WHAT MAKES A GOOD BOTTLE OF WINE?

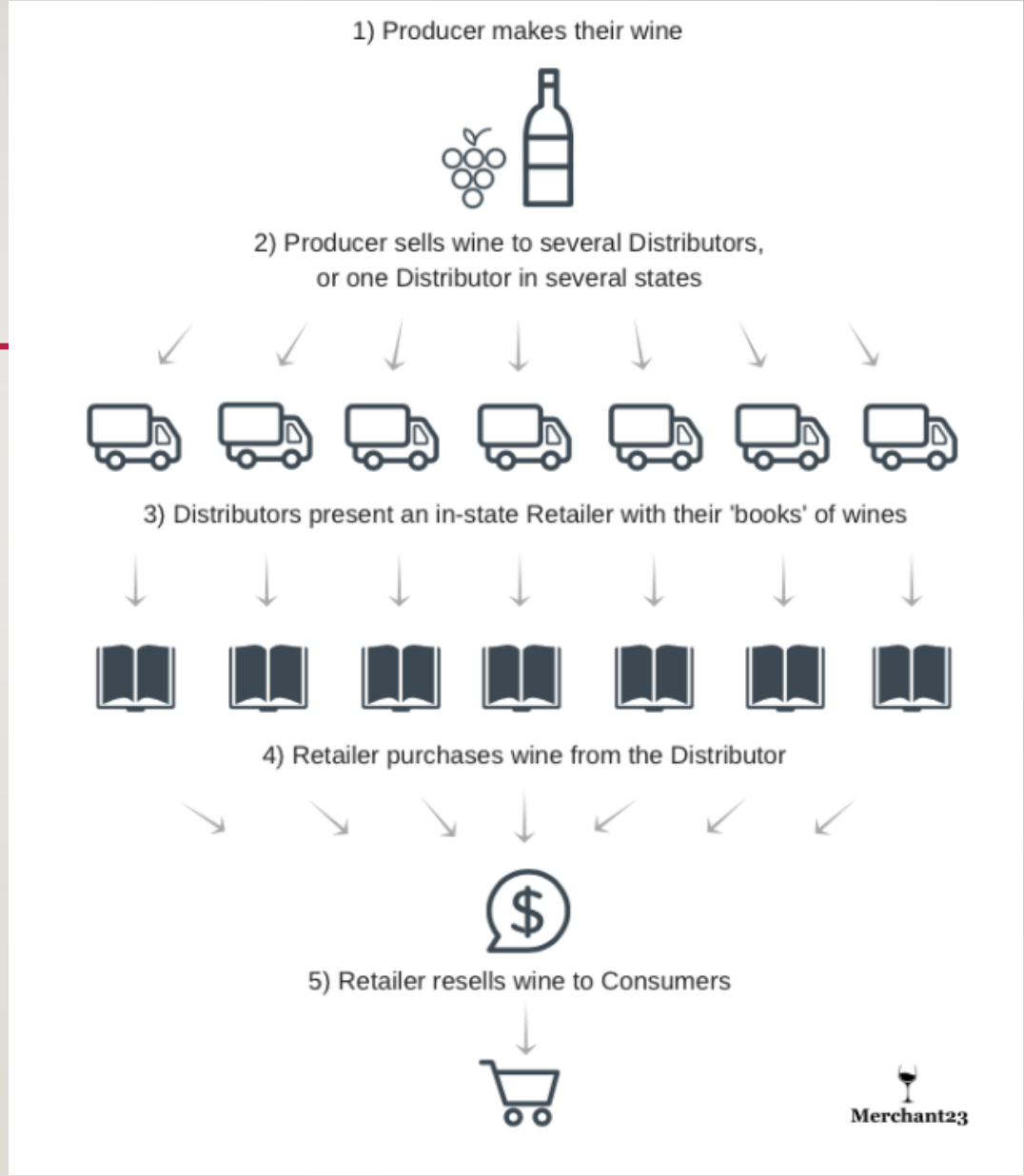
GENERAL ASSEMBLY: DATA SCIENCE REMOTE FINAL PROJECT PRESENTATION

KEY TAKEAWAYS:

- A model to predict wine quality from its chemical properties:
  - Random Forest Classifier Model.
  - An accuracy score of about 70%.
  - 26% higher accuracy score than a null model.
  - Used 100% of the data to train it.
  - Accuracy was calculated with a cross-validation score.
- The accuracy score can be further increased by collecting more data to train the model.

PROBLEM:

- In the wine business, there are two sets of re-sellers between a wine producer and the consumer: distributors and retailers.
- Both distributors and retailers—looking to maximize profits and minimize losses—must purchase inventory that sells well.
- So how can they tell what is a good quality wine and what isn't?



http://blog.merchant23.com/wholesale-wine-how-to-buy-winery-direct

THE DATA:

- Two Wine Quality Datasets were obtained through the following research:

  P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties.
  In *Decision Support Systems*, Elsevier, 47(4):547-553, 2009.

