Estimating Urbanization Rate of the US using Global Sugar Consumption

Tanya Chen & Emily Wang

Intro

This model investigates the relationship between the urbanization rate and sugar consumption in the U.S. The model estimates the U.S. urbanization rate based on the global sugar consuumption dataset which contains U.S. annual sugar consumption and other relevant variables.

	Country	Year	Country_Code	Continent	Region	Population	GDP_Per_Capita	Per_Capita_Sugar_Consumption	Total_S
0	France	1972	FRA	Europe	Western Europe	2.617306e+08	8692.631696	12.827741	
1	Australia	2003	AUS	Oceania	Australia & New Zealand	1.737965e+08	6859.195960	21.362632	
2	Germany	1963	DEU	Europe	Western Europe	1.236366e+08	22075.950575	32.077485	
3	France	1965	FRA	Europe	Western Europe	2.989961e+08	3728.027392	47.648930	
4	Germany	2010	DEU	Europe	Western Europe	7.341531e+06	40420.973962	23.214343	

5 rows × 26 columns

Method

Before model training, we filtered the dataset to only include U.S. data from 2016 to 2023. Irrelevant columns such as country code, continent, campiegn information, etc. are removed. Rows that are missing values are also dropped to ensure the tidiness of the dataset.

We generated two machine learning models for our data: linear regression and decision tree. We tidied data from 2016 to 2023 and applied cross-validation to evaluation the performance of the model. Using data from 2016 to 2020 as the training set, we generate two models and test them with data spanning from 2021 to 2023. We used linear regression and trees. With linear regression, we applied cross validation. We tried bagging, random forest and boosting to find the better model with lower MSE.

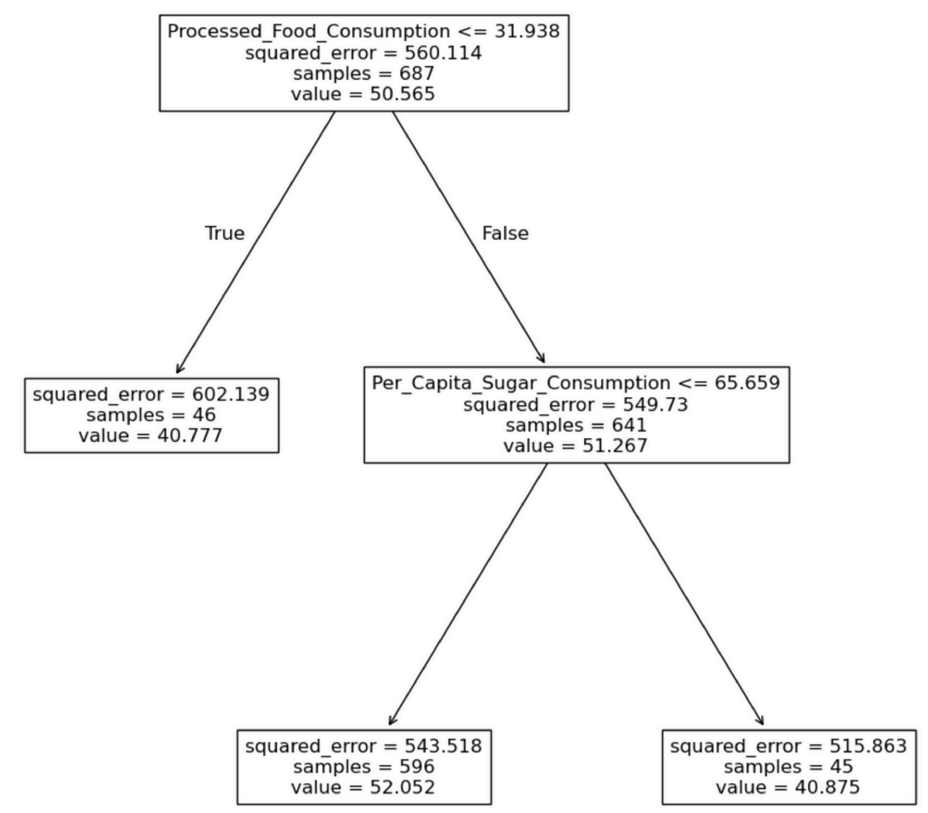
Results

Firstly, we trained a simple linear regression model using data from 1960 to 2023 in the U.S. Ths model has a MSE of 532.6926 and has an r squared of 0.00111. The range of urbanization rate is between 0 -100. MSE (Mean Squared Error) is to divide the sum of the true response subtracted by predicted reposnse variable by the number of response variable. Therefore, we can find an average difference between by finding the squared root of MSE, which is 23.1 and is relatively large. Since the MSE is large and the r squared is extremely small, this linear regression model is not catching the true relationship between the urbanization rate.

