FICO HELOC Interpretable ML

Group ID: 3612202004

Choi Suhyun (3035554658)

CHONG Inbum (3035553355)

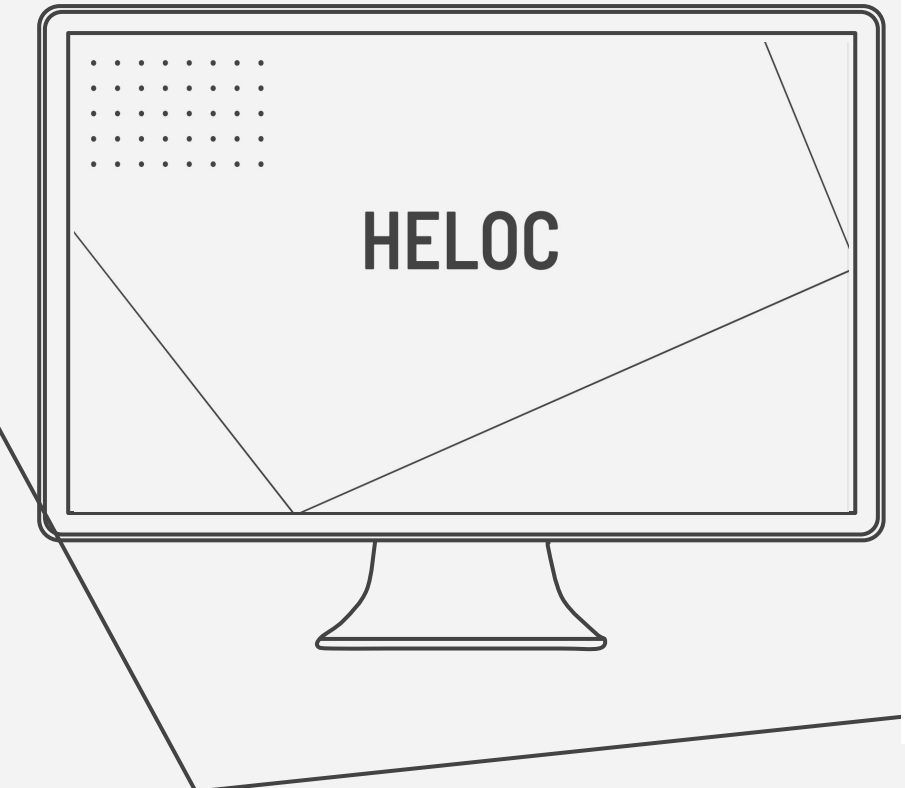
Kathleen Low Zi Yi (3035549017)

Wong Beatrice Chi Kan (3035374436)

Kim Sunghyun (3035603526)

01

Introduction



23 Predictors

1	Consolidated version of risk markers
2	Months Since Oldest Trade Open
3	Months Since Most Recent Trade Open
4	Average Months in File
5	Number Satisfactory Trades
6	Number Trades 60+ Ever
7	Number Trades 90+ Ever
8	Percent Trades Never Delinquent
9	Months Since Most Recent Delinquency
10	Max Delq/Public Records Last 12 Months
11	Max Delinquency Ever
12	Number of Total Trades
13	Number of Trades Open in Last 12 Months
14	Percent Installment Trades
15	Months Since Most Recent Inq excl 7days
16	Number of Inq Last 6 Months
17	Number of Inq Last 6 Months excl 7days
18	Net Fraction Revolving Burden
19	Net Fraction Installment Burden
20	Number Revolving Trades with Balance
21	Number Installment Trades with Balance
22	Number Bank/Natl Trades w high utilization ratio
23	Percent Trades with Balance

‘RiskFlag’: Good or Bad?

Robustness

Fairness

Interpretability

Causality

Trust

Monotonic Constraints

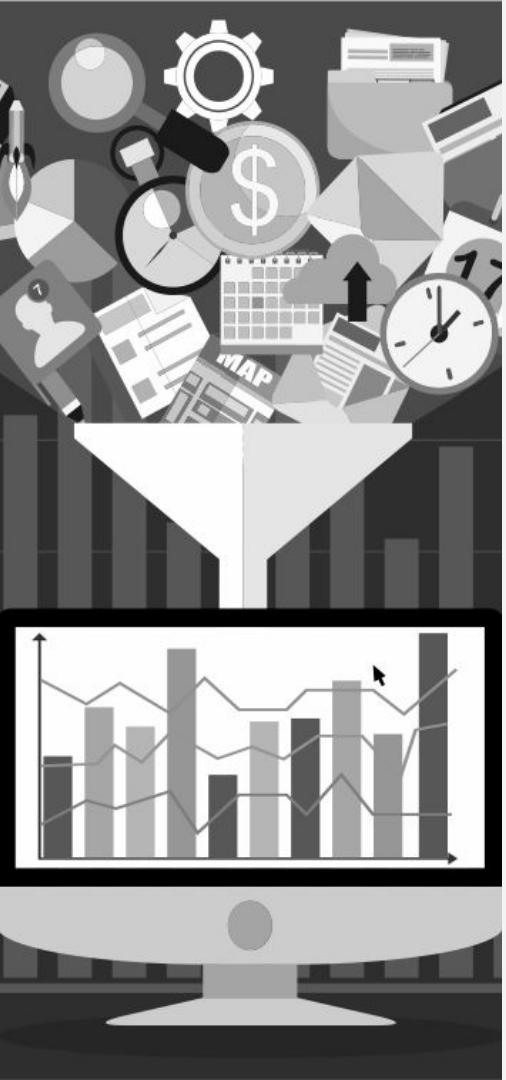
Model Selection

Set 1: Without Monotonicity Constraint

1. Logistic Regression
2. Generalized Additive Model
3. Decision Tree
4. Support Vector Machine
5. Boosting: XGBoost
6. Neural Network