- The first contains information about Red Wine varieties of Vinho Verde wine and the second about White Wine varieties of Vinho Verde.
- This wine is produced in the region of Vinho Verde in northwestern Portugal and is known for its freshness, and fruity and floral notes.

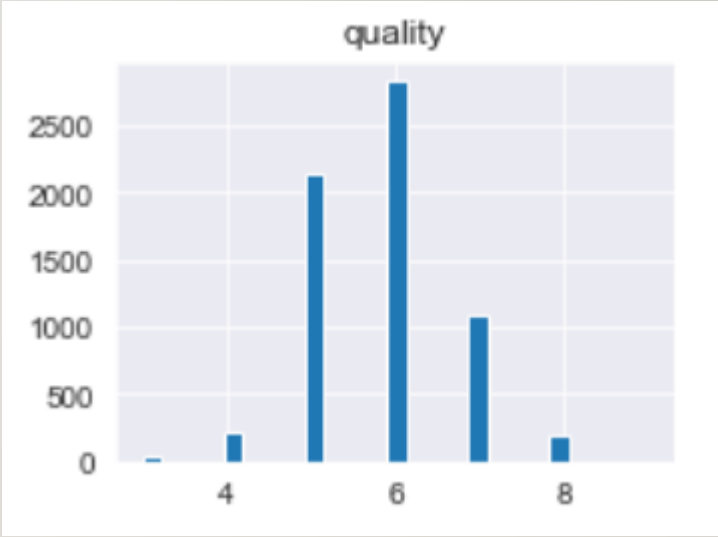
THE DATA:

- Both datasets contained 11 columns with information about the chemistry of each type of Vinho Verde and 1 column providing a quality score for it.
- The quality score is a discrete or categorical variable and is the only integer value column in the dataset. The rest are decimal values.
- The Red Wine and White Wine datasets were combined so as to have the most data possible to train the model with.
- 6497 data points total.

```
fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol
quality
```

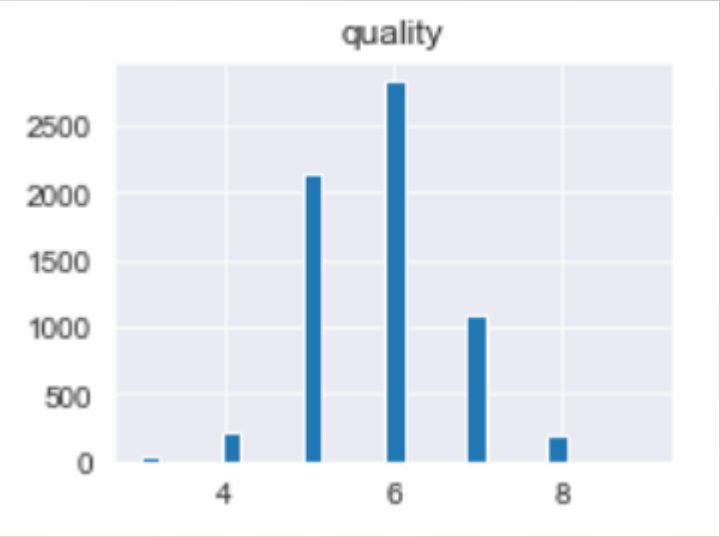
ASSUMPTIONS:

- Since the two datasets were combined, the project assumes that both Red Wine and White Wine quality is based on the same chemical properties, or that any differences are negligible.
- The quality scores go from a low of 3 to a high of 9 in this particular dataset and these were assumed to be the lowest and highest quality scores possible for the purposes of this project.



METRICS:

- Goal: Build a model to predict the quality score of a wine from its chemical properties.
- Initial metric for success:
  - Build a model that is more predictive than simply guessing the most common quality score (6) every time.
  - A null model that predicts a quality score of 6 every time would be accurate 44% of the time.
- Once this metric is met, the goal is to tune the model to be as accurate as possible.



PREPARING THE DATA:

- This was a very clean dataset:
  - No missing values.
  - No null values.
  - All columns were numeric and no categorical variables needed to be transformed.
- The two datasets shared the same columns and were combined into a larger dataset for this project.
- A preliminary exploration of the data did not show any strong correlations between any of the columns and the quality score.

APPROACH AND PROCESS:

- A first-pass linear regression model did very poorly with a 26% accuracy score.
- This is likely because none of the variables are very strongly correlated with the quality score.
- Improving the model:
  - Switching to a classification model because the quality score is a discrete/integer variable.
  - Using a more complex and flexible model to better fit the data: Random Forest Classifier.
- An initial Random Forest Classifier model had a 98% accuracy on the training dataset and a 65-67% accuracy on the test data set.

IMPROVING THE MODEL:

- Tuning the parameters:
  - n_estimators = 100
  - max_features = 1
  - min_samples_leaf = 1
  - max_depth = 140
- Feature engineering did not increase accuracy.
- Training the model on 100% of the data.

