VIVINO MARKETING STRATEGY

Presented By -

Avadhoot jadhav

Ta-Chin (Meek) Ho

Alefiya Lunawadawala

Mark Nevelle Antony

INTRODUCTION

Brief overview

Our project focuses on analyzing a dataset of 1599 variants of the French "Bordeaux" red wine to determine the physicochemical properties that define a "good" quality wine.

Objectives of the analysis

Develop models to predict wine quality based on physicochemical properties Identify top three discriminating physicochemical properties of wine quality

METHODOLOGY

Overview of the analytical approach

Developed two classifier models, the Naive Bayes Model and the Random Forest Model, for comparison.

Description of the dataset and variables analyzed

The dataset includes 1599 observations of French "Bordeaux" red wine variants. It comprises 11 physicochemical variables, such as acidity levels and alcohol content, and a sensory variable indicating wine quality.

SUMMARY STATISTICS

> summary(vivino.df)

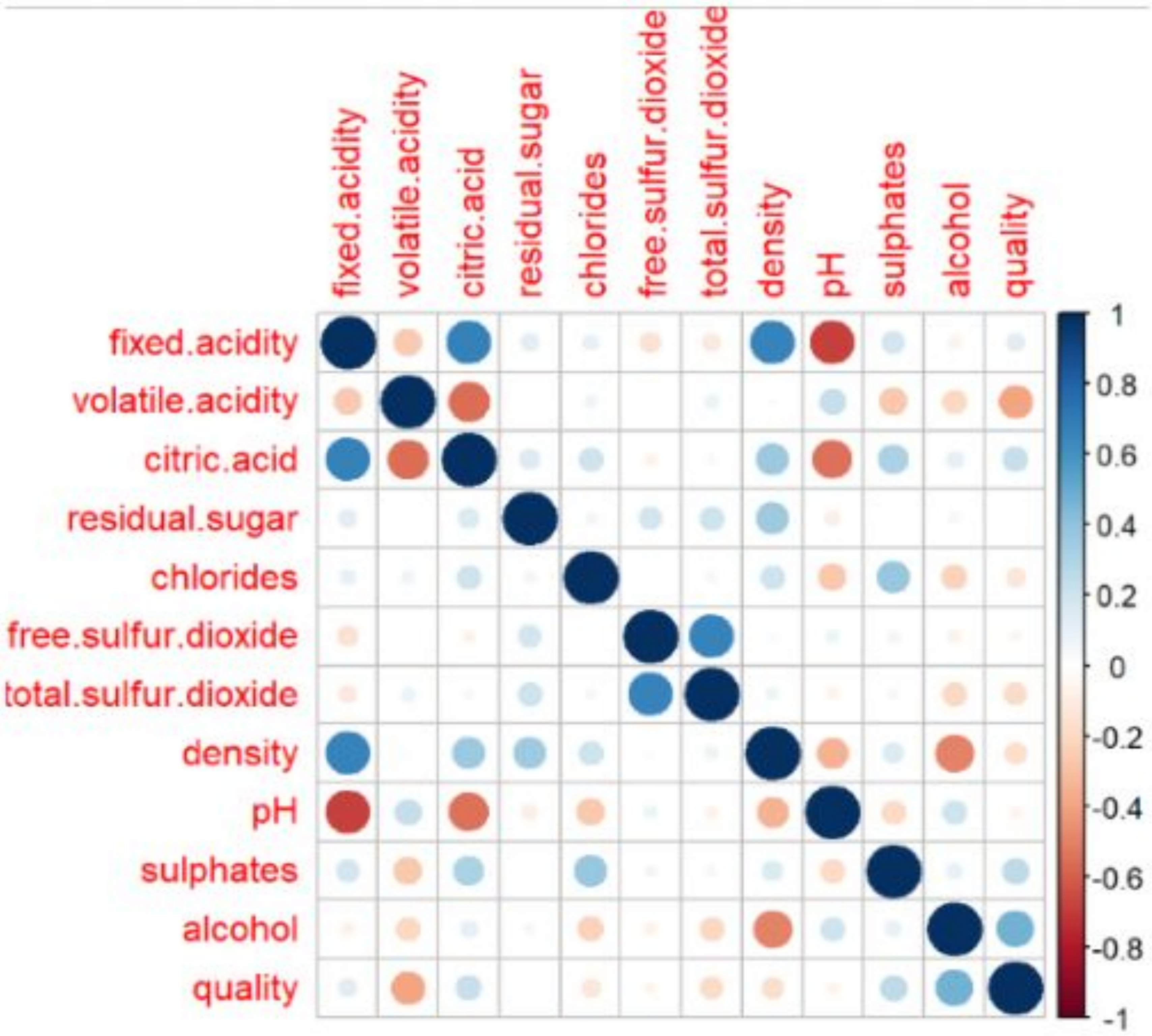
```
 fixed.acidity   volatile.acidity  citric.acid    residual.sugar     chlorides
 Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900   Min.   :0.01200
 1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900   1st Qu.:0.07000
 Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200   Median :0.07900
 Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539   Mean   :0.08747
 3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600   3rd Qu.:0.09000
 Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500   Max.   :0.61100
 free.sulfur.dioxide total.sulfur.dioxide    density             pH          sulphates
 Min.   : 1.00       Min.   :  6.00       Min.   :0.9901   Min.   :2.740   Min.   :0.3300
 1st Qu.: 7.00       1st Qu.: 22.00       1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500
 Median :14.00       Median : 38.00       Median :0.9968   Median :3.310   Median :0.6200
 Mean   :15.87       Mean   : 46.47       Mean   :0.9967   Mean   :3.311   Mean   :0.6581
 3rd Qu.:21.00       3rd Qu.: 62.00       3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300
 Max.   :72.00       Max.   :289.00       Max.   :1.0037   Max.   :4.010   Max.   :2.0000
    alcohol          quality      qualityLabel
 Min.   : 8.40   Min.   :3.000   0:1382
 1st Qu.: 9.50   1st Qu.:5.000   1: 217
 Median :10.20   Median :6.000
 Mean   :10.42   Mean   :5.636
 3rd Qu.:11.10   3rd Qu.:6.000
 Max.   :14.90   Max.   :8.000
```

