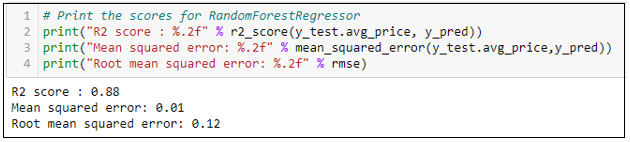
**Summary, Results Description & Addressing the Project’s Questions**

Our goal is to build and train the model that has high accuracy in predicting prices in various US cities and to determine what features play an important role at predicting prices.

**Random Forest Regressor**

Random Forest Regressor ML model showed promising results in predicting avocado prices in various US cities with **R2** score of **0.88**, **mean squared error (MSE)** of **0.01** and **root mean squared error (RMSE)** at **0.12**.

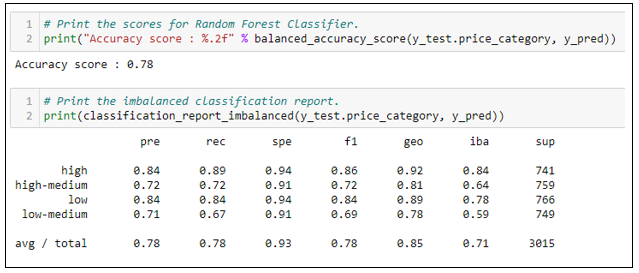
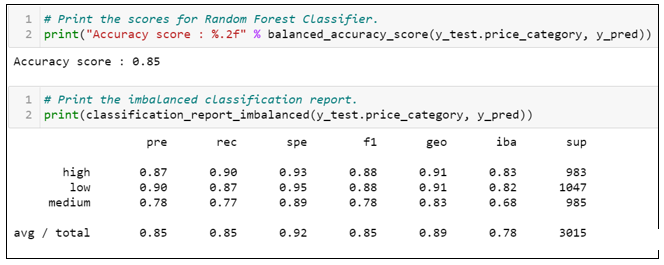
R2 score is defined as “*the proportion of the variance in the dependent variable that is predictable from the independent variable(s)*” (1). MSE is defined as “*the average of the square of the errors. The larger the number the larger the error.*” (1). RMSE is defined as “*the square root of the mean square error”* and “*is directly interpretable in terms of measurement units, so is a better measure of goodness of fit than a correlation coefficient*” (2). High R2 score and low MSE indicates that the model fits the data well.



*Figure 1: Random Forest Regressor scores (prices and production dataset).*

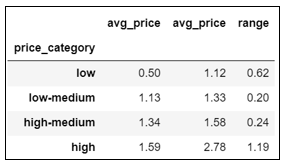
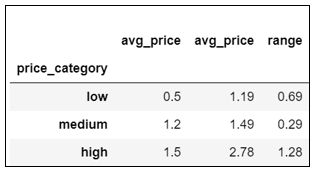
**Random Forest Classifier**

We trained Random Forest Classifier Model on the same dataset, where we categorized y-variable into 3 and 4 categories. As explained in feature engineering section, we decided to train (and keep) the model with 3 categories, because of the better score and more even price range in the bins.



*Figure 2 & 3: Accuracy score and confusion matrix for y-variable categorized into 4 and 3 categories.*

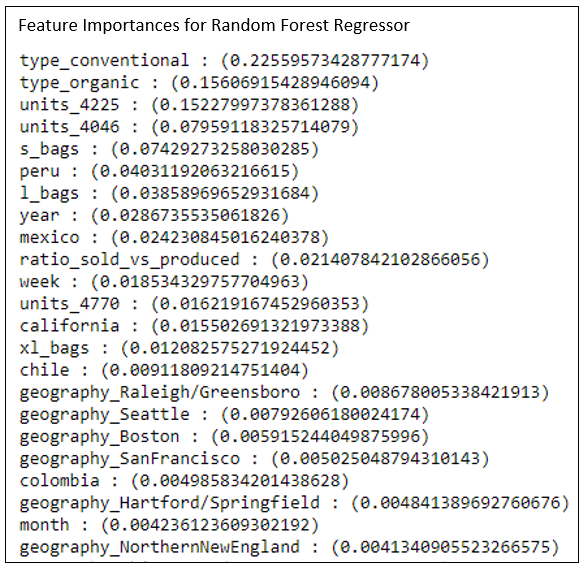
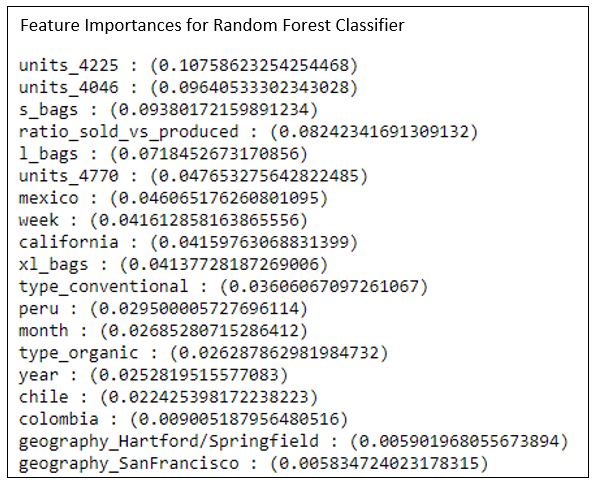
Accuracy score for model with 4 categories cut is 78% and 85% for 3 categories cut, meaning that the model predicts average prices well with 85% accuracy in 3 category cut. Precision and recall scores are high, meaning there is not many false positives and false negatives, respectively. An interesting observation are precision and recall scores for the categories with the narrowest range (“medium” in 3 category cut and “low-medium”, “high-medium” in 4 category cut). Lower precision (some values were marked as false positives) and recall scores (some values were marked as false negatives) might be due to narrow average price range.



*Figure 4 & 5: Bins ranges with qcut for 4 and 3 categories.*

**Feature Importances**

Another question for our project that we are addressing was to find what features play important role when making prediction. Based on the results in the figures below we can see what features are important when predicting prices. An interesting observation is how different features have different weights when making predictions.



*Figure 6 & 7: Features Importances for Random Forest Regressor and for Random Forest Classifier.*

**References**

1. Mean Square Error & R2 Score Clearly Explained, Machine Learning and Big Data Blog, <https://www.bmc.com/blogs/mean-squared-error-r2-and-variance-in-regression-analysis/> , Web, 3 January 2021.
2. What are Mean Squared Error and Root Mean Squared? https://www.vernier.com/til/1014 , Web 3 January 2021.