

Code Issues Pull requests Actions Projects Wiki Security Insights Se

0 stars 0 forks 1 watching 1 Branch 0 Tags Activity

Public repository

1 Branch 0 Tags Go to file t Go to file Add file Code

cellister added readme images 1aceb8b · 3 minutes ago

Development_notebooks	update color scheme	15 minutes ago
Images	update color scheme	15 minutes ago
Project_PDFs	update color scheme	15 minutes ago
.gitignore	initial commit	last month
Brewing_Chemistry_Final.ip...	update color scheme	15 minutes ago
readme.md	added readme images	3 minutes ago

Brewery Operations and Market Analysis Dataset

This project was completed as the final Phase 5, Capstone assessment in the Flatiron School's Data Science Bootcamp.

Analysis by Erin Wasserman, July 2024

Overview

This dataset presents an extensive collection of data from a craft beer brewery, spanning from January 2020 to January 2024. It encapsulates a rich blend of brewing parameters, sales data, and quality assessments, providing a holistic view of the brewing process and its market implications.

Business Problem

In the beer market, demand plays a key role in future industry dynamics. The introduction of new ingredients and innovative flavors into the beer market, combined with a business model that values consumer loyalty, will increase appeal among generations.

The main goal of any brewery is to focus on producing high quality beer. **Quality beer is a key success factor.** The most important thing a brewery can do is keep producing quality beer to stay competitive. As the variety of beers on the market increases, low-quality beers will be eliminated first.

1. Determine best malt and hops types for quality beer.
2. Assess malt-to-hops ratio impact on beer quality.
3. Evaluate ML accuracy in predicting beer characteristics.
4. Visualize beer styles based on quality ratings.

Data Understanding

Dataset Description

Kaggle Dataset

Data Format and Structure: -The dataset is structured in a tabular format, provided in a CSV file for easy integration with various data analysis tools. -It comprises over 10 million records, each representing a unique batch with a comprehensive set of features.

Intended Audience: This dataset is invaluable for data scientists, brewing process engineers, market analysts, supply chain experts, and quality control professionals in the brewing industry. It is also highly relevant for academic research in food technology, fermentation science, and business analytics.

Disclaimer: -The data is synthetic and intended for educational, analytical, and simulation purposes. -Users are advised to apply appropriate data processing and analysis techniques for meaningful insights. -This comprehensive dataset serves as a rich resource for exploring the intricacies of brewing science, market dynamics, and operational efficiency in the craft beer industry.

Highlighted Data Features

Brewing Parameters: Includes crucial brewing factors such as fermentation time, temperature, pH level, gravity, and ingredient ratios. These parameters are pivotal in understanding the brewing process and its impact on the final product.

Beer Styles: The dataset categorizes beers into various styles like IPA, Stout, Lager, etc.

Quality Scores: Each batch is rated for its quality on a scale, offering insights into the success and consistency of different brewing approaches.

Application: Brewing Process Optimization: Ideal for analysis aiming to correlate brewing techniques with beer quality, facilitating the optimization of brewing conditions for superior product quality.

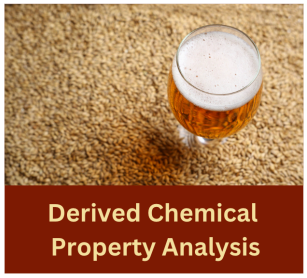
Data Splits

-Use the following code to create a reproducible subset of the larger Kaggle Dataset

- The dataset was randomly split into training (80%) and test (20%) sets to evaluate model performance.

Data Licensing and Usage

Methodology



Data Preparation

Data Cleaning and Preprocessing

1. Correlation Plots
2. Pair plots

Modeling

1. K-Means Clustering
2. Random Forest
3. Gradient Boosted Decision Tree

Evaluation

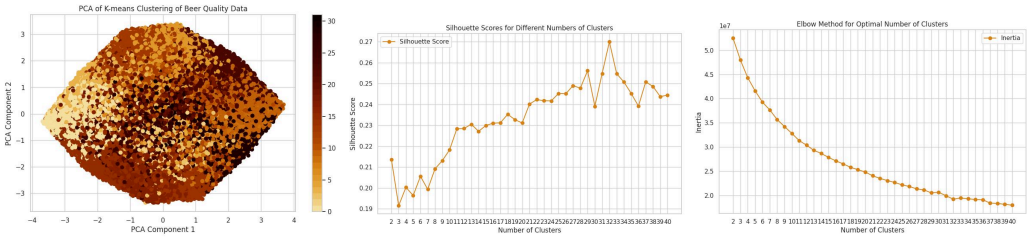
1. Inertia Score
2. Silhouette Score
3. R^2 Score

Limitations

1. Computational Bottlenecks
    - Despite the completeness of the dataset, the sheer size of 10 million data points likely caused significant computational bottlenecks, limiting the depth and breadth of this analysis.
2. Underutilization of Data
    - Due to resource constraints, not all data could be effectively utilized, potentially leading to incomplete insights.
3. Sampling Bias
    - Only a subset of data was analyzed due to resource limits, this introduced sampling bias, affecting the representativeness of these results.
4. Reduced Feature Engineering
    - Limited computational resources constrained the ability to explore and engineer additional features that could improve model performance.

