Modeling Beer Consumption in the USA

What features contribute to beer consumption patterns and can we forecast those patterns?

Introduction

- Goal: Predict a state’s beer consumption in a given year
- Backup goal: Predict a state’s beer production in a given year
  - Consumption would be better for craft breweries looking to expand business, however production is a potentially valuable indicator in a different way
- Process: use machine learning, make predictions, and compare our predictions to our actual results statistically and visually
- Result: models that can explain consumption and production when certain criteria is met

Data

Data Continued - Model Building Variables

- Feature variables were hand selected and hand compiled into a CSV
- Years included: 2010-2021
- All 50 states and Washington, D.C.
- Feature variables included (31 total)
  - Demographic: sex, age, education distributions
  - Economic: employment, income, disposable income, excise tax rates
  - Social: binge drinking rates
  - Geographic: known high tourist/border purchasing states (AK, DC, DE, HI, NH, NV)
- Target variables examined
  - Adult per capita beer consumption
  - Adult per capita (craft) beer production

Diverse Data Sources

- Key data sources
  - Industry
    - Brewers Association
  - Think Tank
    - Tax Policy Center
  - Government
    - US Department of Commerce - US Bureau of Economic Analysis
    - US Census Bureau - Annual American Community Survey
    - National Institutes of Health - National Institute of Alcohol Abuse and Alcoholism
    - Centers for Disease Control and Prevention - Behavioral Risk Factor Surveillance System

Modeling

Where do we begin?

- Ideally, we wanted to use a neural network
  - Robustness
  - Optimizing strategies in place
  - Understood the best (at least for me)
- Ran into issues
  - How do we measure accuracy?
  - Issues with getting it to work at all
  - So much needed to be changed, essentially had to start my code over
- But I wasn’t the only one working on a model…

Gradient boosting Model

Gradient boosting is a machine learning technique used for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. Here's how it works:

Initialization: It starts with a base model (often a simple average of the target variable) to make initial predictions.

Iterative Improvement:

- For each iteration, the algorithm computes the residual, which is the difference between the observed and predicted values from the current model.
- A new model (usually a decision tree) is then trained to predict these residuals.
- This new model is added to the ensemble, with a coefficient called the learning rate (or shrinkage) applied to control the contribution of each new model. This learning rate is a small positive number (e.g., 0.1) that slows down the learning process to make the model more robust.

Additive Modeling:

- The predictions from the new model are combined with the predictions from the existing ensemble to form updated predictions.
- This process is repeated, with each new model focusing on the residuals (errors) left by the previous models.

Stopping Criteria:

- This iterative process continues until a specified number of trees are added or no significant improvement can be made on the prediction accuracy.

The main advantages of gradient boosting are its ability to handle different types of data, robustness to outliers, and effectiveness in capturing complex nonlinear patterns in data. However, it can be prone to overfitting if not carefully tuned, and it may require careful selection of parameters, such as the number of iterations, learning rate, and the depth of the trees.

Predicting Consumption

- Original goal was to predict consumption for a given state in a given year
  - Should mention neural network got a 0.98 r^2, but this was disregarded as I thought it was too high, and therefore, overfit or inaccurate.
- Switching to GBR yielded a r^2 of 0.9968 for total consumption
  - Still suspiciously high, what could be going wrong?
- Implemented grid search and early stopping, r^2 is still 0.99

Why is our r^2 so high, and what can we do?

- What variables were able to explain so much, and why?
  - One reason: target and feature synergy
    - Goal was total consumption. One of our variables was population over 21. More beer is consumed when there are more people to drink it
  - What can we do?
- Changed our target
  - Focus on per capita consumption to avoid this problem
- Analyzed what variables are most important
  - Utilized feature importance to identify what our best predictors were

Brief numbers break

Change	r^2
Initial neural network (all var)	0.98
Change to GBR (all var)	0.99
Early stopping/Grid search (all var)	0.99
Restricted analysis (5 var)	0.99
Further restricted analysis (2 var)	0.99
Further restricted analysis (1 var)	0.98
Linear regression for kicks (1 var)	0.97
Change to per capita (3 var)	0.89
Final model (2 var: FIPS code, percent never married)	0.93

