Pima Indian woman winnowing beans over her head onto a mat on the ground, Gila River Reservation(?), ca.1902[1]

A Machine Learning Approach for Identifying Symptom Clusters that Predict Prediabetes

A study of the Pima Indians

Brian Griner

Presentation at BMS

August 27, 2018

Problem: Social Cost of Diabetes Management

Problem: Rising social cost of managing diabetes.[2]

- Research shows improvements from new therapies but close to half of the diabetes population failed to reach glycemic control targets.[3]

- New research using data driven symptom clustering suggests the existence of multiple categories of diabetes patients with different treatment needs.[4]

- Can machine learning (ML) be used to identify patient subpopulations based on clusters of symptoms with an increased risk of developing diabetes or prediabetes?[5]

- Helping physicians identify patient types with different treatment needs earlier should reduce costs associated with diabetes management and help more patients reach glycemic control targets.

Data Source

The **Pima Indian diabetes data** used for the analysis was obtained from the UCI Machine Learning Repository[6]. This data is no longer available on the UCI repo. An archive of the UCI repository is available on Kaggle[7] at https://www.kaggle.com/uciml/pima-indians-diabetes-database.

Source: National Institute of Diabetes and Digestive and Kidney Diseases.

Population: Females who are 21 years or older of Pima Indian heritage.

Sample Size: 768

Attributes: 9

1. **preg**: Number of times pregnant
2. **gluc**: Plasma glucose concentration at 2 hours in an oral glucose tolerance test
3. **dbp**: Diastolic blood pressure (mm Hg)
4. **skin**: Triceps skin fold thickness (mm)
5. i**nsul**: 2-Hour serum insulin (mu U/ml) - female norm: NA
6. **bmi**: Body mass index (weight in kg/(height in m)^2)
7. **pedi**: Diabetes pedigree function[8]
8. **age**: Age (years)
9. **class**: Class variable (0 or 1) = False/True for onset of diabetes in 1-5 years

Data Quality Assessment

Data quality assessment on this project followed the process of best practices for data science and machine learning projects which included:

- Determining **data types**: nominal, ordinal, numeric, text.

- Identifying **'bad' data elements**: e.g., mixed formats/data types that will generate errors (or not).

- Identify **outliers** using descriptive statistics:
  - **Numbers**: mean, median, mode, min, max, quartiles
  - **Nominal**: class distributions (percentages)
  - **Ordinal**: quartiles, min, max, median, mode

- Identify **missing values**: How are they coded? What percentage are missing, 5% or 50%?

- Plot **distributions**: Normal?, Skewed? Believable?

Data Imputation

Quality issues: Two important model inputs had a significant amount of **missing values**:

- Triceps skin fold thickness (**30% missing**)
- Two hour post glucose insulin levels (**49% missing**)

Remedy: ML models were developed for data imputation.

Transformations:

- A log transform was applied to skin fold thickness and insulin levels to convert positive integers to real numbers.
- All inputs to the models were standardized prior to training.

Model Development: Several algorithms were tested using 10 fold cross validation to assess the out-of-sample predictive accuracy of each algorithm.

Methodology

1. ML Algorithms Tested

- **6 algorithms**: logistic regression, lasso regression, ElasticNet, K Nearest Neighbors (KNN) Regression/ Classifier , Classification and Regression Trees (CART) and Support Vector Machines/Regression (SVR/ SVM).
- **4 ensembles**: AdaBoost, Gradient Boosting Machine (GBM), Random Forest (RF) and the Extra Trees algorithm.

2. ML Best Practice

- Best practice in machine learning is to **test several different algorithms** when developing a ML model and compare them using cross-validation (CV).
- **Why?** ML algorithms vary greatly in design, leading to specific strengths and weakness. General guidelines for different algorithms exist but difficult to know which algorithm will perform best for a particular dataset.

2. Selection Criteria: Global and Local Effects

- **Global** patterns apply to ALL cases in the data; e.g. the impact of age. Example algorithms for **global** effects: Logistic Regression and Support Vector Machines.
- **Local** patterns only apply to a subset of cases; e.g., cases located in NJ. Example algorithms for **local** effects: Classification and Regression Trees and K Nearest Neighbors.