Set 2: With Monotonicity Constraint

1. General Additive Model
2. Support Vector Machine
3. Boosting: XGBoost
4. Neural Network

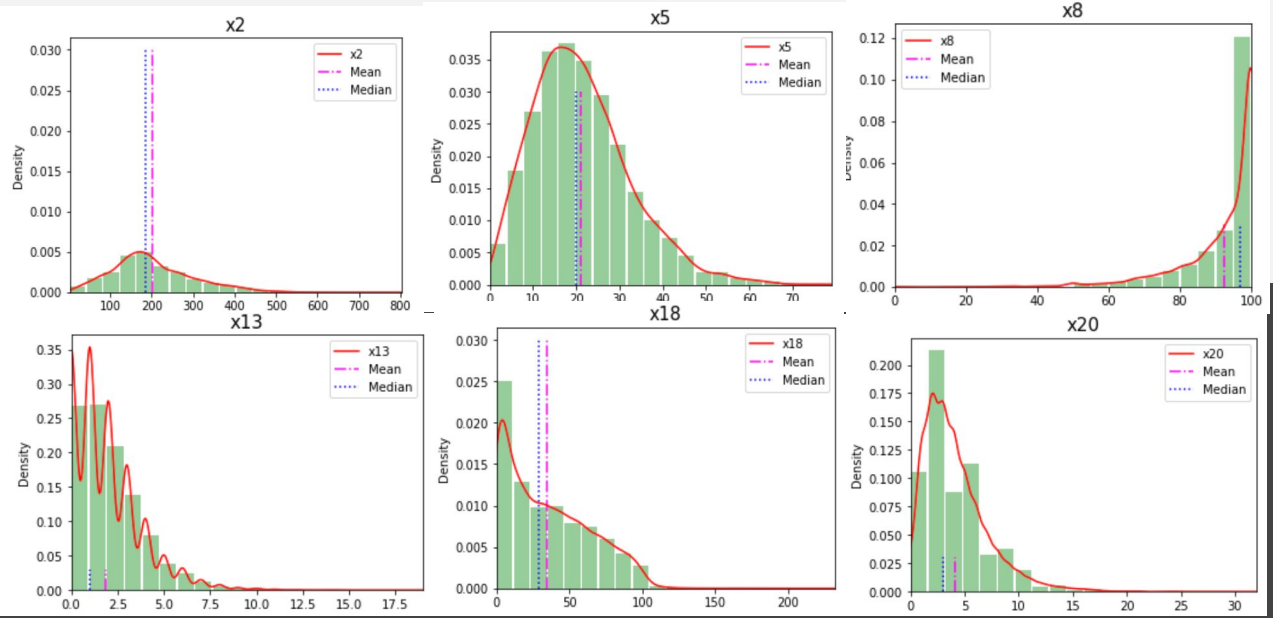


02

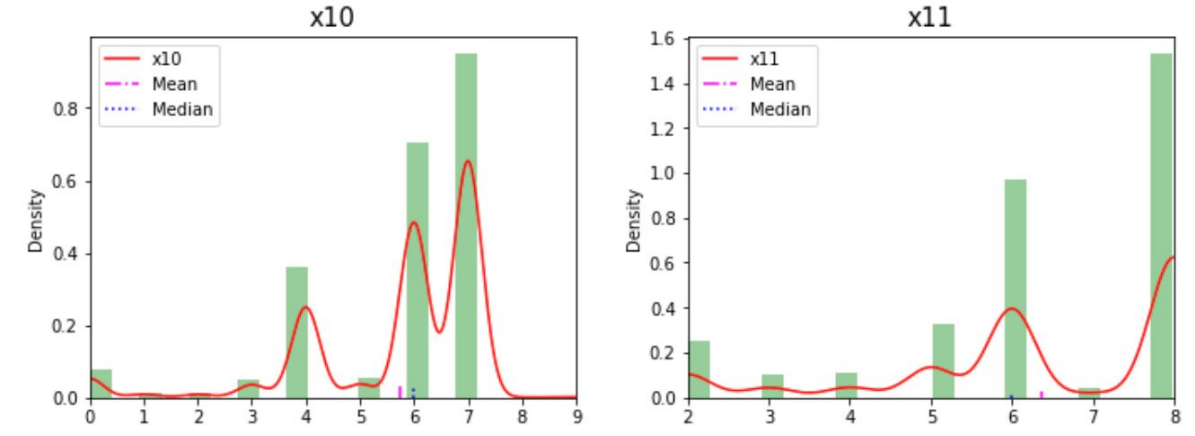
1. Imputation process for missing values
2. Outliers
3. Feature Selection method

EDA and Feature Engineering

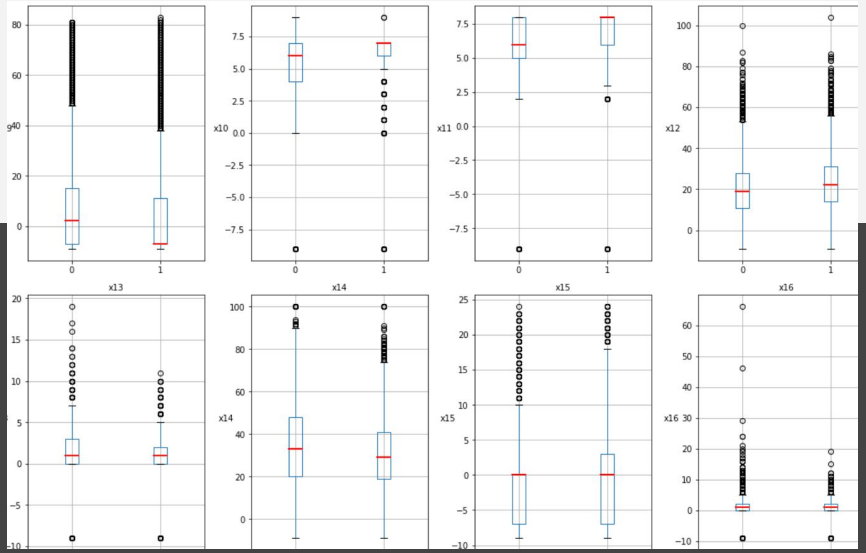
Missing Values Imputation (Median)



Missing Values Imputation (Mode)



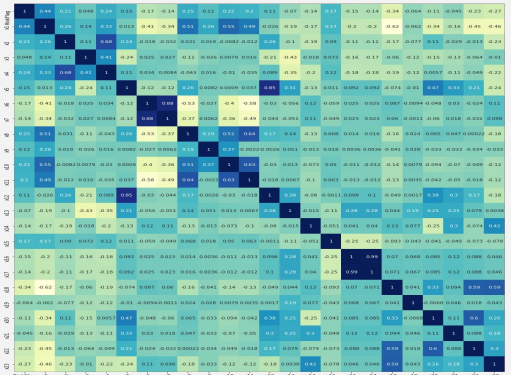
Outlier Analysis



	RiskFlag
7708	0
9501	1
5416	0
2695	1
1436	0
8866	1
1394	1
6389	1
7985	0
5559	0
6705	0
6357	0
6459	0
9078	0
9768	0
4020	0

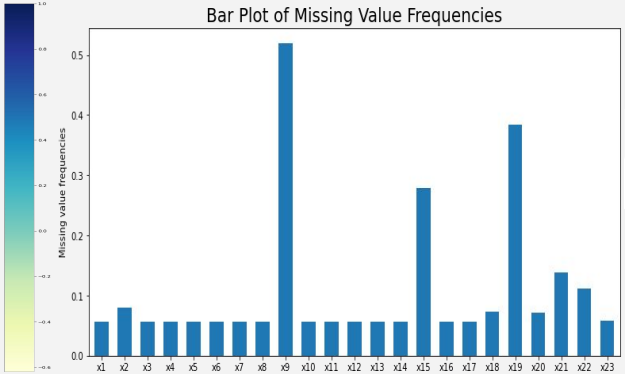
Feature Selection

Multicollinearity



‘X7’, ‘x17’

Missing values



‘X9’, ‘x15’, ‘x19’

Chi-squared test

```
chi_feature

['x1', 'x2', 'x6', 'x7', 'x10', 'x11', 'x15', 'x18', 'x22', 'x23']
```

03

Model Performance And Interpretability

Models without monotonicity constraints

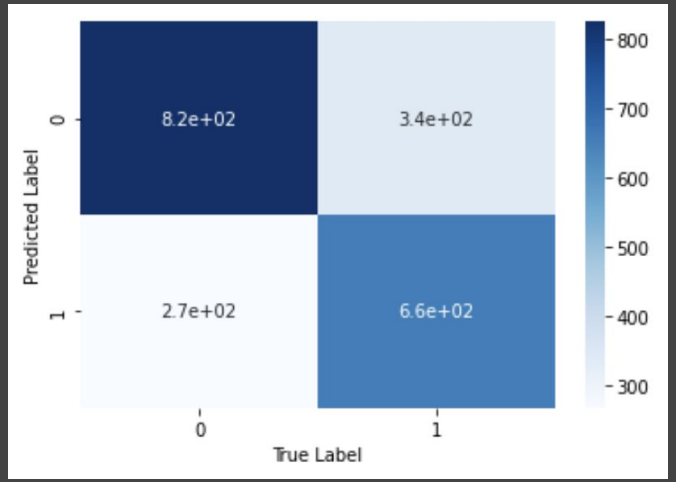
1. Logistic Regression
2. Generalized Additive Model
3. Decision Tree
4. Support Vector Machine
5. Boosting: XGBoost
6. Neural Network

<Logistic Regression>

Prediction Accuracy

```
Accuracy on the training set = 0.7224
Accuracy on the test set = 0.7089
```

Confusion Matrix



Model Interpretability

Logit Regression Results

Dep. Variable:	RiskFlag	No. Observations:	8351
Model:	Logit	Df Residuals:	8319
Method:	MLE	Df Model:	31
Date:	Sun, 29 Nov 2020	Pseudo R-squ.:	0.1964
Time:	16:50:20	Log-Likelihood:	-4645.4
converged:	True	LL-Null:	-5780.7
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-5.6049	0.540	-10.382	0.000	-6.663	-4.547
x1	0.0529	0.006	8.193	0.000	0.040	0.066
x2	0.0009	0.000	2.150	0.032	7.71e-05	0.002
x3	-0.0050	0.002	-2.026	0.043	-0.010	-0.000
x4	0.0076	0.001	5.561	0.000	0.005	0.010
x5	0.0336	0.005	6.296	0.000	0.023	0.044
x6	-0.0437	0.031	-1.404	0.160	-0.105	0.017
x8	0.0097	0.004	2.349	0.019	0.002	0.018
x12	0.0014	0.004	0.349	0.727	-0.006	0.009
x13	-0.0144	0.018	-0.789	0.430	-0.050	0.021
x14	-0.0084	0.002	-4.272	0.000	-0.012	-0.005
x16	-0.1292	0.016	-7.989	0.000	-0.161	-0.098
x20	-0.0577	0.014	-3.984	0.000	-0.086	-0.029
x21	0.0050	0.020	0.251	0.801	-0.034	0.044
x23	0.0033	0.002	1.743	0.081	-0.000	0.007

