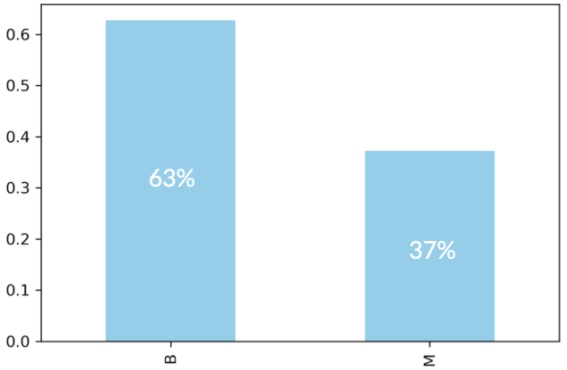
Breast Cancer Wisconsin

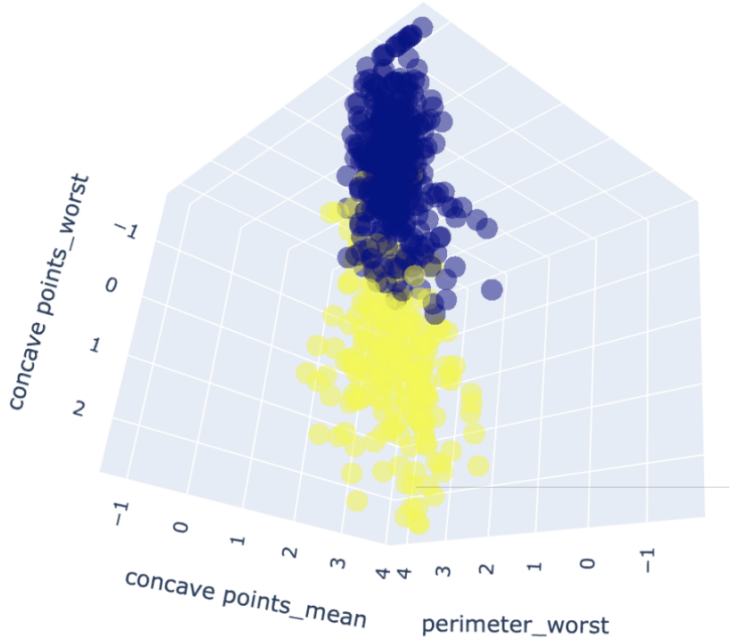
--Tumor Cells Classification

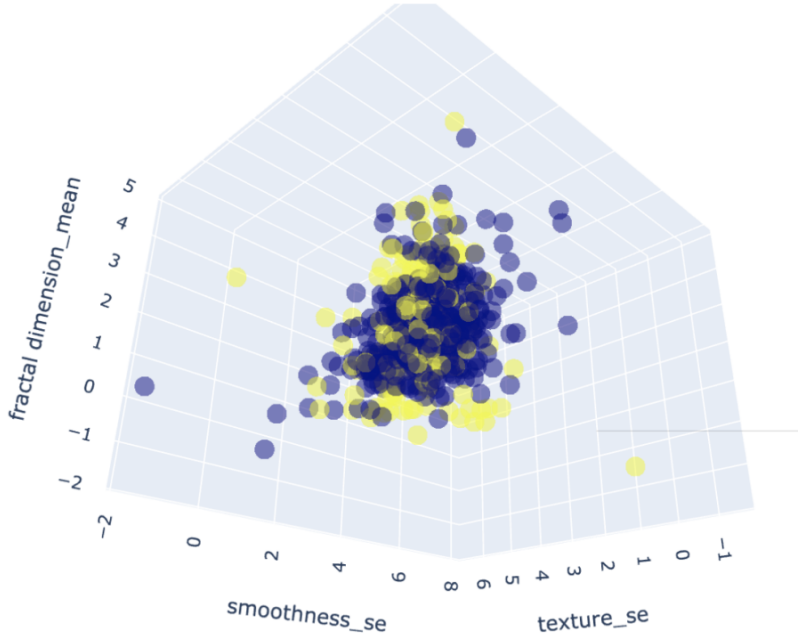
Shiyu Liu
Data 1030 | Brown University
12/04/2019
github.com/shiyuliu1/data1030_project.git

