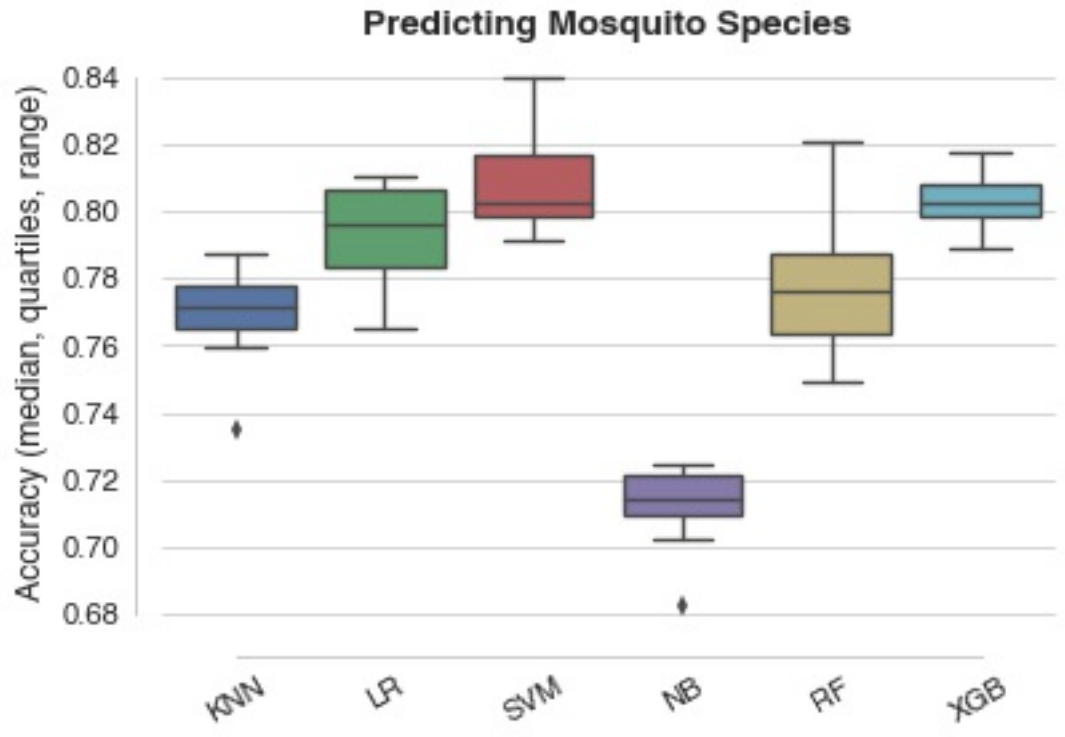
Predicting both age and species of *Anopheles gambiae* and *Anopheles arabiensis* from mid-infrared spectra

Results

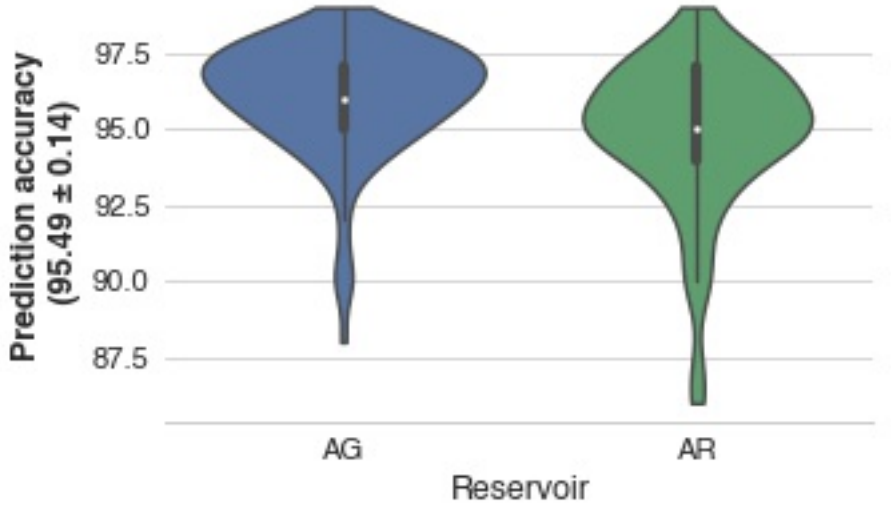
Spot-checking baseline performance of various algorithms

To determine which algorithm may be best suited to identifying the species of a mosquito based on its MIRS and to identify its age age class, we first compared the baseline performance of 6 major algorithms, linear regression, nearest neighbours, ensemble decisions trees, or Naïve Bayes.