	Variable Names	Description
1	x1	Consolidated version of risk markers
18	x18	Net Fraction Revolving Burden. This is revolvi...
4	x4	Average Months in File
2	x2	Months Since Oldest Trade Open
22	x22	Number Bank/Natl Trades w high utilization ratio

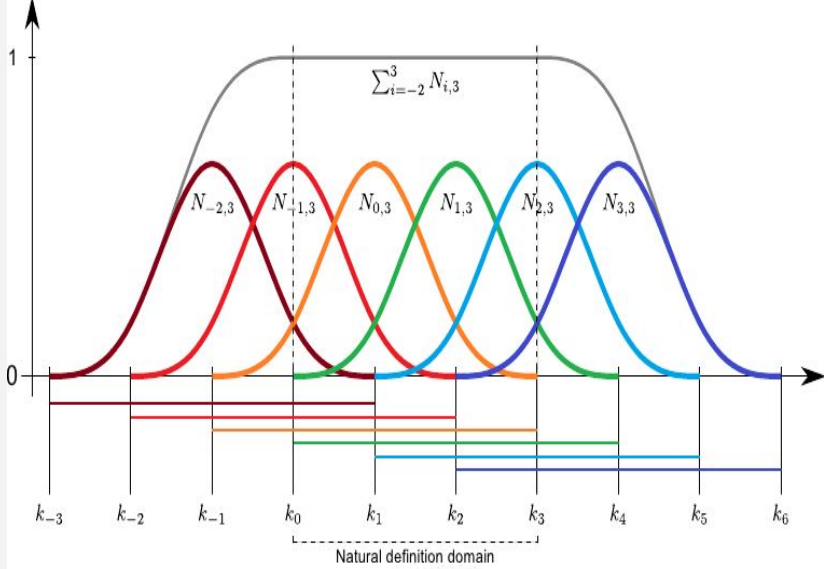
Generalized Additive Model

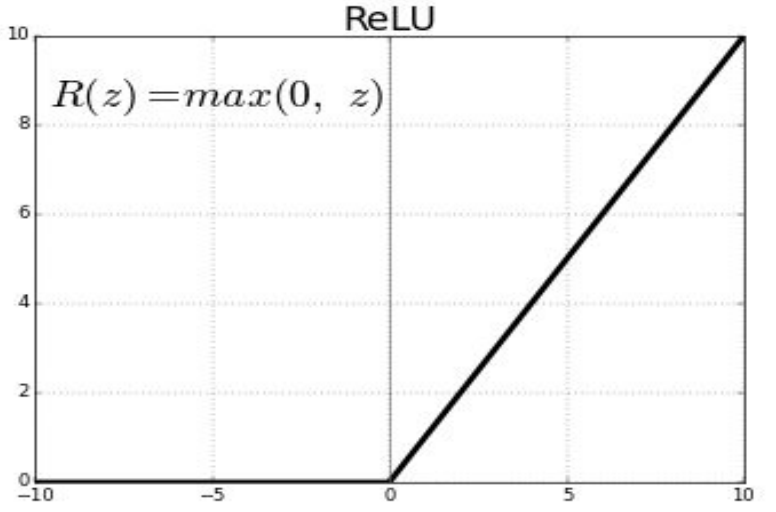
$$g(\mathrm{E}(Y)) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_m(x_m).$$