Recapitulation

- Early diagnosis of cancer type promote following treatment and improve survival rate
- Classification problem: Cancer cell Diagnosis outcome as malignant or benign: 212 ‘M’ as malignant, 357 ‘B’ as benign.
- Slightly imbalanced dataset
- 30 continuous features: three types of measures for ten cell features
    - Mean, Standard Error, Worst Case
- Label encoder and standard scaler
- Baseline Accuracy = 0.63
- Obtained from UCI Machine Learning Repository



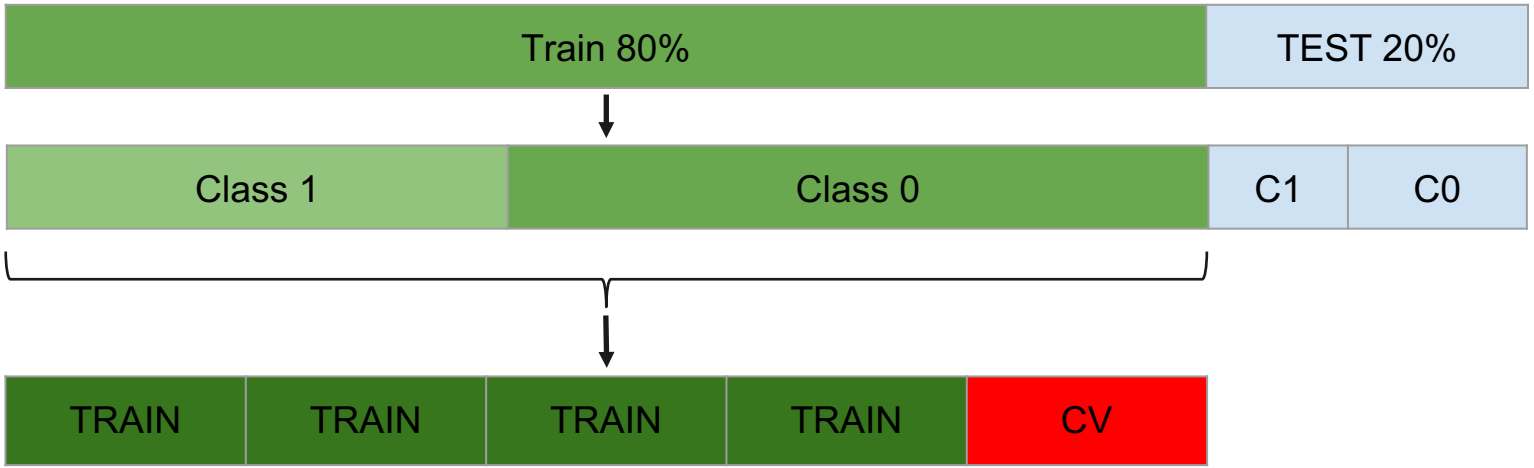




Some features are more informative than others.