Adding a Wine Quality Indicator Column

```
> vivino.df$qualityLabel <- as.factor(ifelse(vivino.df$quality > 6, 1, 0))
```

```
[1] "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "good" "good" "bad"  "bad"  "bad"  "bad"
  [14] "bad"  "bad"  "bad"  "good" "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"
  [27] "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "good" "bad"
  [40] "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"
  [53] "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "good" "bad"  "bad"
  [66] "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"
  [79] "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"
  [92] "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"
 [105] "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"
 [118] "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "good" "bad"
 [131] "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"
 [144] "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"
 [157] "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"  "bad"
```

DATA VISUALIZATION



The term "relatively highly correlated" implies a positive relationship, where one variable's increase corresponds to the other's. Three pairs stand out:

- **Fixed acidity and citric acidity**
- **Density and fixed acidity**
- **Total sulfur dioxide and free sulfur dioxide**

PREDICTING ‘GOOD’ WINE QUALITY

This study seeks to improve data collection and analysis to predict wine quality using physicochemical properties. We will apply and compare the Naive Bayes and Random Forest models to assess wine quality. Additionally, we will determine the three most impactful physicochemical properties affecting wine quality.

PREDICTING ‘GOOD’ WINE QUALITY – Naive Bayes model

- **Naive Bayes model**

```
vivino_nb <- naiveBayes(qualityLabel ~ ., data = vivino.df)
vivino_nb
# Model Evaluation
nb_predictions <- predict(vivino_nb, vivino.df)
nb_accuracy <- mean(nb_predictions == vivino.df$qualityLabel)
nb_confusion_matrix <- table(nb_predictions, vivino.df$qualityLabel)
nb_predictions
print("Naive Bayes Model:")
nb_accuracy
Nb_confusion_matrix
```

Result :

Accuracy

[1] 0.9937461

Confusion Matrix

bad 1382 10

good 0 207

NAIVE BAYES MODEL RESULTS

- Model Choice: Naive Bayes selected for its simplicity and effectiveness in classifying wines based on physicochemical properties.
- Performance: Achieved a high accuracy of approximately 99.37%.
- Classification: Successfully distinguished between 'good' and 'bad' wines, making only 10 misclassifications out of 1392 'bad' wines and correctly classifying all 207 'good' wines.
- Conclusion: The Naive Bayes model performed exceptionally well, demonstrating its suitability for predicting wine quality based on physicochemical properties.