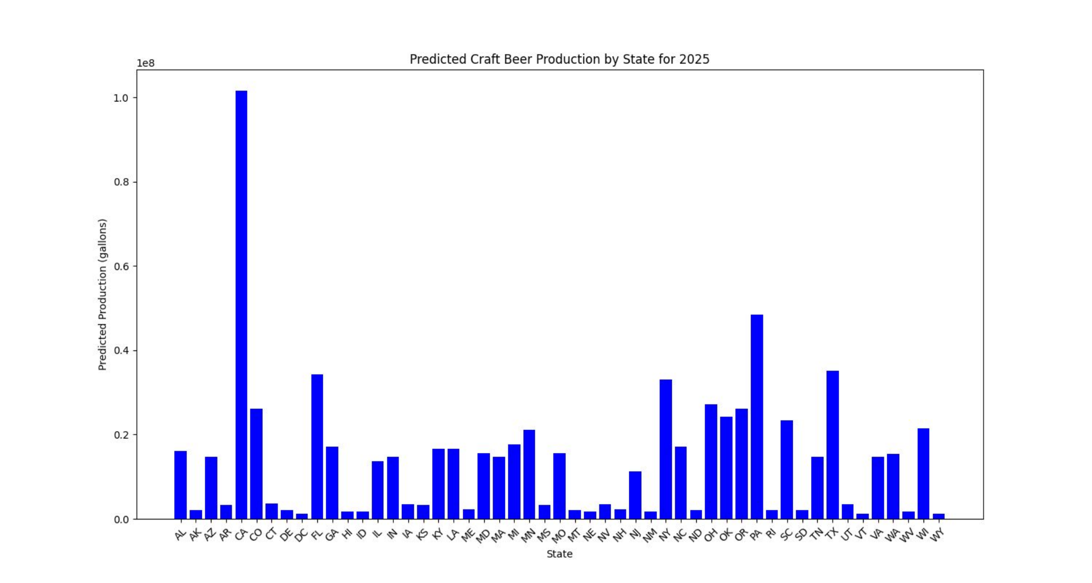
What are we learning?

- Ultimately, we found our model to be different from what we planned
    - We just wanted to predict the consumption of a state in a year, but in order to do that, we would need to have ALL of our other variables
- We need to restrict the model
    - How do we do that?
    - What variables should we use?
- Change in focus of the project: what is the best predictor of beer consumption?
- Changes the question as well: how do we determine that?
    - I used feature importance, but is there another way?
    - “Which variables can we pick out that give us the most predictive power?”

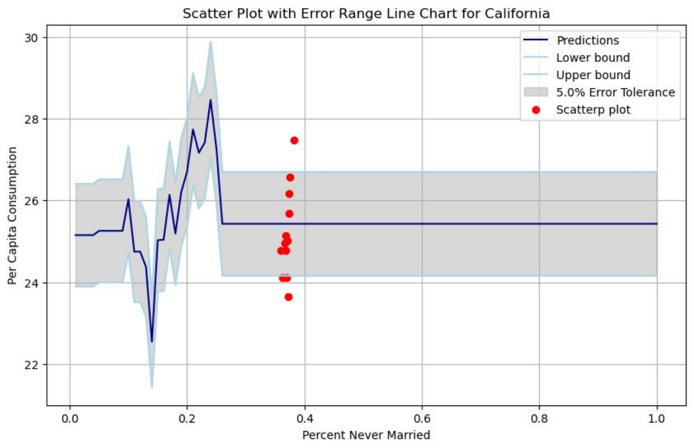
Results & Visualization

Production

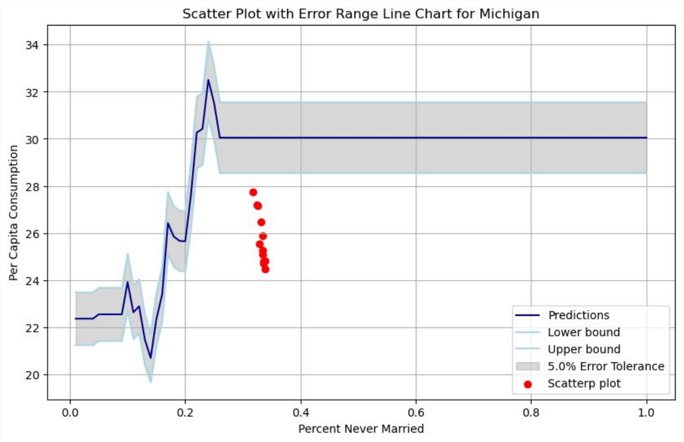
We used a model to predict the production for each state for the year of 2025 using four data points for the x-axis which are (year, census total population, census percent employed, and state id) and we used craft beer produced in gallons for the y-axis. We looped through the code for each state to show us a visualization of production for each state.



