Data Preparation

The data provided for this competition was relatively simple to work with, where no mismatched variable types, no missing values, and no issues requiring data imputation. Because of this simplicity, we tested three different approaches for the data cleaning process. The first was a simple filtering method that selected only predictors with a correlation coefficient greater than 0.1, dropping any variables that could cause multicollinearity problems. The second method automatically dropped variables using Lasso Regression. The third method—which ultimately provided the best results—initially dropped any zero-variance features, followed by the removal of variables exhibiting high multicollinearity.

Modeling Phase

During the modeling phase, two distinct approaches were used: a classical data science approach using XGBoost, and a Bayesian Linear Regression. We will focus mostly on the XGBoost approach since it provided the best results. For the Bayesian Linear Regression, the first data cleaning method discussed provided the best outcomes. Using that filtered data, combined with degree-2 variables and their interactions, the model achieved a best RMSE of 5.63. For this model, we assumed that all priors and the likelihood followed a Gaussian distribution.

Ultimately, XGBoost yielded the best results out of the two methods, with an RMSE of 4.72. To train this model, two approaches were tested to best fine-tune its parameters. The first method was a 5-Fold Cross-Validation strategy, and the second method utilized a Grid Search. Out of the two tuning methods, Grid Search was the most effective at tuning the model. In this tuning method, we utilized Negative-RMSE as the scoring metric we aimed to optimize.

The parameters we used in the Grid Search for tuning were: max depth, learning rate, and number of estimators. Subsample and colsample_bytree were also added to introduce some randomness into the training process.

Conclusion

In the end, we selected XGBoost as our final model for generating predictions. As a gradient-boosted decision tree algorithm, it demonstrated superior capability in handling the non-linear relationships and complex interactions present within the data, significantly outperforming the linear assumptions inherent in the Bayesian approach. The tuning process yielded a highly stable model with a specific configuration: a maximum depth of 4, a learning rate of 0.006, and 1,000 estimators. This combination achieved a best cross-validation Root Mean Squared Error (RMSE) of approximately 4.73, indicating a high level of predictive accuracy.

For an analysis of our code, please visit our GitHub page and look for the file final_model.ipynb; all other files are related to the Bayesian methodology.

https://github.com/GIuliannom/ASA-Florida-Chapter---Data-Competition-2026