Key Findings

1. **K-Means Clustering**



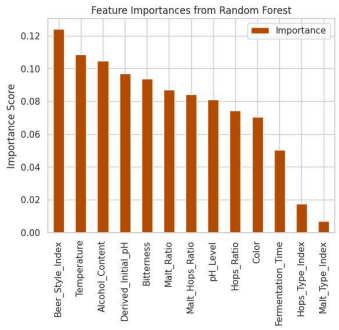
Evaluation Metrics:

- Inertia Score: 19196114.00
- Silhouette Score: 0.2700

Observation (Inertia Score): High inertia score indicates that the clusters are not well-separated, suggesting that the data points within clusters are relatively far from their centroids.

Observation: (Silhouette) A silhouette score of 0.2700 suggests that the clustering may not be well-defined, indicating some overlap between clusters and potential room for improvement in feature selection or clustering parameters.

2. **Random Forest**

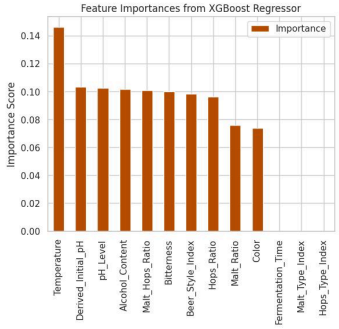


Evaluation Metrics:

- R^2 Score: 1.7314027253467756e-05

Observation: The very low R^2 score suggests poor predictive power, indicating that the model is not capturing the variance in the data well. The feature importance analysis highlights critical factors, but the overall model performance needs further tuning.

3. **Gradient Boosted**



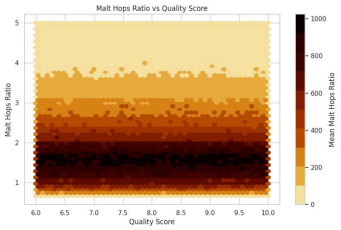
Evaluation Metrics:

- R^2 Score: 6.48369635813939e-06

Observation: Similar to Random Forest, the low R^2 score indicates that the model is not effectively predicting beer quality. The critical features identified provide valuable insights, but further model refinement and feature engineering are necessary.

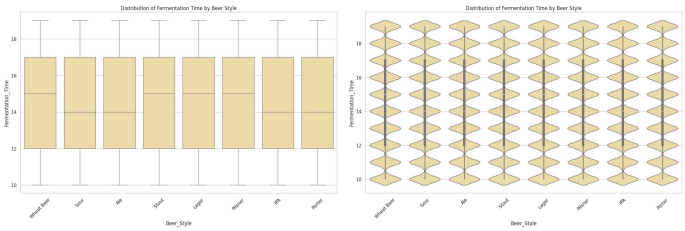
Additional Plots:

4. **Hexbin Plots**



Observation: The Malt-to-Hops ratio shows little correlation with quality scores, suggesting that other factors may play a more significant role in determining quality. This warrants further investigation into other feature combinations.

5. **Box and Violin Plots**



Observation: Consistent median fermentation times across beer styles suggest standardization in the brewing process, while the presence of outliers indicates specific conditions or variations that may impact quality. Exploring these outliers could reveal insights into optimizing the brewing process. ## Actionable Insights

1. Refine Malt-to-Hops ratios.
2. Validate machine learning models.
3. Leverage clustering for beer style differentiation.
4. Analyze malt and hops types combinations.

Next Steps

These steps aim to address the core issues of resource constraints and dataset size, helping to continue this analysis more efficiently and effectively in the next phase of this project.

1. Optimize Resource Allocation:
   - Explore using specialized cloud services like Google Cloud Dataproc for large-scale Spark jobs.
   - Consider AWS Sagemaker for model training to take advantage of managed machine learning services.
   - Investigate high-performance computing (HPC) clusters or collaborate with research institutions for advanced computational resources.
2. Leverage Spark:
   - Continue refining Apache Spark for in-memory processing and distributed computation.
   - Focus on reducing dependency on Pandas to improve processing speed and efficiency.
3. Expand Feature Engineering:
   - Generate a diverse array of features within Spark.

- Aim to maximize dataset utility and model accuracy without the need for data format conversion.

4. Plan for Scalability:
   - Design all solutions with scalability in mind.

README

Author

Name: Erin Wasserman

GitHub: Cellister

Email address: cellister at gmail .com

Repository Structure

- **Notebook**

The Google Colab Notebook, is the key deliverable and contains the details of my data strategy, methodology, data cleaning, visualizations, and actionable insights. notebook PDF

- **Presentation**

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

- **Jupyter Notebook** 100.0%