PREDICTING 'GOOD' WINE QUALITY – Random Forest Model

- **Random Forest Model**

```
set.seed(123)
vivino_rf <- randomForest(qualityLabel ~ ., data = vivino.df, ntree = 500)
vivino_rf
rf_predictions <- predict(vivino_rf, vivino.df)
rf_accuracy <- mean(rf_predictions == vivino.df$qualityLabel)
rf_confusion_matrix <- table(rf_predictions, vivino.df$qualityLabel)
print("Random Forest Model:")
rf_accuracy
rf_confusion_matrix
```

Result:

Accuracy

[1] 1

Confusion Matrix

0	**1382**	**0**
1	**0**	**217**

RANDOM FOREST MODEL RESULTS

- Model Choice: Random Forest selected for its capability to manage complex datasets and its resilience against overfitting.

- Performance: Achieved a remarkable accuracy of 100%.

- Classification: Successfully classified all samples into 'bad' (1382) or 'good' (217) quality wines with no misclassifications.

- Conclusion: The Random Forest model demonstrated outstanding performance, achieving perfect accuracy and effectively classifying wines based on their physicochemical properties.

COMPARISON

Naive Bayes	Random Forest
Advantages: Quick, simple, and ideal for initial experiments.	Advantages: More complex and robust, offering better accuracy.
Performance: Accuracy of approximately 99.37%, with few mistakes in classification.	Performance: Achieved perfect 100% accuracy, showing effectiveness in handling complexity.
Suitability: Best for datasets with few features and when feature independence holds.	Suitability: Suitable for complex datasets with many features, less prone to overfitting.

PREDICTING 'GOOD' WINE QUALITY

We seeks to improve data collection and analysis to predict wine quality using physicochemical properties. We will apply and compare the Naive Bayes and Random Forest models to assess wine quality. Additionally, we will determine the three most impactful physicochemical properties affecting wine quality.

Top 3 factors of Wine Quality – Method 1

- **Feature Importance Display**

```
# Random Forest Classifier
randomForest(qualityLabel ~ ., data = vivino.df)$importance[, "MeanDecreaseGini"]
```

Result :

```
> randomForest(qualityLabel ~ ., data = vivino.df)$importance[, "MeanDecreaseGini"]
      fixed.acidity    volatile.acidity     citric.acid   residual.sugar
         6.016176         13.967547          8.537277        4.164854
        chlorides     free.sulfur.dioxide total.sulfur.dioxide     density
         4.932182          3.803786          6.394951          9.574551
            pH            sulphates           alcohol            quality
         3.727846         14.321793         24.658558        274.336071
```

FEATURE IMPORTANCE DISPLAY

Process:

- It calculate the importance of features in a Random Forest model trained using the Vivino data. (QualityLabel target variable)
- Utilize the **"Mean Decrease Gini"** measure to quantify the importance of each feature.

Purpose:

- Calculate and display the importance of each feature.
- Provide insights into which features impact the model's predictions and decision-making.

Top 3 factors of Wine Quality – Method 2

- **Feature Importance Calculation with names**

```
feature_importance <- randomForest(qualityLabel ~ ., data = vivino.df)$importance
[, "MeanDecreaseGini"]
top_features <- names(sort(feature_importance, decreasing = TRUE))
top_features
```

Result :

```
> top_features <- names(sort(feature_importance, decreasing = TRUE))
> top_features
[1] "quality"        "alcohol"        "sulphates"        "volatile.acidity"
[5] "density"        "citric.acid"     "total.sulfur.dioxide" "chlorides"
[9] "fixed.acidity"   "residual.sugar"   "free.sulfur.dioxide"  "pH"
```

FEATURE IMPORTANCE CALCULATION WITH NAMES

Process:

- Train a Random Forest model using the qualityLabel target variable and all other features in vivino.df.
- Calculate feature importance using the **"Mean Decrease Gini"** measure from the trained model.
- Rank the features by “**importance”** and **return the names of the most important ones.**

Purpose:

- Identify the top features that most influence wine quality.
- Assist in understanding which features contribute significantly to the model's decision-making process.

