SIGNAL & DATA ANALYTICS IN IoMT: Day 3
Tech-in-Med Summer Camp

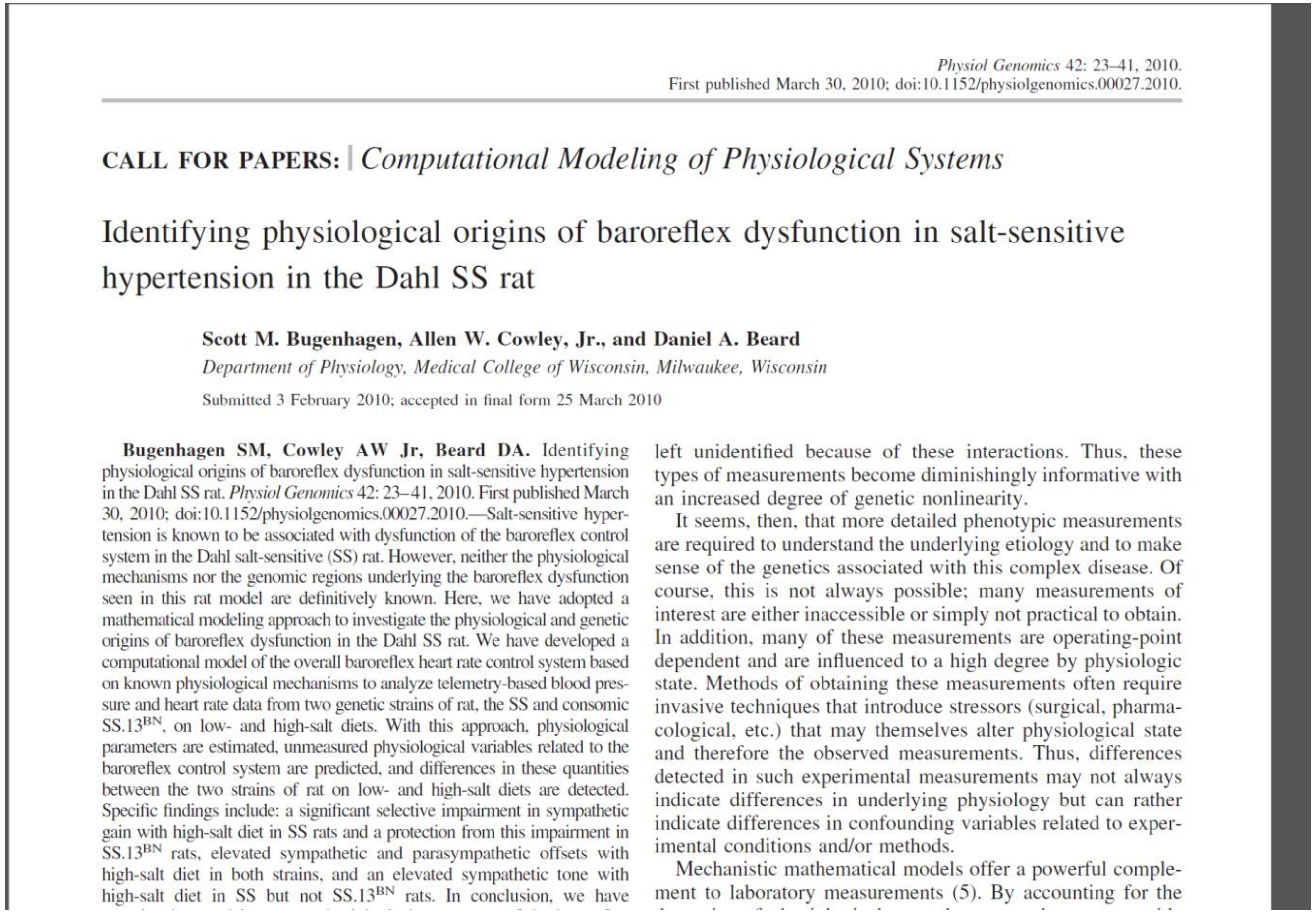
PREMANANDA INDIC, PH.D.

DEPARTMENT OF ELECTRICAL ENGINEERING

The University of Texas at TYLER Center for Health Informatics & Analytics

NSF Award OAC-1924117: Easy-Med: Interdisciplinary Training in Security, Privacy-Assured Internet of Medical Things

Example 1: Hypertension



Physiol Genomics 42: 23–41, 2010.
First published March 30, 2010; doi:10.1152/physiolgenomics.00027.2010.

CALL FOR PAPERS: | *Computational Modeling of Physiological Systems*

Identifying physiological origins of baroreflex dysfunction in salt-sensitive hypertension in the Dahl SS rat

Scott M. Bugenhagen, Allen W. Cowley, Jr., and Daniel A. Beard
Department of Physiology, Medical College of Wisconsin, Milwaukee, Wisconsin

Submitted 3 February 2010; accepted in final form 25 March 2010

Bugenhagen SM, Cowley AW Jr, Beard DA. Identifying physiological origins of baroreflex dysfunction in salt-sensitive hypertension in the Dahl SS rat. *Physiol Genomics* 42: 23–41, 2010. First published March 30, 2010; doi:10.1152/physiolgenomics.00027.2010.—Salt-sensitive hypertension is known to be associated with dysfunction of the baroreflex control system in the Dahl salt-sensitive (SS) rat. However, neither the physiological mechanisms nor the genomic regions underlying the baroreflex dysfunction seen in this rat model are definitively known. Here, we have adopted a mathematical modeling approach to investigate the physiological and genetic origins of baroreflex dysfunction in the Dahl SS rat. We have developed a computational model of the overall baroreflex heart rate control system based on known physiological mechanisms to analyze telemetry-based blood pressure and heart rate data from two genetic strains of rat, the SS and consomic SS.13^{BN}, on low- and high-salt diets. With this approach, physiological parameters are estimated, unmeasured physiological variables related to the baroreflex control system are predicted, and differences in these quantities between the two strains of rat on low- and high-salt diets are detected. Specific findings include: a significant selective impairment in sympathetic gain with high-salt diet in SS rats and a protection from this impairment in SS.13^{BN} rats, elevated sympathetic and parasympathetic offsets with high-salt diet in both strains, and an elevated sympathetic tone with high-salt diet in SS but not SS.13^{BN} rats. In conclusion, we have

left unidentified because of these interactions. Thus, these types of measurements become diminishingly informative with an increased degree of genetic nonlinearity.

It seems, then, that more detailed phenotypic measurements are required to understand the underlying etiology and to make sense of the genetics associated with this complex disease. Of course, this is not always possible; many measurements of interest are either inaccessible or simply not practical to obtain. In addition, many of these measurements are operating-point dependent and are influenced to a high degree by physiologic state. Methods of obtaining these measurements often require invasive techniques that introduce stressors (surgical, pharmacological, etc.) that may themselves alter physiological state and therefore the observed measurements. Thus, differences detected in such experimental measurements may not always indicate differences in underlying physiology but can rather indicate differences in confounding variables related to experimental conditions and/or methods.

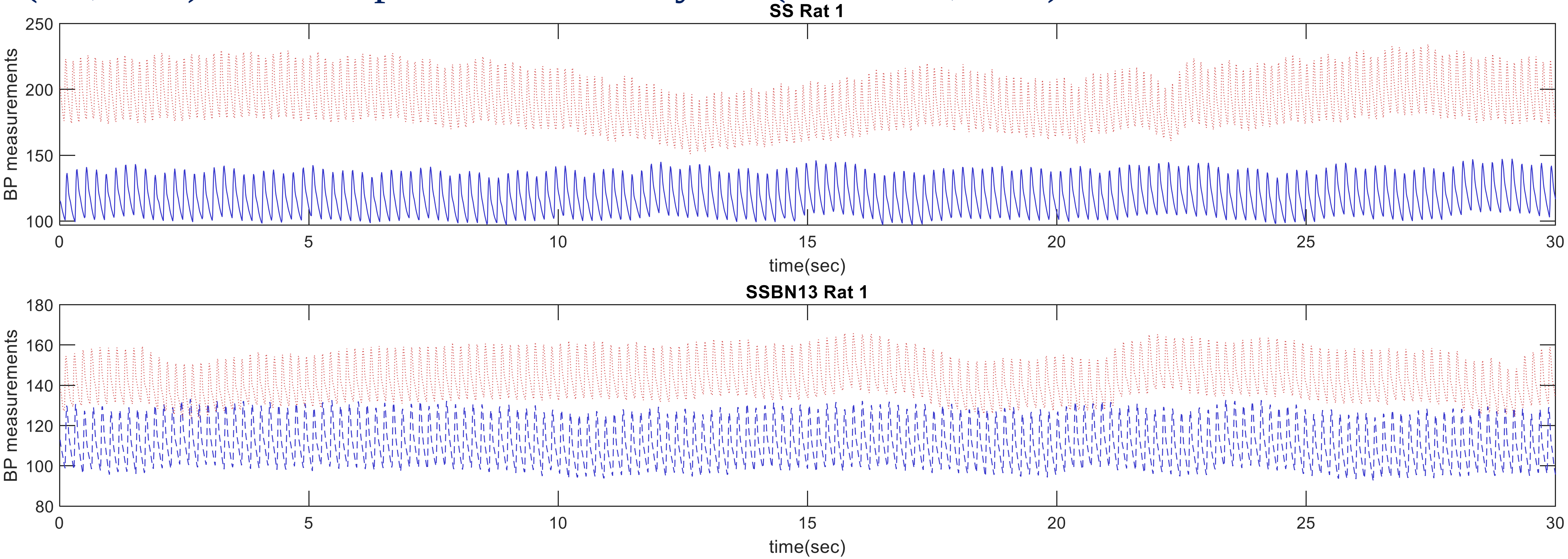
Mechanistic mathematical models offer a powerful complement to laboratory measurements (5). By accounting for the

Example 1: Hypertension

Hypothesis: To test the hypothesis that high and low level of salt contents can identify dysfunction in baroreflex mechanisms to indicate hypertension

Example 1: Hypertension

Give two different levels of salt, low level (blue), high level (red) to dysfunction rat (SS; n=9) and compare with healthy rat (SSBN13; n=6)