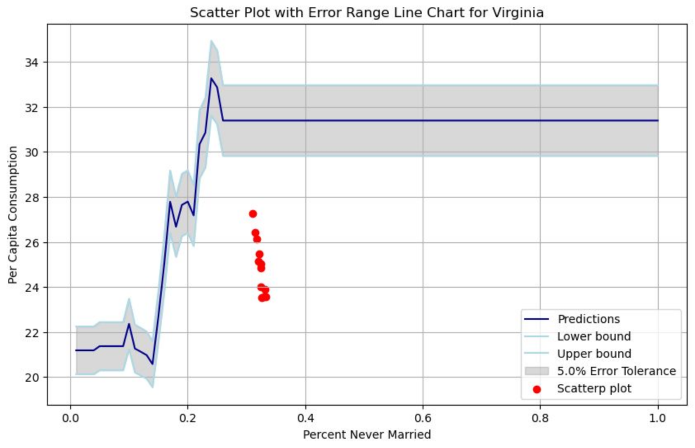
Scatter plot with Error Range Line Charts for California, Michigan and Virginia



Predicted percentage of population never married and number of gallons of beer consumed per capita.

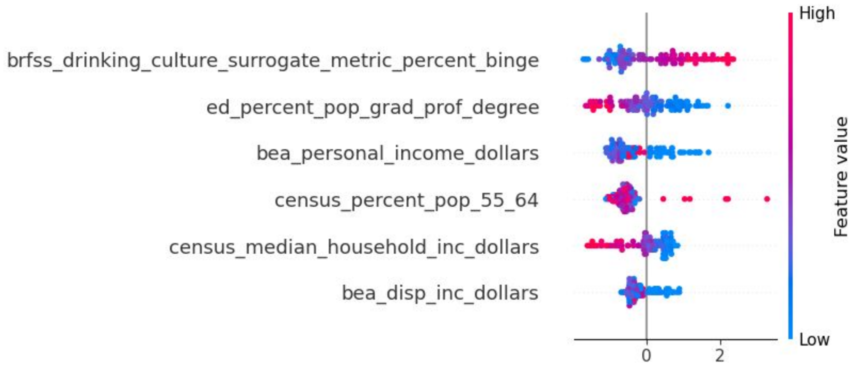


Scatter plot depicts the actual values.

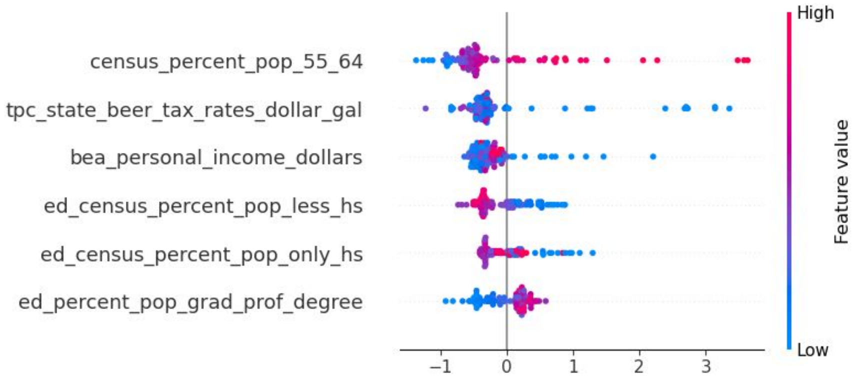


Feature Variable Importance Evaluation

- Model selection (for this): eXtreme Gradient Boosting (decision tree based) regression with Bayesian hyperparameter tuning (all feature variables included)
- Consumption target variable model
  - MSE: 1.759
  - R^2 value: 0.943
- Production target variable model
  - MSE: 0.384
  - R^2 value: 0.953
- SHAP summaries for top 6 most impactful feature variables at right



Consumption target variable



Production target variable

Conclusion

What we learned

- Modeling can be an unpredictable beast
  - Sometimes you get something super accurate, but you can’t explain why
- Being accurate isn’t the be-all end-all
  - If possible, what is your model making decisions on?
  - Should you really be making judgements on that information?
- Some conclusions are obvious once they are stated out loud
  - Total consumption tends to be influenced by the number of people consuming
- Not all conclusions are equally interesting
  - Which is more interesting, what was stated above, or what we made our model on?

Questions?