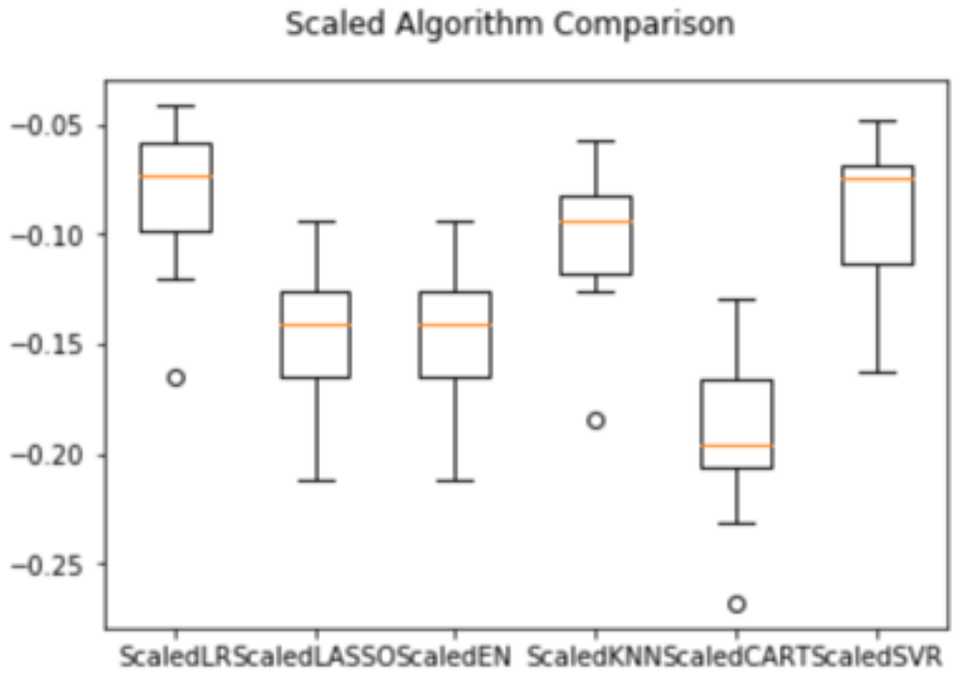
Model Selection Criteria

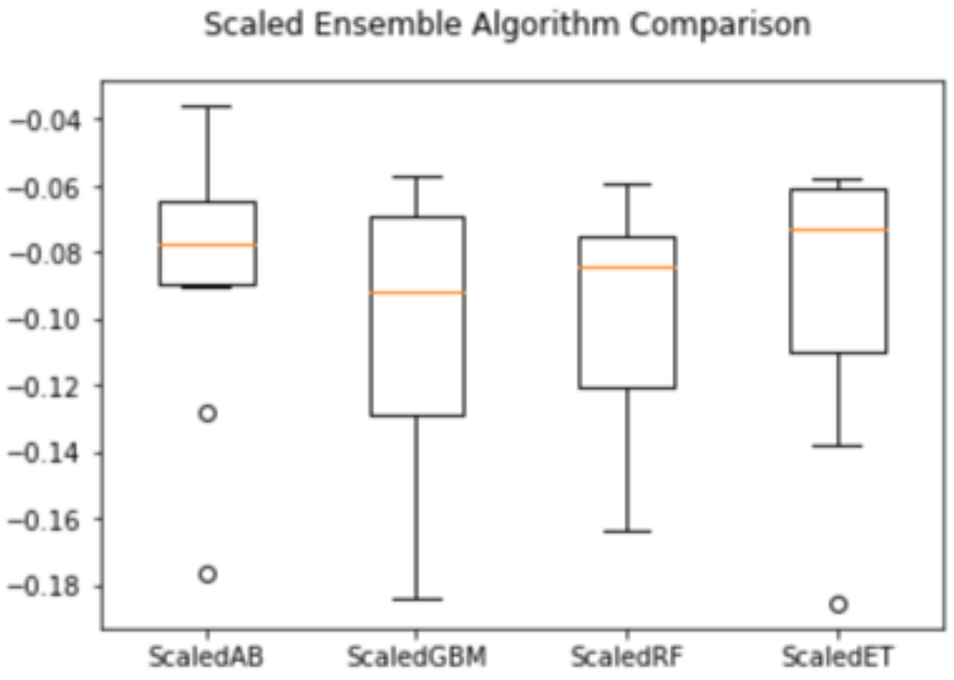
1. **Cross-Validation Tests**
    - ML algorithms were assessed using **10 fold Cross-Validation** on minimizing the mean squared error.[9]

2. **Pre and Post Imputation Histogram Plots**
    - Histogram plots were used to visualize the pre and post imputation distributions from the final models check for face validity.

