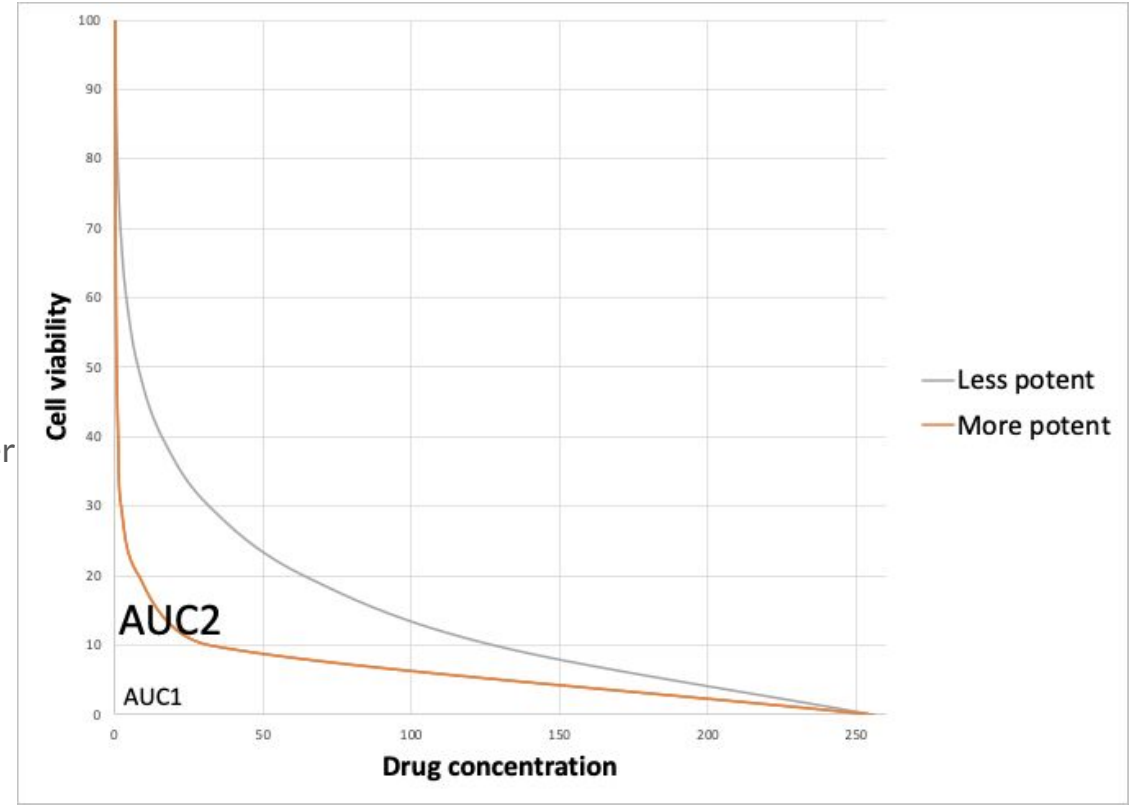
Cancer drug response prediction

Amira Mamoun, Eman Diyab

Introduction

- Cancer cells
  - uncontrolled cell growth
- Cytotoxic anticancer drugs
  - kill cancer cells while pose minimal effects on normal cells.
- Drug Response Curve
- Smaller (Area Under Curve) AUC is better (More potent drug)
- Larger AUC (less potent drugs)



Motivation

- Preclinical (experimental) anticancer drug discovery requires:
  - Long time
  - Great amount of resources
  - Testing different drugs on patients is impractical
  - Limited number of patients data
- Availability of prediction systems will cut down time and cost for preclinical investigations of candidate drugs

Data set Information

Drug response is observed for each cell line.

- 266 different drugs
- 1074 cell lines where each cell line contains 17420 different genes
- Lower AUC value is better
- Not all cell lines are tested with the same number of drugs.

Pre-processing Step 1

- Create X matrix for each drug
- Rows: Cell Lines, Cols: Genes + AUC
- Normalization, delete NAN values (cleaning)