Nonparametric regression

Relax assumption of linearity

Highly complex non-linear relationship



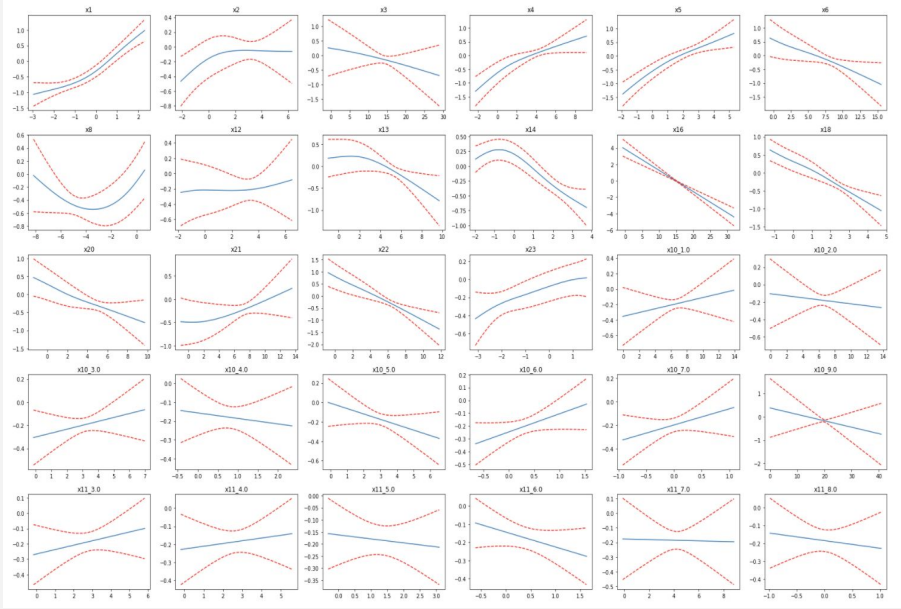


B-Spline

Prediction Accuracy

Training : 71.751% **Test: 70.889%**

Model interpretability

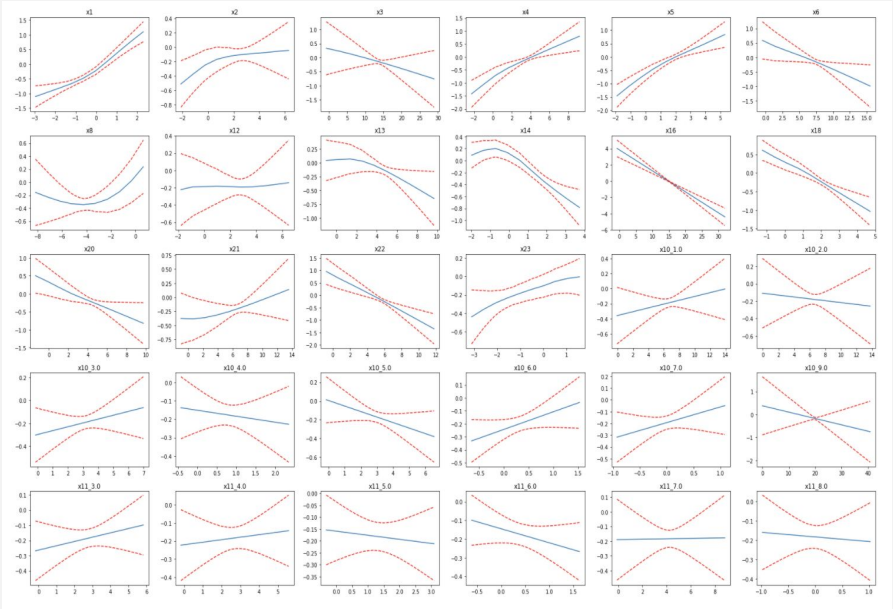


Piecewise ReLU

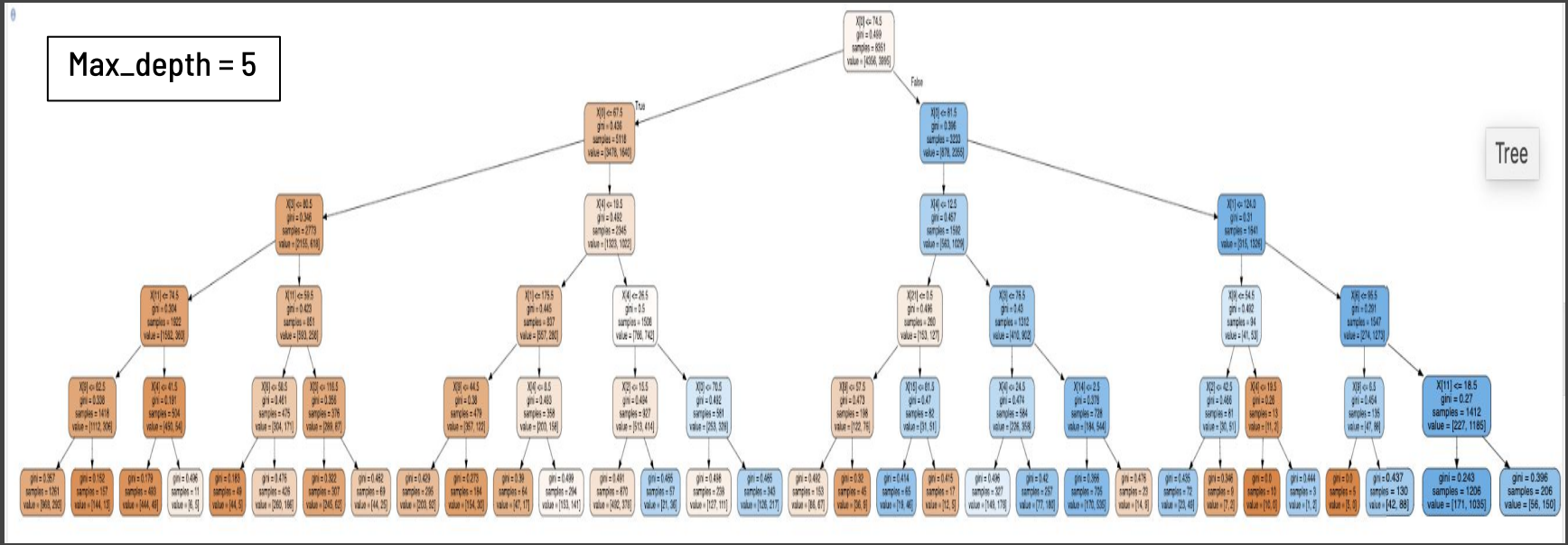
Prediction Accuracy

Training : 71.740% **Test: 70.841%**

Model interpretability



Decision Tree



Accuracy on train set: 72.051%
Accuracy on test set: 70.889%

SVM :Prediction Accuracy

	Linear SVC	Kernel SVC (RBF)	Tuned Kernel SVC (RBF)
Train set	71.608%	71.5%	74.638%
Test set	71.033%	71.08%	71.367%

Hyperparameter tuning

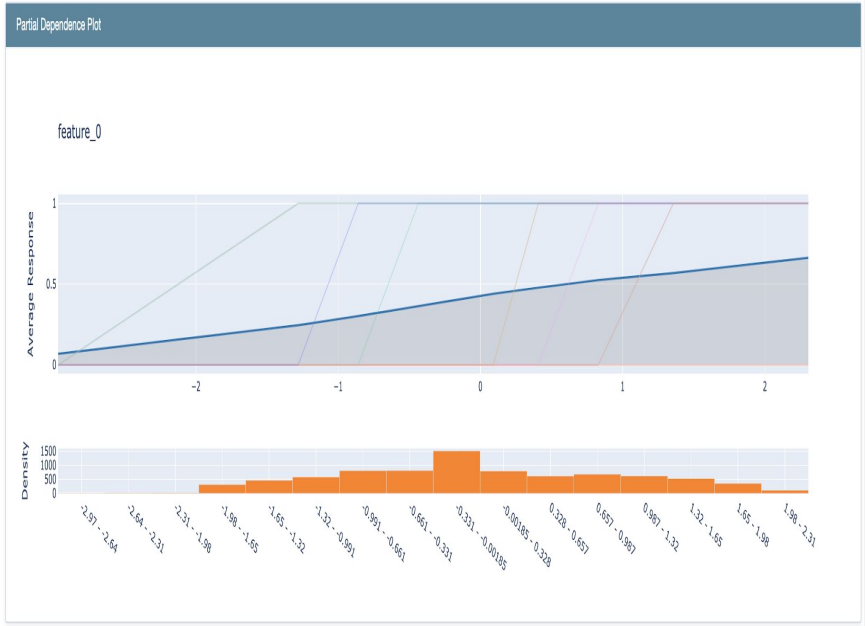
GridSearchCV

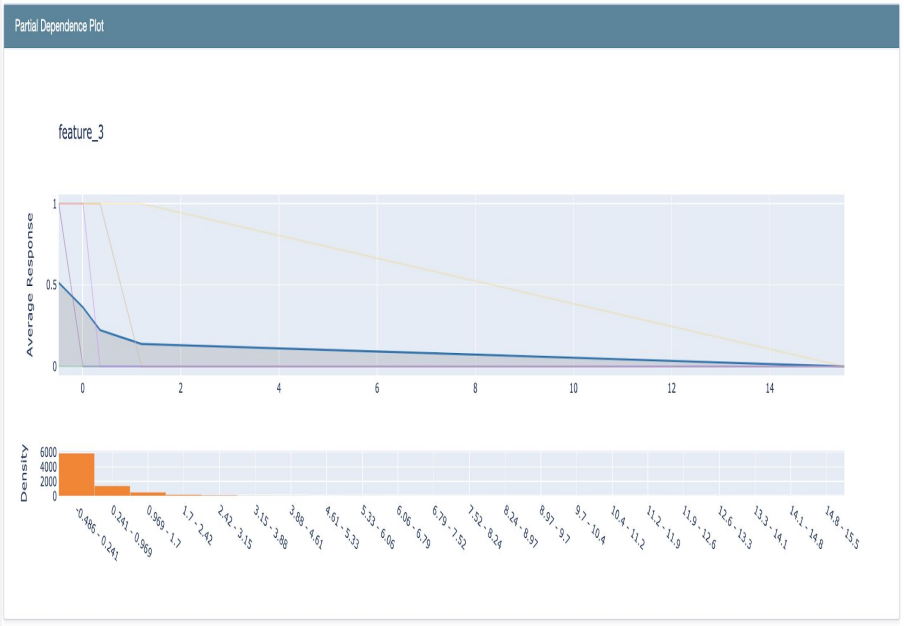
Gamma : [0.001, 0.01, **0.1**, 0.0, 1.0]

C : [**1.0,** 10.0, 100.0]

SVM : Model Interpretability

Partial Dependence Plot





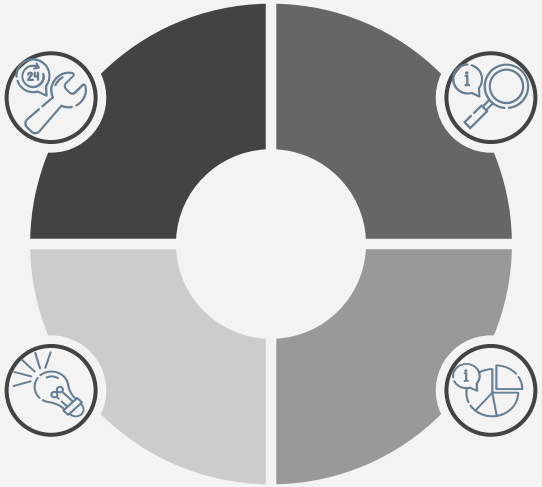
X_1 : Consolidated Version of Risk markers

X_6: Number Trades 60+ Ever

Extreme Gradient Boosting (XGBoost)

Regularization

Weighted Quantile Sketch



Sparsity Awareness

Cross-Validation

Hyperparameter Tuning

`sklearn.model_selection.GridSearchCV`

To tune . . .