Cross-Validation Tests: Imputation Models of **Skin Fold Thickness***

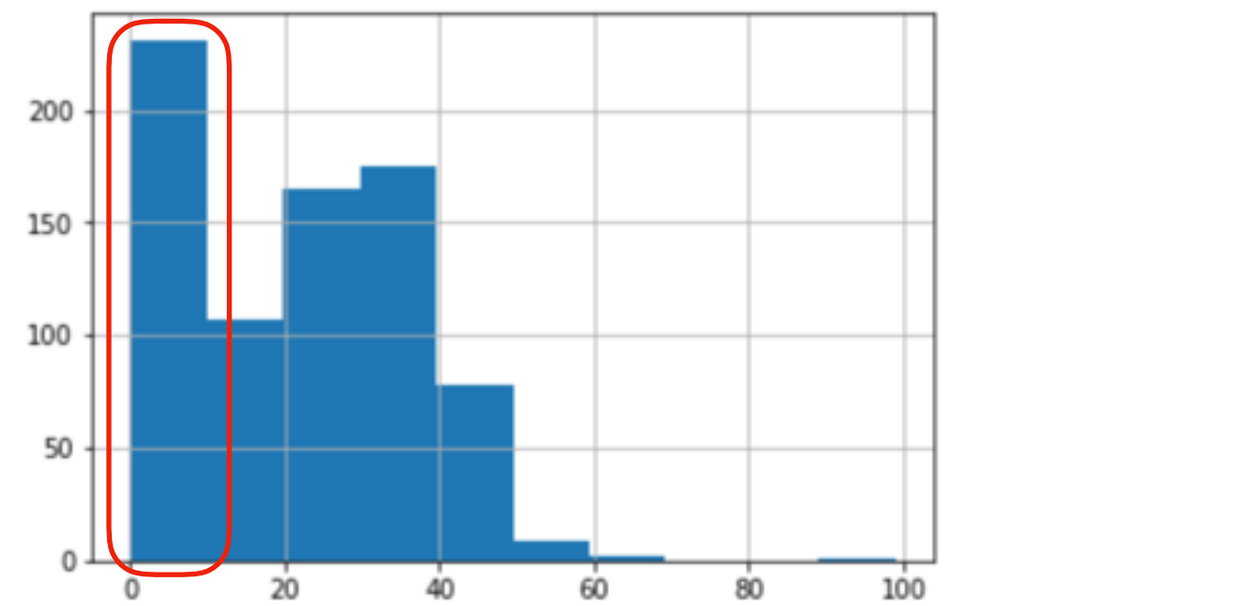
- Linear Regression
- Lasso
- ElasticNet
- Decision Tree Regression
- K Nearest Neighbor (KNN) Regression
- Support Vector Regression (SVR)
- Random Forest Regression
- Gradient Boosting Regression
- Extra Trees Regression
- AdaBoost Regression

