Predictive Analytics – Session 5

Predicting Classifications - Analysis

Associate Professor Ole Maneesoonthorn

Associate Professor in Econometrics and Statistics

Melbourne Business School

O.Maneesoonthorn@mbs.edu

mbs.edu

GLOBAL. BUSINESS. LEADERS.

THE UNIVERSITY OF MELBOURNE | MELBOURNE BUSINESS SCHOOL

Energy Consumer Sentiment Survey

- Semi-annual survey on consumer sentiment on their energy providers
  - Household
  - Business
- Various demographics recorded, e.g. age, income, current financial situations, etc.
- We are interested in the response: Switch
  - Whether the consumer is intending to switch provider
  - Prediction for communications targeting, targeted marketing

Energy Consumer Sentiment Survey

- Let’s look at constructing the predictive assessment of the following models for household consumers
  - Naïve
  - Logistic regression
  - Decision tree
  - Neural network
  - K-NN classification
  - Bagging
  - Random forest
  - Boosting

- Using Dec 2019 survey

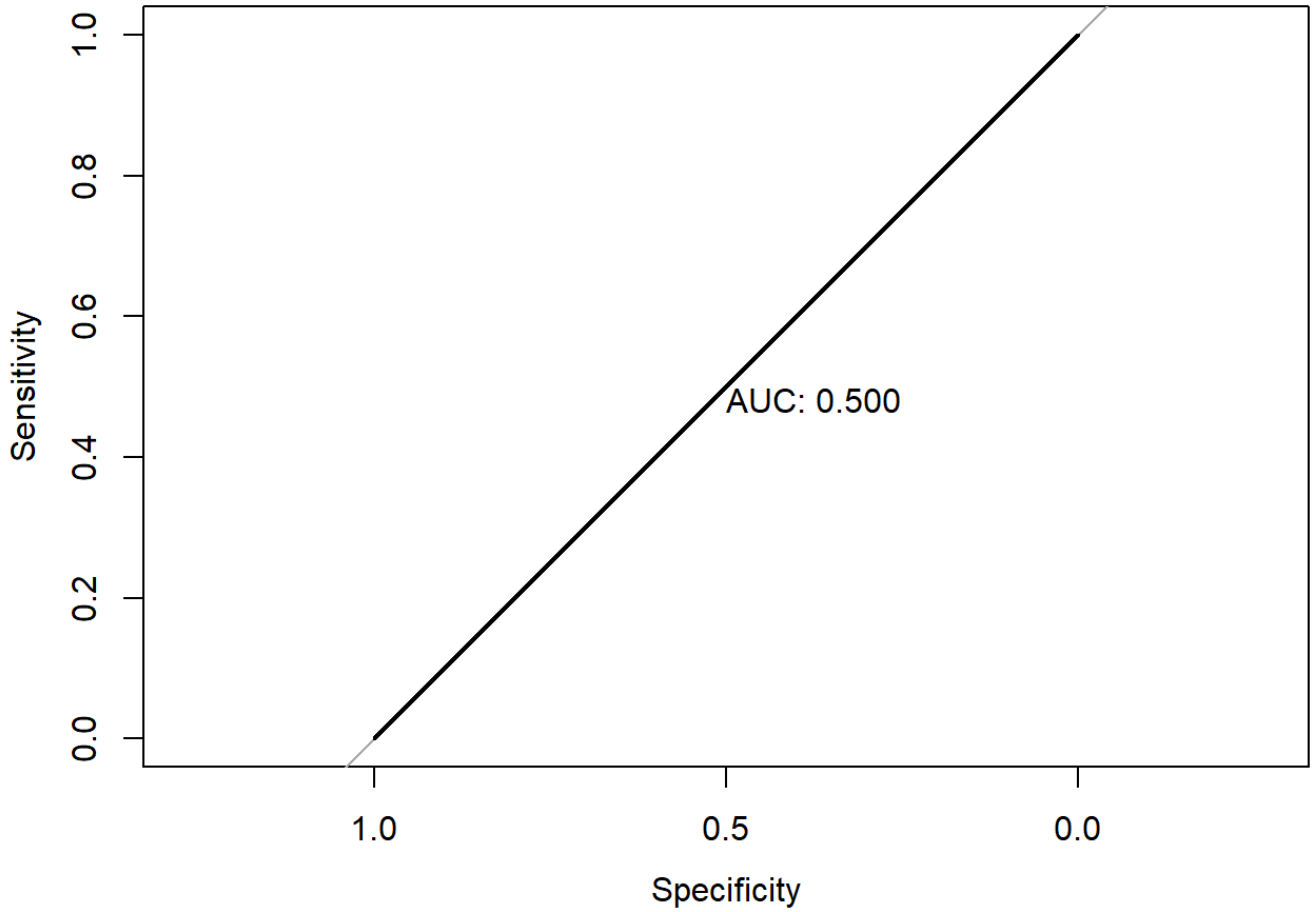
Energy Consumer Sentiment Survey

Naïve prediction – what is it?

- Pretend that no input variables are available
- What would be the best prediction?
- Sample proportion of intention to switch!
    - Training set data indicates that about 14% of respondents intend to switch provider
    - All test data points are allocated a probability of 14% for intention to switch

Energy Consumer Sentiment Survey

- Naïve probability 14%
- Naïve ROC curve →
- Naïve only ever predict “No”
  - At cutoff prob = 0.5
- This is the benchmark case:
  - AUC = 0.5 & Gini = 0



Energy Consumer Sentiment Survey

Logistic Regression: relationships revealed

- Coefficient can only be interpreted as directional impact on probability to switch

Variable	Relationship (* indicates significance)
Employment (others relative to self-employed)	Positive
Gender (female relative to male)	Positive
Annual household income (relative to under 20k)	Mixed, positive for higher income but not significant
Education (relative to not completing yr 12)	Generally positive (* for uni degree)
Dependent children	Positive (*)
House type	Positive for units, Negative for other types
Bill satisfaction	Positive
Bill larger than expected	Positive (*)
Bill pressure	Positive