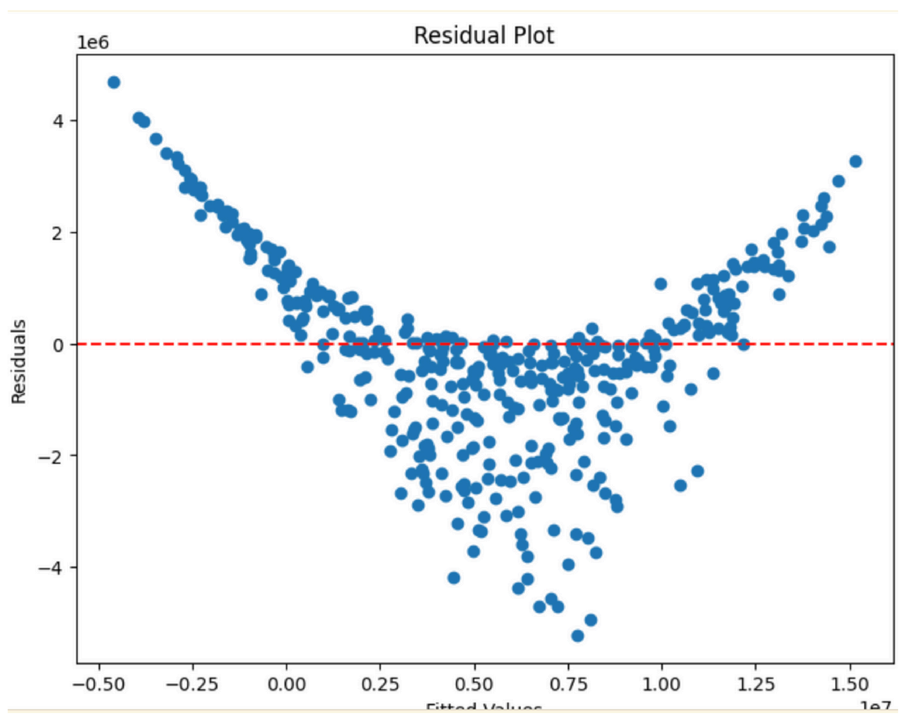
Next, we applied cross-validation into the linear regression model by separating dataset into training data and test data based on whether the data is collected after 2020. According to this model, the MSE is 553.8987, and its r squared value is -0.5834. Comparing with the first simple linear regression model, the MSE increases and there is a stronger relationship between the predator variables and the urbanization rate. However, since r squared is -0.05834, it indicates that this model falls to capture any useful data.

At last, we tested tree-based models, such as pruned decision tree, bagging, random forest and boosting. Looking at the test error of each model, the boosting model preforms the best, with MSE equal to 528.31. We used relatively large number of trees (500), but it captured the complex relationships without overfitting. Although the pruned tree and random forest have higher test error, the difference of between MSE between all the models is small. Bagging slightly improved the basic tree but did not outperformed boosting. There are no model that is too simple or ineffective at generalizing to test data.

Pruned Tree



Extra graph



Residuel plot of total sugar consumption

Discussion

Our data set has some internal issues that causes error in our further analysis. This dataset is a synthesized dataset trying to mimic the data posted by real-world sources, such as WHO and FAO. Since the algorithms of data synthezation is not transparent, the data might be falsely synthesize and cause the poor performance of our model.

All of the models' prediction has a relatively high MSE, which suggest there could be high variance in the data and absence of key predictive variables relative to urbanization.

We are also interested in researching diabetes, obesity, and other health factors relative to sugar consumption and other sugar variables. From our preliminary study of diabetes using linear regression, we find there are many statistically significant variables that would provide us a solid model to improve on. These further research will provide a more comprehensive understanding of the dataset.

Conclusion

In this model, we are trying to investigate the relationship between urbanization rate and other variables in the global sugar consumption dataset such as GDP per capita, total sugar consumption, government subsidies, etc. Based on the MSE in both the linear regression model and random forest model, there is respecively high MSE which implies that there could be high variance in the data, predictive variables for urbanization rate might be missing from the dataset, and the dataset might be noisy that prevents the model from learning the correct relationship.