With output category consisting of ages [1, 3, 5, 7, 9, old], XGB achieved the best prediction accuracy at baseline settings:



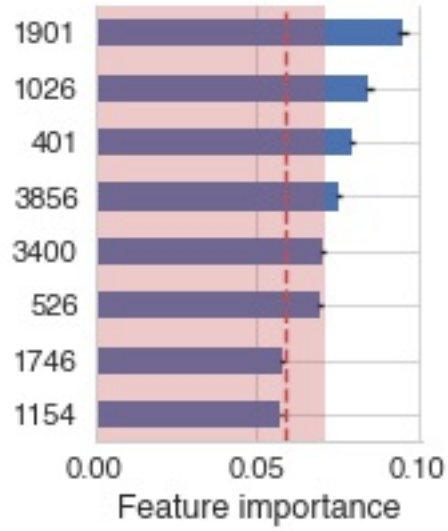
Support Vector machines and eXtreme Gradient Boost achieved highest performance when predicting mosquito species. After tuning using repeated stratified random under-sampling, XGB achieved 95.49% ± 0.14% accuracy on average.



Top features

Three wavelengths stood out as being particularly important to the prediction:['1900.76462', '3855.53371', '1745.50175'].

Ranked by decreasing importance:

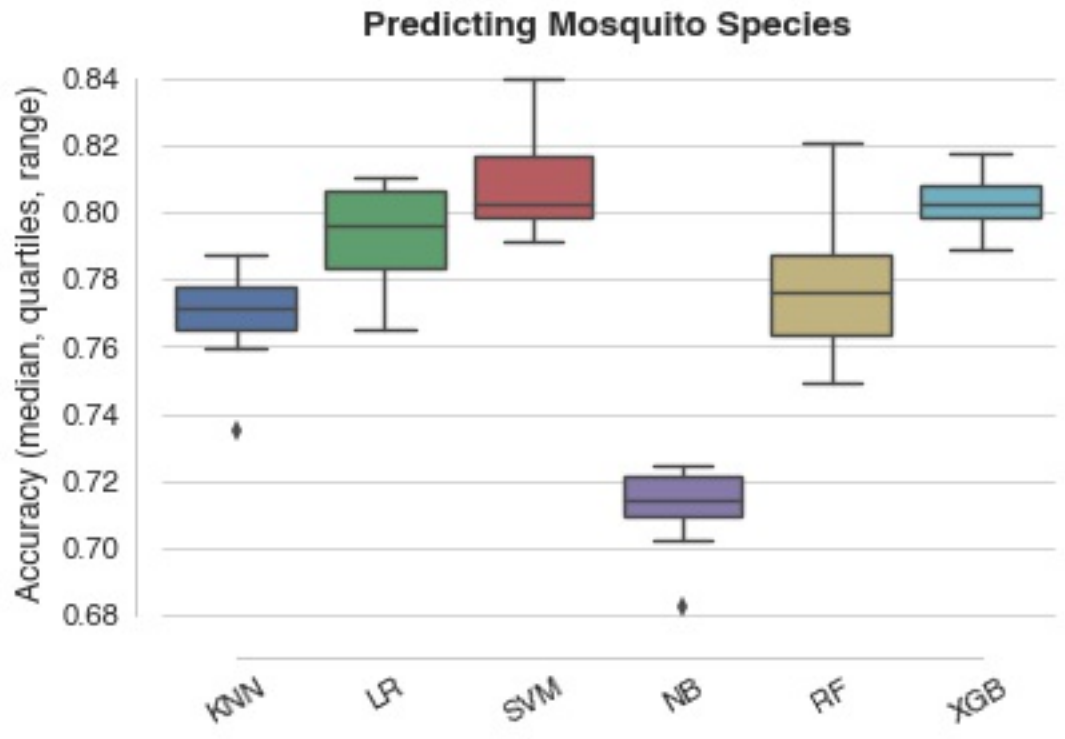


Predicting species only

This uses the binary label for species (AG or AR).

Spot-checking baseline performance of various algorithms

With output category consisting of ages [1, 3, 5, 7, 9, old], XGB achieved the best prediction accuracy at baseline settings:



Both random forest and xgboost performed well here.