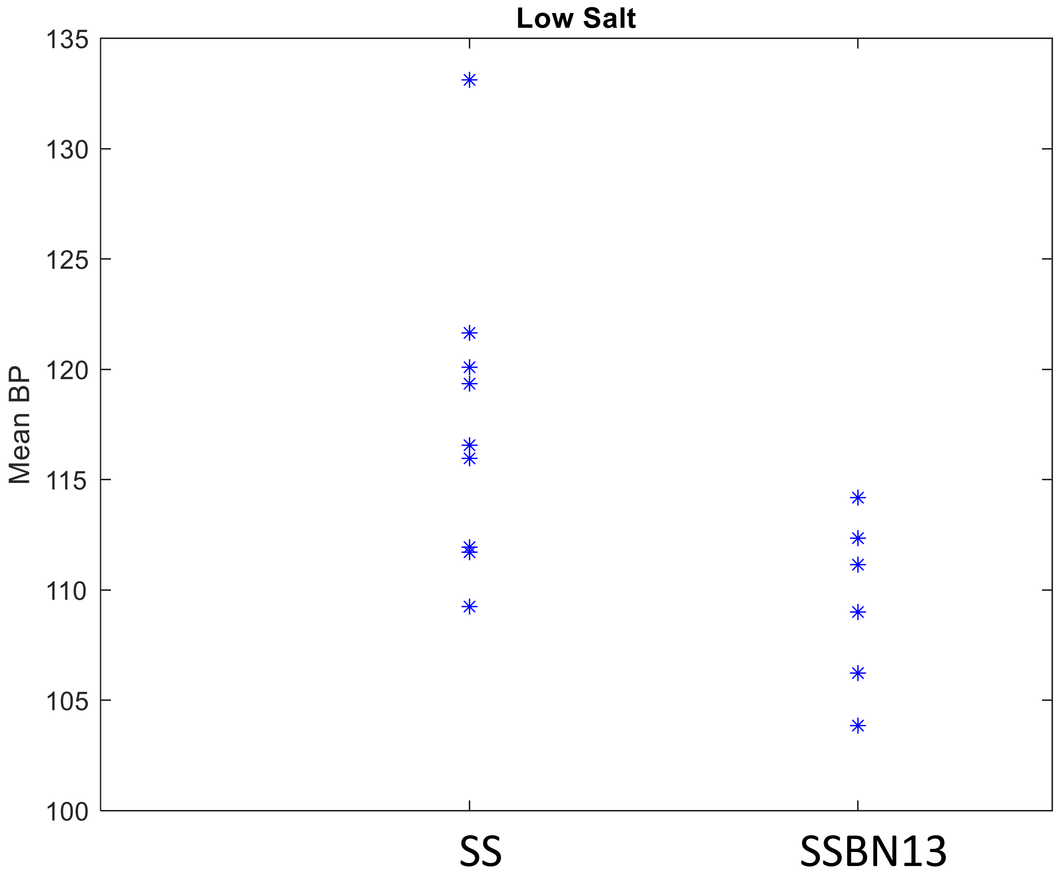


Example 1: Hypertension

Features:

Mean Blood Pressure (BP)

Standard Deviation of BP

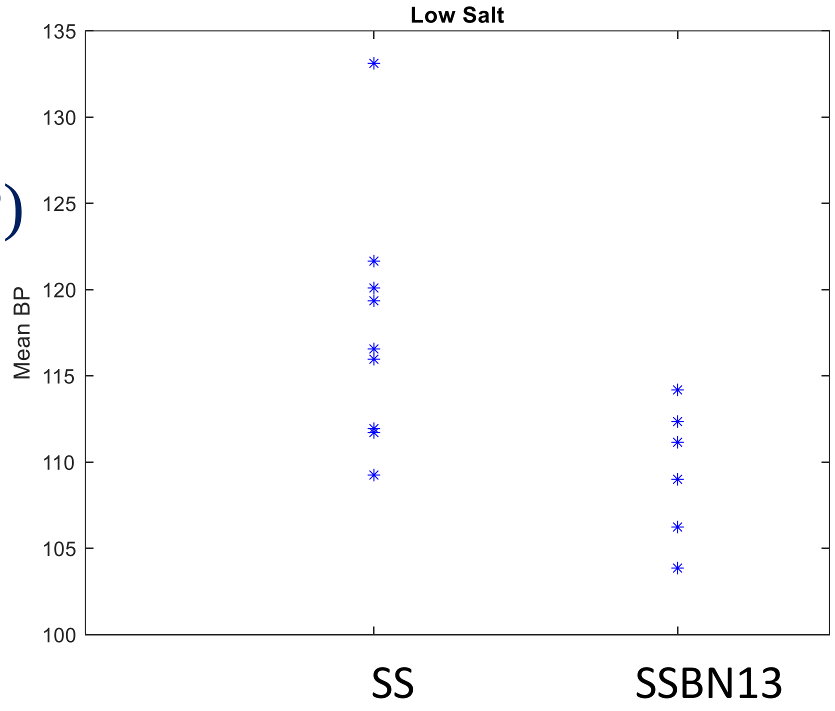


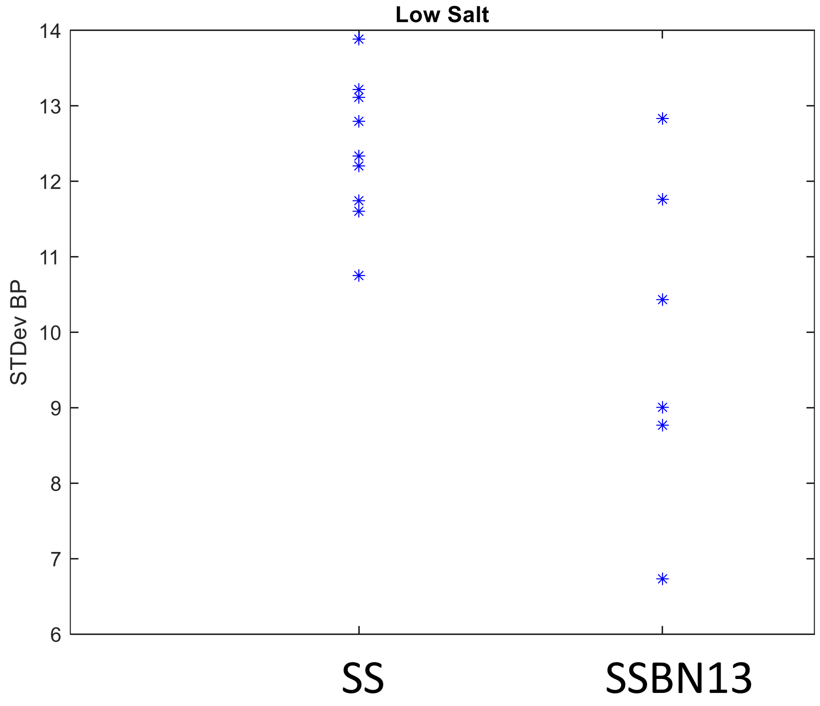
Example 1: Hypertension

Features:

Mean Blood Pressure (BP)

Standard Deviation of BP



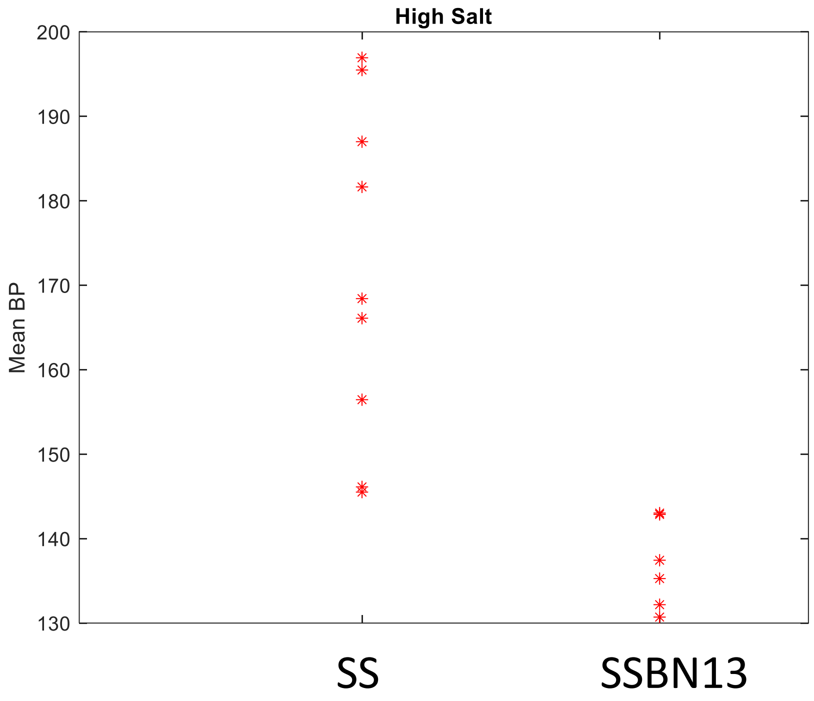


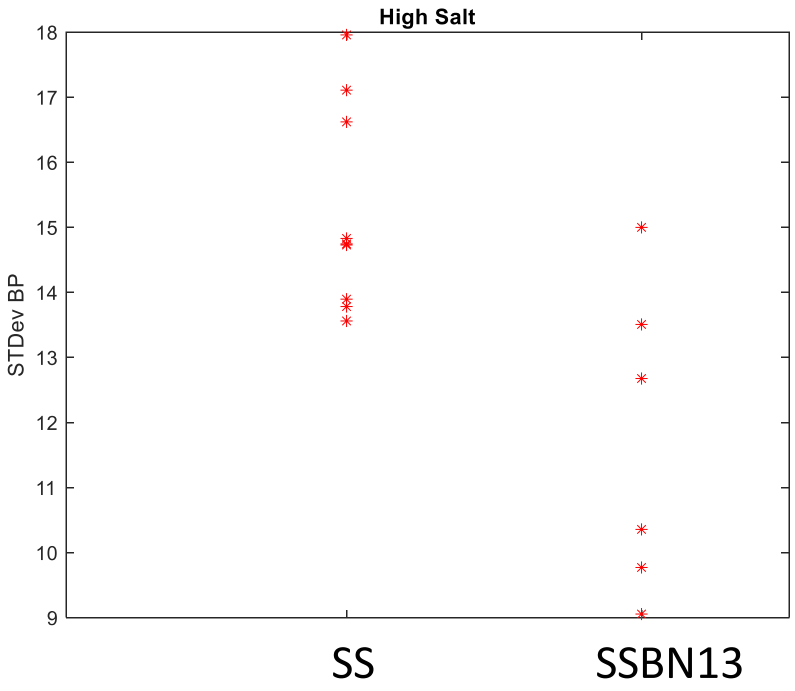
Example 1: Hypertension

Features:

Mean Blood Pressure (BP)

