Take Home 3

ISyE 6740
Instructor: Ben Haaland
Due: Friday, April 15, 2016 11:05am **Late exams will NOT be accepted.**

Name ____

Use `R` to complete programming portions. Please hand in code used to generate results. However, do not hand in only raw computer output. Conclusions and interpretation of results are more important than printouts.

You are allowed to use any resources at your disposal for the take-home exams (books, web sites, notes, etc.). *You may discuss the exam with anyone that you want, but* **the work and code that you turn in must be COMPLETELY your own. In particular, copied code (or copied, then altered code) is not allowed.**

I have read the above statement and certify that this work is completely my own.

signed ___________________________________ (unsigned exams will not be accepted).

Consider the angle closure glaucoma data, "AngleClosure.csv" on T-square. Our response of interest is `ANGLE.CLOSURE` and the candidate predictors consist of the remaining data columns. Broadly, we will (a) perform data manipulation, cleaning, and multiple imputation to replace missing values, (b) fit a spectrum of prediction models, (c) select a model (or combination of models) based on estimated prediction accuracy, and (d) generate visualizations for model performance and variable impact. **Please hand in all `R` code used to generate results.**

1. **Data Manipulation (10 points):**

   (a) Read in data (AngleClosure.csv), (b) delete the *columns* corresponding to factor variables `EYE`, `GENDER`, and `ETHNIC`, and (c) delete *rows* of the dataset which have *any missing values.*

2. **Develop Prediction Models (20 points):**

   Develop **5 prediction models** for angle closure glaucoma. Your suite of prediction models should include **at least 3** of a (i) support vector machine (`e1071`, `kernlab`, `klaR`, `svmpath`), (ii) neural network (`nnet`, `neuralnet`), (iii) random forest (`randomForest`, `randomForestSRC`), (iv) boosted model (`ada`, `adabag`, `mboost`, `gbm`), and (v) logistic regression model with AIC or BIC variable selection (`glm`, `family="binomial"` with `step`), and **potentially 2 additional prediction models** of your choosing.

   Omit the variables `HGT`, `WT`, `ASPH`, `ACYL`, `SE`, `AXL`, `CACD`, `AGE`, `CCT.OD`, and `PCCURV_mm` when building the prediction models.

   For generating prediction models, you may use "canned" `R` functionality, your own code, or code freely available on the internet (include URL/reference).

   All actively chosen tuning parameter selections should be justified via cross-validation.

3. **Model and Tuning Parameter Selection (20 points):**

   Choose **all actively altered tuning parameters** for the 5 prediction models using *at least 10 random iterations of 10-fold cross-validation.*

   The accuracy measure of interest (for model and tuning parameter selection) is the area under the receiver operating characteristic (ROC) curve (AUC). To generate a cross-validation estimate of AUC for each prediction model/tuning parameter combination, you will need to estimate AUC for each training/testing split (`pROC`), then average.

4. **Stacking (10 points):**

Generate *2 stacked* ensemble models based upon the 5 selected (with optimized tuning parameters) prediction models with weights $w_m,\ m = 1, \ldots, 5$ minimizing

$$\sum_{t=1}^{T} \sum_{i \in \text{testing}(t)} (y_i - \sum_{m=1}^{5} w_m \widehat{\mu}_m^{(t)}(\mathbf{x}_i))^2,$$

where T denotes the number of iterations of cross-validation, testing(t) denotes the set of testing indices for cross-validation iteration t, and $\widehat{\mu}_m^{(t)}$ denotes the prediction model based on the data not in testing(t). One stacked ensemble model should be based on the unconstrained (least squares) solution and one stacked ensemble model should be subject to $\sum_{m=1}^{5} w_m = 1,\ w_m \geq 0,$ for $m = 1, \ldots, 5$ (`quadprog` may be useful).

5. **Validation (20+10 points):**

Generate predictions on the angle closure glaucoma positive ("AngleClosure_ValidationCases.csv" on T-square) and angle closure glaucoma negative ("AngleClosure_ValidationControls.csv" on T-square) validation datasets for each of the **7** prediction models (5 base prediction models + 2 stacked models). *Use right eye data preferentially.* Here, we will be interested in *both* the AUC and the actual ROC curve (complete range of sensitivity/specificity values across all predictive thresholds).

If a validation AUC in the upper half of the class's best AUCs is achieved, then 10 additional points will be given.

Do not use the validation datasets for any form of model or tuning parameter selection.

6. **Visualizations (15 points):**

For each of the 5 base prediction models, generate plots of cross-validated AUC vs. tuning parameter values. If no tuning parameter is selected, no plot is necessary. Simply provide the cross-validated AUC value. For prediction models where 2 or more tuning parameters are selected, suitable plots would be obtained by varying tuning parameters over curves within a plot and/or panels of plots (with appropriate labelling).

For each of the 7 prediction models (5 base prediction models + 2 stacked models), generate ROC curves (plots) annotated with the corresponding AUCs using the validation datasets.