Max_depth: [3, **5**, 10]

Min_child_weight: [1, **4**, 8]

Gamma: [0.0, 0.1, **0.2**, 0.3, 0.4, 0.5]

Subsample: [0.6, 0.7, **0.8**, 0.9, 1.0]

Colsample_bytree: [**0.6**, 0.7, 0.8, 0.9, 1.0]

```
XGBClassifier(colsample_bytree=0.6,colsample_bylevel=0.5,gamma=0.2,learning_rate=0.01,max_depth=5,
              min_child_weight=4,missing=-9,n_estimators=130,n_jobs=3,objective='binary:logistic',
              reg_alpha=0.02,reg_lambda=0.05,subsample=0.8,colsample_bynode=0.6,nthread=4,
               scale_pos_weight=1,seed=1)
```

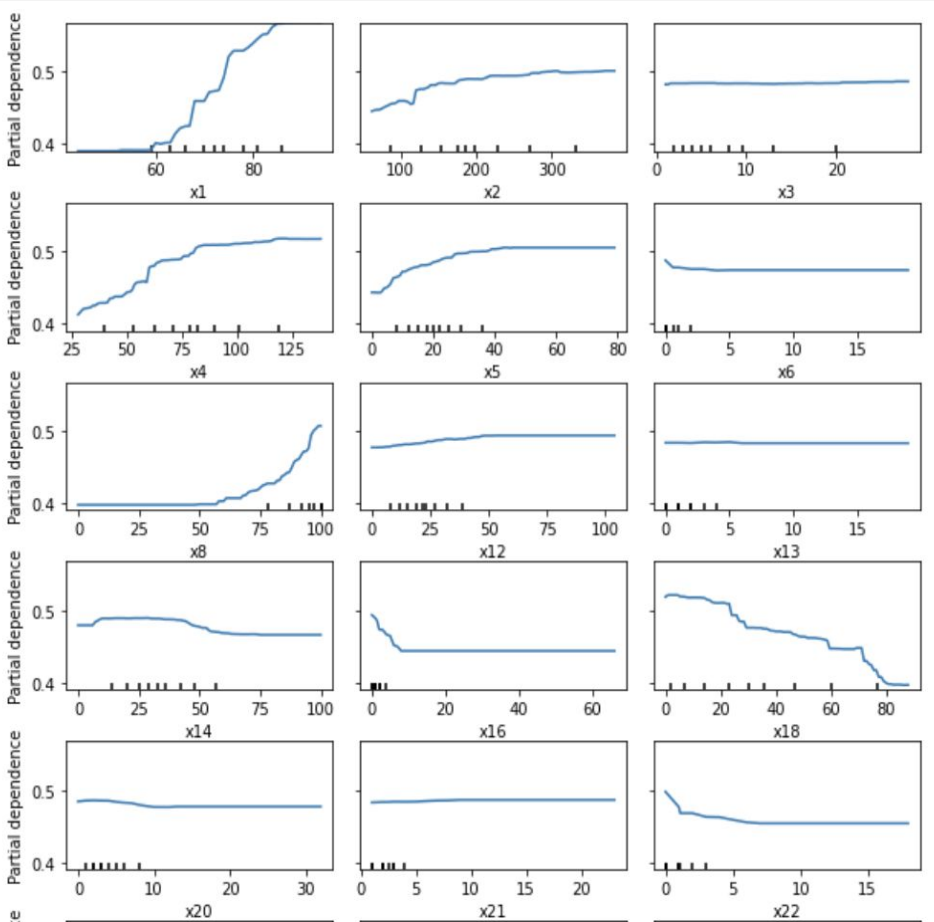
Prediction Accuracy

* Best ROC AUC : 0.7977

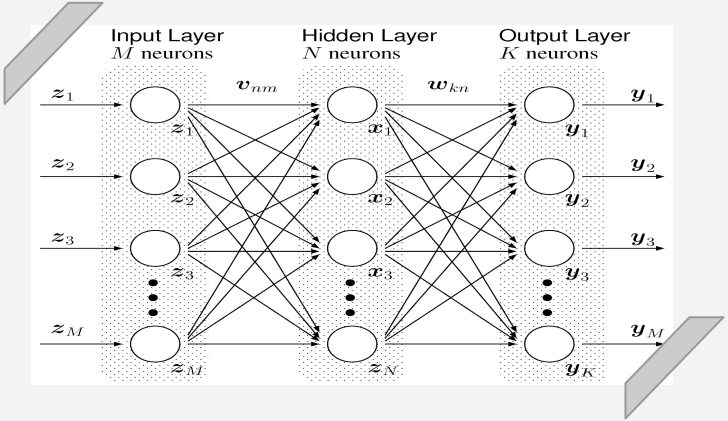
Train set : 73.512%

Test set : 71.941%

Model Interpretability



Neural Network : MLP



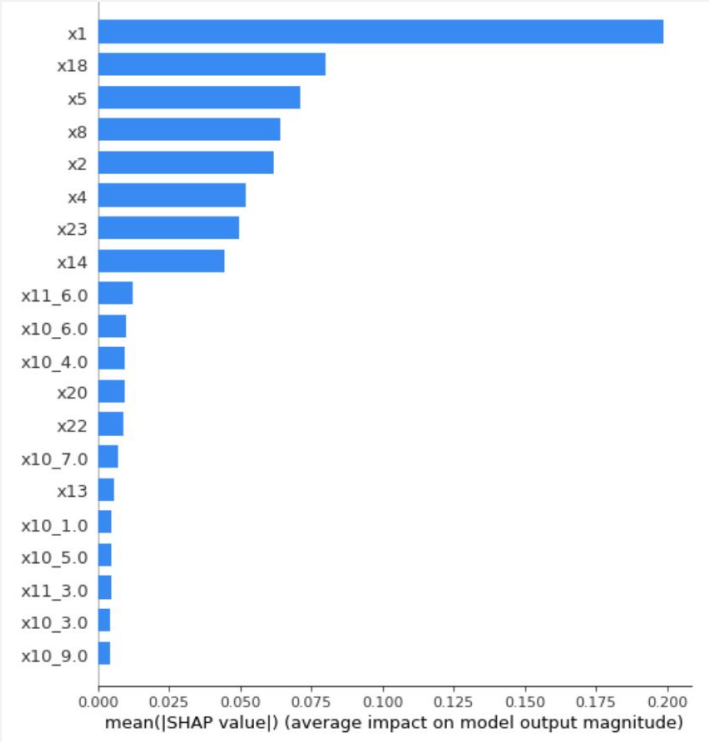
Prediction Accuracy

Hyperparameter Tuning: RandomizedSearchCV

```
'alpha': 10.0 ** -np.arange(-4, 4,4)
```

