Melissa Morgan

Blake Skinner

Godwin Thomas

Julia Thompson

Rebekah Vinson

THE EFFECT OF PHYSICOCHEMICAL ON THE WINE QUALITY

Why **Wine?**

the vino, the grape, plonk, vin du pays, port, burgundy, vin de table, vin ordinaire, walla walla, splishy splashy, slap the belly

when wine
goes in
wisdom
comes out

- It's data science

the

Data Dorks

can help out the

Cork Dorks

PROJECT DATA

THE SOURCE

Paulo Cortez, University of Minho, Guimarães, Portugal

THE DATA

Our two datasets include:

- Sample set of approximately 1,600 red and 4,800 white wines
- 12 numerical attributes / features for each sample wine

OBJECTIVE

- Develop a machine learning model using the chemical makeup of wine to predict its quality.

Data Attributes

Input variables (based on physicochemical tests):

1 - fixed acidity
2 - volatile acidity
3 - citric acid
4 - residual sugar
5 - chlorides
6 - free sulfur dioxide
7 - total sulfur dioxide
8 - density
9 - pH
10 - sulphates
11 - alcohol

Output variable (based on sensory data):

12 - quality (score between 0 and 10)

MACHINE LEARNING

Data Exploration

- **Pulled in data**
  - *Separate csv files for red and white wine*
- **Looked at the datatypes**
  - *All floats except for quality*
  - *Changed datatype to float for quality*
- **Looked at data shape**
  - *Quality not evenly distributed amongst classifications*
  - *Skewed heavily to quality of 5 & 6*
  - *Understand that this is heavily going to affect our model*

quality	
3	20
4	163
5	1457
6	2198
7	880
8	175
9	5

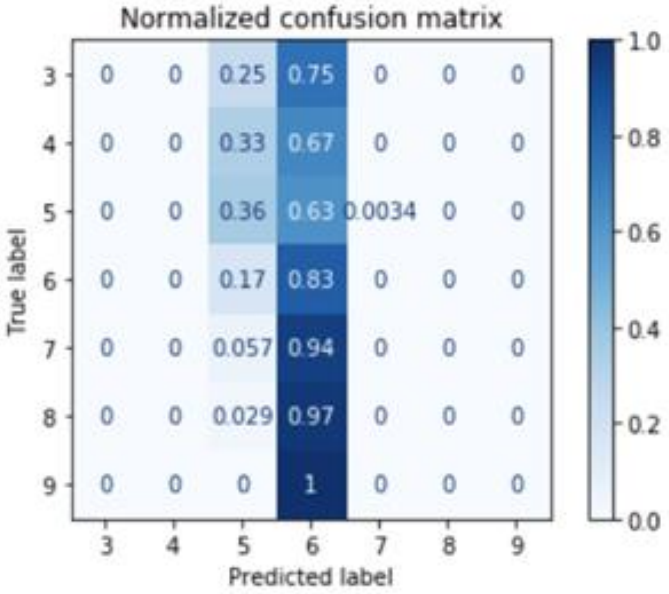
The Models

Model	Accuracy
RandomForestClassifier	0.707142857
XGBoost	0.66122449
DecisionTreeClassifier	0.582653061
KNeighborsClassifier	0.558035714
LogisticRegression	0.481632653
NaivesBayes	0.448341837

- Created x and y variables
  - *output (y) = quality*
  - *x = other features*
- Train test split from x and y variables
- Ran multiple models to test and compare accuracies
  - *linear regression r2 score = 0.281361*

The Models Continued

- Confusion matrix to see where there were inaccuracies - shows that model chose 6 far too often



- Tried two different types of data scaling – feature scaling and PCA. Neither helped the shape of our data enough to impact our model.
- Tried two different types of hypertuning to improve our accuracy. Randomized Search & GridSearch. When the optimal params from those searches were added to the model – accuracy decreased.
- Used pickle to save our model down into a callable sav file

LIVE DEMONSTRATION

Project Final Materials & Demonstration

Project Link

bc633.pythonanywhere.com

- Route to different html files using render_template.
- Includes the home page, imbedded tableau visualization, and site for users to submit data and receive a prediction on wine quality.
- The route doShit() retrieves the input from the user, and runs the prediction with the ML.

ADDITIONAL VISUALIZATIONS

Tableau Interactive Dashboards and Visualizations

https://public.tableau.com/views/WineProject

https://public.tableau.com/views/RandomForestClassifier-WineAttributeImportance

DATA SCIENCE HAPPENS… WINE HELPS

ANY QUESTIONS?