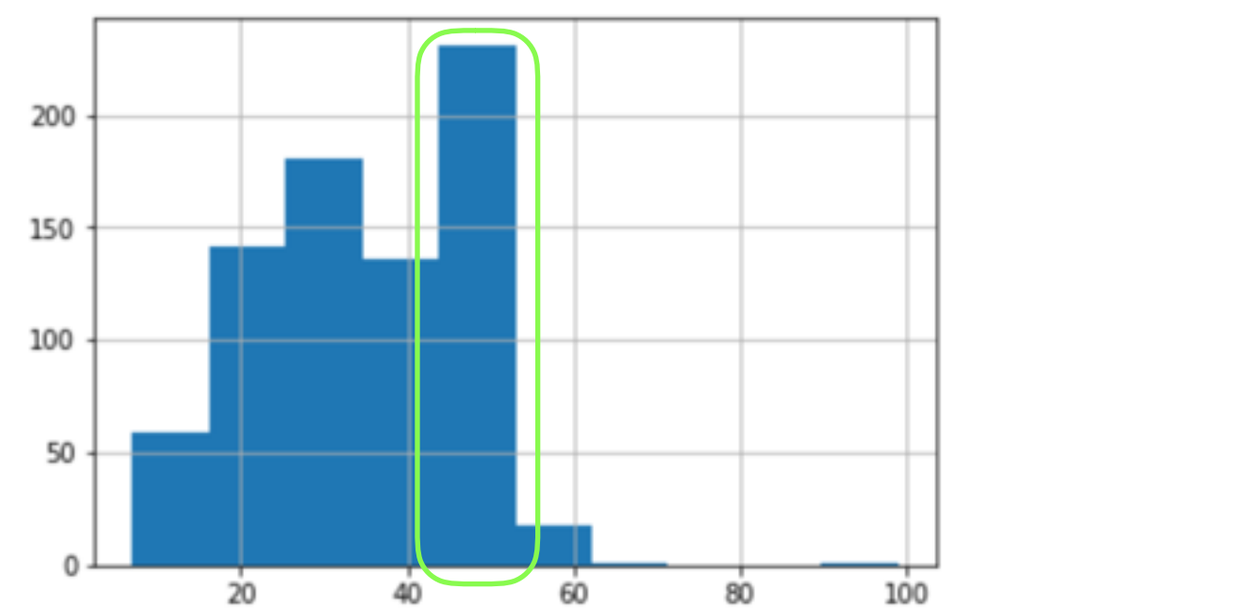




**** All algorithms trained using standardized data. Hyper parameters tuned using Grid Search Cross-Validation.***

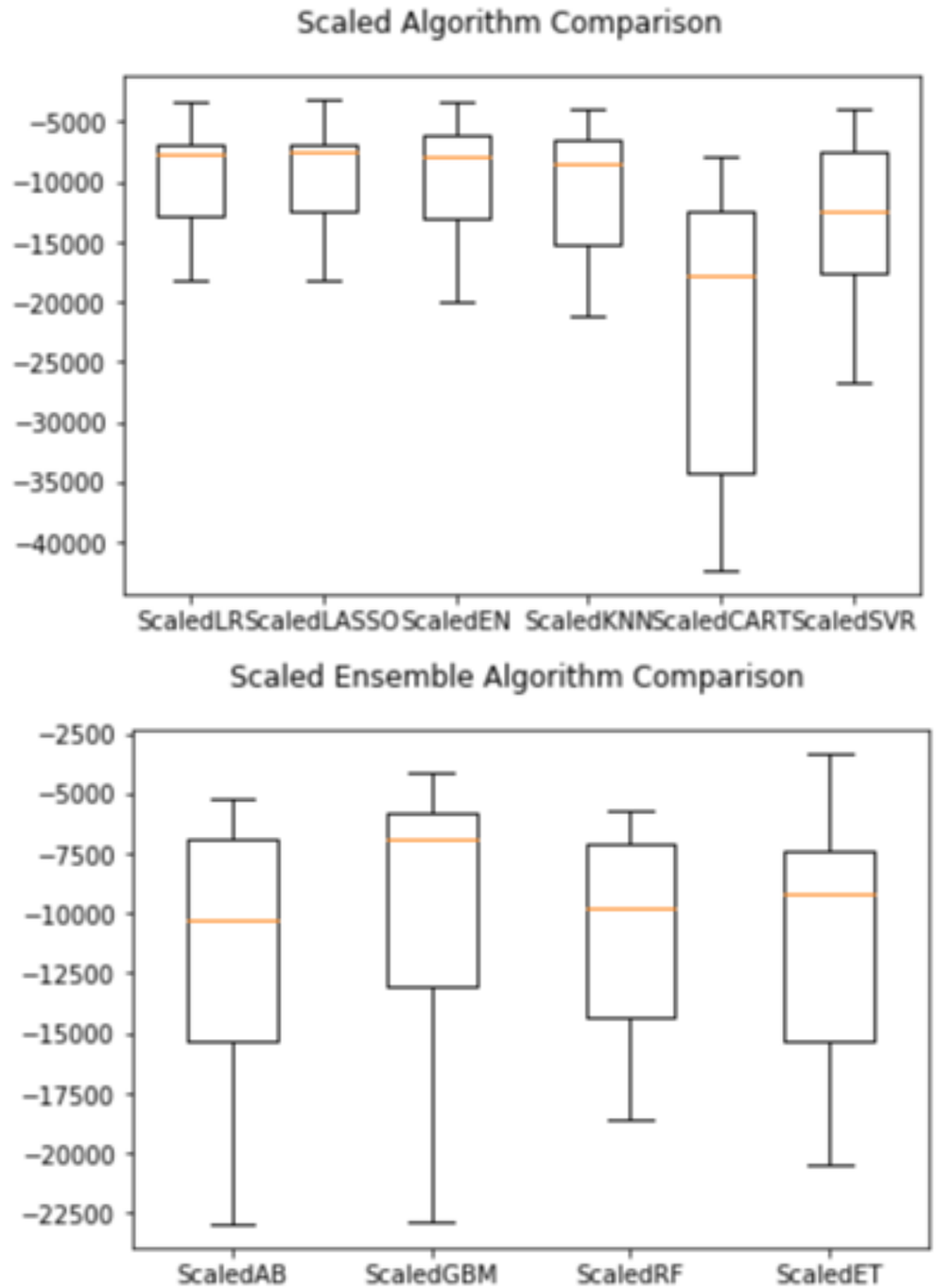
Pre & Post Imputation Distributions of **Skin Fold Thickness**





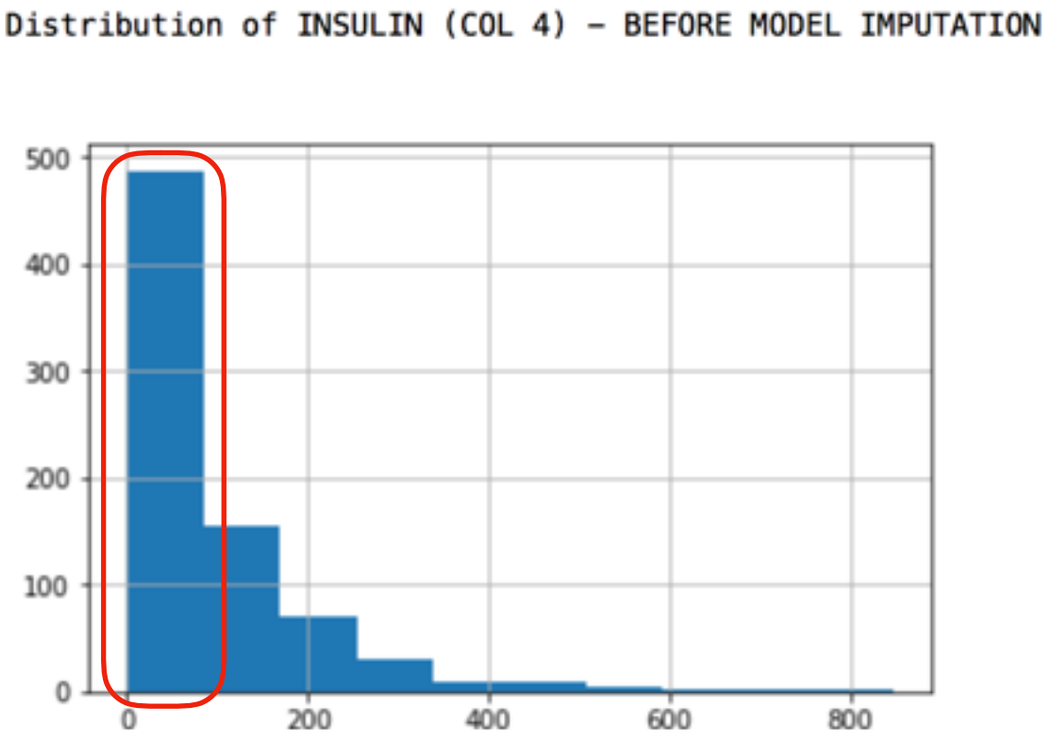
Cross-Validation Tests: Imputation Models of **Insulin Levels***