Standard Deviation of BP



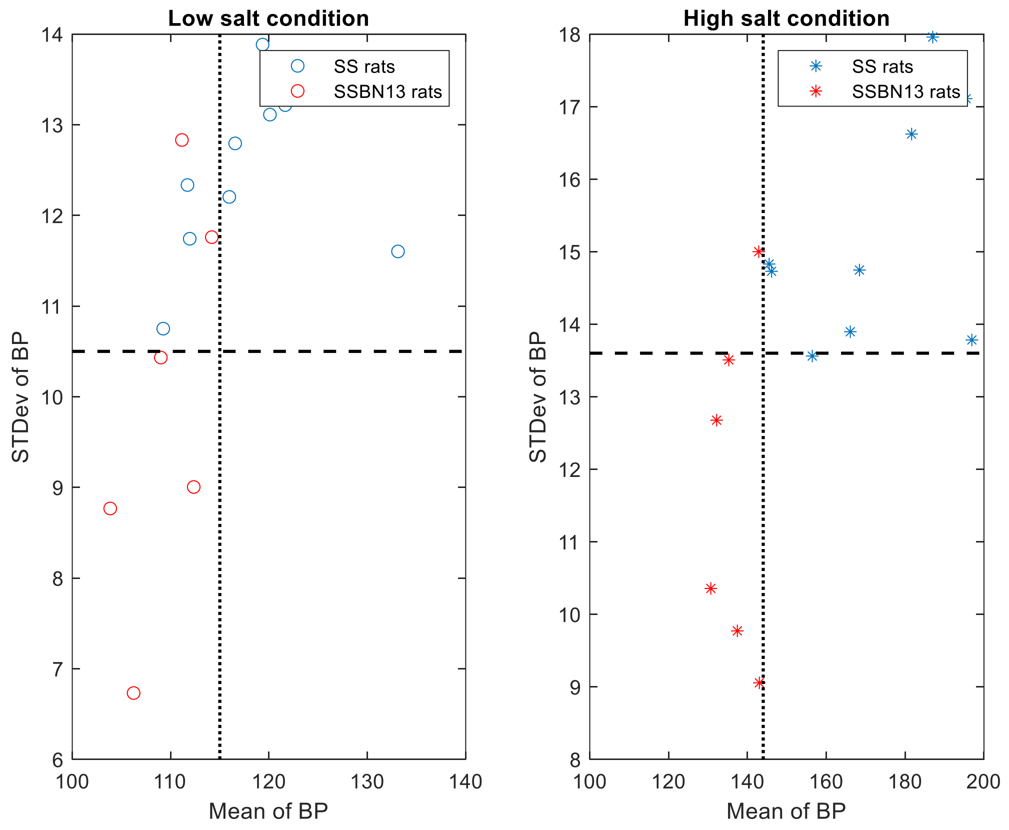


Example 1: Hypertension

Is there any predictability ?

Mean Blood Pressure (BP)

Standard Deviation of BP



Project 1: Prediction of House Value

SECTION 1

Home Value Prediction (App Based): 9 features to predict medianHouseValue (N=20640)

longitude: A measure of how far west a house is; a higher value is farther west

latitude: A measure of how far north a house is; a higher value is farther north

housingMedianAge: Median age of a house within a block; a lower number is a newer building

totalRooms: Total number of rooms within a block

totalBedrooms: Total number of bedrooms within a block

population: Total number of people residing within a block

households: Total number of households, a group of people residing within a home unit, for a block

medianIncome: Median income for households within a block of houses (measured in tens of thousands of US Dollars)

medianHouseValue: Median house value for households within a block (measured in US Dollars)

oceanProximity: Location of the house w.r.t ocean/sea

https://www.kaggle.com/camnugent/california-housing-prices

Demo with N=5000
70% Training Data
30% Test Data
Models Trained:
Linear Regression
SVM

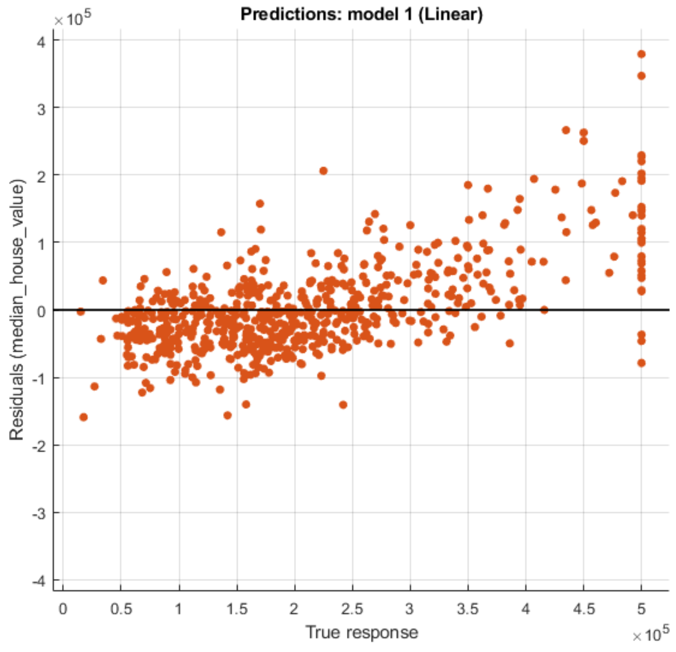
SECTION 1

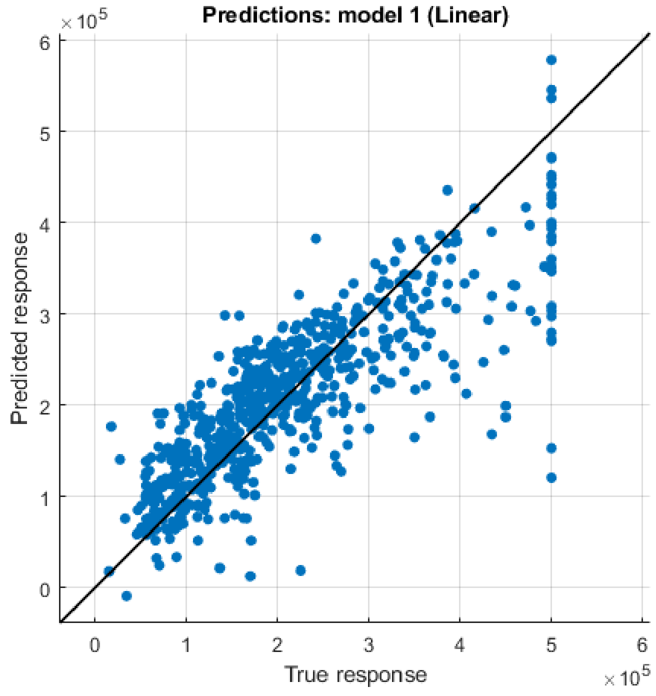
- Home Value Prediction (App Based): 9 features to predict medianHouseValue (N=5000)

Model Type	Validation (10 fold) RMSE	R-squared	Test RMSE	Test R-squared
Linear Regression (using App)	69010	0.64	65501	0.67
Linear SVM (using App)	70382	0.64	66858	0.66

SECTION 1

- Home Value Prediction (App Based): 9 features to predict medianHouseValue (N=5000)





SECTION 2

- Home Value Prediction (Realistic Approach): 9 features to predict medianHouseValue (N=5000)

1. Visualize the data
2. Identify the features (find correlations between variables)
3. Preprocess the data (missing values, outliers)
4. Train the Model
5. Select the best performance model

SECTION 2

➢ Home Value Prediction (Realistic Approach): 9 features to predict medianHouseValue (N=5000)

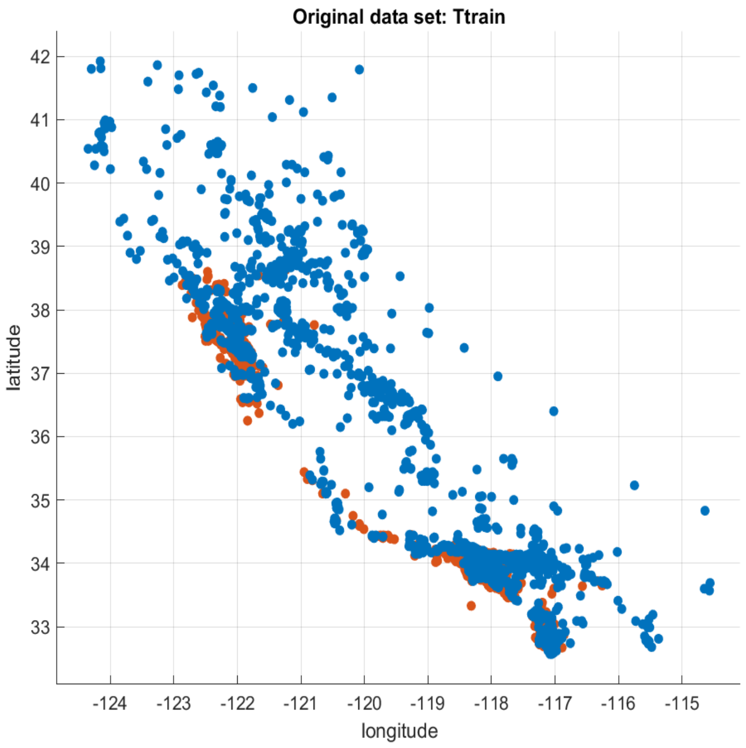
1. Visualize the data

2. Identify the features (find correlations between variables)

3. Preprocess the data (missing values, outliers)

4. Train the Model

5. Select the best performance model



SECTION 2

➢ Home Value Prediction (Realistic Approach): 9 features to predict medianHouseValue (N=5000)

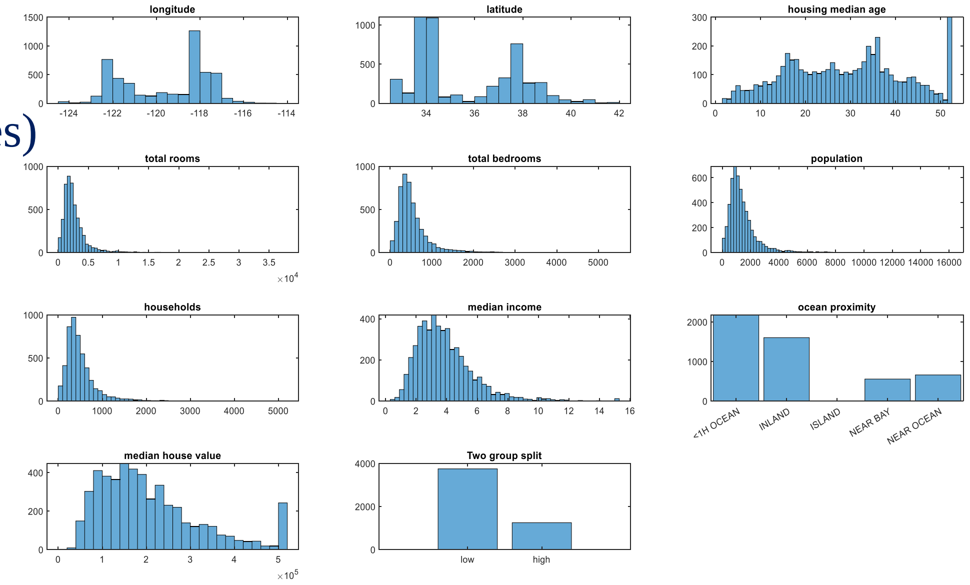
1. Visualize the data

2. Identify the features (find correlations between variables)

3. Preprocess the data (missing values, outliers)

4. Train the Model

5. Select the best performance model



Visualize the data, Summarize variables, data cleaning, pre-processing if needed

SECTION 2

- Home Value Prediction (Realistic Approach): 9 features to predict medianHouseValue (N=5000)