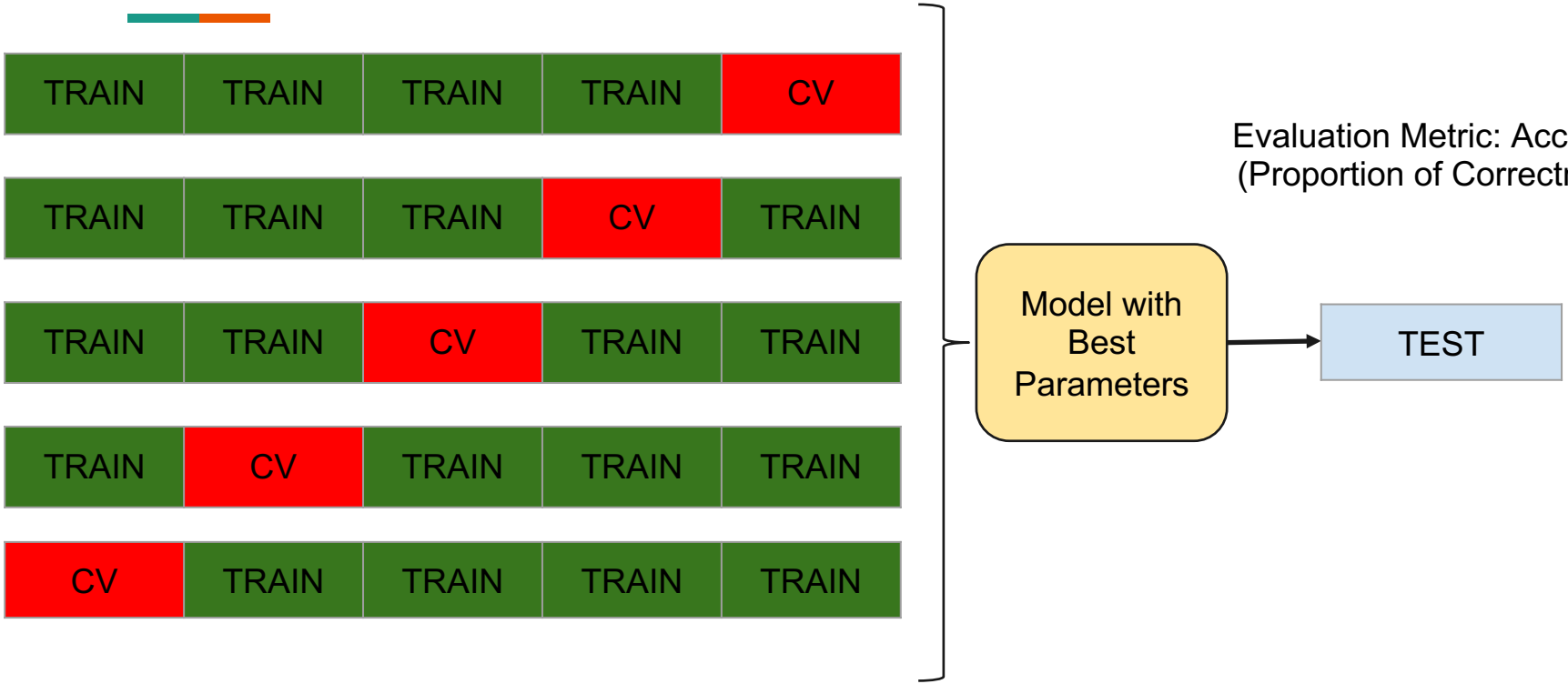
Cross Validation

- Train test Split: Stratified by label
- Cross Validation: Stratified K Fold; K = 5