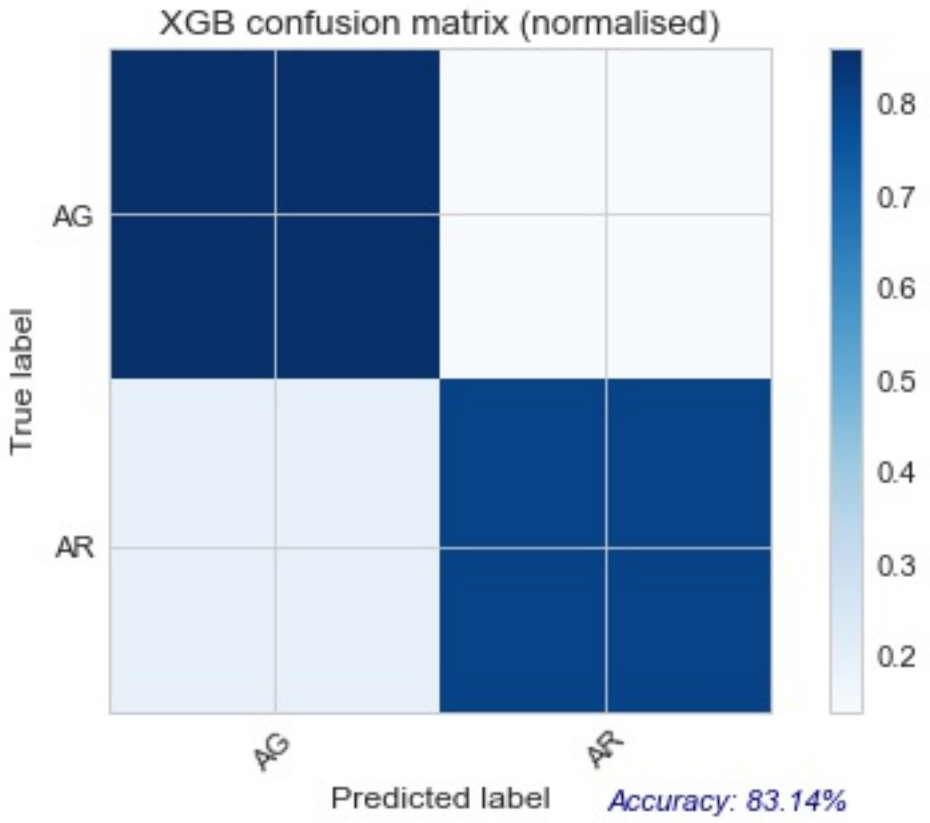
After tuning Random Forest parameters

Accuracy on test set:84.67%

Classification report:

	precision	recall	f1-score	support
AG	0.82	0.88	0.85	255
AR	0.88	0.82	0.84	267
avg / total	0.85	0.85	0.85	522

Confusion matrix



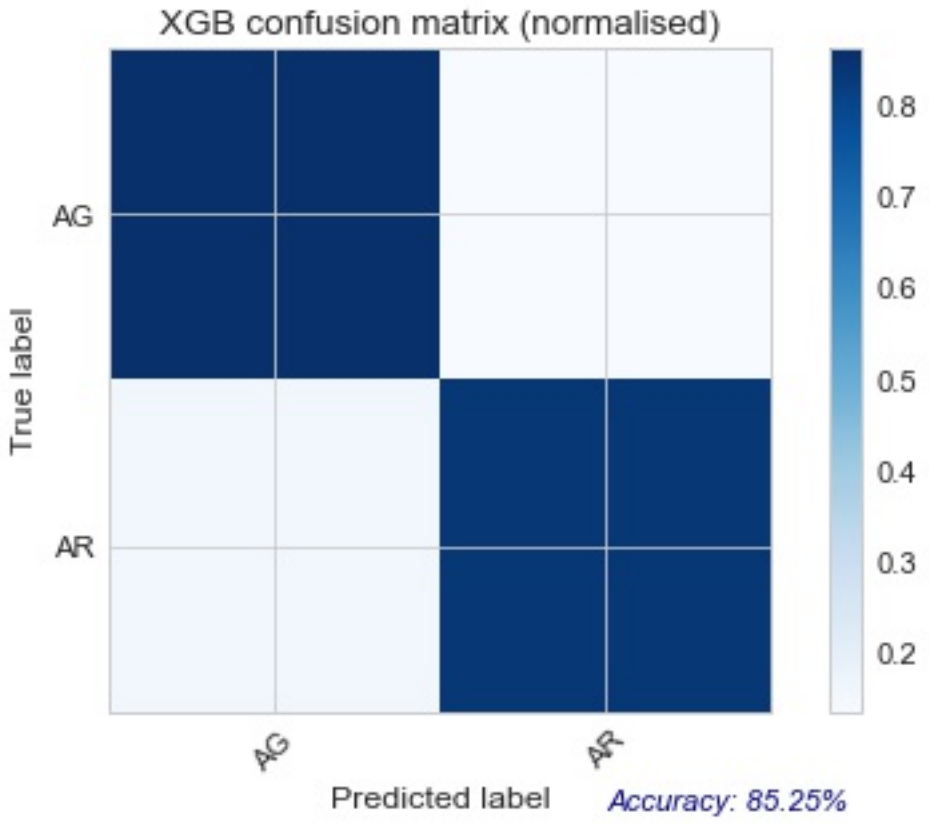
After tuning XGBoost parameters

Accuracy on test set:85.25%

Classification report:

	precision	recall	f1-score	support
AG	0.84	0.86	0.85	255
AR	0.87	0.84	0.85	267
avg / total	0.85	0.85	0.85	522

