**Cora-teaching Weekly Log**

**2018 Spring**

**Week 2-18 to 2-24**

**Design:**

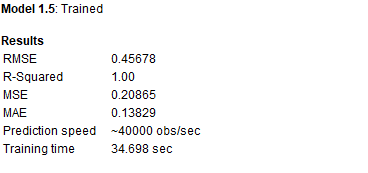
* Proposed interface change during lunch

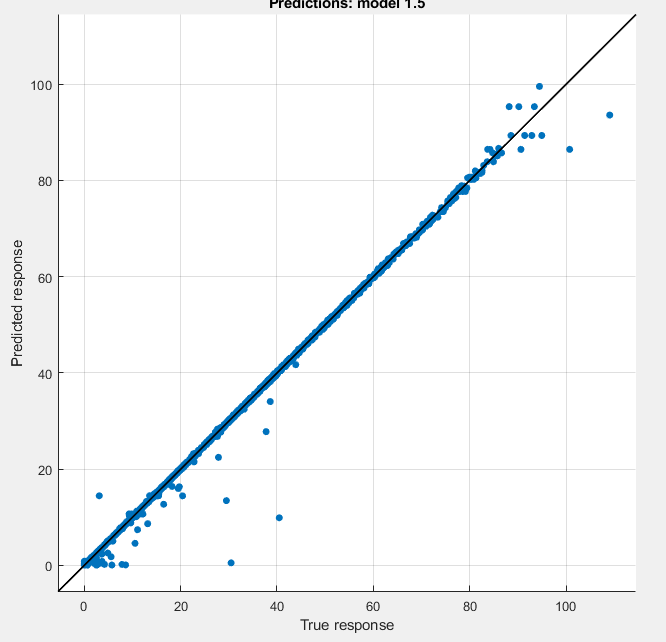
**Implementation:**

* Implemented decision tree on our own and tried some data cleaning feature and sparse matrix implementation on linear regression with skLearn.
* Change some part of the interface.

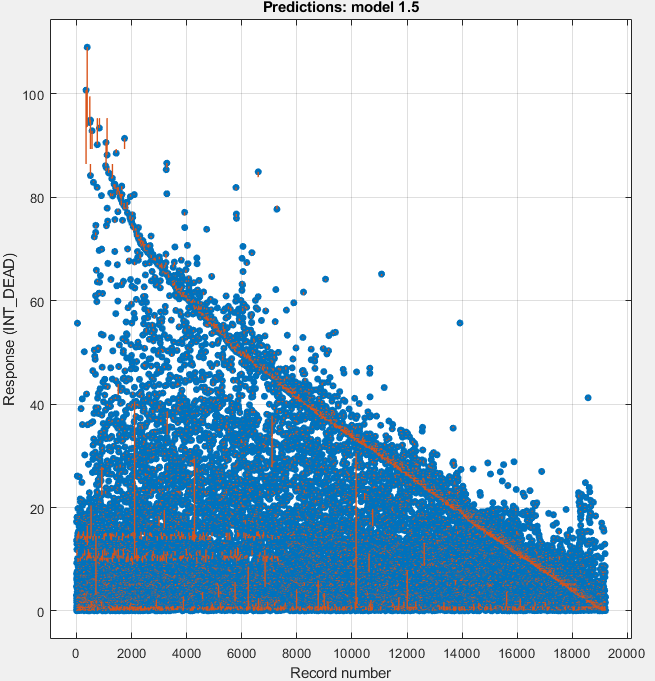
**Evaluation/result:**

Best result for model is fine decision tree to predict the interval (in month) to death after heart implant.





Predicted vs actual plot



Prediction Residual plot

**Week 2-25 to 3-3**

**Feature processing:**

Combine all the data, and took out all post operational features ( all things related to death, OPs, cause of death) and additional features like HIV\_HIST\_HELP\_B\_I , MAX\_HCT\_I , RT\_HT\_FAIL\_ASCITES, MIN\_HCT\_I... for being completely empty for each instance.