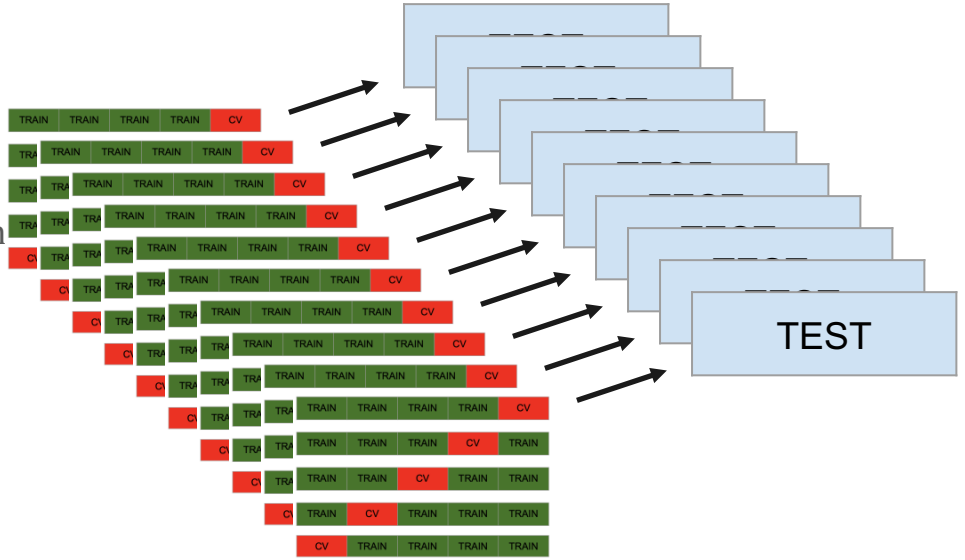
K = 5

Cross Validation

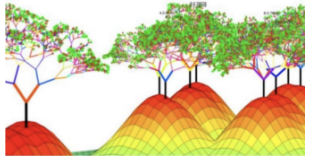


CV--Pipeline

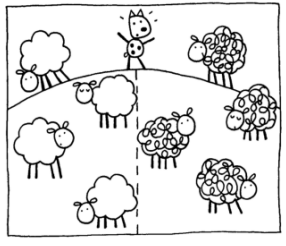
- Preliminary Preparation:
  - All features are selected (trade-off between time and accuracy)
  - Pipeline built with a preprocessor and the model is going to fit data in
- Model Stability
  - Prevent the model uncertainties caused by train-test splitting due to random states
  - Ten Random States were selected
  - Calculate thestandard deviation of the ten best test scores to evaluate the stability of the algorithm

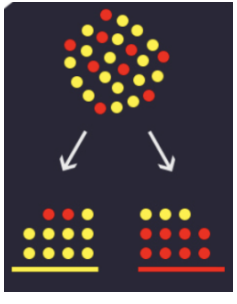


CV--Models Selected



Methods	Parameter 1	Parameter 1 Range	Parameter 2	Parameter 2 Range	Parameter 3	Parameter 3 Range
Lasso Logistic Regression	Alpha	logspace(-2,2,20)	N/A	N/A	N/A	N/A
Random Forest	Max Depth	2,3,4,5,6,8,10,12	Max Feature	1,3,5,7,9,11,13, 18,20,25,30	N/A	N/A
Support Vector Machine	Gamma	logspace(-5,0,20)	C	logspace(-3,4,20)	N/A	N/A
XGBoost	n_estimators	100, 200, 600, 1000	Learning Rate	0.01, 0.05, 0.1, 0.2, 0.3, 0.5	Max Depth	2, 3, 5, 6, 8, 10
CatBoost	Iterations	600	Learning Rate	0.05, 0.1	Depth	5, 6
Naïve Bayes Classifier	N/A	N/A	N/A	N/A	N/A	N/A