- ➔ **Accuracy on train set: 71.20%**
- ➔ **Accuracy on test set: 71.89%**

Model Interpretability

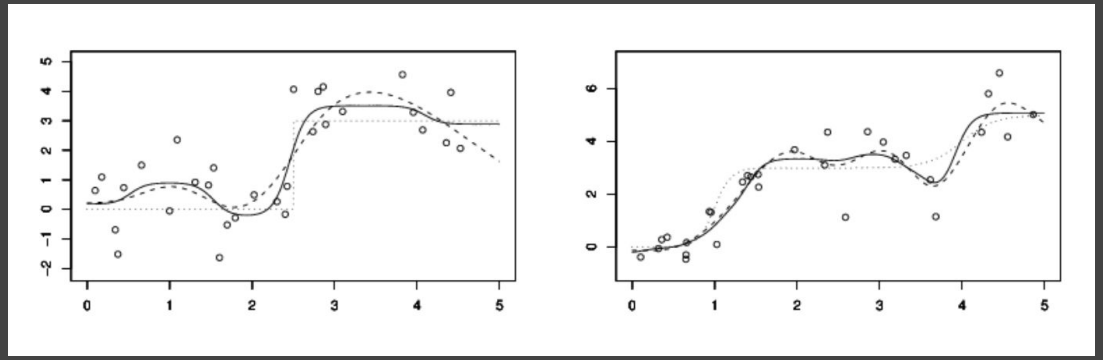


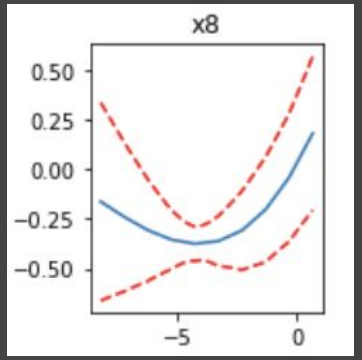
Models with monotonicity constraints

1. Generalized Additive Model
2. Extreme Gradient Boosting
3. Artificial Neural Network

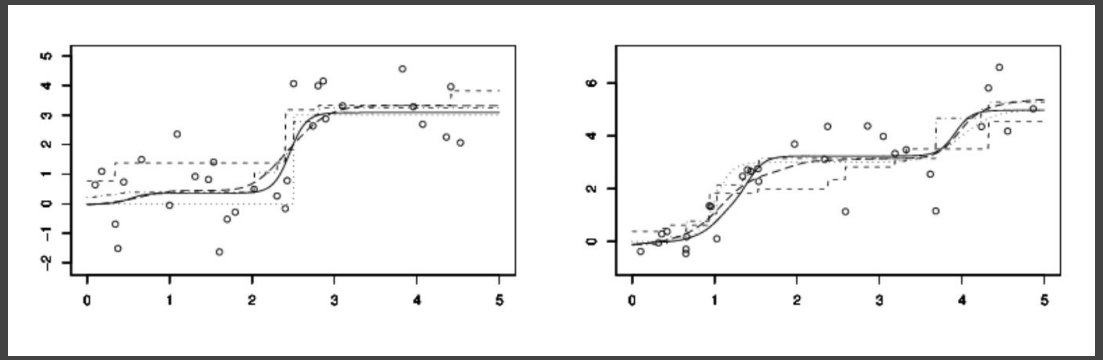
GAM with constraints

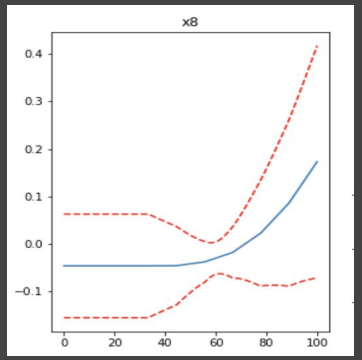
PDP without Monotonicity



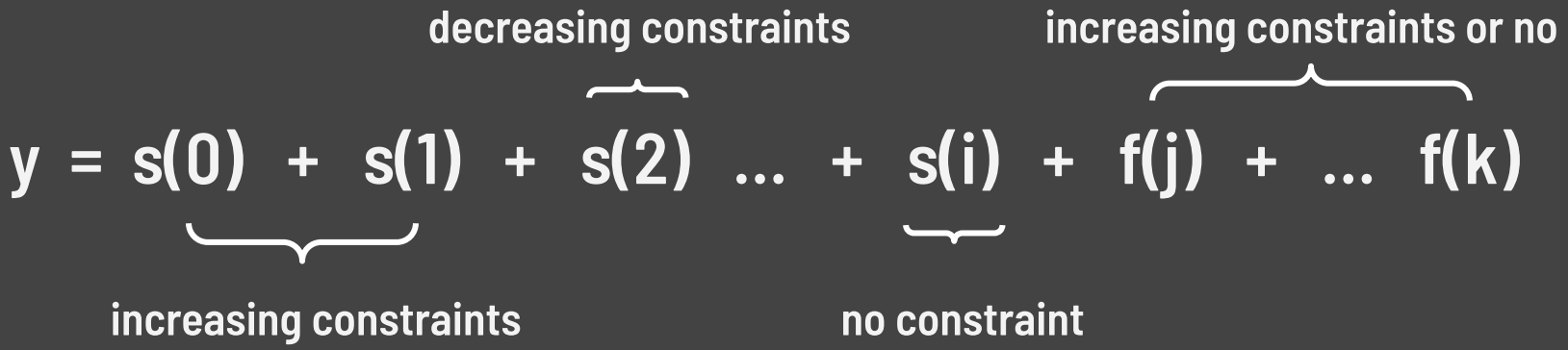


PDP with Monotonicity





GAM with constraints



Accuracy = 0.7032

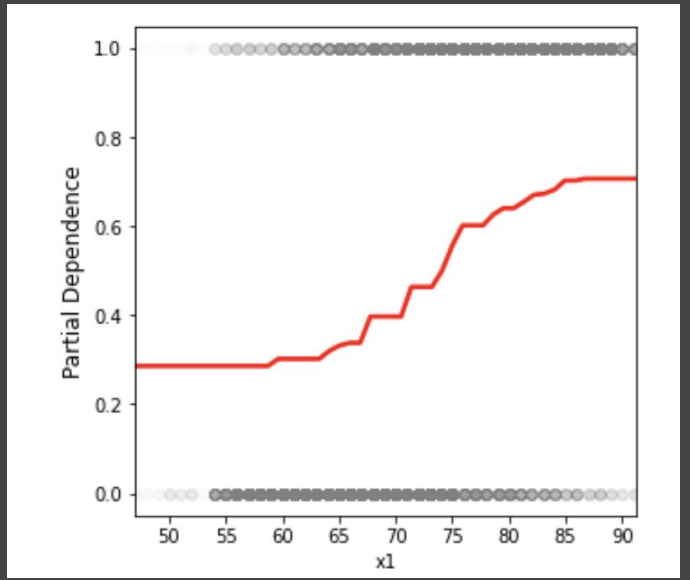
XGBoost

Increasing constraints on:

$$f(x_1, x_2, \ldots, x, \ldots, x_{n-1}, x_n) \leq f(x_1, x_2, \ldots, x', \ldots, x_{n-1}, x_n)$$

Decreasing constraints:

$$f(x_1, x_2, \ldots, x, \ldots, x_{n-1}, x_n) \geq f(x_1, x_2, \ldots, x', \ldots, x_{n-1}, x_n)$$



Accuracy = 0.7141

Monotone MLP

Constraints on all x_j such that:

partial (y) / partial (x_j) > 0 for increasing constraints

partial (y) / partial (x_j) < 0 for decreasing constraints

Outputs regression line as in sigmoid

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 731 234
         1 361 766

               Accuracy : 0.7156
                 95% CI : (0.6957, 0.7348)
                 Mcnemar's Test P-Value : 2.398e-07