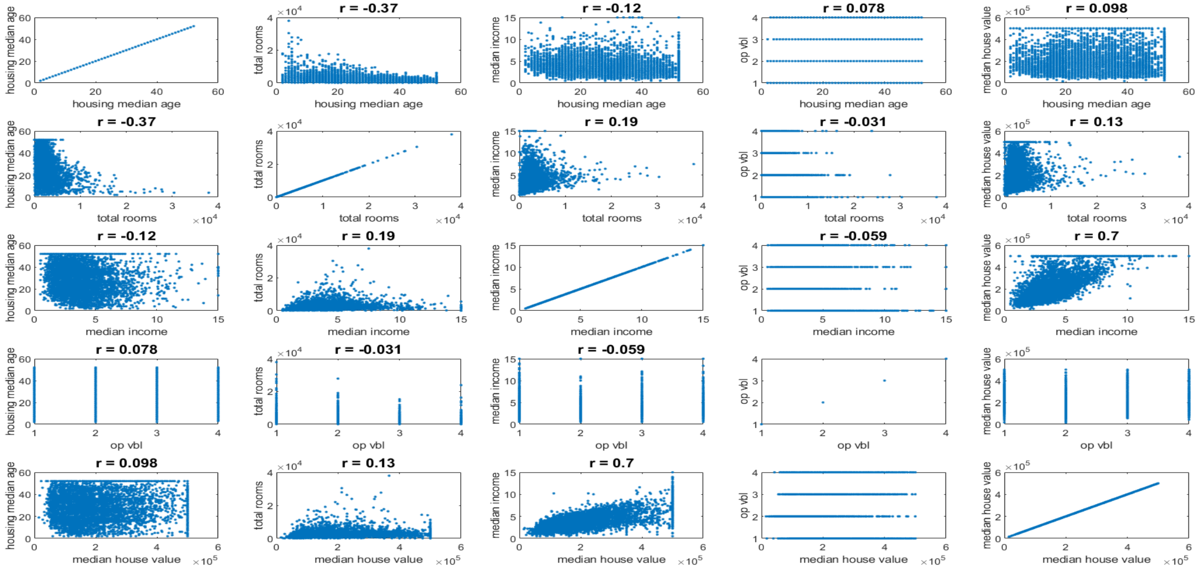
    1. Visualize the data

    2. Identify the features (find correlations between variables)

    3. Preprocess the data (missing values, outliers

    4. Train the Model

    5. Select the best performance model



FIND VARIABLE CORRELATIONS TO EACH OTHER AND THE MEDIAN_HOUSE_VALUE

SECTION 2

- Home Value Prediction (Realistic Approach): 9 features to predict medianHouseValue (N=5000)

  1. Visualize the data
  2. Identify the features (find correlations between variables)
  3. Preprocess the data (missing values, outliers)
  4. Train the Model
  5. Select the best performance model

207 Missing values, replace with median values

ocean_proximity: 20636×1 categorical
Values:

<1H OCEAN	9135
INLAND	6550
ISLAND	5
NEAR BAY	2289
NEAR OCEAN	2657

```
Visualize the data, Summarize variables, data cleaning, pre-processing if needed
```

SECTION 2

- Home Value Prediction (Realistic Approach): 9 features to predict medianHouseValue (N=5000)

1. Visualize the data
2. Identify the features (find correlations between variables)
3. Preprocess the data (missing values, outliers)
4. Train the Model
5. Select the best performance model

Linear Regression Fewer Variables RMSE ~69100

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	-2.3266e+06	2.011e+05	-11.57	2.0947e-30
longitude	-27661	2340.9	-11.816	1.2823e-31
latitude	-26535	2321.7	-11.43	9.9957e-30
housing_median_age	1014	104.58	9.6958	5.9307e-22
total_rooms	-3.6077	1.7753	-2.0322	0.042206
total_bedrooms	101.37	16.167	6.2701	4.0505e-10
population	-42.973	2.7491	-15.632	2.7235e-53
households	44.258	18.03	2.4547	0.014149
median_income	38847	799.97	48.56	0
op_inland	-38746	4137.6	-9.3641	1.3342e-20

Number of observations: 3500, Error degrees of freedom: 3490
Root Mean Squared Error: 6.91e+04
R-squared: 0.645, Adjusted R-Squared 0.644
F-statistic vs. constant model: 704, p-value = 0

SPLIT INTO TRAINING AND TEST DATA AND FIT REGRESSION MODELS

SECTION 2

- Home Value Prediction (Realistic Approach): 9 features to predict medianHouseValue (N=5000)

  1. Visualize the data
  2. Identify the features (find correlations between variables)
  3. Preprocess the data (missing values, outliers)
  4. Train the Model
  5. Select the best performance model

SECTION 2

- Home Value Prediction (Realistic Approach): 9 features to predict medianHouseValue (N=5000)

Model Type	Validation RMSE	Test RMSE
Lin regression	70071	65501
Lin. Regression – fewer variables	69031	65357
SVM –linear kernel	116370	116130
SVM –Gaussian Kernel	60099	57708

LASSO REGRESSION

- Linear Regression

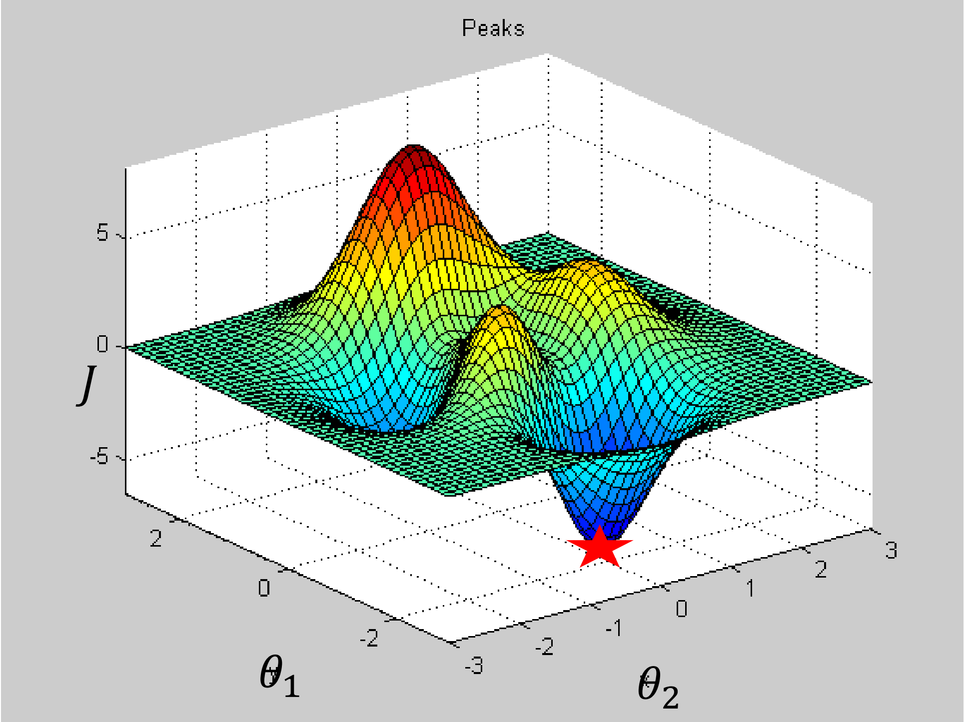
$$\hat{y}^i = \theta_0 + \theta_1 x_1^i + \theta_2 x_2^i + \cdots \ldots \ldots . + \theta_n x_n^i$$

$$\hat{Y} = \Theta^T X$$

- Gradient Descent by **Louis Augustin Cauchy** in 1847

Cost Function to Minimize

$$J = \left\langle \left(\hat{y}^i - y^i\right)^2 \right\rangle = (\hat{Y} - Y)^T (\hat{Y} - Y)$$

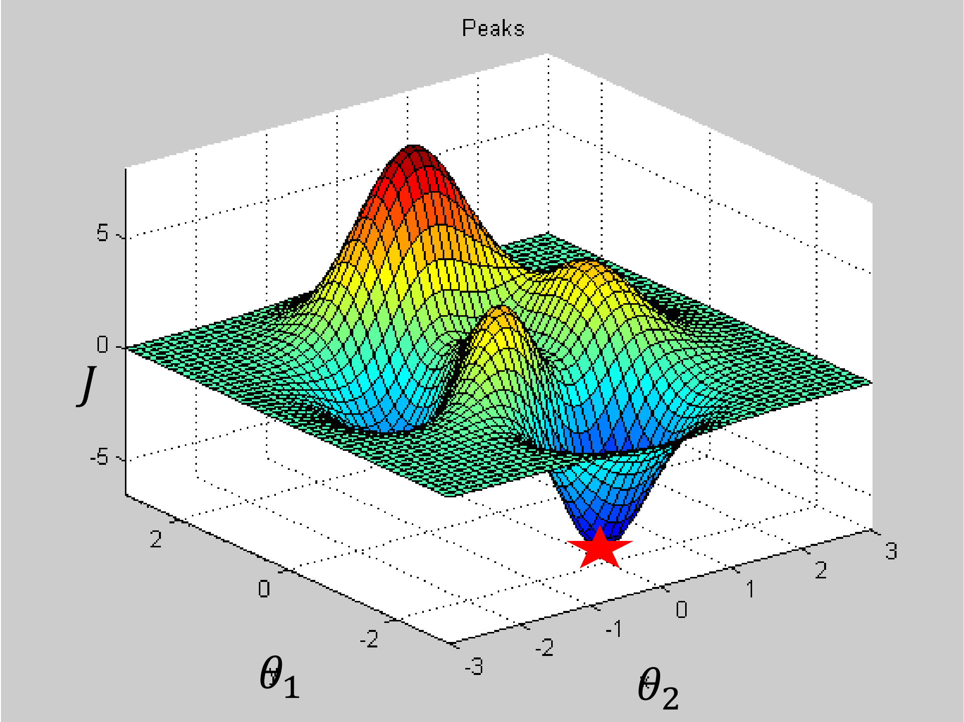


LASSO REGRESSION

- Linear Regression with Lasso

$$\hat{y}^i = \theta_0 + \theta_1 x_1^i + \theta_2 x_2^i + \cdots \ldots \ldots . + \theta_n x_n^i$$

$$\hat{Y} = \Theta^T X$$

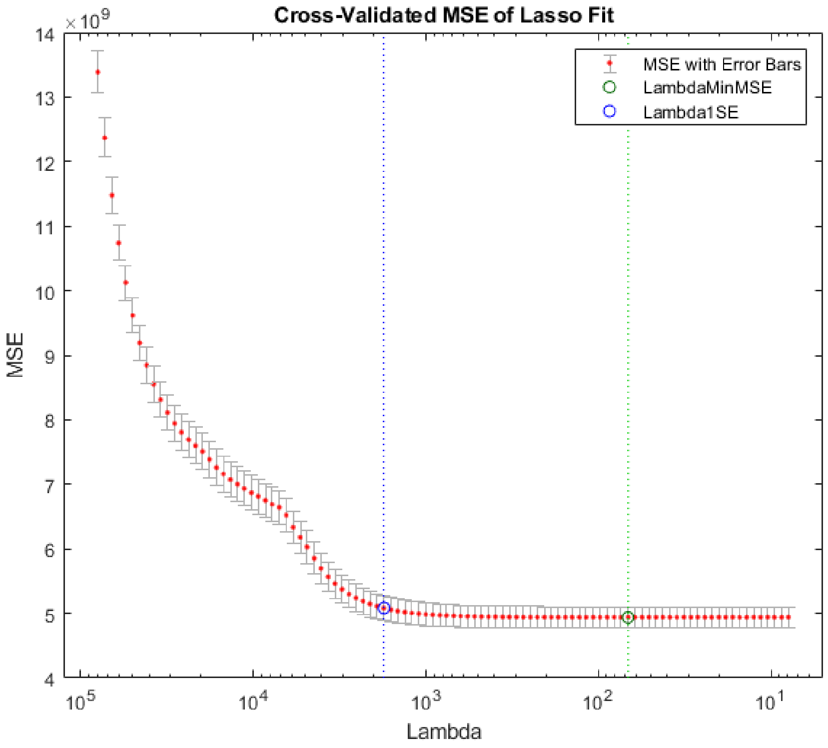


Cost Function to Minimize

$$J = \left\langle \left(\hat{y}^i - y^i\right)^2 \right\rangle = \left(\hat{Y} - Y\right)^T \left(\hat{Y} - Y\right) + \lambda \sum_{j=1}^{n} |\theta_j|$$