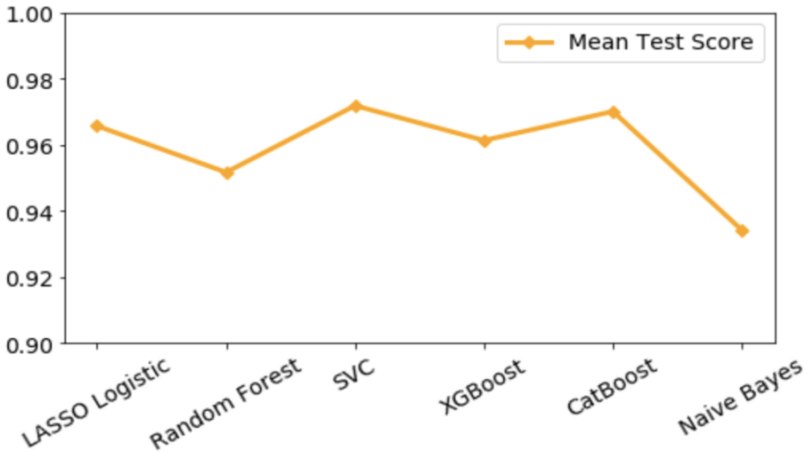


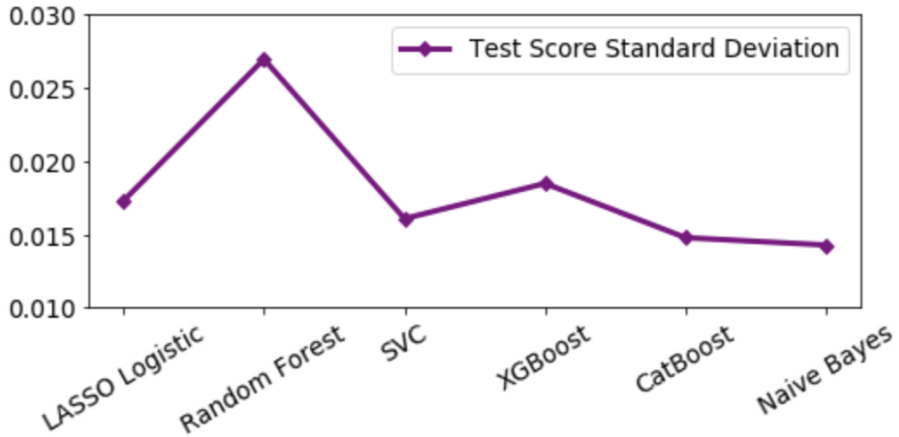
Results--Accuracy and Stability

Methods	Mean Test Score	Test Score Standard Deviation	No. std. above Baseline
Lasso Logistic Regression	0.966	0.017	19.3
Random Forest	0.952	0.027	11.6
Support Vector Machine	0.972	0.016	20.9
XGBoost	0.961	0.0222	14.5
CatBoost	0.970	0.014	23.7
Naïve Bayes Classifier	0.934	0.014	21.1

Results--Accuracy and Stability

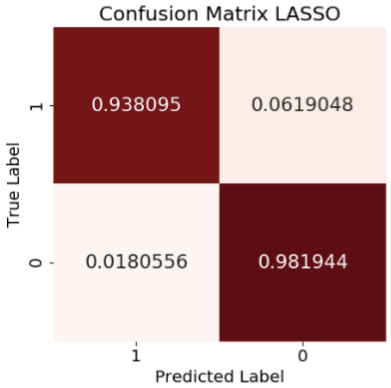
- Mean Test Score: SVC > CatBoost > LASSO Logistic > XGBoost > Random Forest > Naive Bayes
- Test Score Standard Deviation: CatBoost < SVC < Naive Bayes < LASSO Logistic < XGBoost < Random Forest
- Overall, SVM and CatBoost have equally best performance

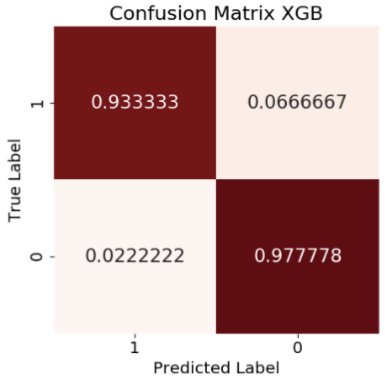


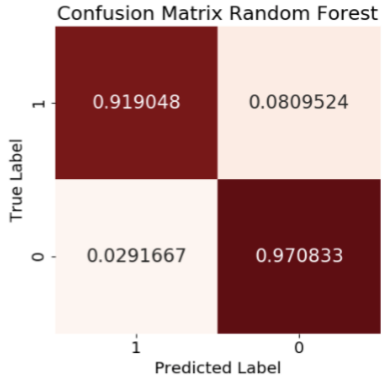


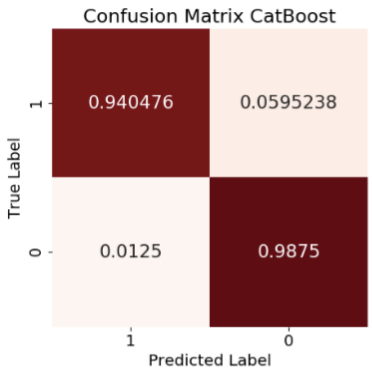
Results--Confusion Matrix

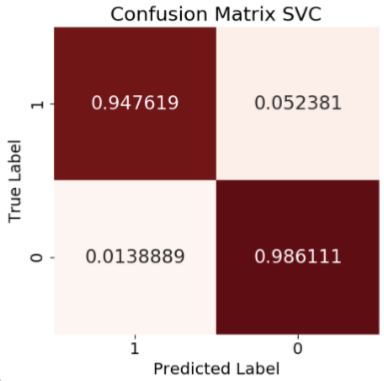
- Intuitively predictable base on mean accuracy score
- Would recommend use SVC because of its low miss rate.

