Sea Turtle Rescue: Error Detection Challenge solution notes (final score ~ 0.0442291)

János Sávoly (a.k.a. CacoS)

05/03/2019

Short summary

My solution was written in R. I treated this problem as **25** (number of columns in the database, except Rescue_id) separate **binary classifcation** tasks (error - no error). The best performance was obtained with xgboost models. The workhorse in my approach was the **mlr library**. I used hyperparamter tuning (Bayesian optimization, 5-times CV stratified by the year of the bycatch) to find the best parameters for the individual models which maximized accuracy or minimzed the logarithmic loss of the classification problems.

Important note

During the competition I ran my code both on Linux and Windows. I noticed that the feature **tchar_wrong** differed on these platforms, it had **1 nonzero element** on Windows and **3** on Linux. My code is intended for **Linux**.

Scripts

The following 3 scripts make up the solution:

1. **01_sea_turtle_error_detection.R**: contains the model generation for the *Researcher*, *CaptureSite*, *CaptureMethod*, *Fisher*, *LandingSite*, *Species*, *ReleaseSite*, *Tag_1*, *Tag_2*, *Lost_Tags*, *CCL_cm*, *CCW_cm*, *Weight_Kg*, *Sex*, *Tag_3*, *T_Number*, *PCVNumber*, *Expenditure*, *Date_Release*, *Status*, *SpecialRemarks* columns.
2. **02_sea_turtle_error_detection.R**: contains the model generation for the *Release_Admiss_Notes*, *Date_Caught*, *ForagingGround* and *TurtleCharacteristics* columns.
3. **03_combinator.R**: creates the solution csv file.

The first 2 scripts have the same structure, they only contain some feature engineering and metric differences (accuracy vs log loss).