Top 3 factors of Wine Quality – Method 3

- **Random Forest Model Training and Importance**

```
set.seed(123)
vivino_rf <- randomForest(qualityLabel ~ ., data = vivino.df, ntree = 500)
importance(vivino_rf)
```

Result :

```
> set.seed(123)
> vivino_rf <- randomForest(qualityLabel ~ .,
data = vivino.df, ntree = 500)
> importance(vivino_rf)
```

	MeanDecreaseGini
fixed.acidity	5.686958
volatile.acidity	14.108792
citric.acid	8.602808
residual.sugar	3.918395
chlorides	5.327812
free.sulfur.dioxide	3.638993
total.sulfur.dioxide	6.080499
density	8.903310
pH	3.804800
sulphates	***15.105785***
alcohol	***26.365500***
quality	***273.078107***

RANDOM FOREST MODEL TRAINING AND IMPORTANCE

Process:

- Sets a seed for random number generation to ensure reproducibility of results.
- Trains a Random Forest model using the dataset vivino.df with qualityLabel as the target variable and 500 trees (ntree = 500).
- After training, it retrieves the importance of each feature in the model.

Purpose:

- It provides a trained model and its feature importance for predicting qualityLabel.
- The model trained with 500 trees may be slightly different from previous 2 methods in terms of randomness and model complexity.

DIFFERENCE BETWEEN METHODS

Similarities:

- All three methods use Random Forest to calculate feature importance.
- Identify the top features that influence wine quality successfully.

Differences:

- Method 1 and Method 2 directly calculate feature importance after training the Random Forest model, but method 2 contains the names of the features that have the highest importance in predicting wine quality.

- Method 3 involves training a Random Forest model with a specified seed and number of trees before calculating feature importance.

Identify The Top 3 Factors

- The top three factors influencing wine quality are **alcohol content, sulphates, and overall quality rating.**

- While there might be slight variations in results due to randomness in training and the number of trees used, these factors consistently stand out as the most important for predicting wine quality by using Random Forest model.

CONCLUSION - PREDICTING ‘GOOD’ WINE QUALITY

The Random Forest and Naive Bayes models achieved high accuracy in classifying wines as 'bad' or 'good'.

Random Forest achieved 100% accuracy, while Naive Bayes reached 99.37%.

Based on the result of most of the wine (1382) belong in “bad” quality, we provide following solution :

- **Quality Improvement:** Wineries can analyze factors like grapes, fermentation, and storage to improve wine quality.
- **Market Segmentation:** Retailers can discount or promote 'bad' wines differently to manage consumer expectations.
- **Customer Education:** Educate consumers about wine faults to reduce dissatisfaction.
- **Recycling or Repurposing:** Repurpose 'bad' wines for cooking or other products.

CONCLUSION – Top 3 Factors

The top 3 discriminating factors of wine quality based on importance in the Random Forest model are **alcohol, sulphates, and quality.**

These features have the highest importance values, indicating the model's decision-making process.

Vivino should focus on the top 3 discriminating factor of wine quality (alcohol, sulphates, quality) to improve their product and grow their business. Therefore, we provide following solution :

- **Quality Verification Service:** Offer a premium service where Vivino verifies the alcohol content, sulphate levels, and overall quality of wines before they are listed on the platform. This would ensure that users can trust the authenticity and quality of the wines.
- **Collaborations with Wineries:** Partner with wineries that produce high-quality wines with optimal alcohol content and sulphate levels. Vivino could feature these wines on the platform and offer exclusive deals to users to attract more customers and driving sales.
- **Customized Wine Clubs:** Create customized wine clubs where members receive monthly shipments of wines selected based on their preferences for alcohol content, sulphate levels, and overall quality.

Thank you !