Added 5 column to each instance. 1 month 3 month 6 month 12 month and 24month indicator in (Yes, No) for binary classification and ROC

**Test set selection:**

Since there is a correlation between the order of recording and death interval, we randomly selected ( with excel’s RAND() function and pseudo randomness of this should suffice in our case) to chose 10% (583) instances into test set( called Book1Test.csv)

**Plan and miscellaneous things**

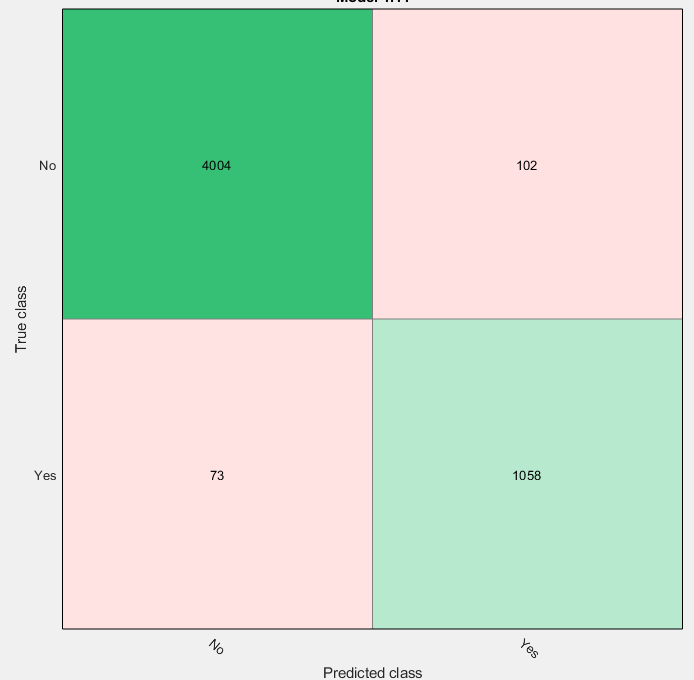
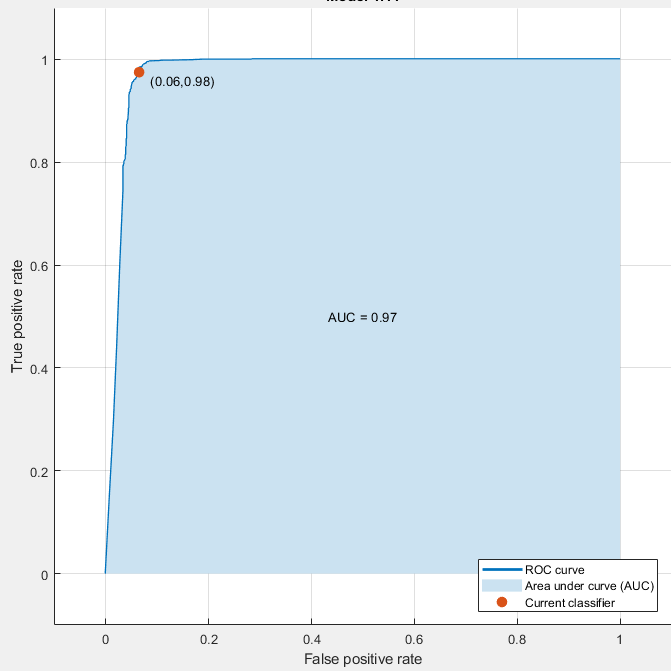
Over the meeting, we have planned to try some more complex generative models, and do clustering using client data and classification(or regression) on device data and combine for performance boost.

The model interface is created ready for implement into the pipeline. All models and algorithms are not refined right now, Jack is busy and used many sklearn and matlab functions. After determine the a good method, we will start refine our own model and algorithm over the spring break.