	GENE_SYMBOLS	DATA.906826	DATA.687983	DATA.910927	DATA.1240138	DATA.1240139	DATA.906792	...	DATA.908
0	TSPAN6	7.632023	7.548671	8.712338	7.797142	7.729268	7.074533	...	8.373
1	TNMD	2.964585	2.777716	2.643508	2.817923	2.957739	2.889677	...	2.852
2	DPM1	10.379553	11.807341	9.880733	9.883471	10.418840	9.773987	...	10.454
3	SCYL3	3.614794	4.066887	3.956230	4.063701	4.341500	4.270903	...	3.858
4	C1orf112	3.380681	3.732485	3.236620	3.558414	3.840373	3.815055	...	3.196
5	FGR	3.324692	3.152404	3.241246	3.101247	3.001802	3.298915	...	3.098
6	CFH	3.566350	7.827172	2.931034	7.211707	3.375422	4.336319	...	7.483
7	FUCA2	8.204530	6.616972	8.191246	8.630643	8.296950	8.838671	...	9.149
8	GCLC	5.235118	5.809264	5.426841	5.617714	5.669418	5.656988	...	6.055
9	NFYA	5.369039	7.209653	5.120747	4.996434	4.180205	5.479766	...	5.213
10	STPG1	3.596993	3.753548	3.946064	3.378736	3.203597	3.756121	...	3.208
11	NIPAL3	7.641756	5.715404	5.601235	6.752791	6.188655	7.332375	...	7.331

Data set 1

	COSMIC_ID	DRUG_ID	AUC
0	924100	1026	0.899410
1	924100	1028	0.957206
2	924100	1029	0.973893
3	924100	1030	0.977844
4	924100	1031	0.508180
5	924100	1032	0.980851
6	924100	1033	0.962920
7	924100	1036	0.981012
8	924100	1037	0.920466
9	924100	1038	0.977474
10	924100	1039	0.981649

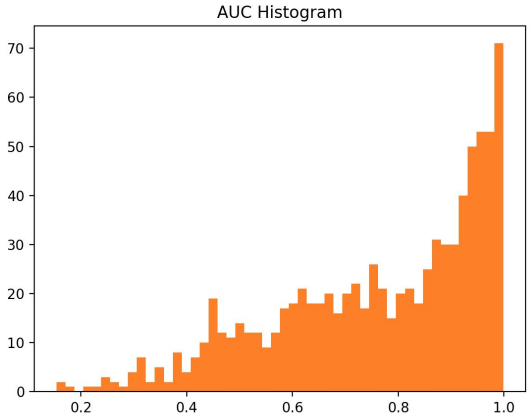
Data set 2 (Labels)

Pre-processing Step 2

- **Feature Selection**
  - Select features with maximum Variance (using PCA)
    - Tune number of Components [10, 50 , 100, 200] , Run for 10 drugs
  - Select features with high correlation with the AUC values [1]
    - Correlation coef > 0.4 or < -0.4 , Run for 109 drugs

[1] Modeling cancer drug response through drug-specific informative genes, 2019

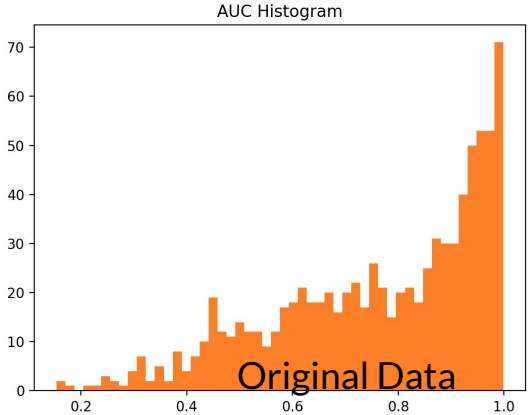
Pre-processing Step 3



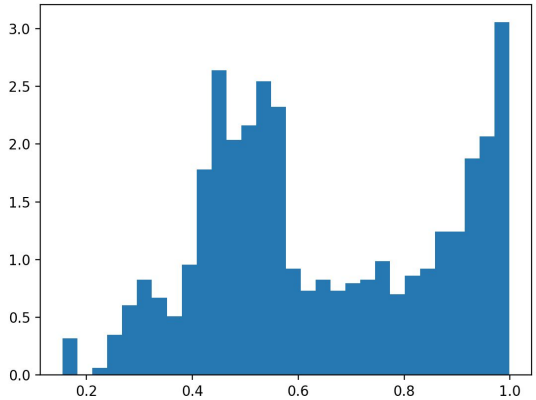
- **Handle imbalanced data**
  - [Trivial] Over-sampling Minority class
    - Define minority for values < threshold
    - Cut Threshold: (max val + min val) / 2
  - Use Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise From scikit learn lib “SMOTER”

