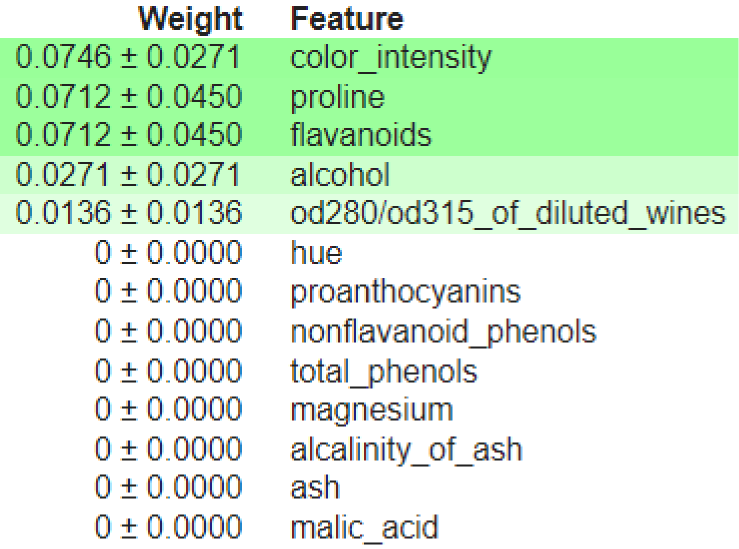
Homework 5

Task 2

Random forest with 100 estimators and no limit for tree depth was chosen as a base model. It achieved 100% accuracy on the test set. Here is its permutational feature importance:



Weight	Feature
0.0746 ± 0.0271	color_intensity
0.0712 ± 0.0450	proline
0.0712 ± 0.0450	flavanoids
0.0271 ± 0.0271	alcohol
0.0136 ± 0.0136	od280/od315_of_diluted_wines
0 ± 0.0000	hue
0 ± 0.0000	proanthocyanins
0 ± 0.0000	nonflavanoid_phenols
0 ± 0.0000	total_phenols
0 ± 0.0000	magnesium
0 ± 0.0000	alcalinity_of_ash
0 ± 0.0000	ash
0 ± 0.0000	malic_acid

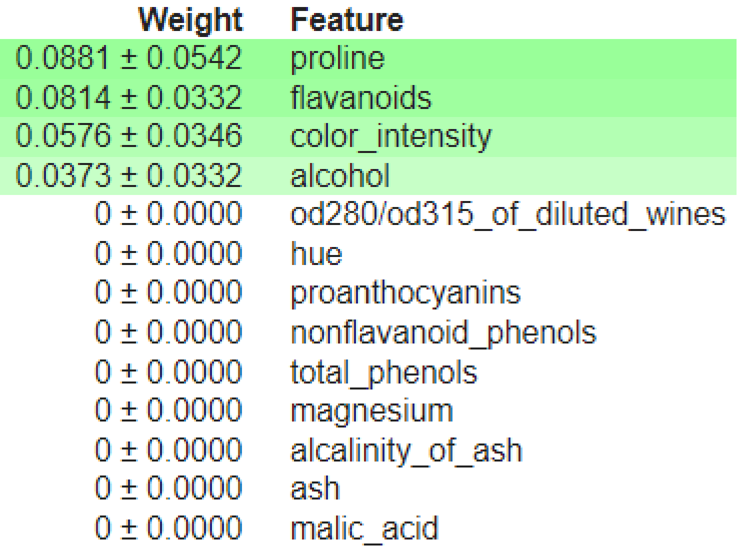
There are three features that have a major impact on model's accuracy: *color_intensity*, *proline* and *flavanoids*. Next two features have only a minor impact: *alcohol* and *od280/od315_of_diluted_wines*. Rest of the features are not important for this model. What is interesting here is that accuracy decrease during reshuffling of *proline* and *flavanoids* varied a lot. For both of them the average accuracy decrease was 0.0712, but that number varied from one reshuffling to the next by 0.045 (which is around 63%).

Task 3

Since premutational feature importance depends on the model, changing model may change importances. This section presents this phenomenon.

Candidate 1

First of the candidate models was the same model with two times less estimators. This model still achieved 100% accuracy on the test set. Here is its permutational feature importance:

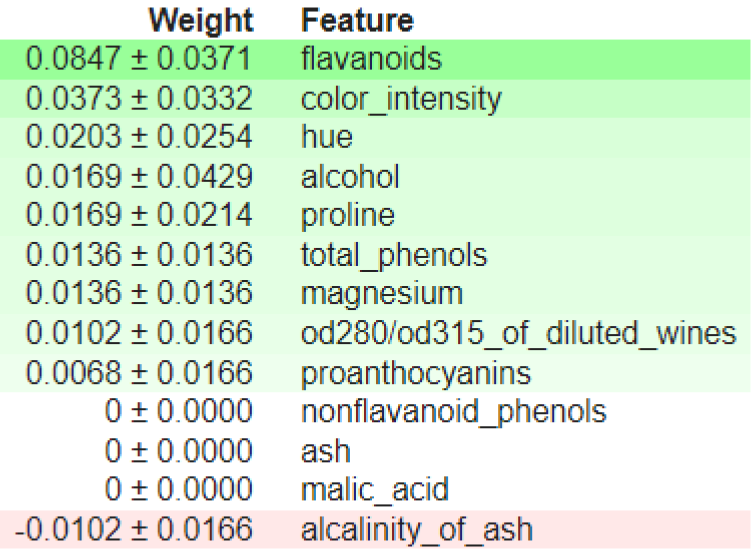


Weight	Feature
0.0881 ± 0.0542	proline
0.0814 ± 0.0332	flavanoids
0.0576 ± 0.0346	color_intensity
0.0373 ± 0.0332	alcohol
0 ± 0.0000	od280/od315_of_diluted_wines
0 ± 0.0000	hue
0 ± 0.0000	proanthocyanins
0 ± 0.0000	nonflavanoid_phenols
0 ± 0.0000	total_phenols
0 ± 0.0000	magnesium
0 ± 0.0000	alcalinity_of_ash
0 ± 0.0000	ash
0 ± 0.0000	malic_acid

Important features changed their order a little bit, but most important ones stayed on top. *alcohol* gained some importance, while *od280/od315_of_diluted_wines* lost all of its.

Candidate 2

Second candidate model was even more limited. It had only 20 estimators, and all of them had a depth limit of 2. That caused lower accuracy of around 97%. Here is its permutational feature importance:

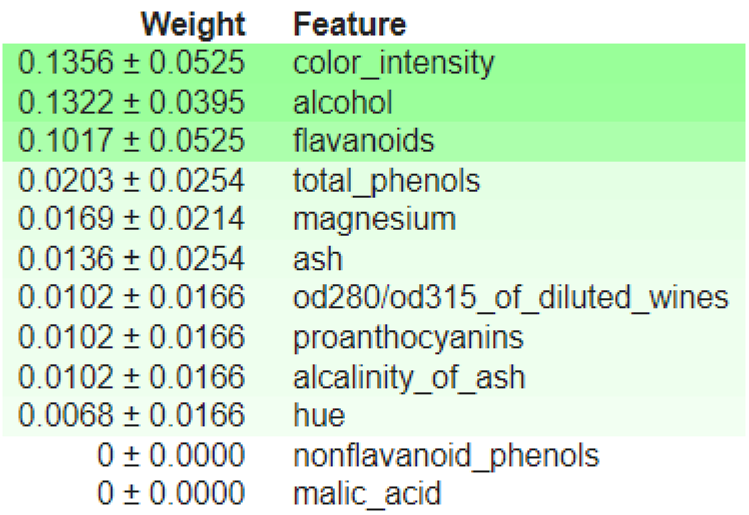


Weight	Feature
0.0847 ± 0.0371	flavanoids
0.0373 ± 0.0332	color_intensity
0.0203 ± 0.0254	hue
0.0169 ± 0.0429	alcohol
0.0169 ± 0.0214	proline
0.0136 ± 0.0136	total_phenols
0.0136 ± 0.0136	magnesium
0.0102 ± 0.0166	od280/od315_of_diluted_wines
0.0068 ± 0.0166	proanthocyanins
0 ± 0.0000	nonflavanoid_phenols
0 ± 0.0000	ash
0 ± 0.0000	malic_acid
-0.0102 ± 0.0166	alcalinity_of_ash

Suddenly more features became important, though their importance was low. *flavanoids* became by far the most important feature. It seems that the model's limited resources made it explore the dataset more, and use more features. Funny thing here is that randomisation of *alcalinity_of_ash* made it a better predictor, but that's of course because of randomness. In reality it's importance should be 0.

Candidate 3

Third candidate model was exactly the same as the base one. The thing that changed here was the dataset - a very important feature (*proline*) was removed. That caused lower accuracy of around 98%. Here is its permutational feature importance:



Weight	Feature
0.1356 ± 0.0525	color_intensity
0.1322 ± 0.0395	alcohol
0.1017 ± 0.0525	flavanoids
0.0203 ± 0.0254	total_phenols
0.0169 ± 0.0214	magnesium
0.0136 ± 0.0254	ash
0.0102 ± 0.0166	od280/od315_of_diluted_wines
0.0102 ± 0.0166	proanthocyanins
0.0102 ± 0.0166	alcalinity_of_ash
0.0068 ± 0.0166	hue
0 ± 0.0000	nonflavanoid_phenols
0 ± 0.0000	malic_acid

Removing such an important feature made other important features more important, and many of not important features a little important. Here importance of top features is the highest from all of the models.