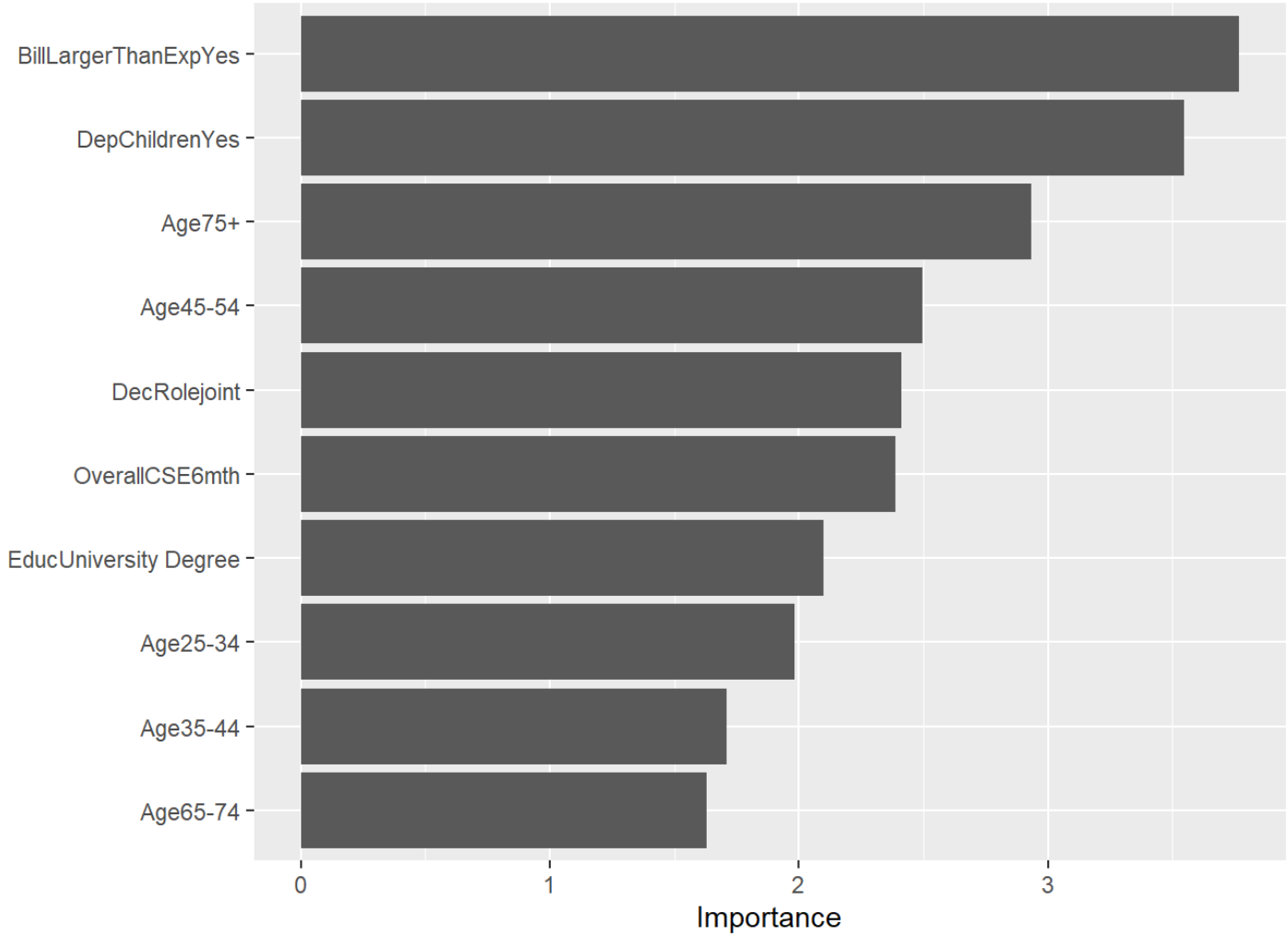
Energy Consumer Sentiment Survey

Logistic Regression: relationships revealed

Variable	Relationship (* indicates significance)
Decision role (joint relative to main)	Negative (*)
Age (others relative to 18-24 age group)	Negative (*)
Number of people in household	Negative
Value for money	Negative
Overall satisfaction	Negative (*)
Blackout satisfaction	Negative
Own home	Negative (very slight)