SECTION 3

- Home Value Prediction (Lasso Regression): 9 features to predict medianHouseValue (N=5000)



$$J = \left\langle \left(\hat{y}^i - y^i\right)^2 \right\rangle = \left(\hat{Y} - Y\right)^T\left(\hat{Y} - Y\right) + \lambda \sum_{j=1}^{n} |\theta_j|$$

Lambda

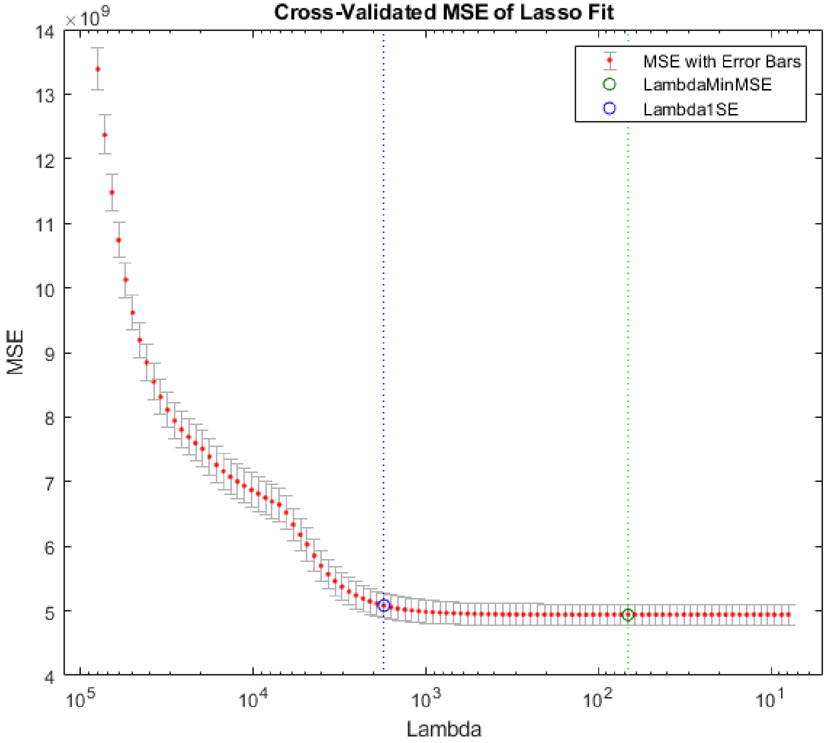
Lasso removes the 'total_rooms' and 'Ocean Proximity_inland' variables as least important.

RMSE on test data with 7 features = 66443

DEMO

SECTION 3

➢ Home Value Prediction (Lasso Regression): 9 features to predict medianHouseValue (N=5000)



{'longitude’	-3.2643
'latitude’	-3.2856
'housing_median_age’	0.1177
'total_rooms’	0
'total_bedrooms’	0.0074
'population’	-0.0028
'households’	0.0014
'median_income’	3.8702
'op_vbl'}	0

All coefficients multiplied by 10.^4

Project 2: Classification of House Value

SECTION 1: Learner App

- **Home Value Classification:** 9 features to classify high vs low medianHouseValue

longitude: A measure of how far west a house is; a higher value is farther west

latitude: A measure of how far north a house is; a higher value is farther north

housingMedianAge: Median age of a house within a block; a lower number is a newer building

totalRooms: Total number of rooms within a block

totalBedrooms: Total number of bedrooms within a block

population: Total number of people residing within a block

households: Total number of households, a group of people residing within a home unit, for a block

medianIncome: Median income for households within a block of houses (measured in tens of thousands of US Dollars)

medianHouseValue: Median house value for households within a block (measured in US Dollars)

oceanProximity: Location of the house w.r.t ocean/sea

https://www.kaggle.com/camnugent/california-housing-prices

Demo with N=5000
70% Training Data
30% Test Data
Models Trained:
Logistic Regression
SVM

SECTION 1: Learner App

- Prediction of House Price Classification Problem

Confusion Matrix

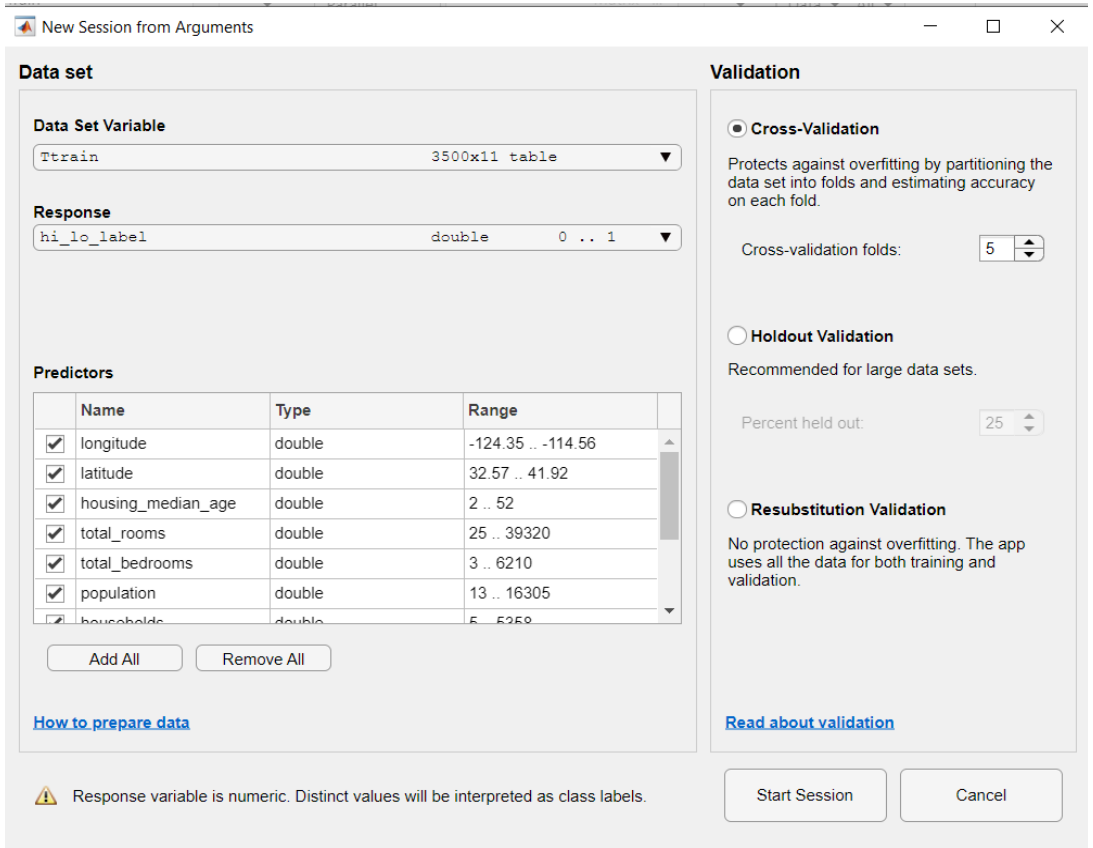
		Predicted Class: 1	Predicted Class: 0	
True Class	1	True Positive	False Negative	→ Total Positive
	0	False Positive	True Negative	→ Total Negative

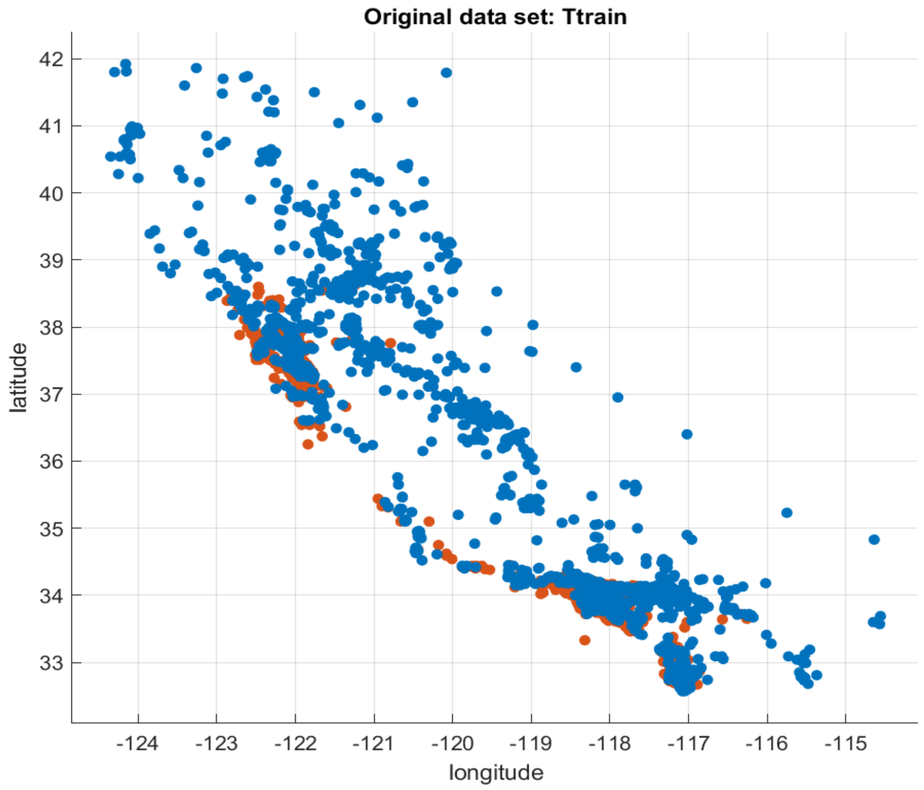
True Positive Rate = True Positive / Total Positive

