Final Project Write-up: Estimating U.S. Urbanization Rate and Total Sugar Consumption based on the global sugar consumption dataset

Tanya Chen, Emily Wang

Introduction

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 consumption dataset which contains U.S. annual sugar consumption and other relevant variables.

Methods

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

We performed different training methods for the model predicting urbanization rate and the model predicting total sugar consumption.

To estimate the urbanization rate, we generated two machine learning models: linear regression and decision tree. Firstly, we created a linear regression model that takes all predictor variables in all time into consideration. Next, we separated the dataset into two, a training set and a testing set to apply cross-validation to evaluate 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 create a residual plot to catch potential patterns in error and determine whether transformation can reduce MSE. Lastly, we conducted a model of the decision tree to catch errors that our cross-validated linear regression did not handle. For the decision tree model, we tried bagging, random forest and boosting to find the better model with lower MSE.

To estimate the total sugar consumption, we also fitted a simple linear regression model to identify whether there is any relationship between total sugar consumption and other variables in the global sugar consumption dataset. Then, we also separated the dataset based on year and performed a cross-validated linear regression model. We create a residual plot to catch potential error patterns that might be underrepresented by the linear regression model. Lastly, we transformed the linear regression model using log transformation and polynomial transformation with different degrees.

Results

Estimating Urbanization Rate

Firstly, we trained a simple linear regression model using data from 1960 to 2023 in the U.S. This model has a Mean Squared Error (MSE) of 532.6926 and has an r squared of 0.00111. The range of urbanization rate is between 0 - 100. MSE is to divide the sum of the true response subtracted by the predicted response variable by the number of response variables. Therefore, we can find an average difference between by finding the square 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 the 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. Compared 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. To identify whether there are any patterns in error that are unable to be captured by a linear regression model, we created a residual plot. Residual means the difference between the estimated value and the actual value. The residual plot does not have any apparent patterns. However, we suspect that the residuals are symmetric along the diagonal.

At last, we wanted to explore other results beyond what our initial regression analysis provided, so we tested tree-based models, such as pruned decision trees, bagging, random forest and boosting. The prune decision tree predicted urbanization by splitting the data based on variables such as processed food consumption and per capita sugar consumption. Looking at the test error of each model, the boosting model performs the best, with MSE equal to 528.31. We used a 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 MSE between all the models is small. Bagging slightly improved the basic tree but did not outperform boosting. While all trees are effective at generalizing to test data, the results in our case show high error rates and poor predictive accuracy. Therefore, we may conclude that decision trees are not the most suitable methods for predicting urbanization in this context.

Estimating Total Sugar Consumption

First, we trained a simple linear regression model using data from 1960 to 2023 to gain an overall understanding of the relationship between the response variable, total sugar consumption, and the other predictors in the dataset. This initial model yielded a mean squared error (MSE) of 8.9421×10^22, which is significantly larger than the maximum observed value of total sugar consumption. The R-squared value was 0.8735, indicating that approximately 87.35% of the variance in total sugar consumption is explained by the predictor variables in the global sugar dataset.

To assess the model's generalizability, we then performed cross-validation by splitting the data into a training set (years 2016–2020) and a test set (years 2021–2023). After cross-validation, the model achieved an MSE of 2.8426*10^12 and an R-squared value of 0.8464. We also examined the residual plot, which revealed a parabolic pattern in the residuals. This suggests that the true relationship between total sugar consumption and the predictor variables is non-linear. To better capture this underlying structure, we applied two transformations: a log transformation of the target variable and polynomial feature expansion of the predictors.

For log transformation, it yields a MSE = 3159824116868.2944, which is still significantly greater than the average value of total sugar consumption in the dataset, and a R Squared = 0.8293. The residual plot for the regression model after log transformation is highly right-skewed. For polynomial transformation, when degree = 2, we obtain the lowest MSE, 4.531529257352467e-07, and highest R-squared = 1.0. As the degree increases, the MSE increases sharply and R-squared decreases. Based on the residual plot for each polynomial regression model with different degrees, though the residuals are different, for all of the residual plots, most points scatter around 0.

Discussion

Our data set has some internal issues that cause 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 synthesization are not transparent, the data might be falsely synthesized and cause the poor performance of our model.

All of the models' prediction has a relatively high MSE, which suggests 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 respectively 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.