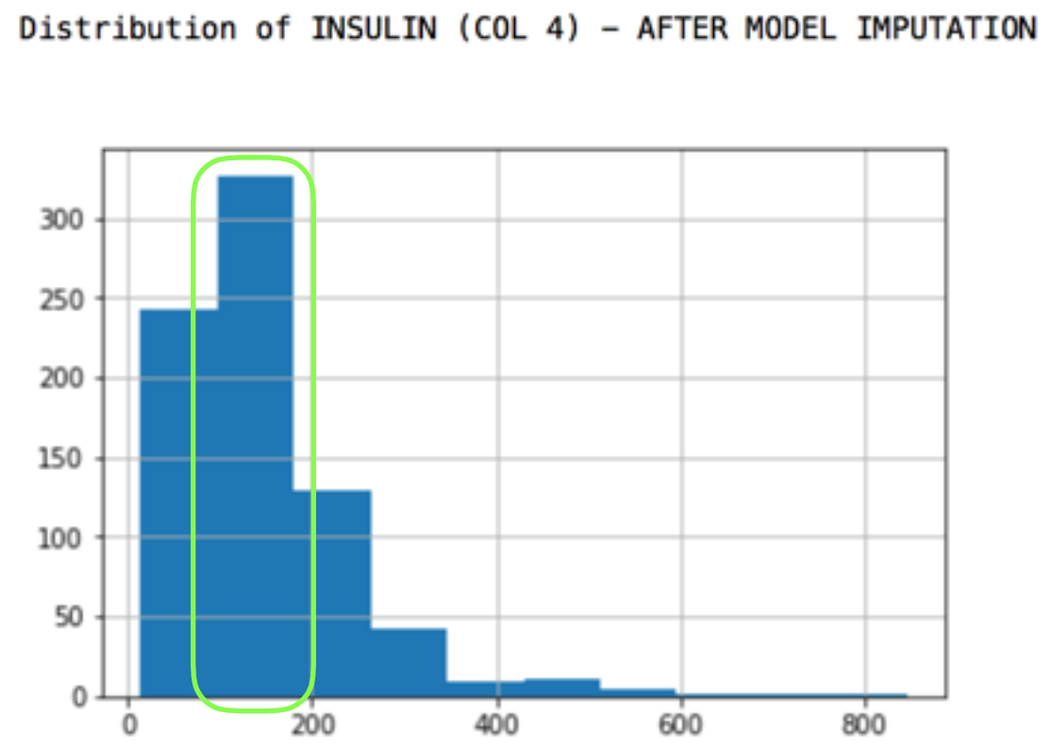
- Linear Regression
- Lasso
- ElasticNet
- Decision Tree Regression
- K Nearest Neighbor (KNN) Regression
- Support Vector Regression (SVR)
- Random Forest Regression
- Gradient Boosting Regression
- Extra Trees Regression
- AdaBoost Regression



**** All algorithms trained using standardized data. Hyper parameters tuned using Grid Search Cross-Validation.***

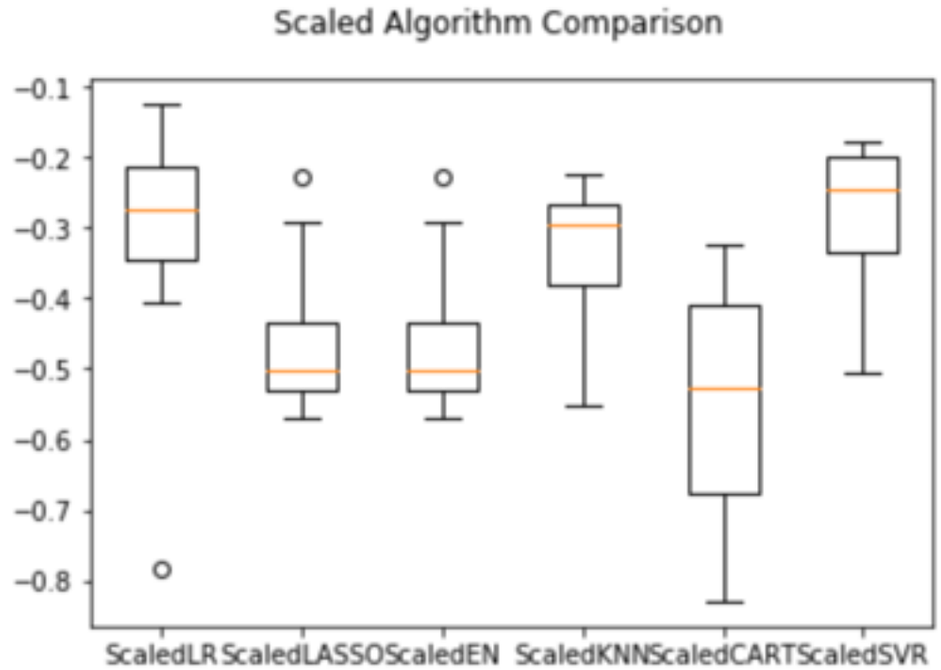
Pre & Post Imputation Distributions of **Insulin Levels**