True Negative Rate = True Negative / Total Negative = 1 – False Positive Rate

SECTION 1: Learner App

➢DATA IMPORT & CLASSIFICATION LEARNER INITIALIZATION





SECTION 1: Learner App

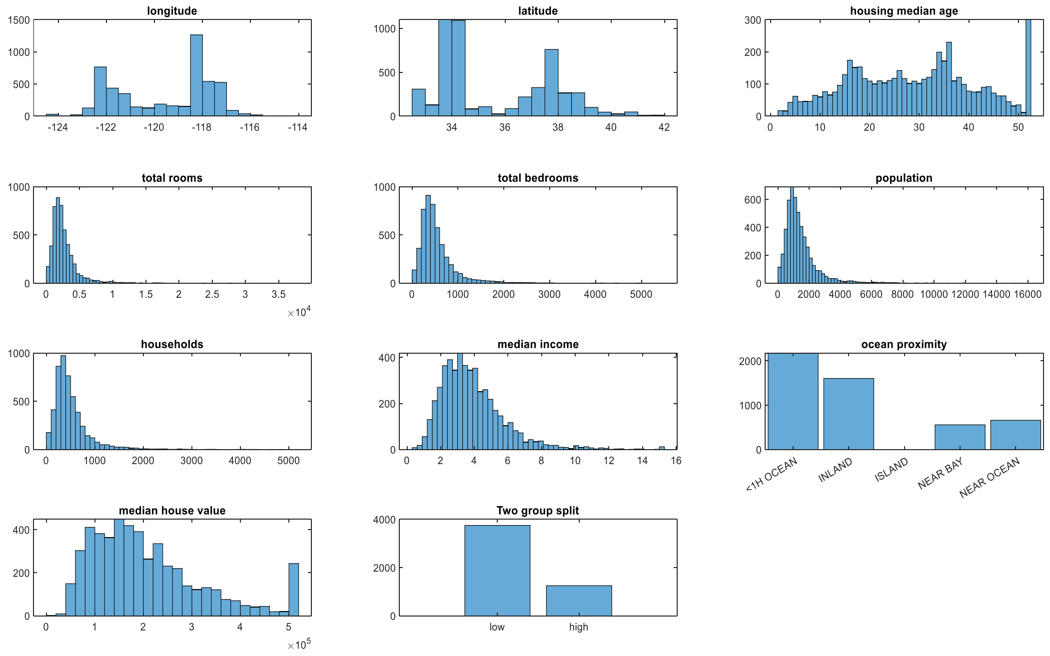
➢ DATA IMPORT & CLASSIFICATION LEARNER INITIALIZATION

```
classificationLearner(Ttrain,'hi_lo_label');
```

Demo with logistic regression and linear SVM

SECTION 2: Raw Data Analysis

Visualize the data, Summarize variables, data cleaning, pre-processing if needed



207 Missing values, replace with median values

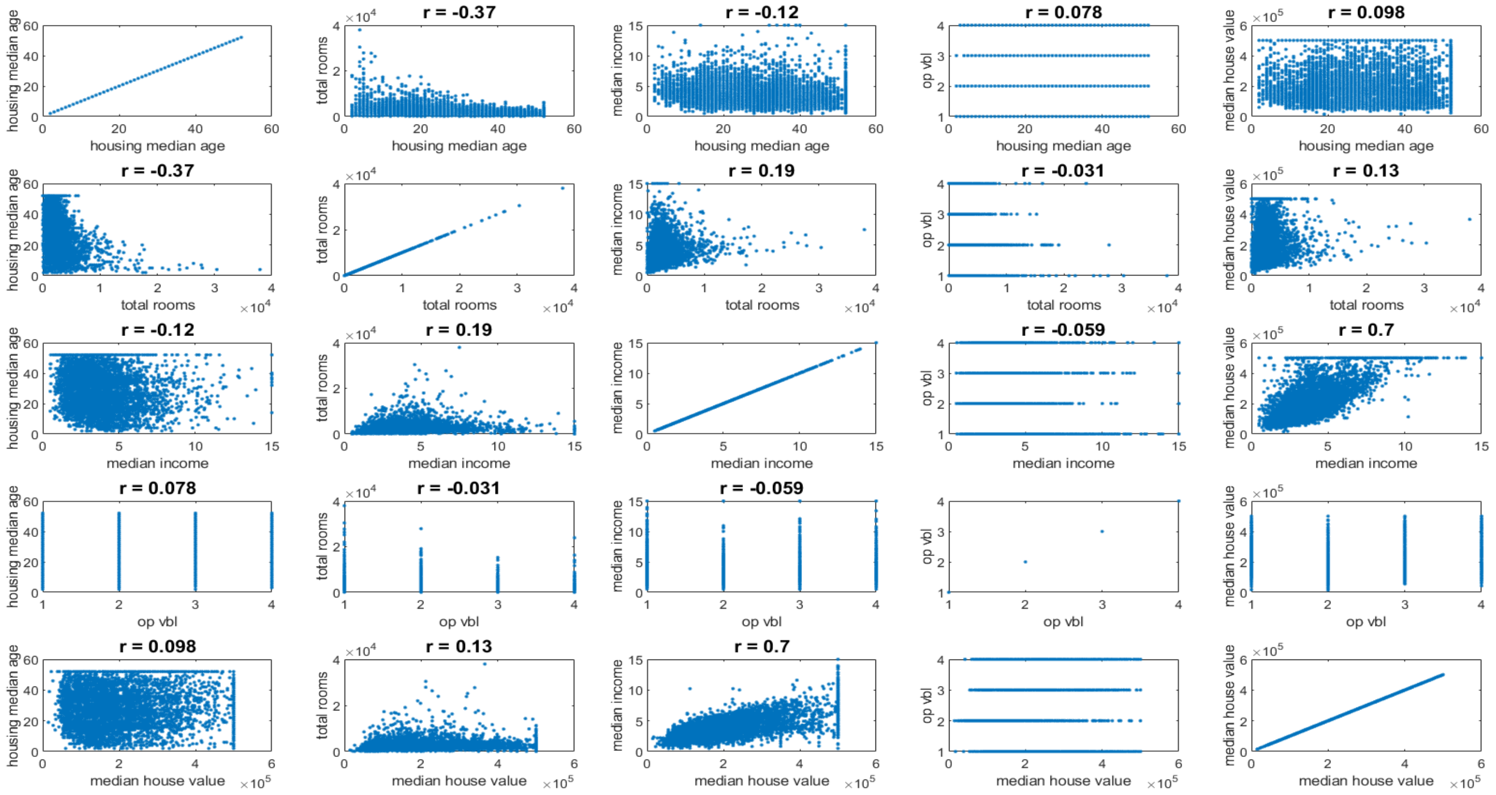
ocean_proximity: 20636×1 categorical

Values:

<1H OCEAN	9135
INLAND	6550
ISLAND	5
NEAR BAY	2289
NEAR OCEAN	2657

SECTION 3: Correlation Analysis

FIND VARIABLE CORRELATIONS TO EACH OTHER AND THE MEDIAN HOUSE VALUE




```
[R,pp] = corr(table2array(T1(:,select_vars)));
```

SECTION 4: Logistic Regression

SPLIT INTO TRAINING AND TEST DATA AND FIT LOGISTIC REGRESSION MODEL

```
Estimated Coefficients:
```

	Estimate	SE	tStat	pValue
(Intercept)	-154.19	14.421	-10.692	1.1065e-26
longitude	-1.7683	0.17448	-10.135	3.8752e-24
latitude	-1.8133	0.18885	-9.6018	7.8546e-22
housing_median_age	0.044239	0.0051484	8.5928	8.4901e-18
total_rooms	0.0003444	9.7387e-05	3.5364	0.00040561
total_bedrooms	0.00080298	0.00084259	0.95299	0.3406
population	-0.0023529	0.00020995	-11.207	3.7737e-29
households	0.0039573	0.00094559	4.185	2.8517e-05
median_income	1.0172	0.053904	18.87	2.0101e-79
ocean_proximity_INLAND	-0.053285	0.24937	-0.21368	0.8308
ocean_proximity_ISLAND	0	0	NaN	NaN
ocean_proximity_NEAR BAY	-0.10616	0.19861	-0.53449	0.593
ocean_proximity_NEAR OCEAN	0.11076	0.15948	0.6945	0.48737

```
3500 observations, 3488 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 1.83e+03, p-value = 0
```

```
mdl = fitglm([Ttrain(:,1:9) table(y)],'Distribution','binomial');
```