              Precision : 0.7575
                 Recall : 0.6694
                     F1 : 0.7107
             Prevalence : 0.5220
         Detection Rate : 0.3494
   Detection Prevalence : 0.4613
      Balanced Accuracy : 0.7177
```

	GAM	XGBoost	Monotone MLP
Accuracy	0.7032	0.7141	0.7156
Interpretability	Highest	Second Highest	Lowest

04

Conclusion

Accuracy

XGBoost - low learning rate Without monotonicity	**0.7194**
Neural Network Without monotonicity	**0.7189**
Neural Network With monotonicity	**0.7156**
XGBoost With monotonicity	**0.7141**

Interpretability

Logistic

Able to see effect of features

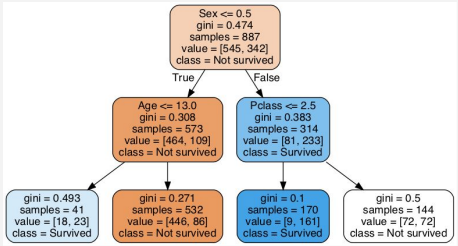
Logit Regression Results

Dep. Variable:	RiskFlag	No. Observations:	8351
Model:	Logit	Df Residuals:	8319
Method:	MLE	Df Model:	31
Date:	Sun, 29 Nov 2020	Pseudo R-squ.:	0.1964
Time:	16:50:20	Log-Likelihood:	-4645.4
converged:	True	LL-Null:	-5780.7
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-5.6049	0.540	-10.382	0.000	-6.663	-4.547
x1	0.0529	0.006	8.193	0.000	0.040	0.066
x2	0.0009	0.000	2.150	0.032	7.71e-05	0.002
x3	-0.0050	0.002	-2.026	0.043	-0.010	-0.000
x4	0.0076	0.001	5.561	0.000	0.005	0.010
x5	0.0336	0.005	6.296	0.000	0.023	0.044
x6	-0.0437	0.031	-1.404	0.160	-0.105	0.017
x8	0.0097	0.004	2.349	0.019	0.002	0.018