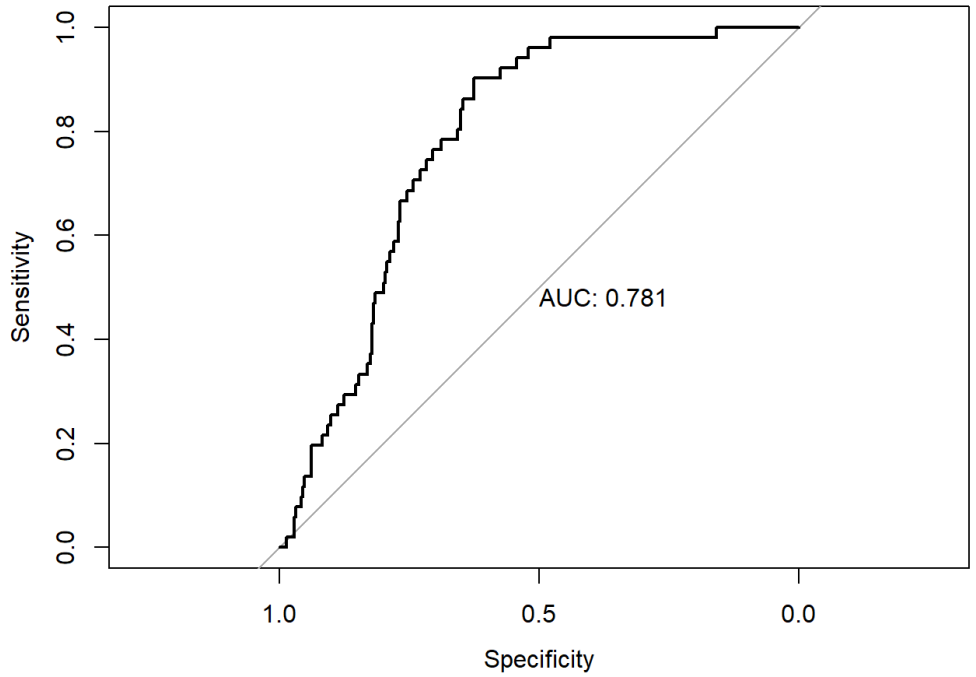
Energy Consumer Sentiment Survey

Logistic Regression: relationships that are significant