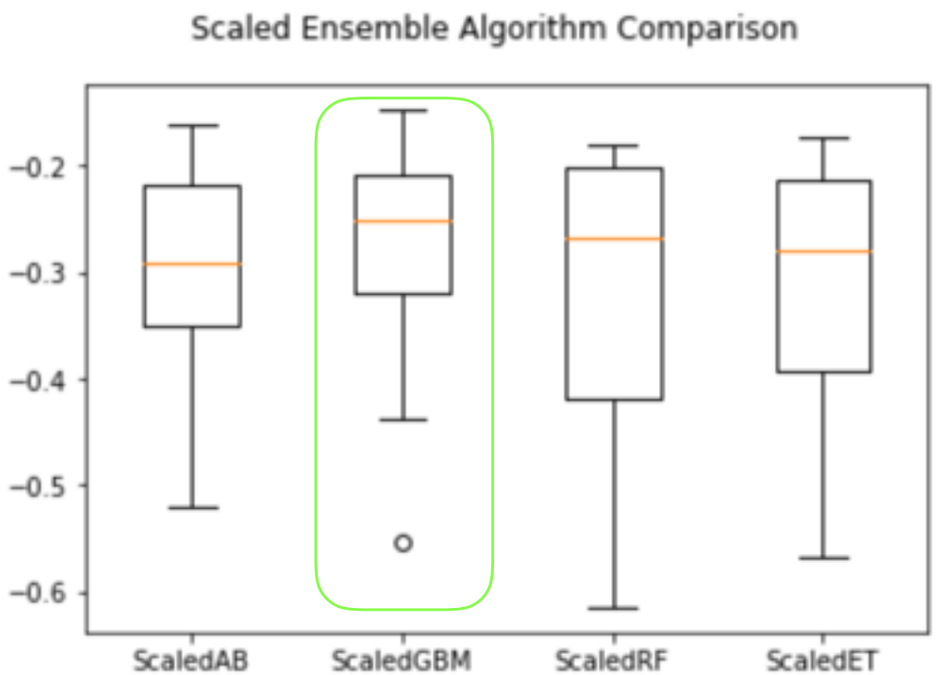




Cross-Validation Tests: **Prediabetes Classification Models***

- Linear Regression
- Lasso
- ElasticNet
- Decision Tree Regression
- K Nearest Neighbor (KNN) Regression
- Support Vector Regression (SVR)
- Random Forest Regression
- Gradient Boosting Regression
- Extra Trees Regression
- AdaBoost Regression