x12	0.0014	0.004	0.349	0.727	-0.006	0.009
x13	-0.0144	0.018	-0.789	0.430	-0.050	0.021
x14	-0.0084	0.002	-4.272	0.000	-0.012	-0.005
x16	-0.1292	0.016	-7.989	0.000	-0.161	-0.098
x20	-0.0577	0.014	-3.984	0.000	-0.086	-0.029
x21	0.0050	0.020	0.251	0.801	-0.034	0.044
x23	0.0033	0.002	1.743	0.081	-0.000	0.007

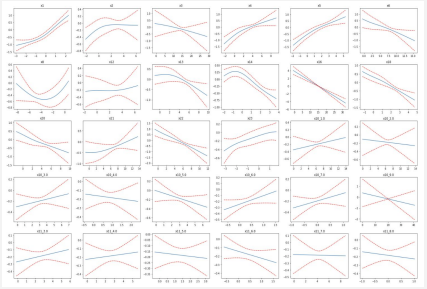
Decision tree

Partitioning
Clear Visualisation



GAM

Partial Dependence Plot



SVM

Partial Dependence Plot

XGBoost

Partial Dependence Plot
Depth of trees make it harder to interpret

Neural Network

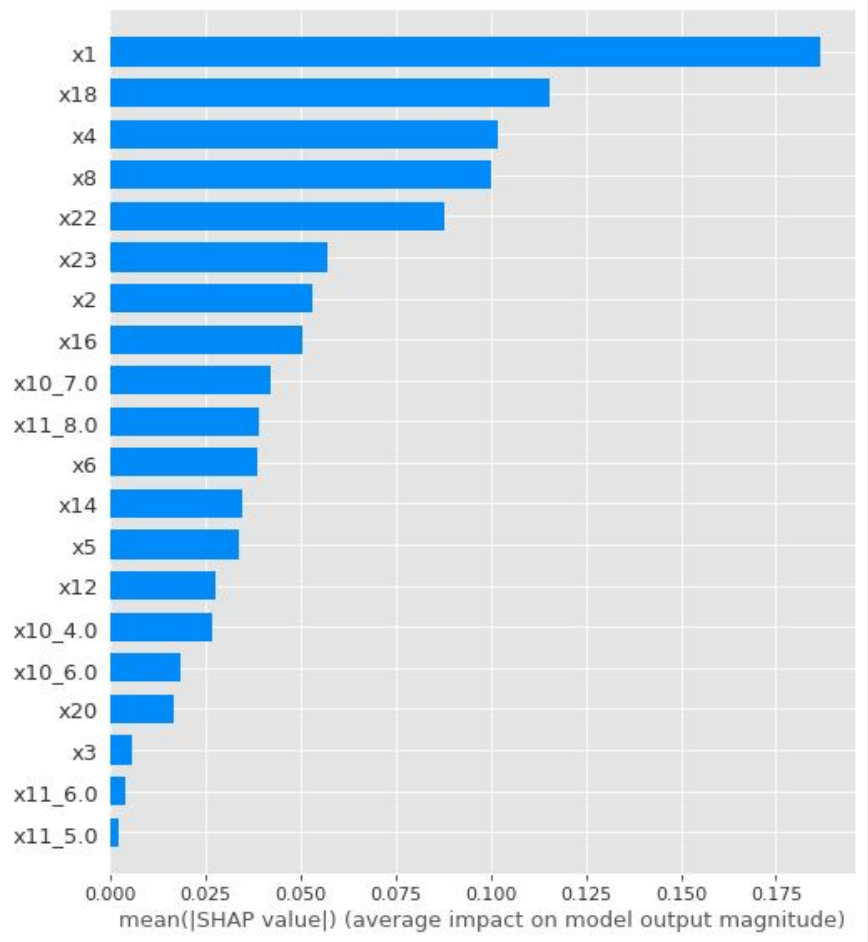
Hidden Layers make it harder to interpret

XGBoost

Low learning rate

Without monotonicity constraints

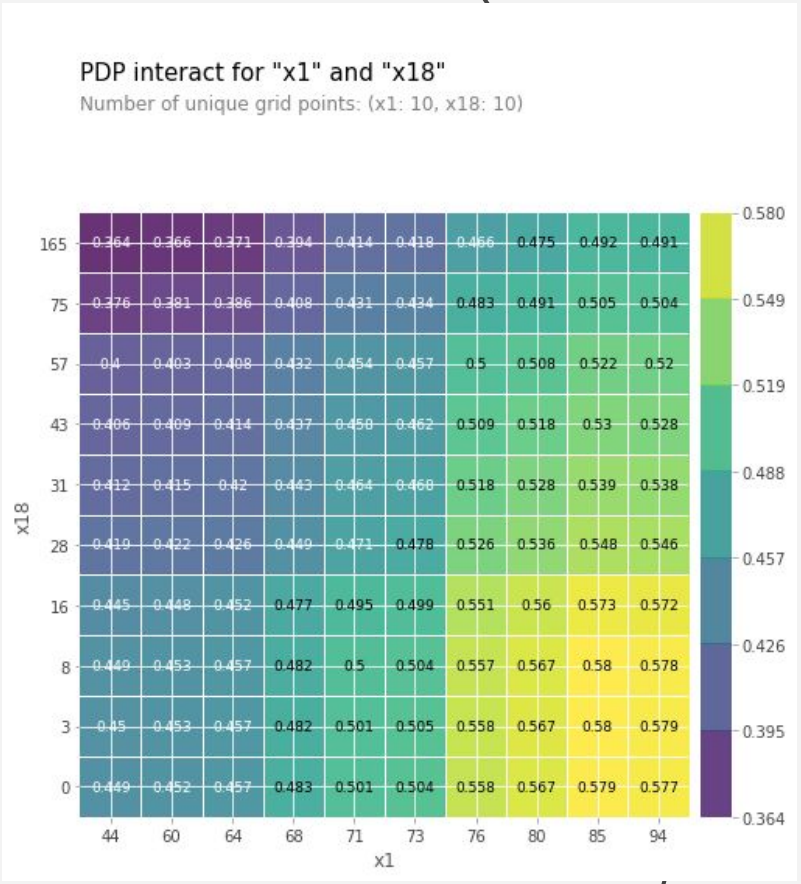
0.7194



XGBoost

Low learning rate

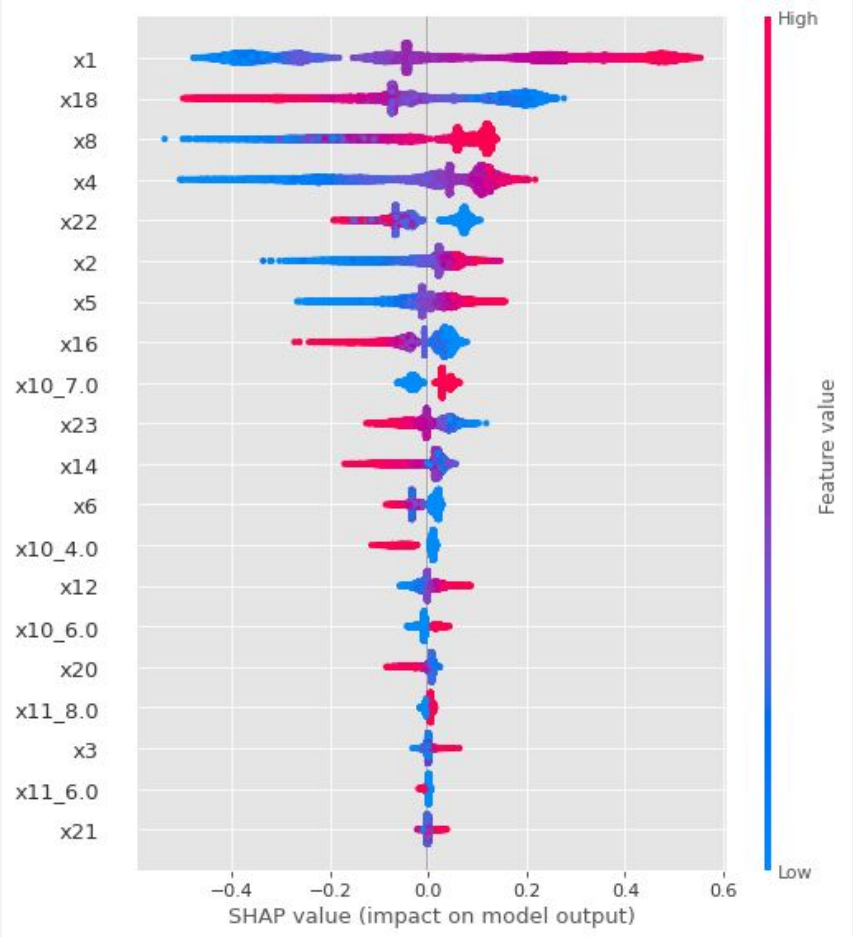
Without monotonicity constraints



XGBoost

Low learning rate

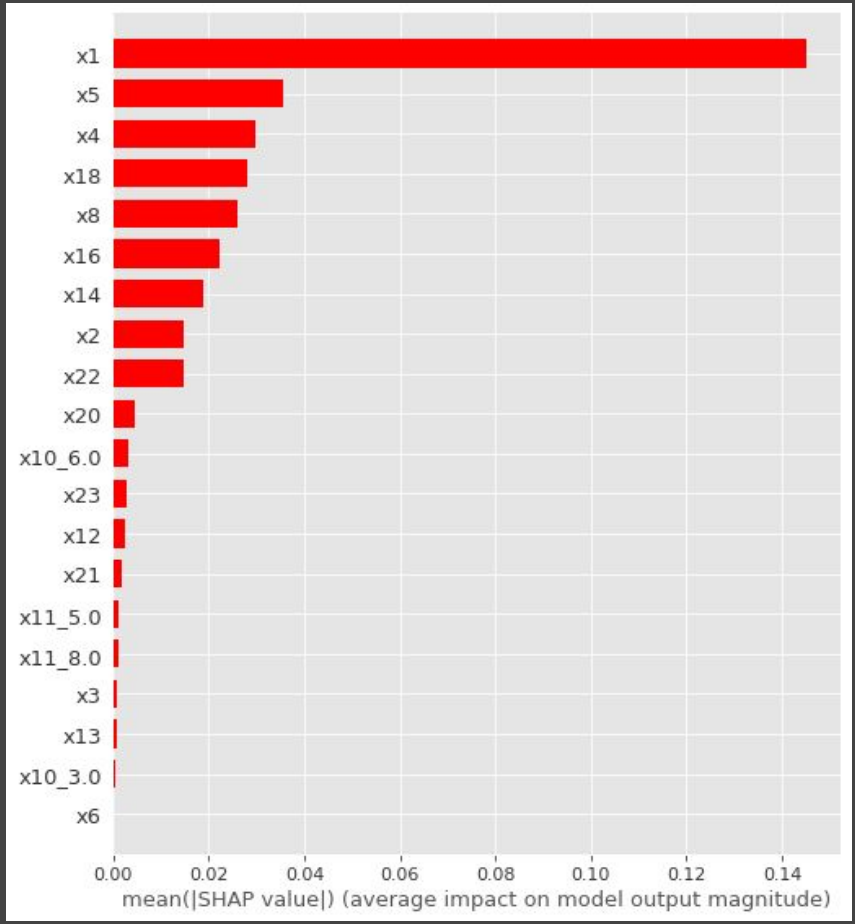
Without monotonicity constraints



XGBoost

With monotonicity constraints

0.7141

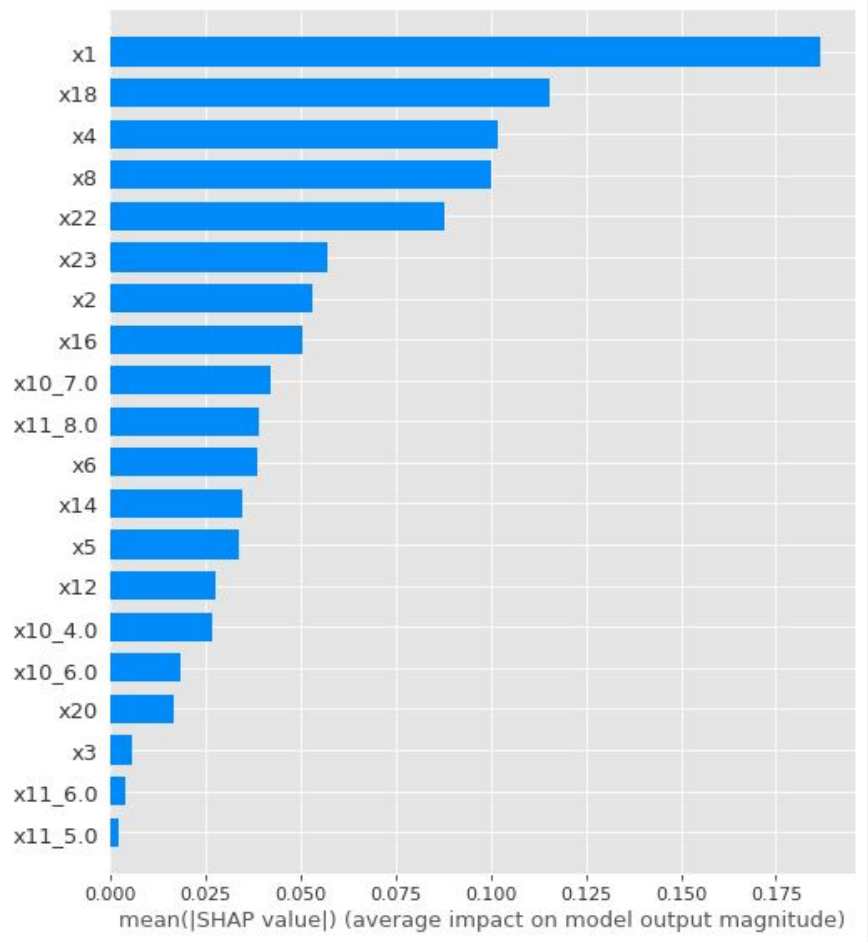


XGBoost

Low learning rate

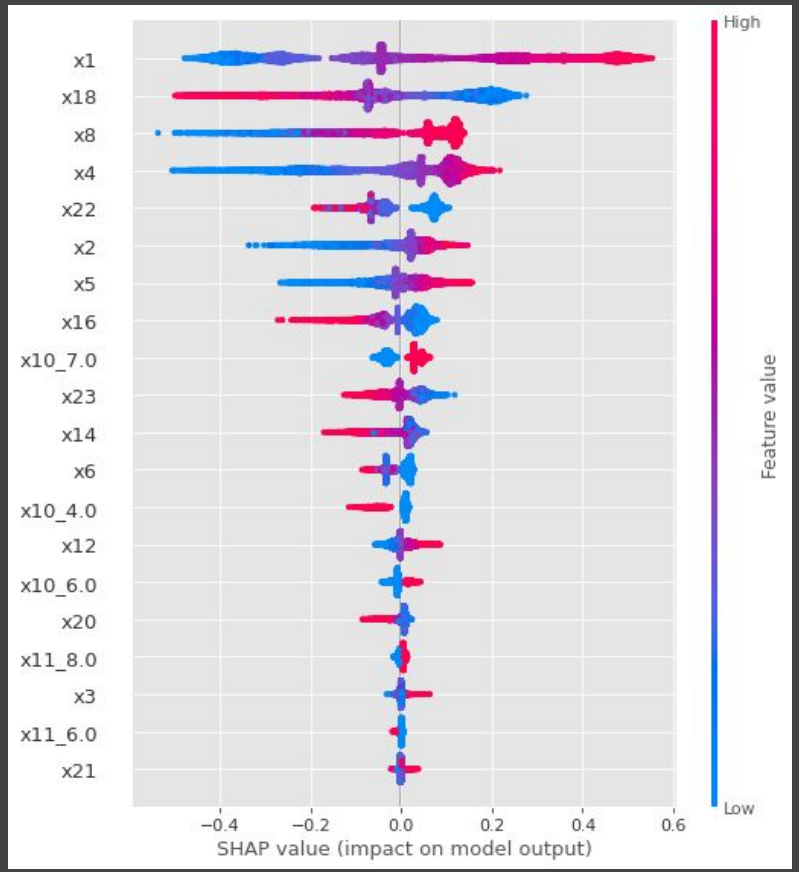
Without monotonicity constraints

0.7194



XGBoost

With monotonicity constraints



Final Models

Without Monotonicity Constraints

XGBoost

Monotonicity Constrained

XGBoost

QnA