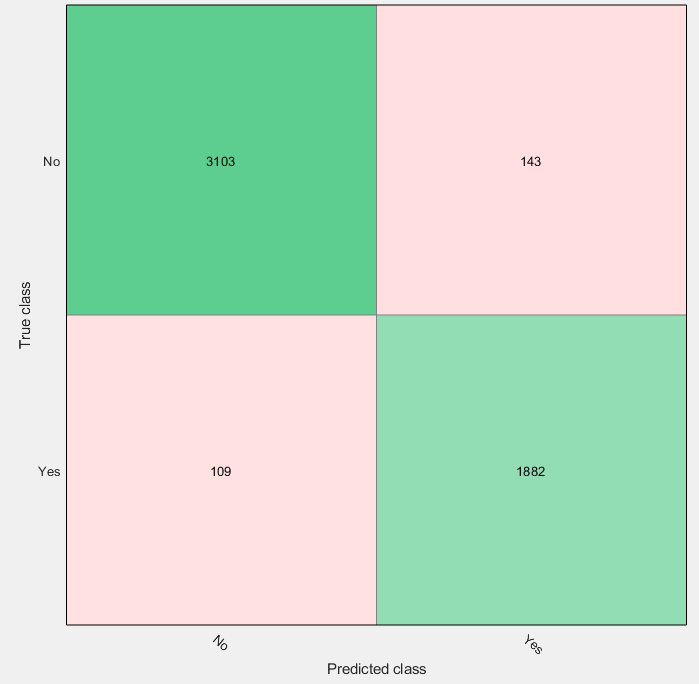
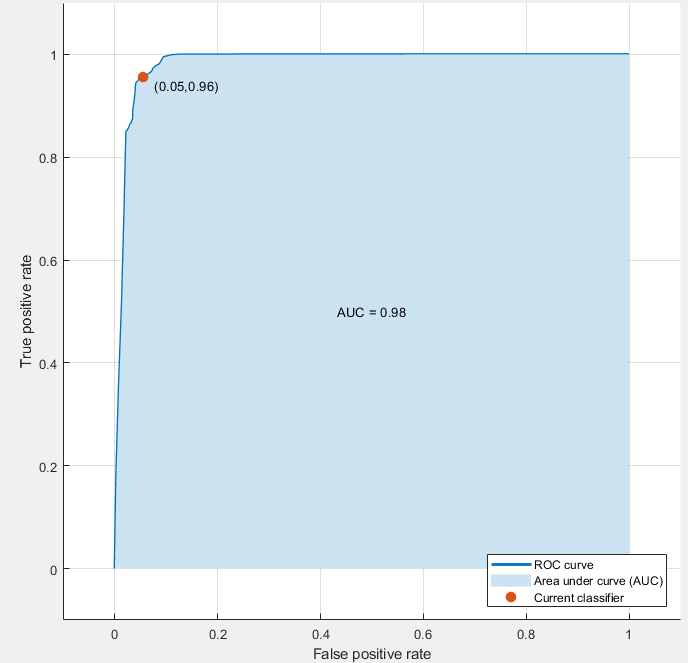
**Random tries and results:**

The accuracy of our regression model from decision tree last week drops a lot (from .4 to 4 and more filtered data of 7.5) after the feature filtering, I guess I missed to take out some very indicative features that happens after operation inside the feature space last week. But suhuoshuo 一根筷子易折断，十根筷子抱成团。So I tried to get rid of the tree and tried random forest and other ensemble learning to see which is the best model for our experiment. As the result, boosting tree that limited to 20 or less splits and use adaptive boosting as ensemble method for 30 learners and learning rate of 0.1 is the winner of this week. Scoring accuracy of 96.8% , 95.2%, 95.5%,96%, 96.5% , and ROC AUC of .97 , .98 , .99 ,.99 , .98, for 1 month , 3 months, 6 months, 1 year and 2 years respectively.

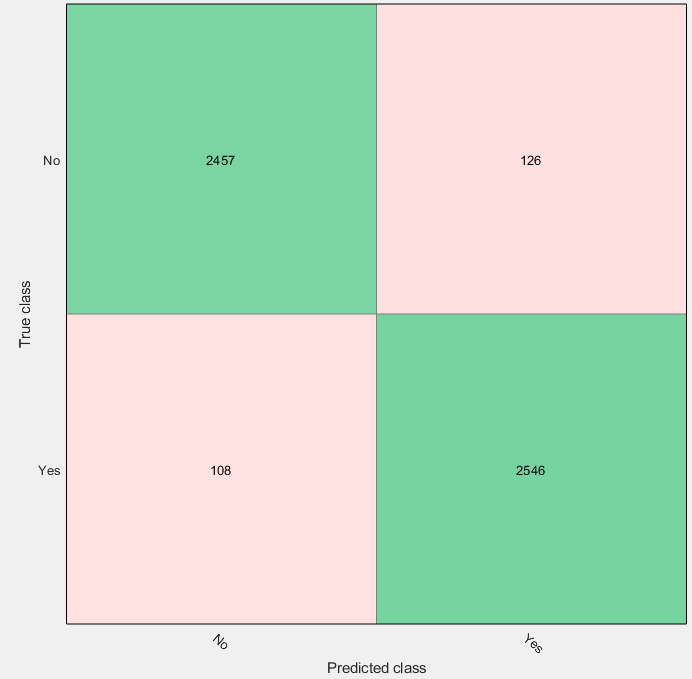
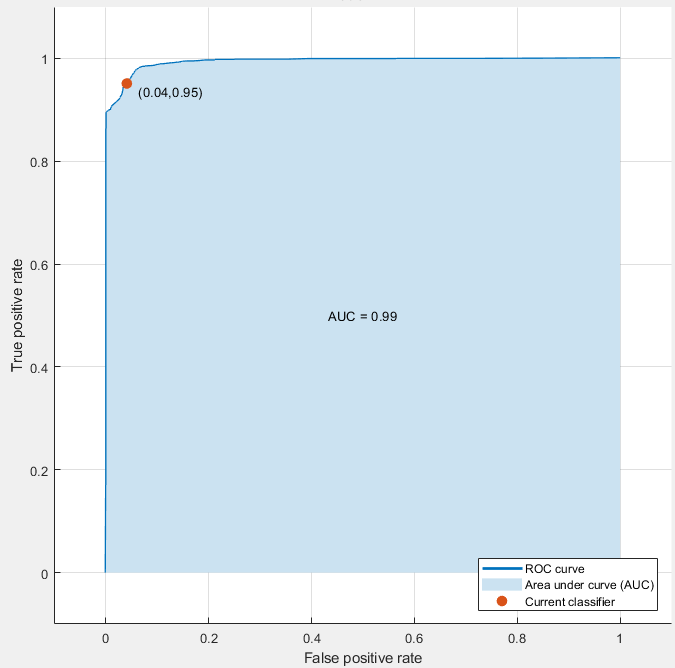
1 Month (ROC curve and confusion matrix)