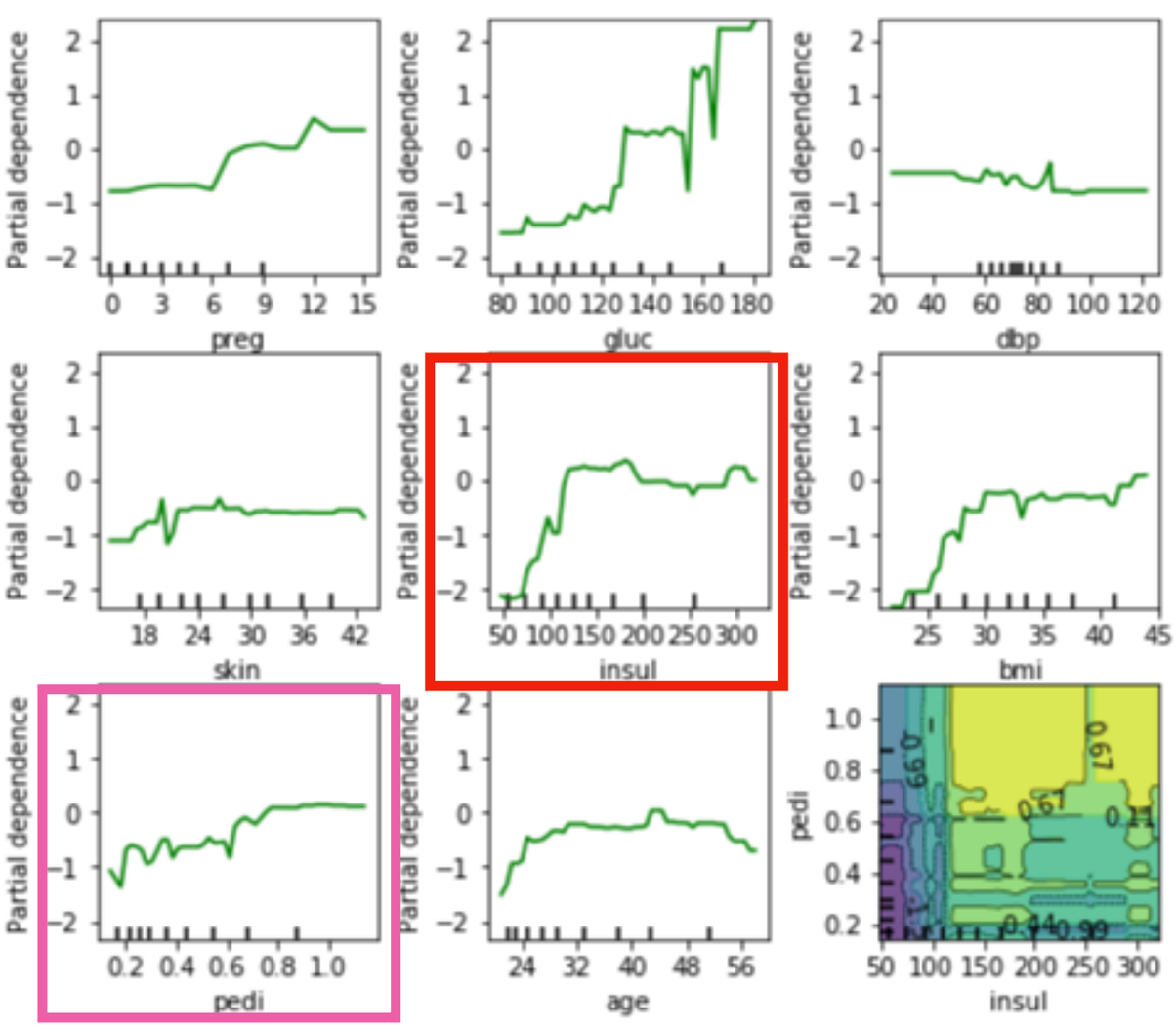


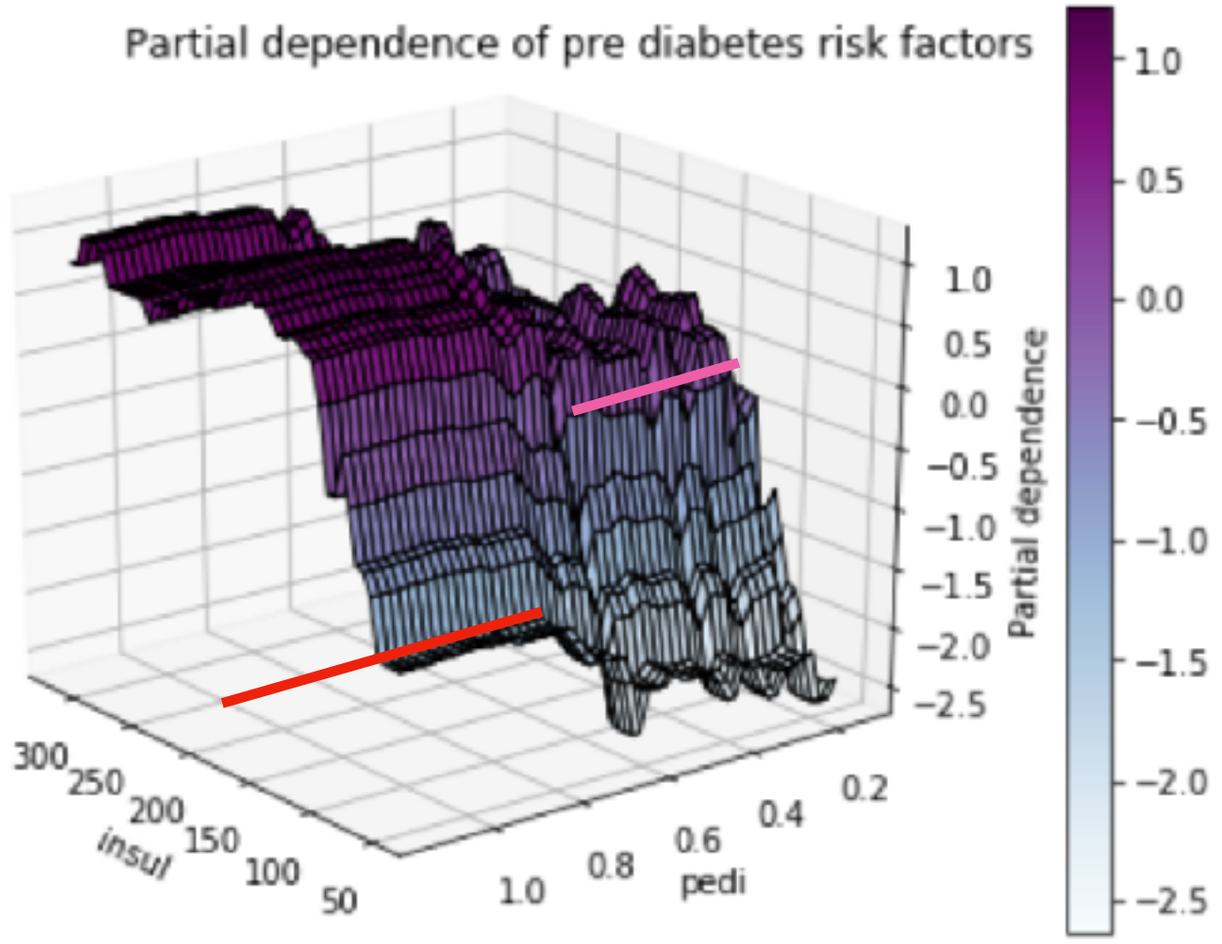


**** All algorithms trained using standardized data. Hyper parameters tuned using Grid Search Cross-Validation.***

Partial Dependence Plots to Identify High Risk Subpopulations

Subpopulation with pedigree function > .5 have higher insulin levels





Analysis of High Risk Subpopulation

High risk subpopulation (pedigree function >= .6, normal glucose, high insulin levels) has a ***2/3 greater relative risk of developing prediabetes*** *compared to a control population (pedigree < .6 + normal glucose + high insulin)*

age_y	0								1							
pedi_h	0				1				0				1			
insul_h	0		1		0		1		0		1		0		1	
gluc_h	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
class																
0	0.702	0.345	0.524	0.138	0.6	0.143	0.2	0.364	0.883	0.4	0.897	0.471	0.84	0.308	0.667	0.417
1	0.298	0.655	0.476	0.862	0.4	0.857	0.8	0.636	0.117	0.6	0.103	0.529	0.16	0.692	0.333	0.583

Age < 30
Pedi < .6
Insulin > 166
Glucose < 140
Prediabetes = 48%

Age < 30
Pedi >= .6
Insulin > 166
Glucose < 140
Prediabetes = 80%

Of the ***women under 30 who develop prediabetes****, the prevalence of high risk subpopulation is* ***18%*** *compared to* ***8.5%*** *in the control population*

What was the actual or expected impact of the model?

- The expect impact of this project is:

    - Provide further evidence to support the findings of Ahlqvist, Storm and Karajamaki that multiple subpopulations exist that require different treatment and monitoring (e.g., the insulin-resistance cluster (15% of patients studied) has high risk of ESRD implying monitoring of kidney function)

    - To encourage healthcare regulators to consider broaden the guidelines for diagnosis and treatment of type 2 diabetes to start to include testing for different types of diabetes like the insulin-resistant cluster whose disease is not diagnosed with glucose monitoring until it is too late.