Energy Consumer Sentiment Survey

Logistic Regression: ROC curve

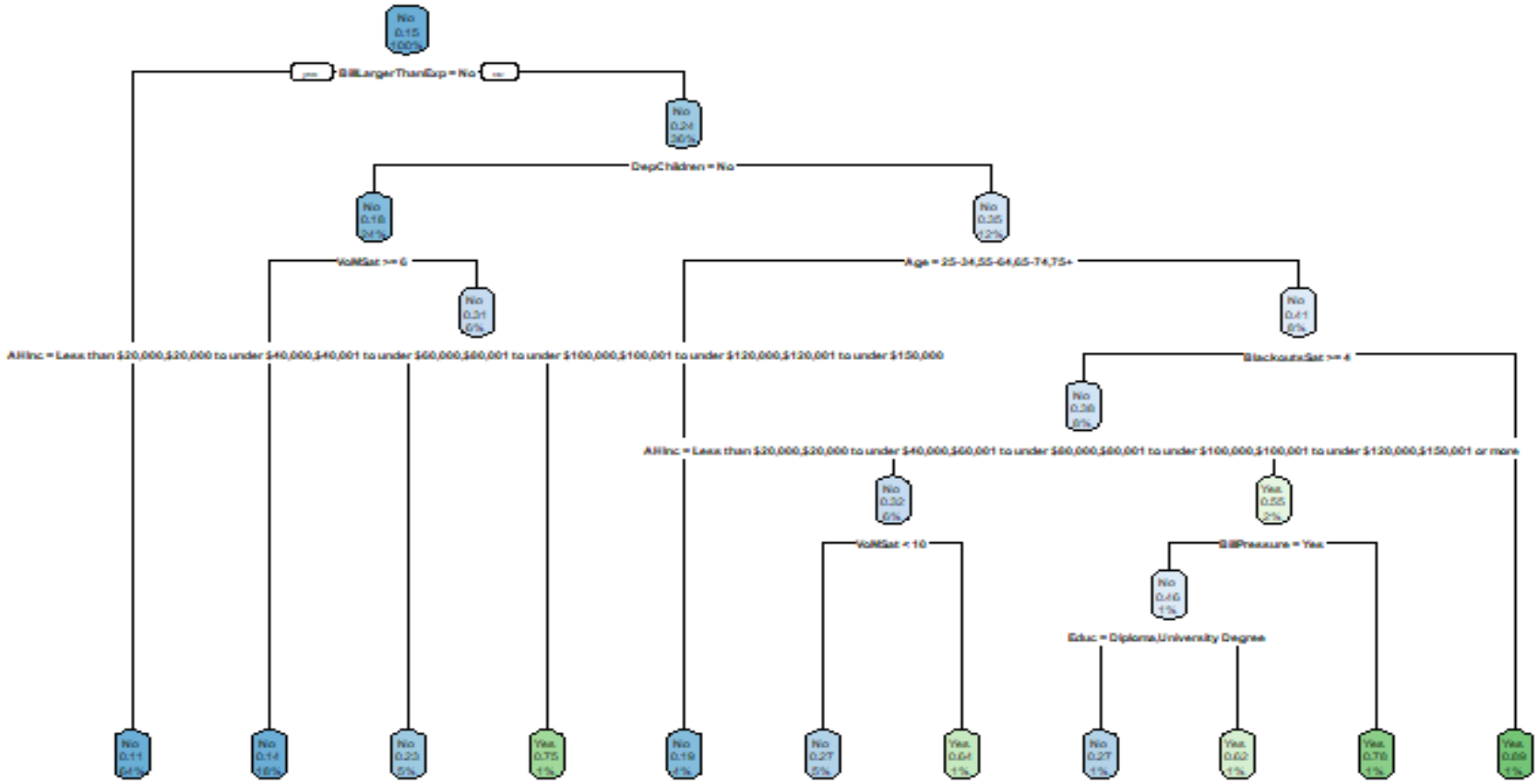


- Hit and miss table (@ cutoff probability of 0.5)