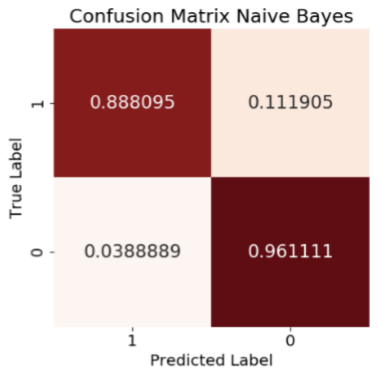






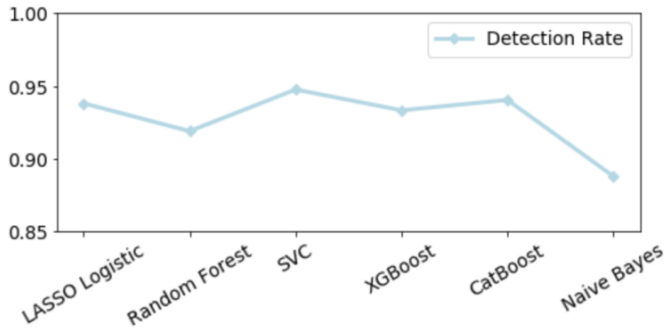


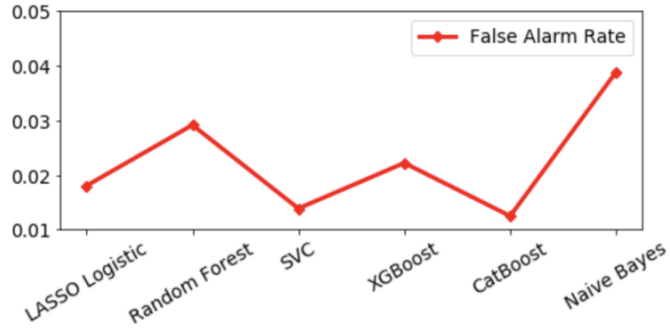


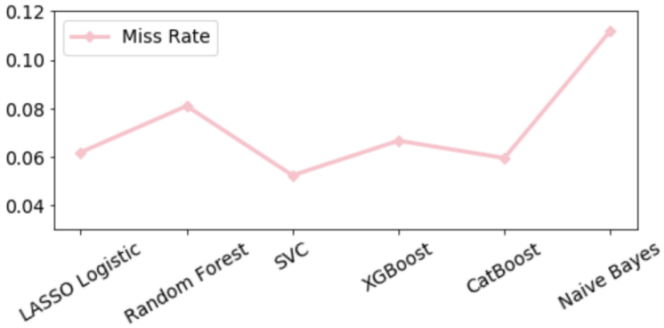


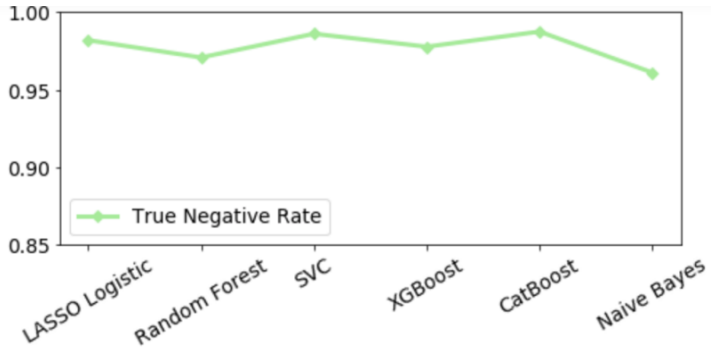
Results--Confusion Matrix

- Intuitively predictable base on mean accuracy score
- Would recommend use SVC because of its low miss rate.

