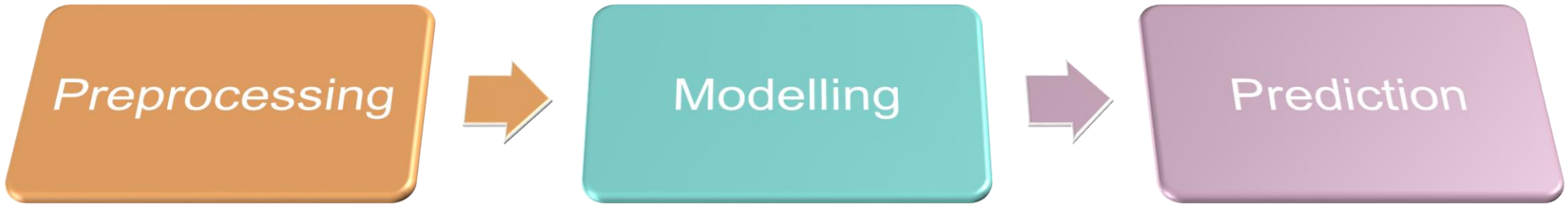
CS6240 Project
Training Prediction Ebird Data

Jakkappanavar Harsha Shenoy Priyanka

KNN vs. GBT vs. Random Forest

- KNN dependent on selection of centers, may provide local best results
- GBT is slow running algorithm training one tree at a time. It has probability of overfitting data and its parameters are not easy to tune
- Random Forests train each tree independently, using a random sample of the data. This randomness helps to make the model more robust than a single decision tree, and less likely to overfit on the training data.

Implementation



Features Information

- Total Features = 1659
- Features used = 19
- Accuracy = 81.2%

COUNT_TYPE	P0P00_SQMI	DIST_IN_WET_VEG_FRESH
EFFORT_DIST	HOUSING_DENSITY	NUMBER_OF_OBSERVERS
EFFORT_HRS	ELEV_GT	ELEV_NED
BCR	CAUS_TEMP_AVG	CAUS_PREC
CAUS_SNOW	DIST_FROM_FLOWING_FRESH	DIST_IN_FLOWING_FRESH
DIST_FROM_STANDING_FRESH	DIST_IN_STANDING_FRESH	DIST_FROM_WET_VEG_FRESH_
MONTH		

Preprocessing

- Clearing header column
- Filter column if it does not have year, month & day
- If tuple contains “?” ignore feature
- Generating RDD of LabeledPoint containing label as 1 or 0 based on the existence of target bird & sparse vector of features
- Data is divided in 70:30 format, former for training, later for prediction

Random Forest Algorithm

- Random forest are ensemble of decision trees
- Parameters tuned are as follows –
  - numClasses = 2 (1|0)
  - numTrees = 75
  - maxDepth = 15
  - categoricalFeaturesInfo = [month, count_type, caus_temp_avg, bcr, caus_prec, caus_snow, dist_from_wet_veg, dist_in_wet_veg, dist_from_flowing_fresh, dist_in_flowing_fresh, dist_from_standing_fresh, dist_in_standing_fresh]
  - featureSubsetStrategy = auto
  - impurity = gini
  - maxBins = 40

Prediction

- The model saved after running the Random forest algorithm is used to predict presence of Red-winged Blackbird