	Actual "Yes"	Actual "No"
Predicted "Yes"	0.000	0.000
Predicted "No"	0.126	0.874

Energy Consumer Sentiment Survey

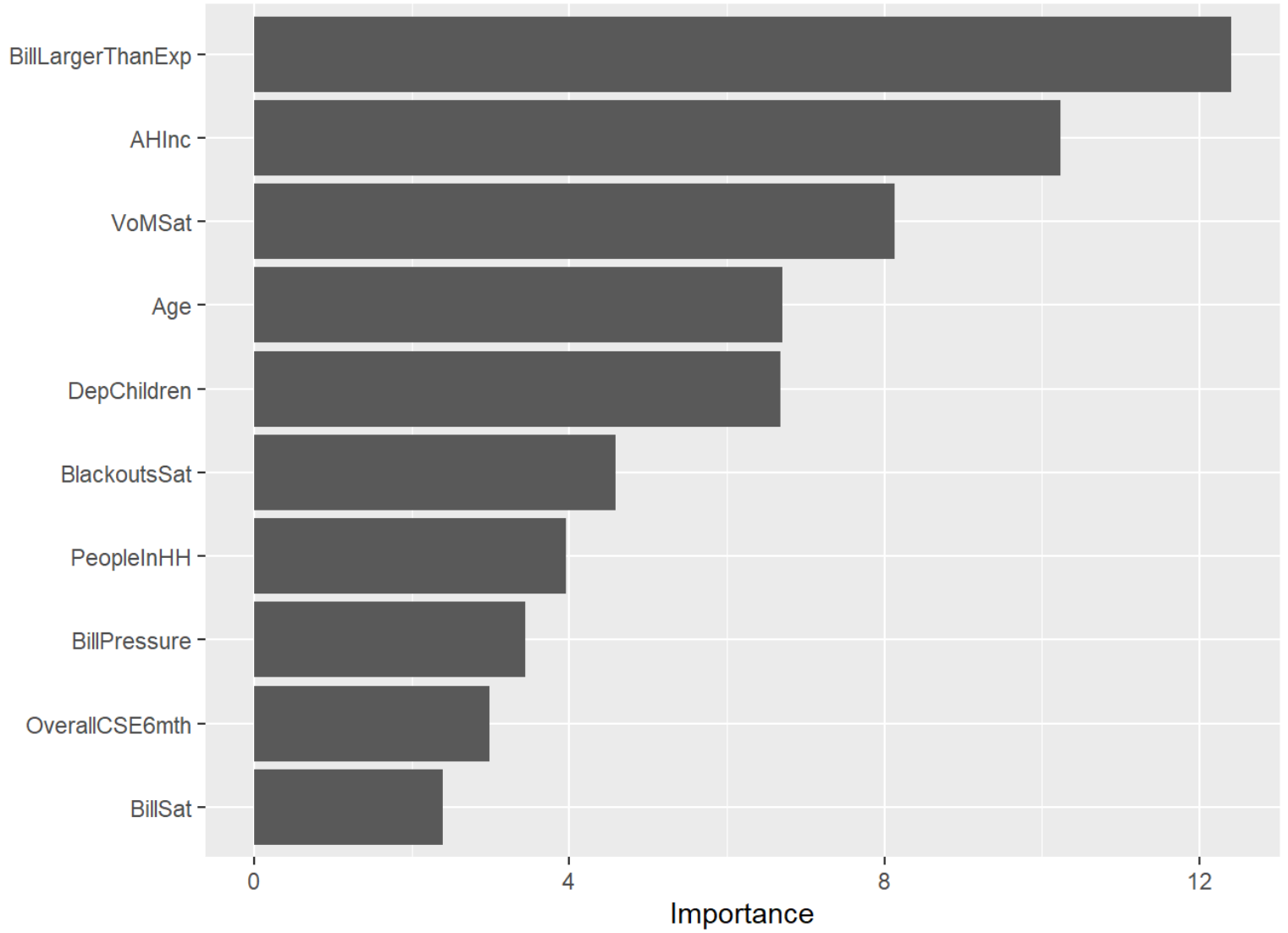
- Decision tree



**A little hard to read!

Energy Consumer Sentiment Survey

- Decision tree



Energy Consumer Sentiment Survey

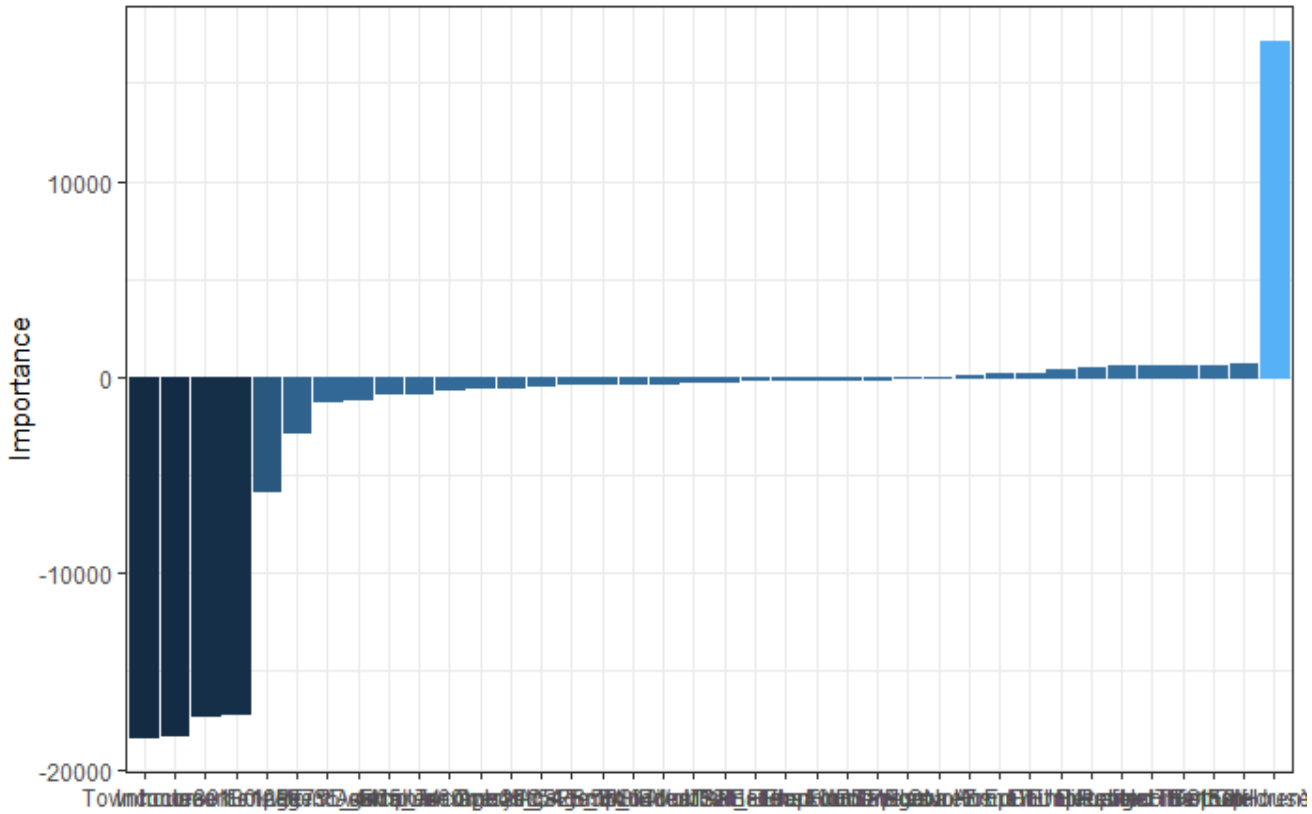
Neural network

- Remember…
  - All factor variables must be converted to 0/1 numeric coding
  - All numeric variables must be scaled
- Note that neural network considers is modelling problem as “multiple output” problem
  - i.e. producing probabilities for both the “Yes” and “No” responses
  - Only Olden variable importance can handle multiple output cases

Energy Consumer Sentiment Survey

Neural network

- Olden variable importance
  - Plots not legible with many variables
  - Table results can also be generated
  - See work through document
  - Summary provided below