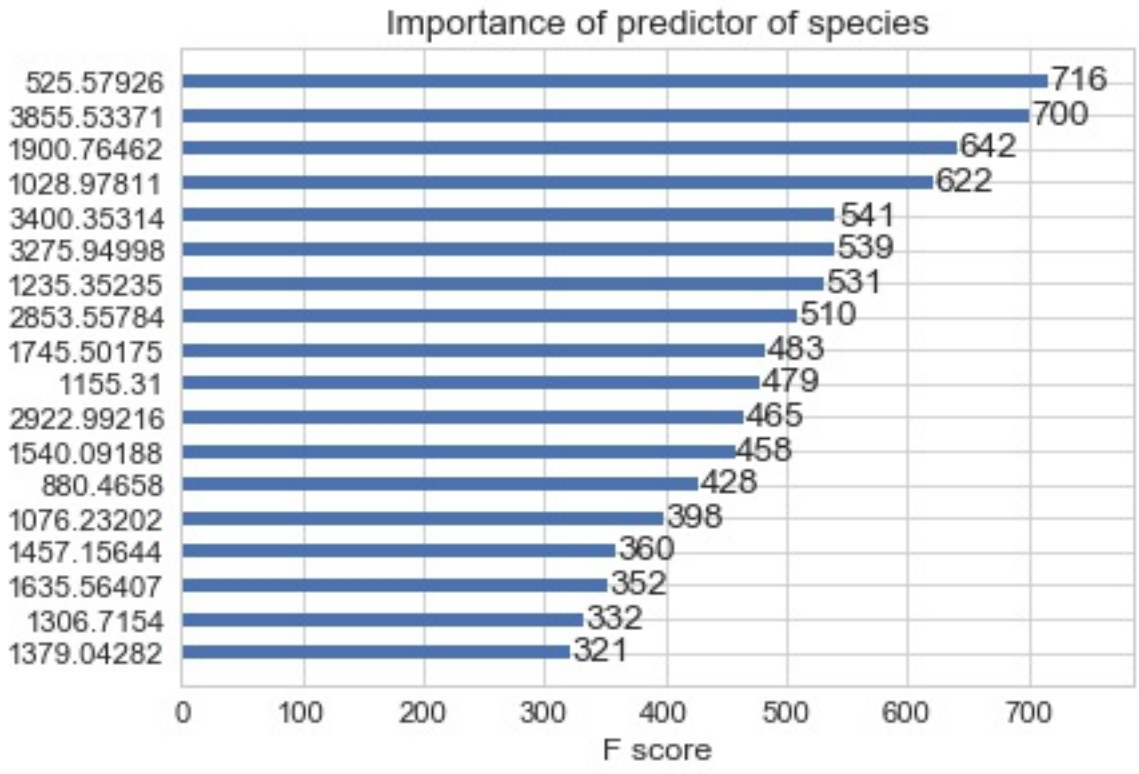
Confusion matrix



Top features

Four wavelengths stood out as being particularly important to the prediction: '525.57926', '3855.53371', '1900.76462', '1028.97811'

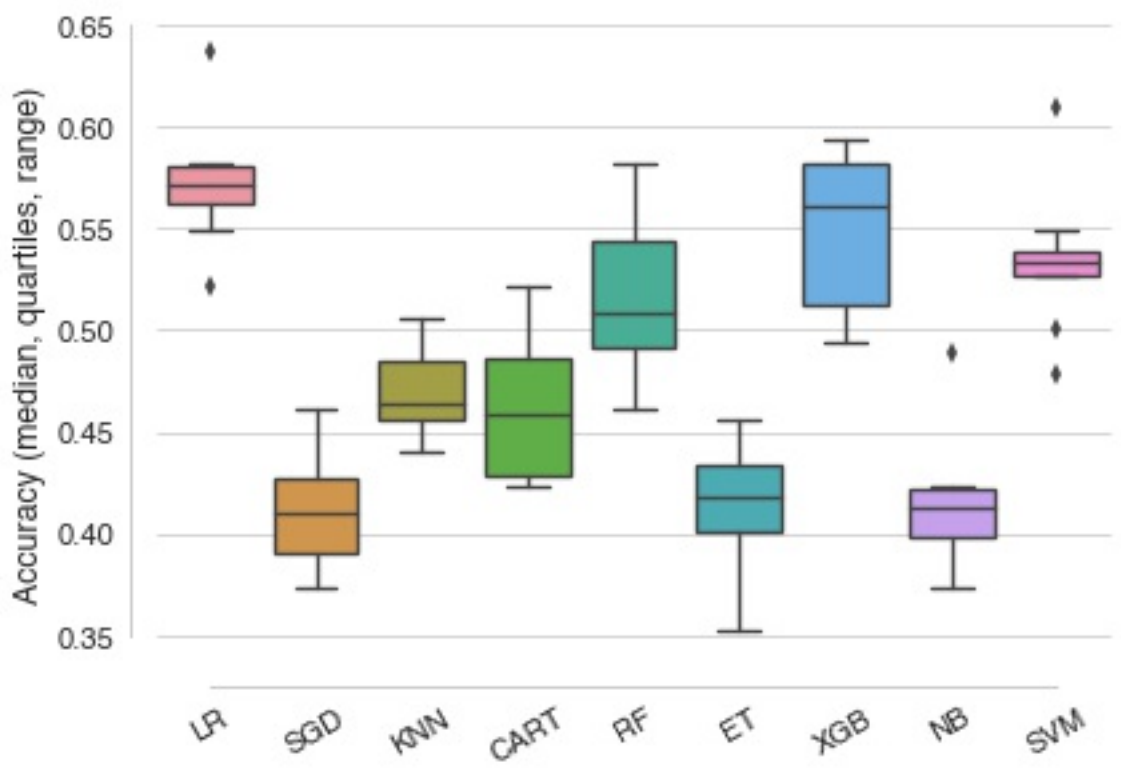
Ranked by decreasing importance:



Predict age from both species

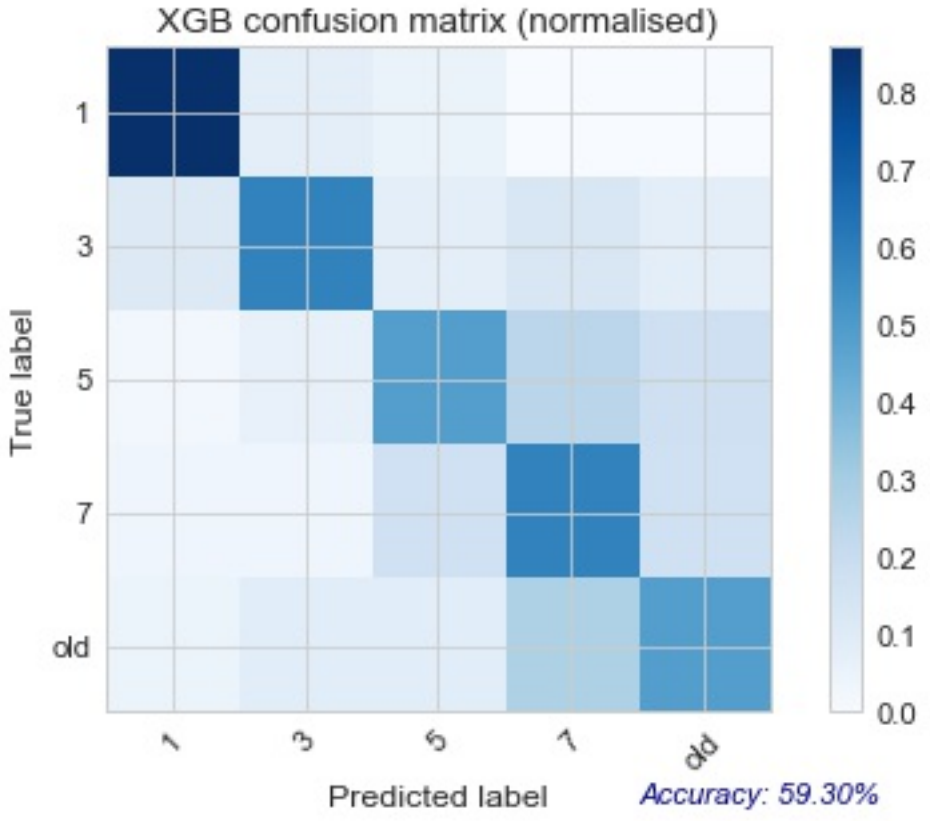
Spot-checking baseline performance of various algorithms

With output category consisting of ages [1, 3, 5, 7, 9, old], XGB achieved the best prediction accuracy at baseline settings:



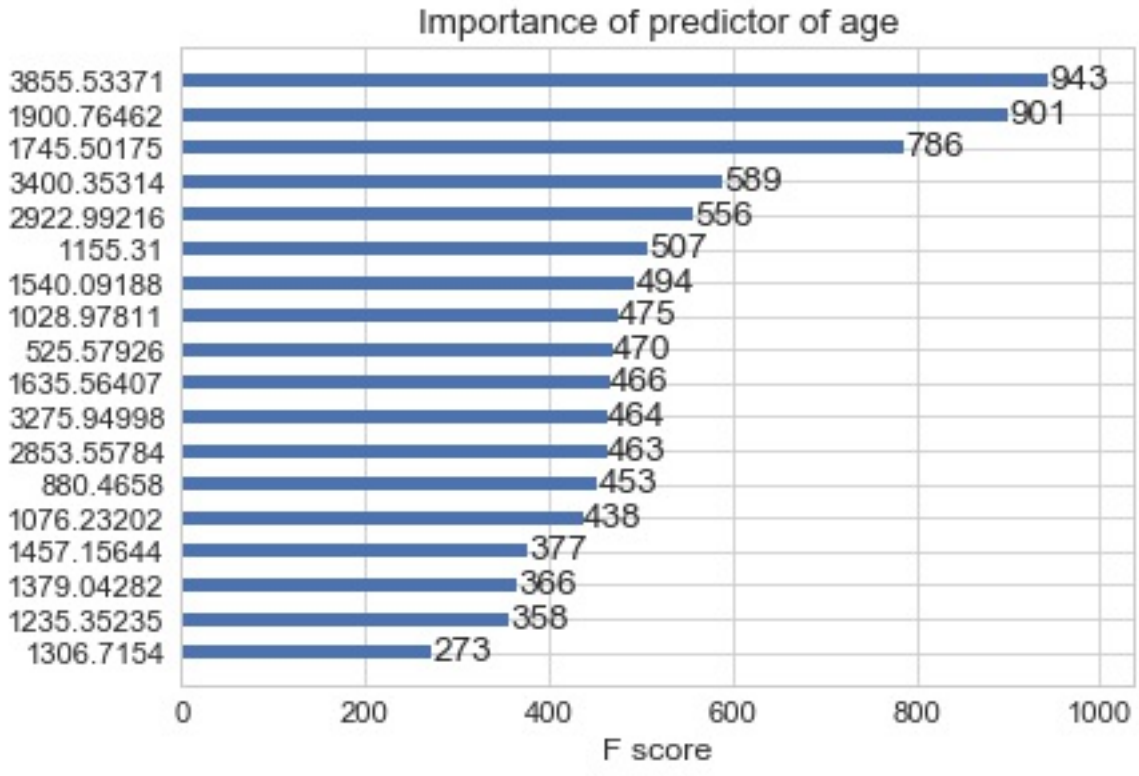
After tuning XGBoost parameters

Confusion matrix



Top features

Ranked by decreasing importance:



Four wavelengths stood out as being particularly important to the prediction: '3855.53371', '1900.76462', '1745.50175', '2922.99216'

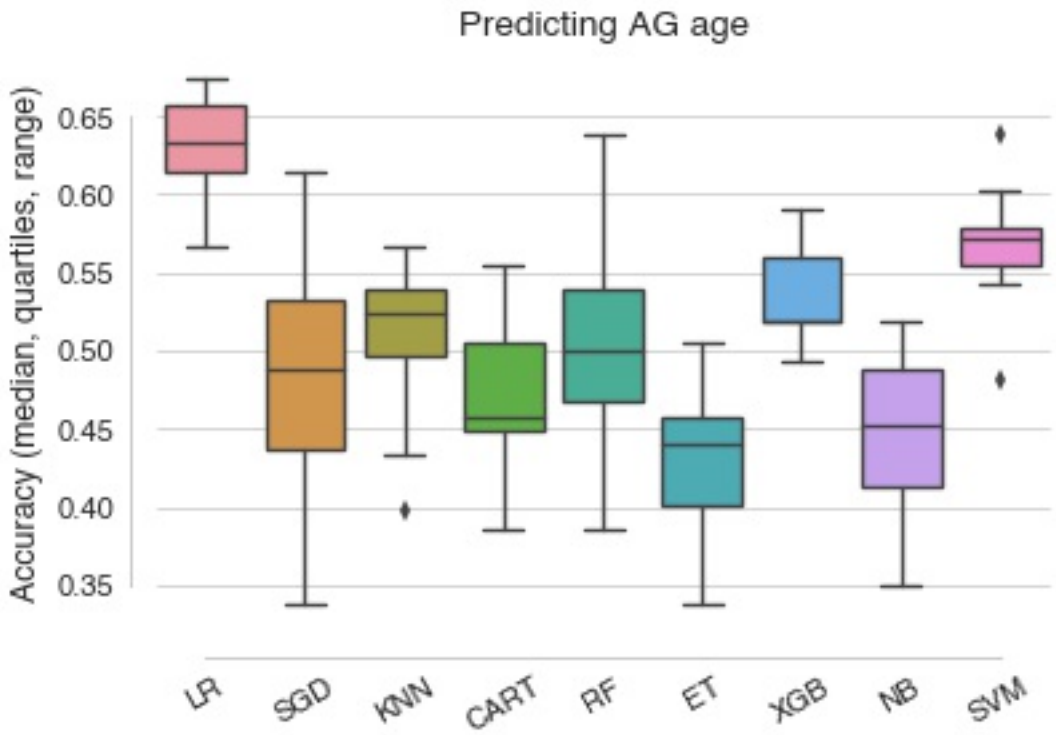
Predicting age separately for AG ad AR

I then built 2 separate models of age: one selecting only AG and the other with only AR.

Predicting age of AG

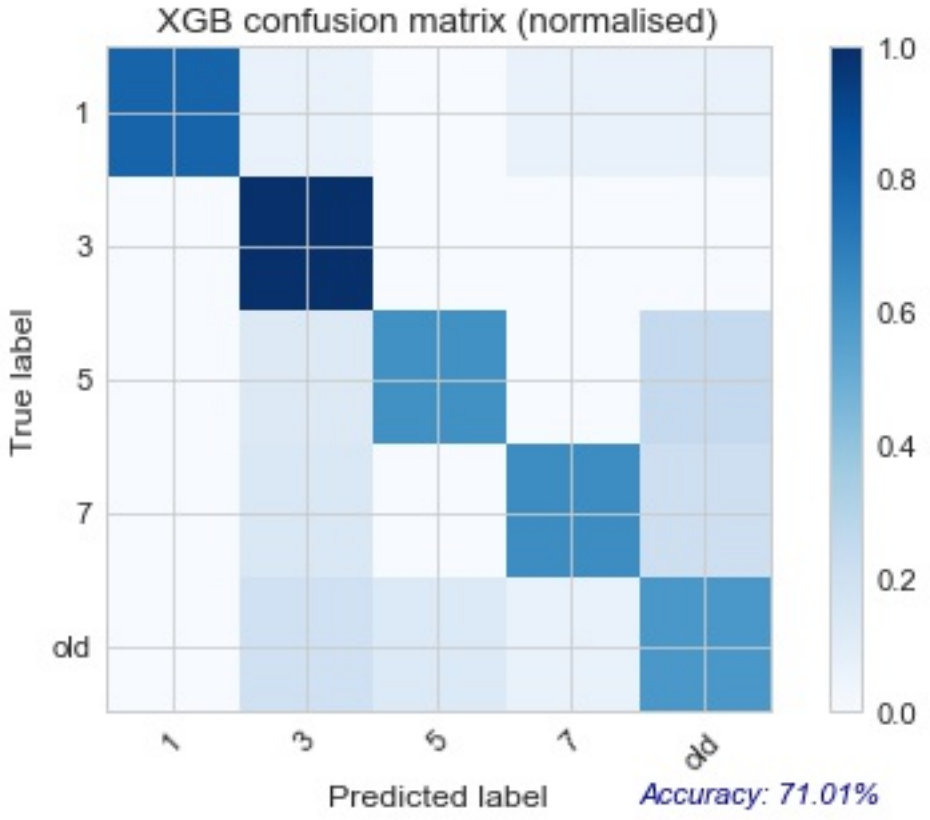
Spot-checking baseline performance of various algorithms