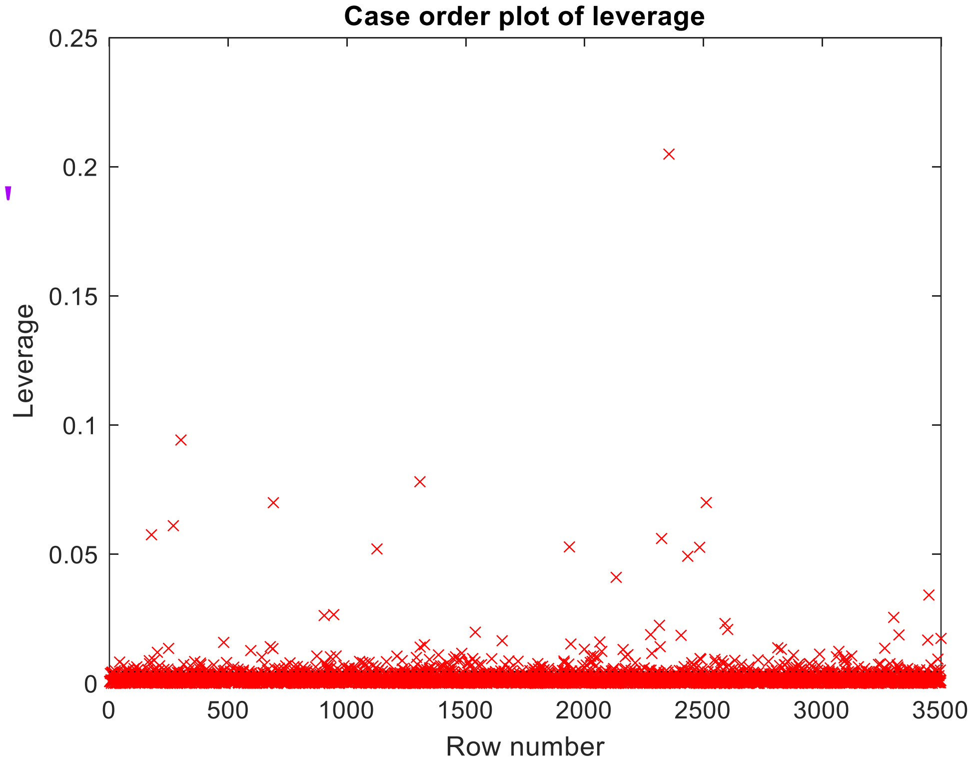
Remove Insignificant features

SECTION 5: Outliers

DIAGNOSTICS OF MODELS- IDENTIFY OUTLIERS

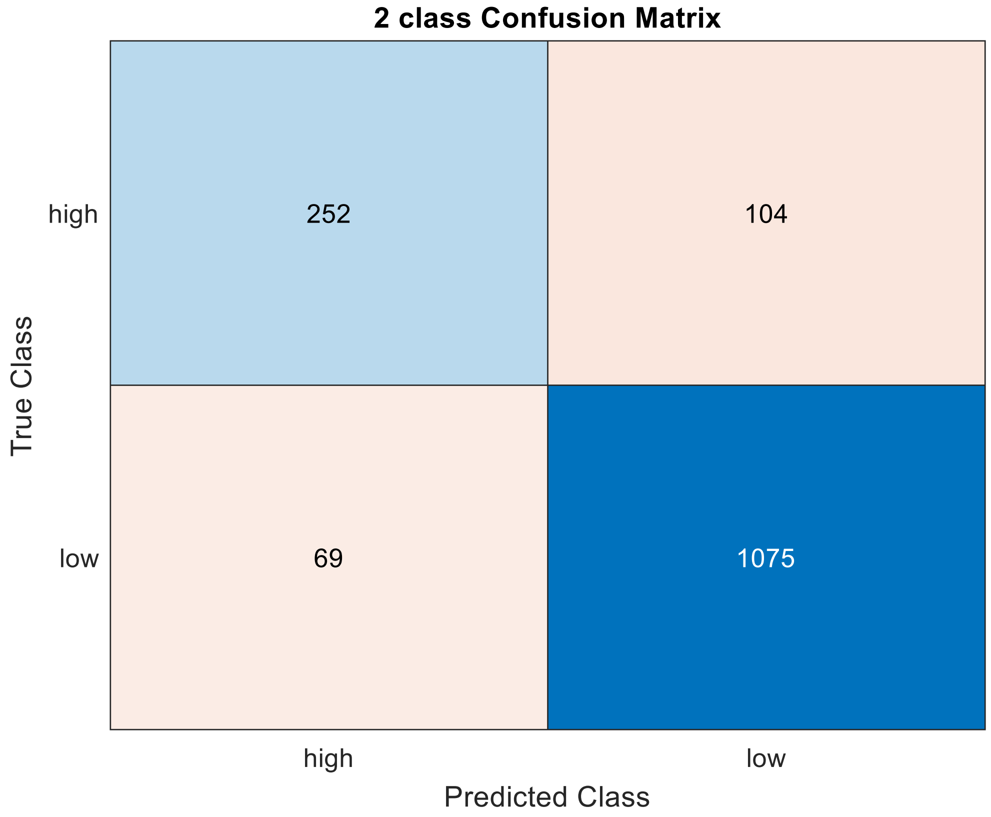
```
mdl1 = fitglm([Ttrain(:,[1:4 6:8])
table(y,'variablenames',{'Hi_lo_label'})],'
Distribution','binomial');
```

```
plotDiagnostics(mdl1,'leverage')
```



SECTION 6: Classification (Clean Data)

TEST MODEL FOR TWO CLASS CLASSIFICATION (Logistic Regression)

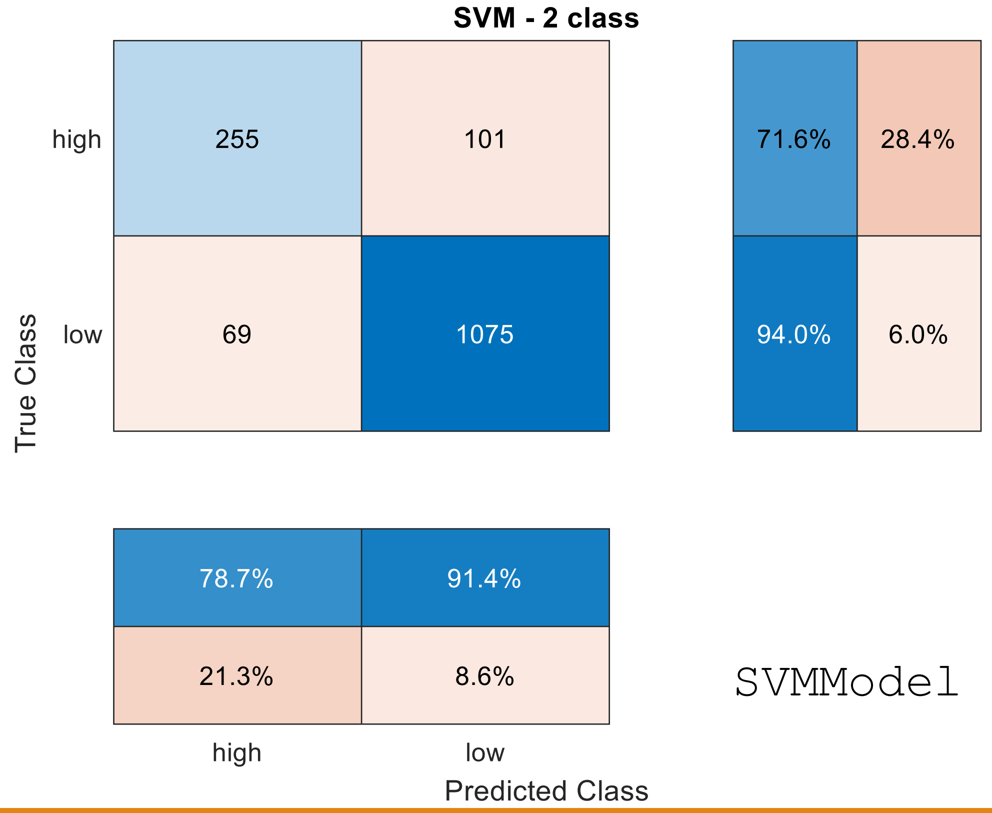


Test Data N = 1500
(30% of 5000)

Missing Values
Insignificant Features
Outliers

SECTION 7: SVM Classification

REGULARIZATION OF VARIABLES DONE AUTOMATICALLY, NO NEED TO CHOOSE FEATURES SEPARATELY AS WAS DONE EARLIER FOR LOGISTIC REGRESSION



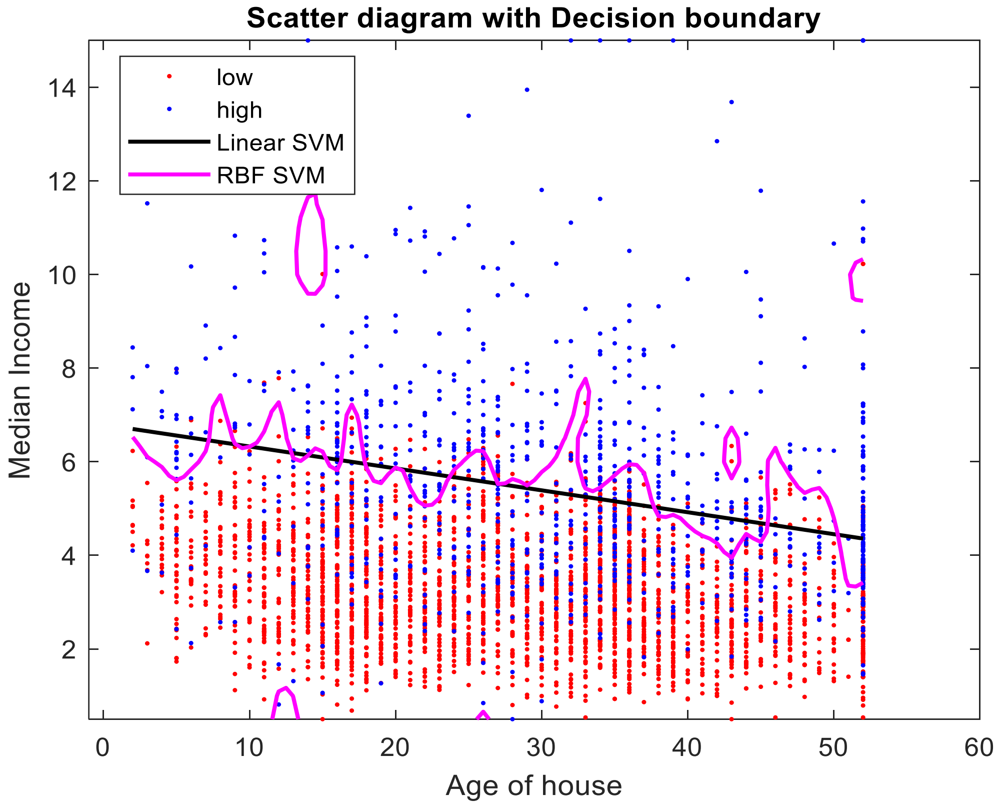
Test Data N = 1500
(30% of 5000)

Linear SVM

```
SVMModel = fitcsvm(Ttrain(:,1:9),y,'standardize',true);
```

SECTION 8: SVM Classification

LINEAR vs RADIAL BASIS FUNCTION (RBF) KERNEL



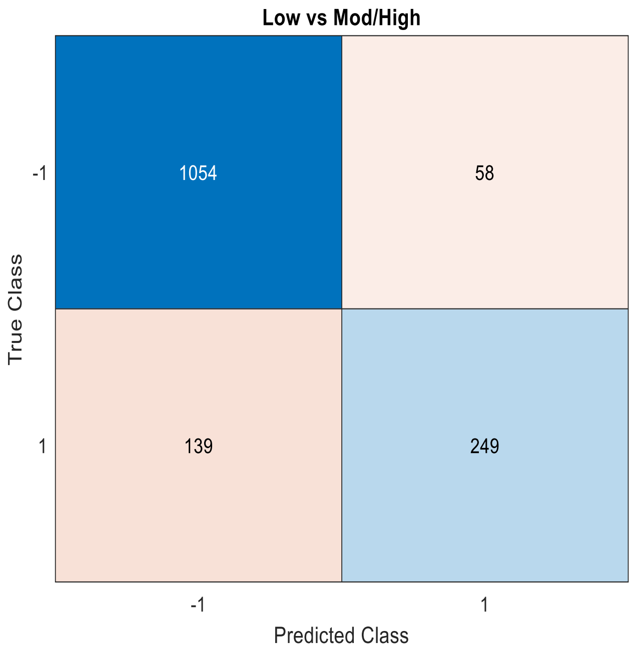

```
fitcsvm([x1 x2],y1);

fitcsvm([x1 x2],y1,'KernelFunction','rbf');
```

