RandomForest.

Previous research[1] and alert detection studies[2] suggests trees perform well for classification with mixed data. RandomForests were selected. Randomforest benefits over a bagging and boosted tree, which creates sub-sets of data (with replacement) and trains models and aggregates predictions. The RandomForest additionally samples N (n<M, where M is the number of features in a data set). Given a large data set, with many features, the additional of random feature sampling as well as bagging is beneficial in reducing overall error, specifically variance. Due to bagging and feature sampling, RandomForests have the disadvantage of being blackbox; therefore hard to visualise compared to a single tree. Additionally

K-NN

K-NN was selected for its relative simplicity compared to other blackbox models (Neural-Network and RandomForest). The algorithm works by grouping datapoints, based on other datapoints which are 'close' to one another'. As such it does not make any underlying statistical assumptions about the data. The base model calculates distance/closeness via Euclidean distance. The disadvantage of Euclidean distance, is the data set contains categorical data which could be ordinal or labelled. As such Euclidean distance would be statistically incorrect as we cannot measure the vector distance between a giraffe and cat). Measuring distance using Gower would be an area of future improvement [3]. The initial model was selected with K=5 (the number of neighbours a datapoint is categorised by). Where 5, is sqroot(N). N number of features = 20 [4]

Binary Logistic Regression

Simple, easy to understand results. Easy to simplify.

Neural-Network

Talk blackbox. Very flexible. Computationally expense. Does not require memory for this task.

SVM

RBF kernel vs Linear kernel. Gamma and Cost. Relative ease of understanding. Less prone to overfitting (due to transformations). Pretty good on data with large number of features due to the transformations. Can be difficult to refine due to selection of correct kernel, gamma and cost.