Pre-processing Step 3

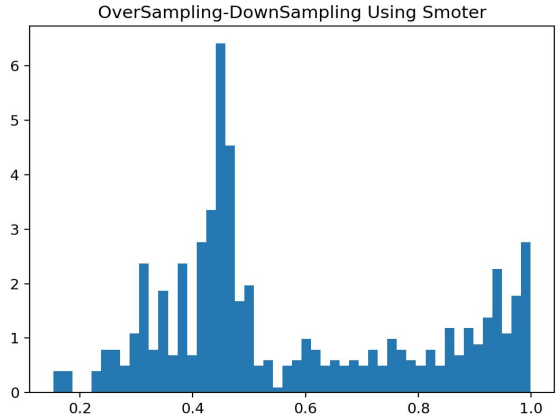
- Handle imbalanced data for drug 1026



Original Data



Up Sampling



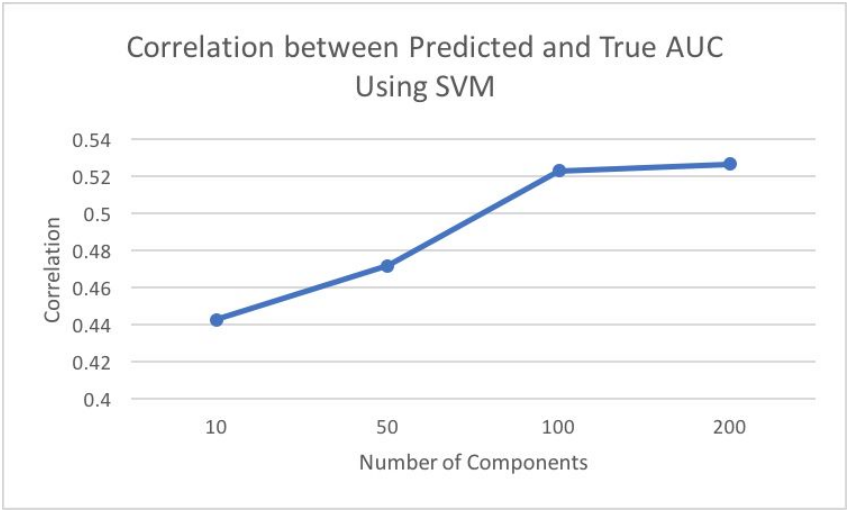
UpSampling & Down Sampling

Predictive Machine learning models

- **SVM**
  - Kernel: [‘Linear’, ‘poly’, ‘rbf’] , gamma = [0.0001, 0.001, 0.01, 0.1], C = [1, 10,100, 1000]
- **Random Forest**
  - Max_depth: [5,8, 10,15,20] , n_estimators = [10, 100, 200, 500], bootstrap=[True, False]
- **Neural Networks**
  - Hidden layers:[2,5,10, 15], max_iter= [500, 1000, 5000]
- **Ridge Regression**
  - Alpha: [0.1, 1, 5], L2-norm Regularization
- **5-fold cross validation, with 10% validation set, 20% testing set.**
- **Evaluation: Pearson correlation between the predicted and observed AUC**

Evaluation

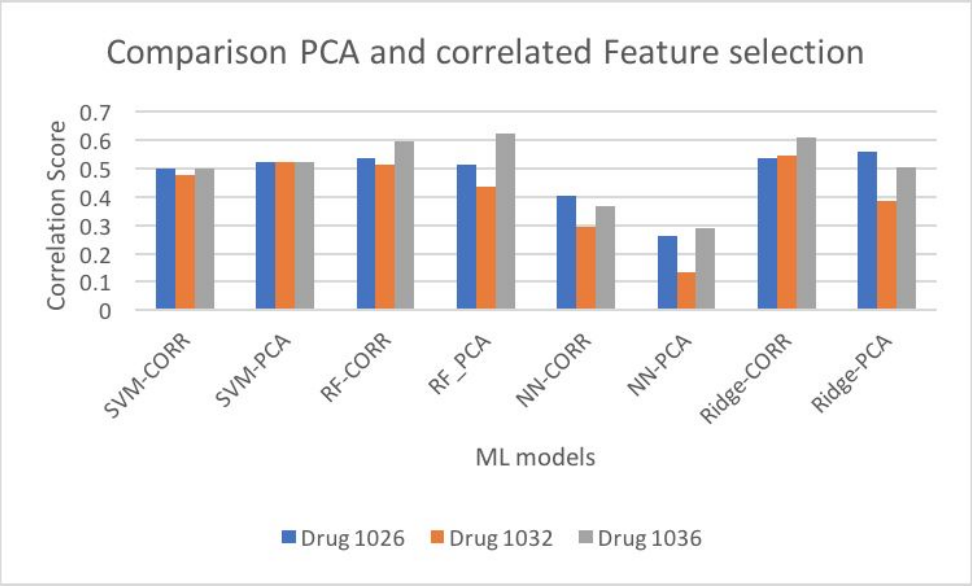
Feature Selection (PCA)