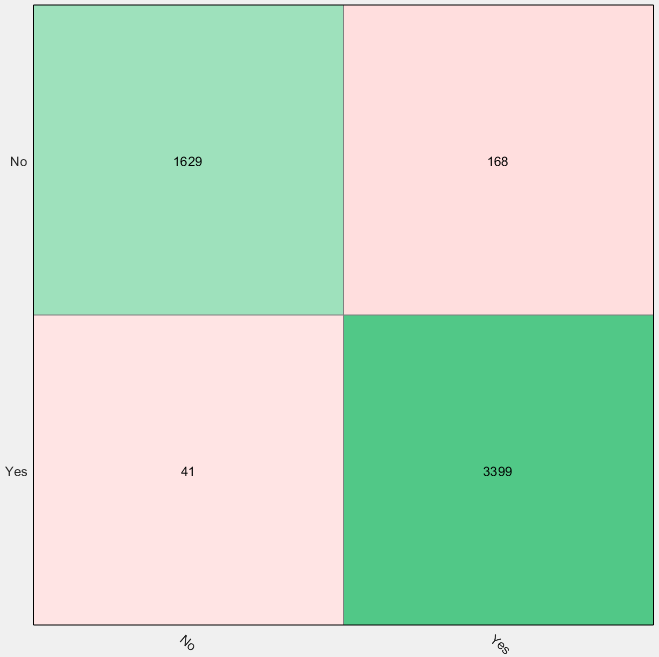
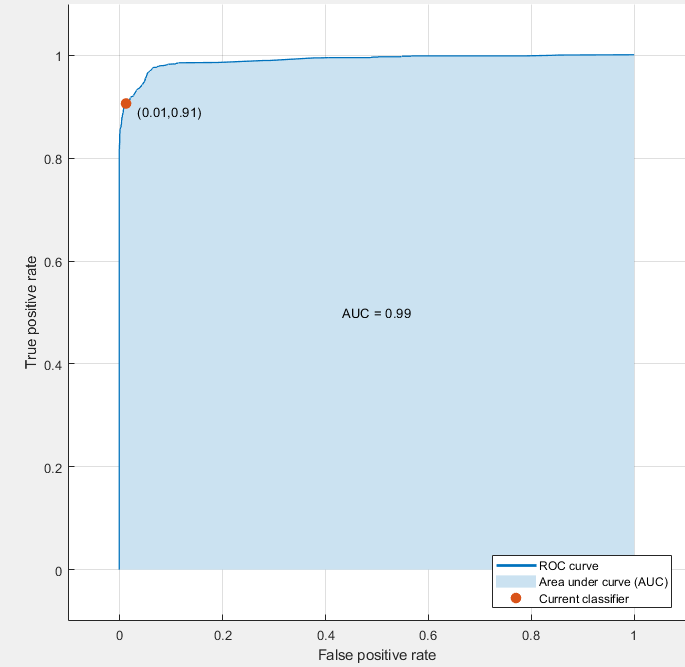
3 Months



6 months



12 Months



24 months