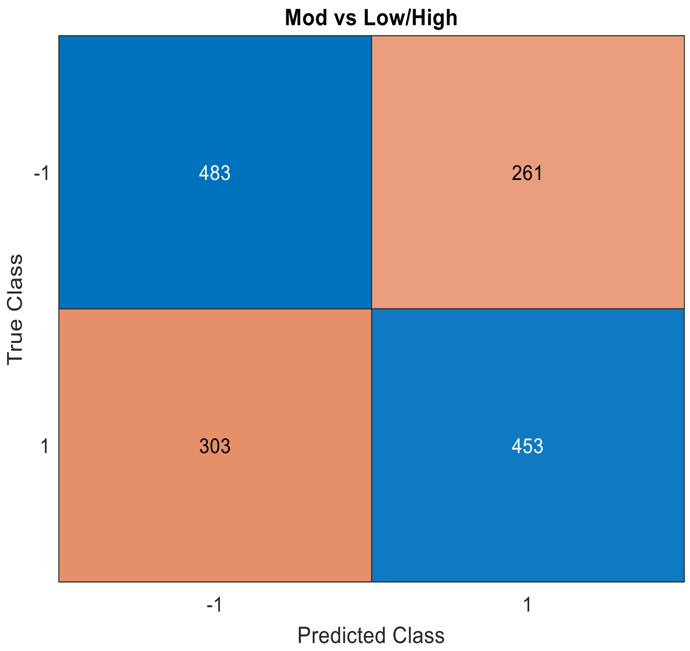
x1: Age of House
X2: Median Income

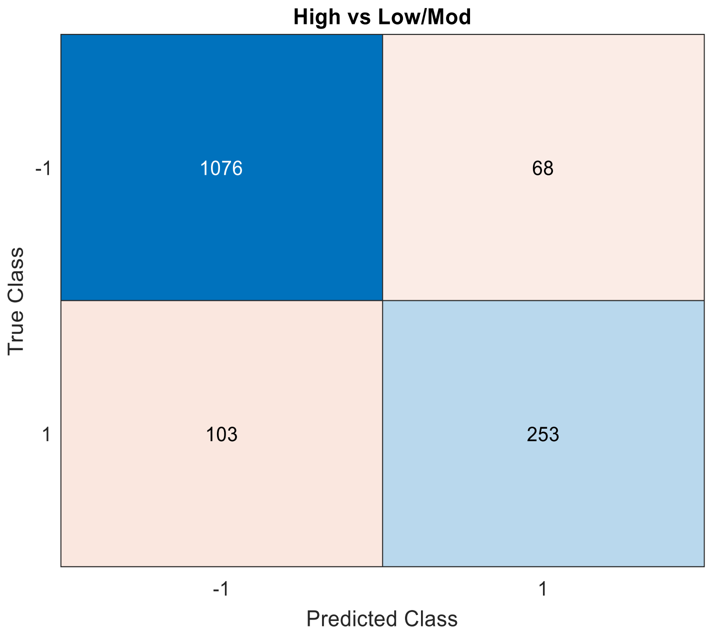
SECTION 9: Multiclassification (SVM)

ONE CLASS vs REST

Also perform one to one class





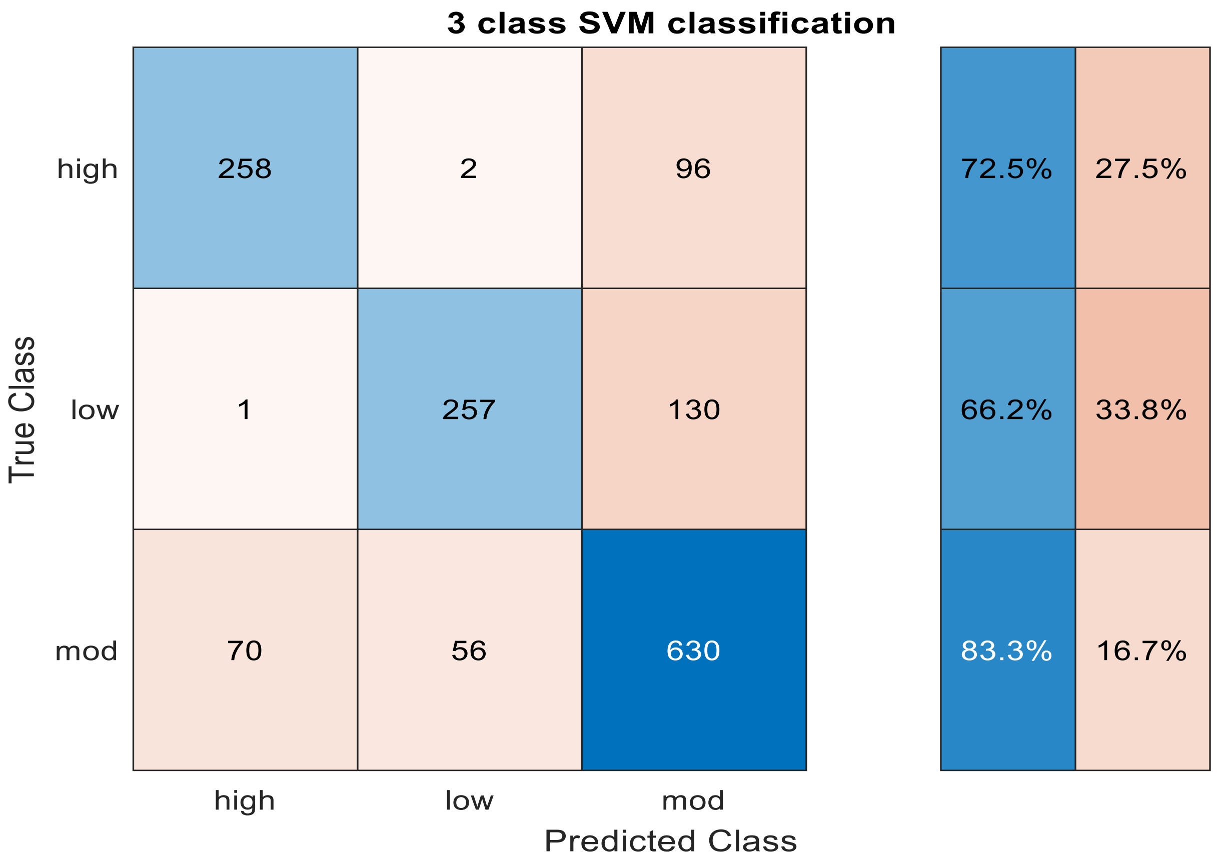



```
Mdl =
fitcecoc(Ttrain(:,1:8),y,'Learners',t,'Coding',coding,'ResponseName',responseName,...
    'PredictorNames',predictorNames,'ClassNames',classNames);
```

SECTION 10: Multiclassification (SVM)

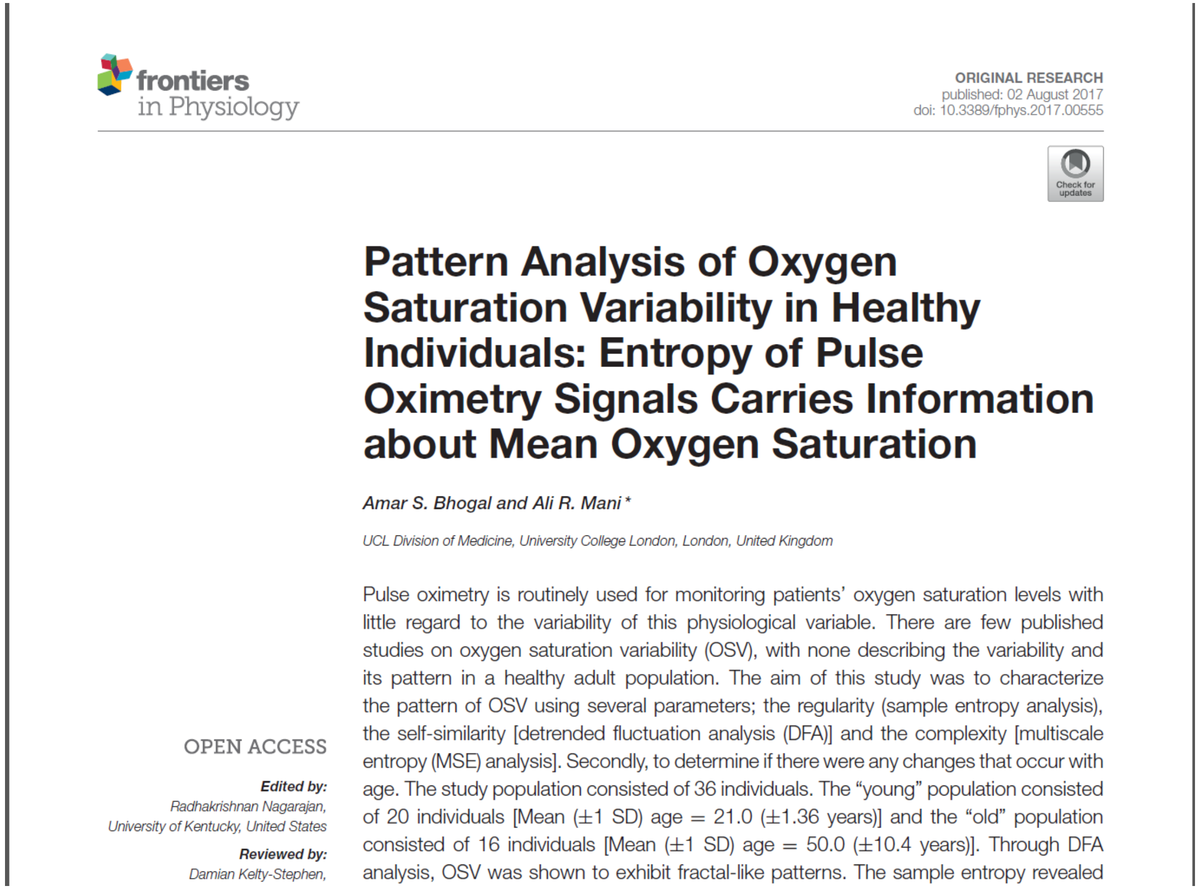
LOW vs MOD vs HIGH CLASS

```
Mdlp =
fitcecoc(Ttrain(:,1:8),y,'Learner
s',t,'FitPosterior',true,...

'ClassNames',{'low','mod','high'}
,...
    'Verbose',2);
```



Project 3: Oxygen desaturation

Differentiating Smokers vs Non-Smokers



frontiers in Physiology

ORIGINAL RESEARCH
published: 02 August 2017
doi: 10.3389/fphys.2017.00555

Check for updates

Pattern Analysis of Oxygen Saturation Variability in Healthy Individuals: Entropy of Pulse Oximetry Signals Carries Information about Mean Oxygen Saturation

*Amar S. Bhogal and Ali R. Mani**

UCL Division of Medicine, University College London, London, United Kingdom

Pulse oximetry is routinely used for monitoring patients' oxygen saturation levels with little regard to the variability of this physiological variable. There are few published studies on oxygen saturation variability (OSV), with none describing the variability and its pattern in a healthy adult population. The aim of this study was to characterize the pattern of OSV using several parameters; the regularity (sample entropy analysis), the self-similarity [detrended fluctuation analysis (DFA)] and the complexity [multiscale entropy (MSE) analysis]. Secondly, to determine if there were any changes that occur with age. The study population consisted of 36 individuals. The "young" population consisted of 20 individuals [Mean (±1 SD) age = 21.0 (±1.36 years)] and the "old" population consisted of 16 individuals [Mean (±1 SD) age = 50.0 (±10.4 years)]. Through DFA analysis, OSV was shown to exhibit fractal-like patterns. The sample entropy revealed

OPEN ACCESS

Edited by:
Radhakrishnan Nagarajan,
University of Kentucky, United States

Reviewed by:
Damian Kelty-Stephen,