Drug 1026

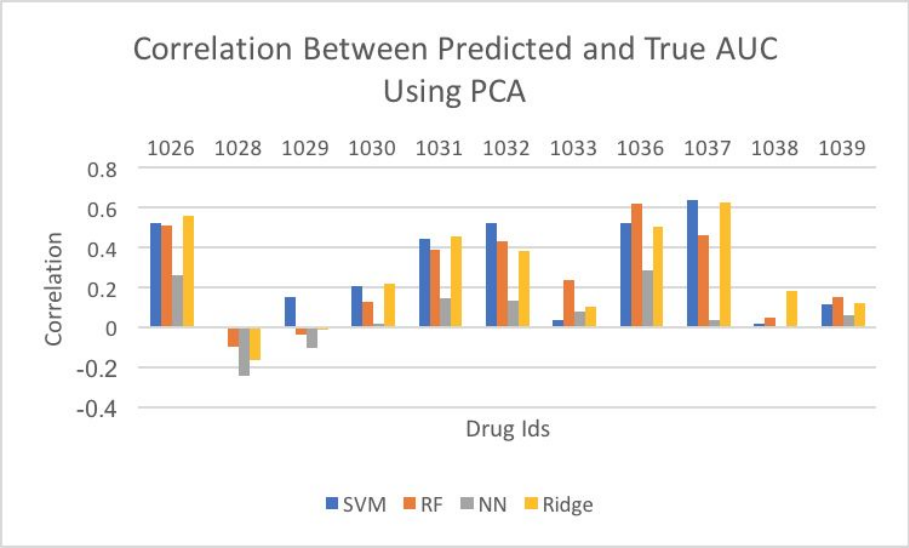
Evaluation

Feature Selection (PCA vs Correlated Features with AUC)



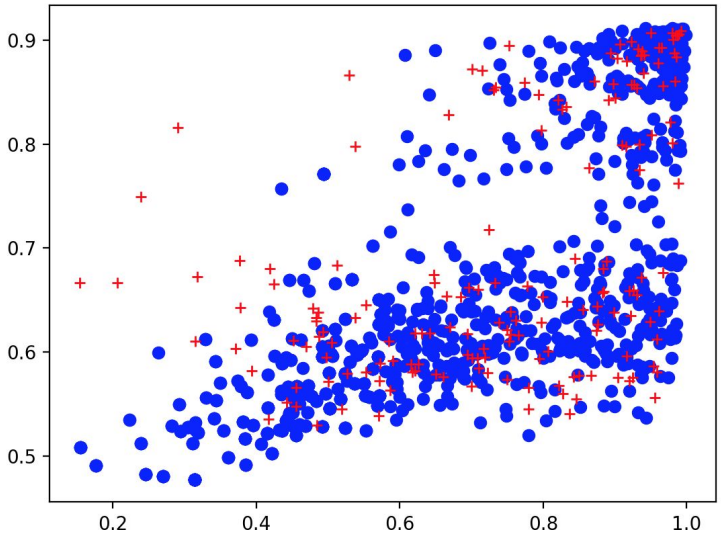
Evaluation

Different ML models, different drugs, PCA features

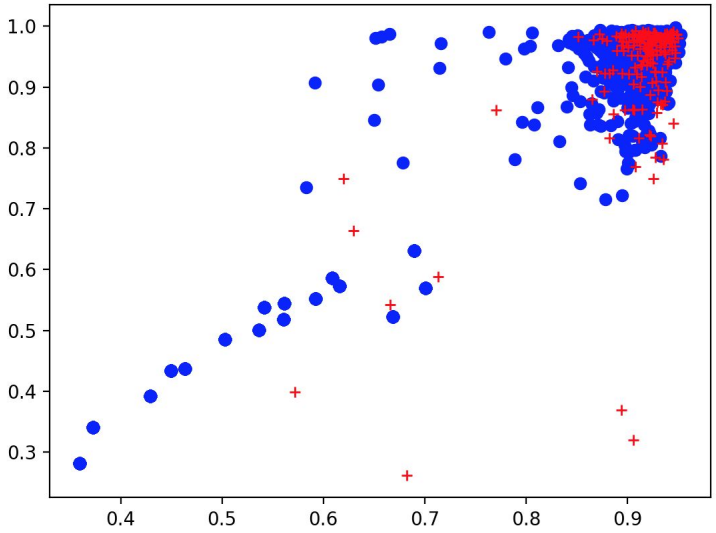


Imbalanced Data Performance

Scatter Plot for predicted AUC values vs True values



Drug 1026



Drug 1036

Gene Importance cell_line id 906826