Presented By:

Paridhi Talwar

Brewery Operations and Market Analysis

LIST OF CONTENTS

INTRODUCTION

The craft beer industry is characterized by its focus on quality and innovation, making it essential for brewers to optimize production processes while maintaining high-quality standards. This project leverages machine learning techniques to enhance the quality control and efficiency of craft beer production. By analyzing comprehensive brewing data, we aim to identify key factors influencing quality and efficiency, enabling brewers to make data-driven decisions and improve their operations.

DATASET

The dataset used in this project is titled "Comprehensive Analysis of Brewing Parameters, Sales Trends, and Quality Metrics in Craft Beer Production (2020-2024)." It includes a rich set of features, such as brewing parameters, ingredient ratios, production volumes, sales trends, and quality scores, collected over five years. This data provides a holistic view of the brewing process and its outcomes, allowing for in-depth analysis and modeling.

PREPROCESSING

- Handling Missing Values
- Duplicate Check
- Data Transformation

FEATURE ENGINEERING

- Ingredient Ratio Splitting: The Ingredient_Ratio column was split into separate components, allowing for detailed analysis of individual ingredient contributions.
- Brew Ratio Calculation: Calculated the brew ratio to assess the balance of ingredients used in each batch.
- Sales Efficiency: Computed sales efficiency as the ratio of total sales to volume produced, providing insights into production effectiveness.

EXPLORATORY DATA ANALYSIS

- Frequency Distribution
- Summary Statistics
- Trend Analysis
- Rolling Statistics
- Correlation Analysis
- Control Limits

Number of distinct values in 'beer_style': 8
Frequency distribution for 'beer_style':

beer_style	count
Ale	1251002
Porter	1250773
Sour	1250307
Stout	1250296
IPA	1249603
Lager	1249570
Pilsner	1249426
Wheat Beer	1249023

summary	fermentation_time	temperature	pH_Level	gravity
count	10000000	10000000	10000000	10000000
mean	14.500898	19.999898511019087	4.999940543893469	1.0550028700788614
stddev	2.8720060965182497	2.887029712032824	0.28863762894103623	0.014434649211836932
min	10	15.000001163771435	4.500000005935603	1.0300000027891478
max	19	24.999998289887966	5.499999818305633	1.0799999980323736

Brew_Date	sum(total_sales)
2020-01-01 00:00:19	2664.7593448382822
2020-01-01 00:00:31	9758.801062471319
2020-01-01 00:00:40	11721.087016274963
2020-01-01 00:01:37	12050.177463190277
2020-01-01 00:01:43	5515.0774647529615
2020-01-01 00:01:48	6278.389850288936
2020-01-01 00:01:49	14362.653665879505
2020-01-01 00:01:51	1082.3549117830858
2020-01-01 00:02:16	7392.644809330611
2020-01-01 00:02:32	7648.2830929155325
2020-01-01 00:02:36	10634.431632453308
2020-01-01 00:02:55	11852.097970324789
2020-01-01 00:03:04	11651.628703689214
2020-01-01 00:03:09	1665.1818593163712
2020-01-01 00:03:12	17473.8825235854
2020-01-01 00:03:17	2557.880022111579
2020-01-01 00:03:22	13527.919120822819
2020-01-01 00:03:26	6824.2626567713705
2020-01-01 00:03:31	5925.879914754564
2020-01-01 00:03:33	6972.8499301042675

Brew_Date	total_sales	rolling_mean_sales	rolling_std_sales
2020-01-01 00:00:19	2664.7593448382822	2664.7593448382822	null
2020-01-01 00:00:31	9758.801062471319	6211.7802036548	5016.245004558584
2020-01-01 00:00:40	11721.087016274963	8048.215807861521	4764.330399759633
2020-01-01 00:01:37	12050.177463190277	9048.70622169371	4374.527065264268
2020-01-01 00:01:43	5515.0774647529615	8341.98047030556	4104.835163794392

Correlation between ingredient2 and ingredient3: -0.0002594275209472055
Correlation between ingredient2 and Brewhouse_Efficiency: -0.00032305355560068163
Correlation between ingredient3 and Brewhouse_Efficiency: 0.00023379403558401138

Quality Score Control Limits:
Upper Control Limit: 7.507318879418029
Lower Control Limit: 7.492302400577395

Brewhouse Efficiency Control Limits:
Upper Control Limit: 80.0423767254251
Lower Control Limit: 79.9594585796965

MODEL IMPLEMENTATION

- Data Splitting: Divided the dataset into training and testing sets to evaluate model performance.
- Feature Selection: Identified relevant features that contribute significantly to the prediction of quality scores.
- Model Training: Trained the Random Forest model, leveraging its ability to handle complex, nonlinear relationships within the data.
- Model Evaluation: Calculated RMSE and MAE values.

CLUSTERING

```
Silhouette with squared euclidean distance = 0.7526
Cluster Centers:
[1.45007875e+01 1.99995142e+01 5.00008955e+00 1.05500222e+00
 1.88981216e+00 3.92189562e+00 5.24941715e+00 3.94940786e+01
 1.19989311e+01 2.74623885e+03 3.00140315e+00 3.00049181e+00
 2.99950687e+00 3.49907948e-01 2.00026483e-01 8.00017045e+01]
[1.45005756e+01 2.00002963e+01 4.99979909e+00 1.05500992e+00
 1.88991460e+00 2.49806246e+00 5.25008621e+00 3.94952642e+01
 1.19988859e+01 4.24768663e+03 2.99906394e+00 2.99975036e+00
 3.00068498e+00 3.49930838e-01 2.00011701e-01 8.00031377e+01]
[1.45013322e+01 1.99998840e+01 4.99993334e+00 1.05499645e+00
 1.88957850e+00 9.71490755e+00 5.24962263e+00 3.94992612e+01
 1.20002226e+01 1.24797001e+03 2.99977968e+00 2.99976548e+00
 3.00028307e+00 3.50045352e-01 1.99974015e-01 7.99979074e+01]
```

RESULTS

```
Test RMSE after cross-validation: 1.1182
Test MAE after cross-validation: 1.0002
```

quality_score	prediction
7	7.500510685424583
8	7.498263186830121
9	7.501571453117759
8	7.501333654983706
9	7.503690781533601
9	7.501796059108134
7	7.5007777243879685
9	7.500596589534146
8	7.502048945497348
8	7.502096544064179

```
Feature Importances:
fermentation_time: 0.0326
temperature: 0.0611
pH_Level: 0.0909
gravity: 0.0520
brew_ratio: 0.0511
sales_efficiency: 0.0707
Alcohol_Content: 0.0642
Bitterness: 0.0762
Color: 0.0423
Volume_Produced: 0.0600
Loss_During_Brewing: 0.0682
Loss_During_Fermentation: 0.0809
Loss_During_Bottling_Kegging: 0.0806
ingredient2: 0.0580
ingredient3: 0.0490
Brewhouse Efficiency: 0.0623
```

CONCLUSION

The model achieved a test Root Mean Square Error (RMSE) of 1.1182 and a Mean Absolute Error (MAE) of 1.0002, indicating a high level of accuracy in predicting quality scores. Feature importance analysis highlighted key drivers such as pH level, fermentation time, and sales efficiency. KMeans clustering was also applied, successfully segmenting the data into distinct groups with unique characteristics, as evaluated by the Silhouette score.

THANK YOU!!!