For evaluation metrics, we knew that false negatives and false positives are not equally costly. We knew that since we are dealing with time series data that we would be using two different evaluation metrics to judge our models. Therefore, we chose to use an F0.5 score for high season. This is because we want to reduce false positives, meaning that we want to reduce overbooking. On the contrary, for low season we used a F1.5 score because false negatives are more costly.

Feature selection methods

We used a variety of feature selection methods to get a consensus on what features are the most valuable to our models. We found using a Decision Tree Regressor with Mean Squared Error as the criteria was the most important method in calculating what features are the best for our modeling. We also compared these results to other criteria such as Friedman and Mean Absolute Error with a bar plot. Given our previous experience with MSE, and knowing that it is using an estimated measure of the average squared differences between the estimated values and true values, we felt confident in its results.

XGBoost proved far and away to be the best model during K-fold cross validation testing. While we achieved better F scores with the shuffle parameter enabled to True, we discovered that this is falsely inflated.