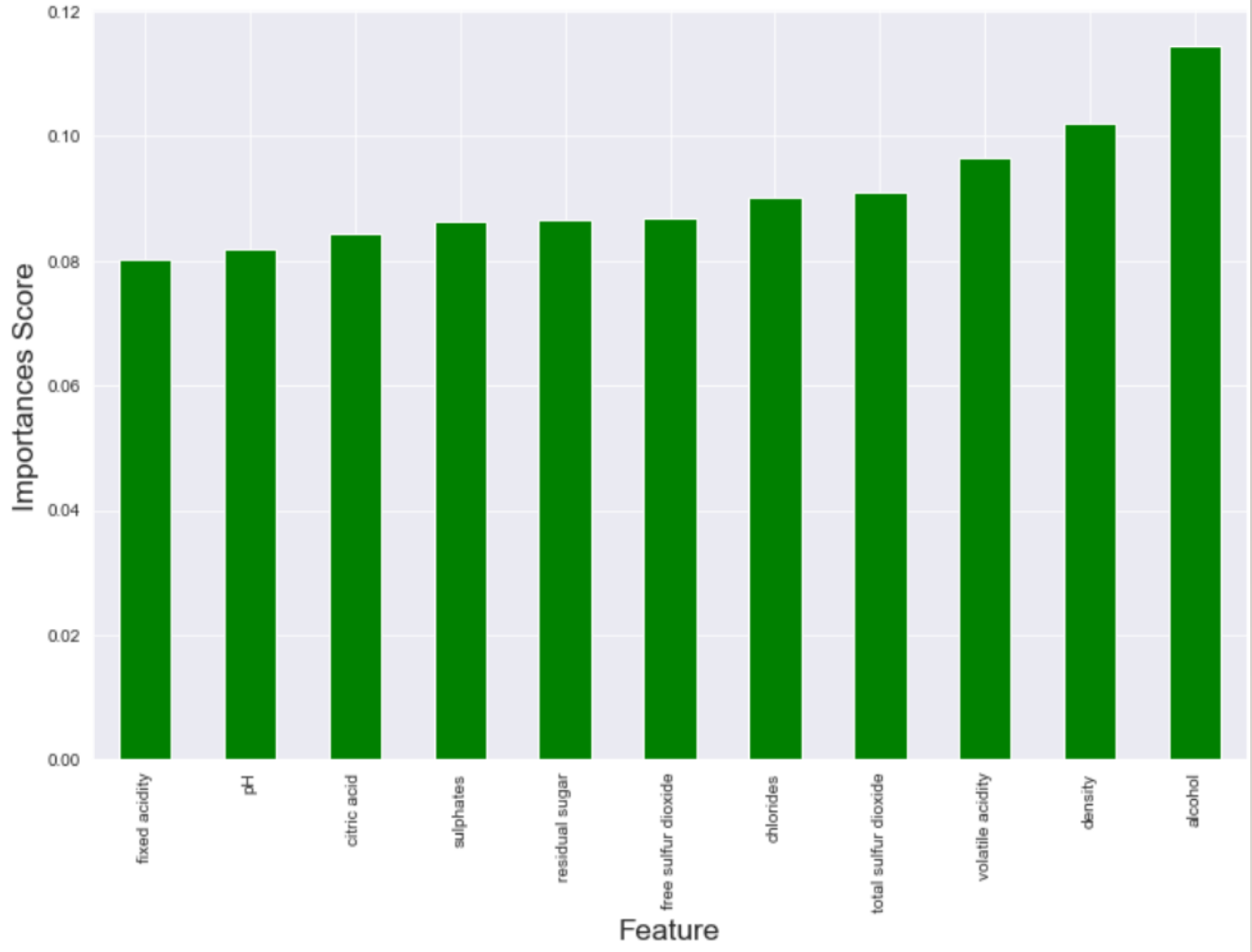
THE FINAL MODEL:

- OOB Score = 70%
- Cross-validation Score = 69-70%

```
# Splitting columns into target and features
target_col = 'quality'
feature_cols = df_all.columns.drop('quality')  # using all features optimized the accuracy scores

# Creating a dataframe of features (X) and a series of target values (y)
X = df_all.loc[:,feature_cols]
y = df_all.loc[:,target_col]

rfc_final = RandomForestClassifier(n_estimators = 100, max_features = 1, min_samples_leaf = 1, max_depth = 140, oob_score = True)
rfc_final.fit(X, y)

# Evaluating the model
print('Full Dataset Accuracy:', rfc_final.score(X, y))
print('Full Dataset Out of Bag Error Score:', rfc_final.oob_score_)

# Doublechecking model performance with a 5Fold Cross-Validation Score
scores = cross_val_score(rfc_final, X, y, cv = kf)
print('Cross-Validation With 10-Fold Split Scores:', scores)
print('Cross-Validation With 10-Fold Split Average Score:', scores.mean())
```

IMPACT AND NEXT STEPS:

- Distributors and retailers can predict with 70% accuracy the quality score of a Vinho Verde wine from its chemical properties.
- The most important chemical properties seem to be alcohol content and density.
- More data must be collected to further improve the model.



THE END