With output category consisting of ages [1, 3, 5, 7, 9, old], XGB achieved the best prediction accuracy at baseline settings:



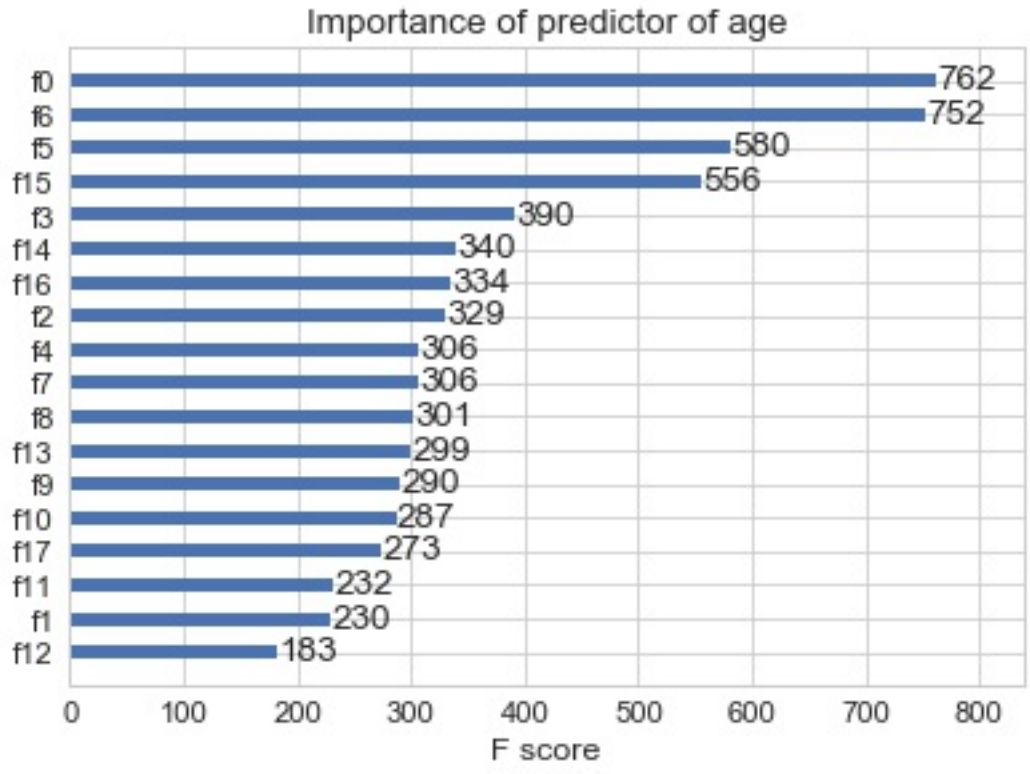
After tuning XGBoost parameters

Confusion matrix



Top features

Ranked by decreasing importance:

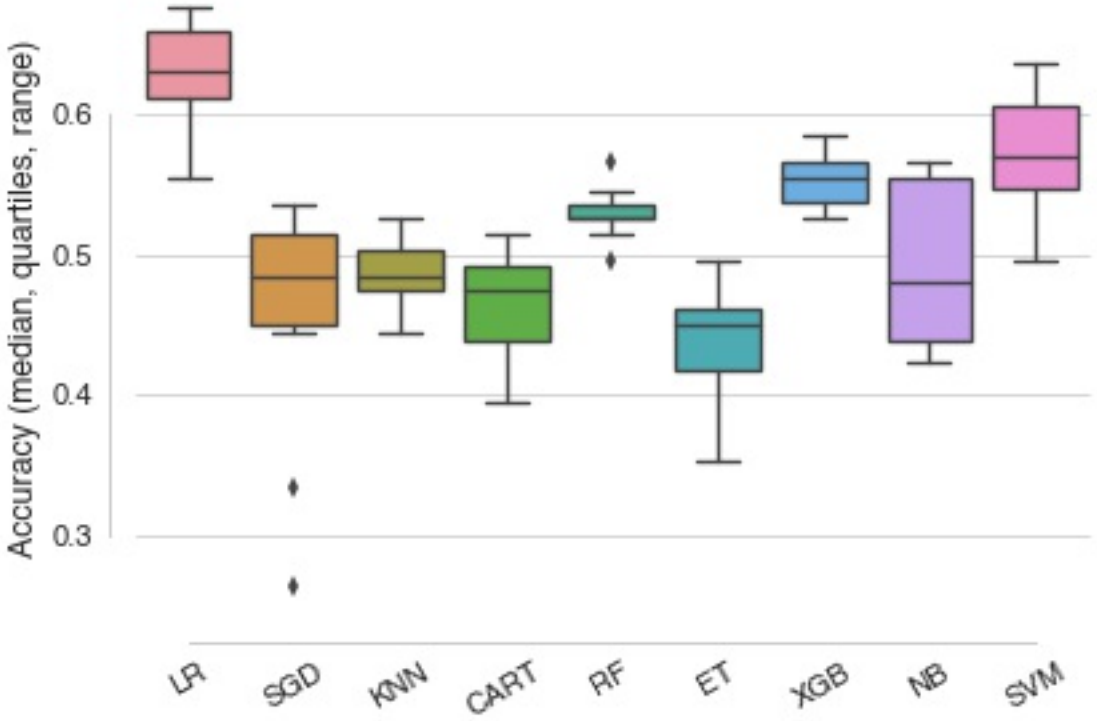


Four wavelengths stood out as being particularly important to the prediction:

Predicting age of AR

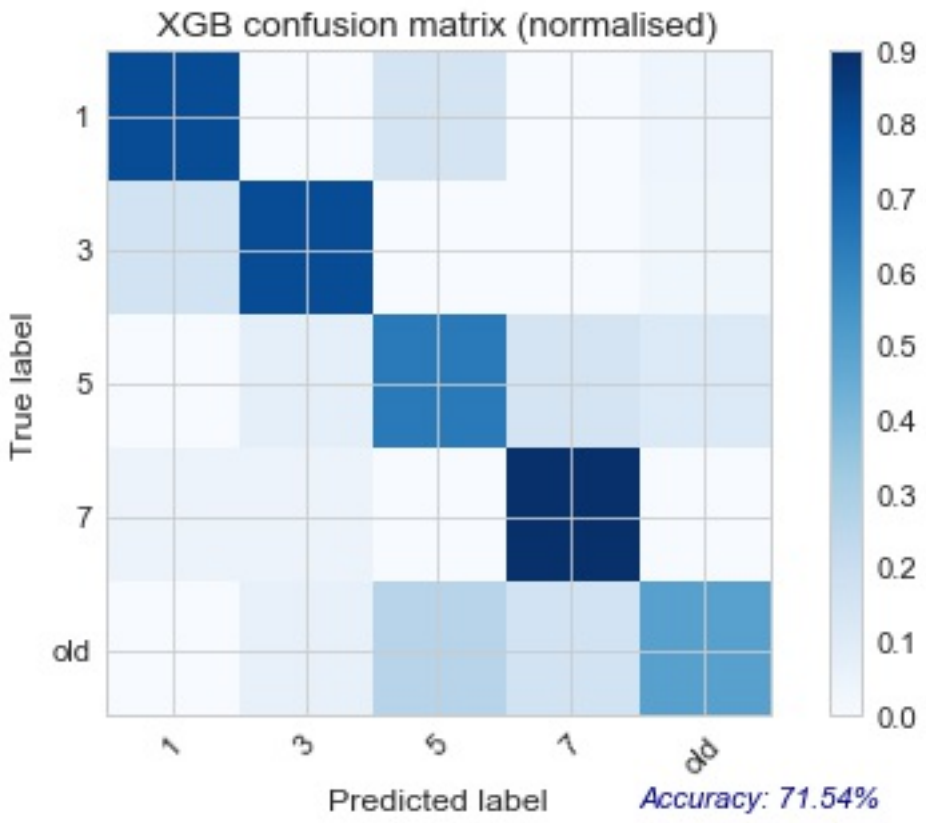
Spot-checking baseline performance of various algorithms

With output category consisting of ages [1, 3, 5, 7, 9, old], XGB achieved the best prediction accuracy at baseline settings:



After tuning XGBoost parameters

Confusion matrix



Top features

Ranked by decreasing importance:

Three wavelengths stood out as being particularly important to the prediction: '1900.76462', '1745.50175', '3855.53371'

Conclusions

1. Predicting age and species at the same time yields an accuracy of **47%**.
2. However, using the full dataset (which includes *Anopheles gambiae* and *Anopheles arabiensis*), to predict species alone achieves **85.25%** accuracy (xgboost)
3. predicting age using both AG and AR achieves **52%** accuracy
4. predicting age using AG only achieves **71%** accuracy
5. predicting age using AR only achieves **71.5%** accuracy