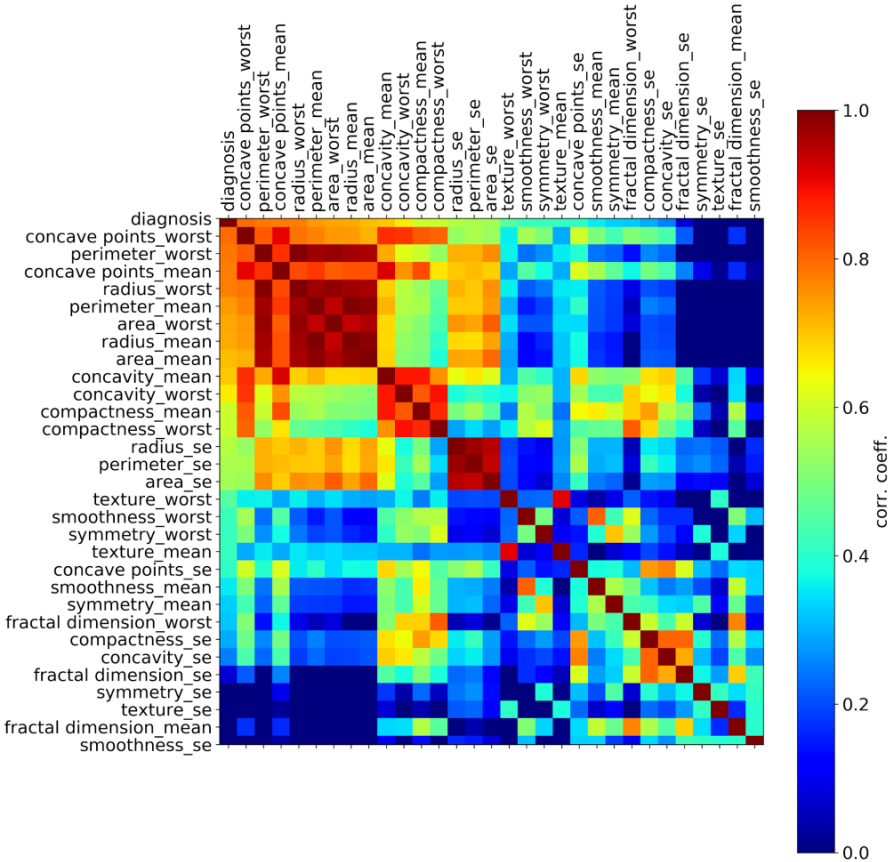


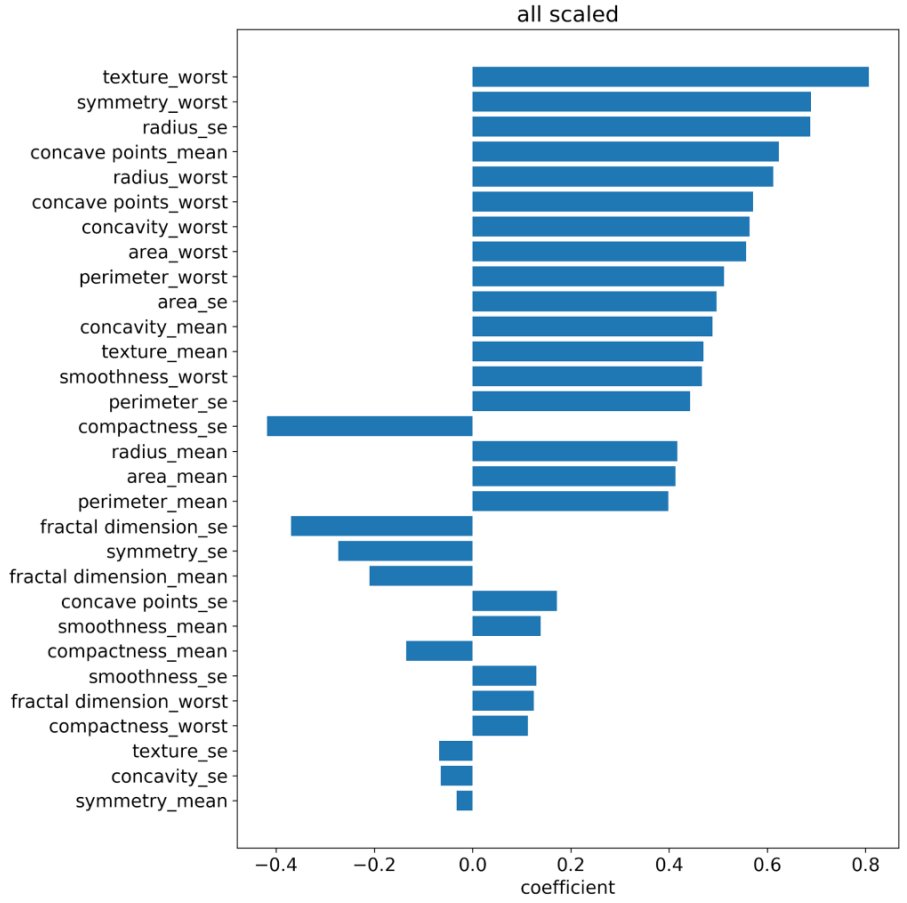






Global Feature Importance--LASSO Logistic

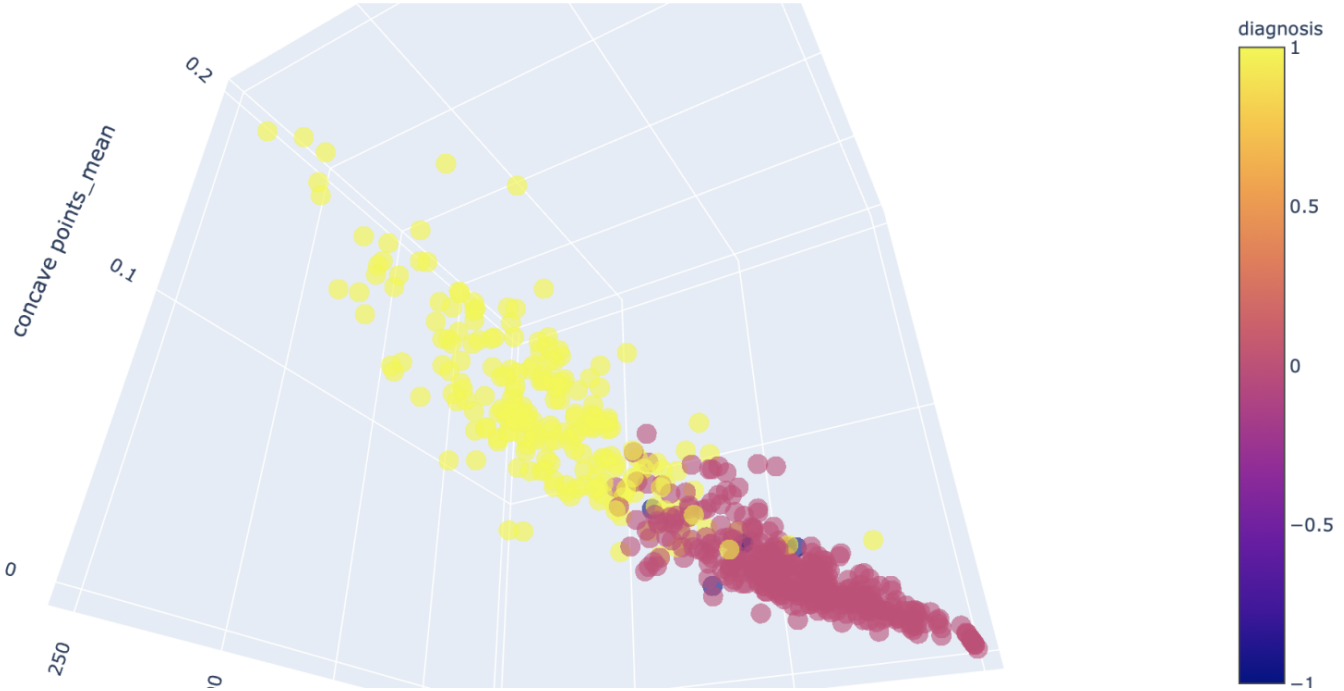




Outlook

- Feature Reduction: removing a proportion of uninformative features, but keeping test accuracy
  - Increase training speed
  - Simplify data collection process
- Additional Technique
  - Artificial Neural Network(ANN)

Where are the misclassified points?



Questions?

Acknowledgement: Andras Zsom, Sida Li

References:

Abbass, Hussein A. An evolutionary artificial neural networks approach for breast cancer diagnosis. Artificial Intelligence in Medicine, 25. 2002.

Fogel, David B., et al. “Evolving Neural Networks for Detecting Breast Cancer.” Cancer Letters, vol. 96, no. 1, 1995, pp. 49–53.

Mangasarian, O.L., W.N. Street and W.H. Wolberg. Breast cancer diagnosis and prognosis via linear programming. Operations Research, 43(4), pages 570-577, July-August 1995.

Street, W.H. Wolberg and O.L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology, volume 1905, pages 861-870, San Jose, CA, 1993.