    - To encourage the use of machine learning tools with relative smaller dataset with complex causal relationships in the data, dig deeper into the causal relationships in the data using tools like PDP graphs and work with domain experts to help with interpretation

References

1. File:Pima Indian woman winnowing beans over her head onto a mat on the ground, Gila River Reservation(?), ca.1902 (CHS-771).jpg. *Wikimedia Commons, the free media repository*. Published March 8 2015. Retrieved 22:30, August 24, 2018 from https://commons.wikimedia.org/w/index.php?title=File:Pima_Indian_woman_winnowing_beans_over_her_head_onto_a_mat_on_the_ground,_Gila_River_Reservation(%3F),_ca.1902_(CHS-771).jpg&oldid=152417023.

2. Peter P, Lipska K. The rising cost of diabetes care in the USA. Lancet Diabetes Endocrinol. 2016 June;4(6): 479-480.

3. Ali M, Bullard KM. Achievement of Goals in U.S. Diabetes Care, 1999-2010. N Engl J Med 2013; 368:1613-1624.

4. Ahlqvist E, Storm P, Karajamaki A, et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. Lancet Diabetes Endocrinol. 2018;6(9):361-369.

5. Griner B. Decoding Health with Data Science and Machine Learning. Data Science & Learning Systems. https://briangriner.github.io/decoding-health-risk-factors-pre-diabetes-ML-3.5.18.html. Published March 5, 2018.

6. Sigillito V. UCI Machine Learning Repository [https://archive.ics.uci.edu/ml/machine-learning-databases/pima-indians-diabetes/]. Irvine, CA: University of California, School of Information and Computer Science.

7. Pima Indians Diabetes Database. kaggle [https://www.kaggle.com/uciml/pima-indians-diabetes-database].

8. Smith JW, Everhart JE, Dickson WC, Knowler WC, Johannes RS. Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. *Proceedings of the Annual Symposium on Computer Application in Medical Care*. November 1988:261-265.

9. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. 12th printing. New York, NY. Springer. 2017.

10. Esparza-Romero J, Valencia M. Differences in Insulin Resistance in Mexican and U.S. Pima Indians with Normal Glucose Tolerance. J. Clin. Endocrinol. Metab. 2010;95(11):358–E362. https://doi.org/10.1210/jc.2010-0297.