Changes made in the code:

1. There are no categorical entries in the data, so “OneHotEncoder” is unnecessary. Though I am not sure how this would affect the performance. In the manual, it is mentioned that preprocessing (which also includes one-hot encoding) is done automatically. After this change the accuracy of Random Forest itself jumped from 0.65 to 0.67. Furthermore, after one-hot encoding is applied y_train and y_test are pandas series which I think is the problem. It is mentioned in reference [3] that auto-sklearn has not been tested with pandas, however it is an old thread, so it might not actually be the case. In the same thread, inputs are recommended to be numpy arrays.

2. I think the biggest improvement happened when resampling strategy is set to “cv” for cross validation. In addition, it is mentioned in the API reference that if aforementioned option is used for resampling, then certain arguments should be passed such as the number of folds, shuffle, and train size.

Sources

1. https://automl.github.io/auto-sklearn/master/api.html
2. https://automl.github.io/auto-sklearn/master/manual.html
3. https://github.com/automl/auto-sklearn/issues/923