Energy Consumer Sentiment Survey

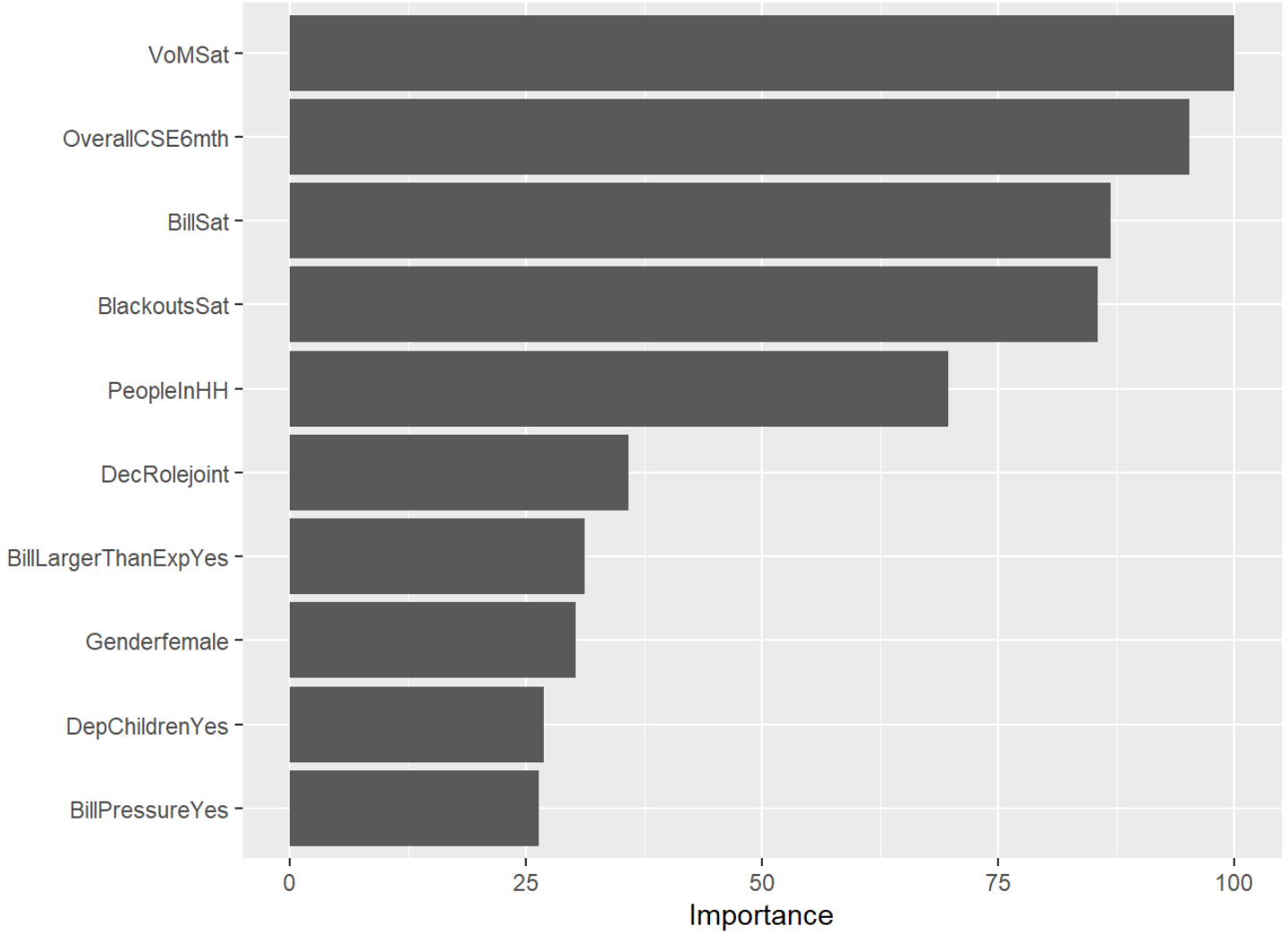
K-NN Classification

- No relationship diagnostics available per se
- Predictive assessment is key to assessing K-NN algorithms (as in the numeric case)

Energy Consumer Sentiment Survey

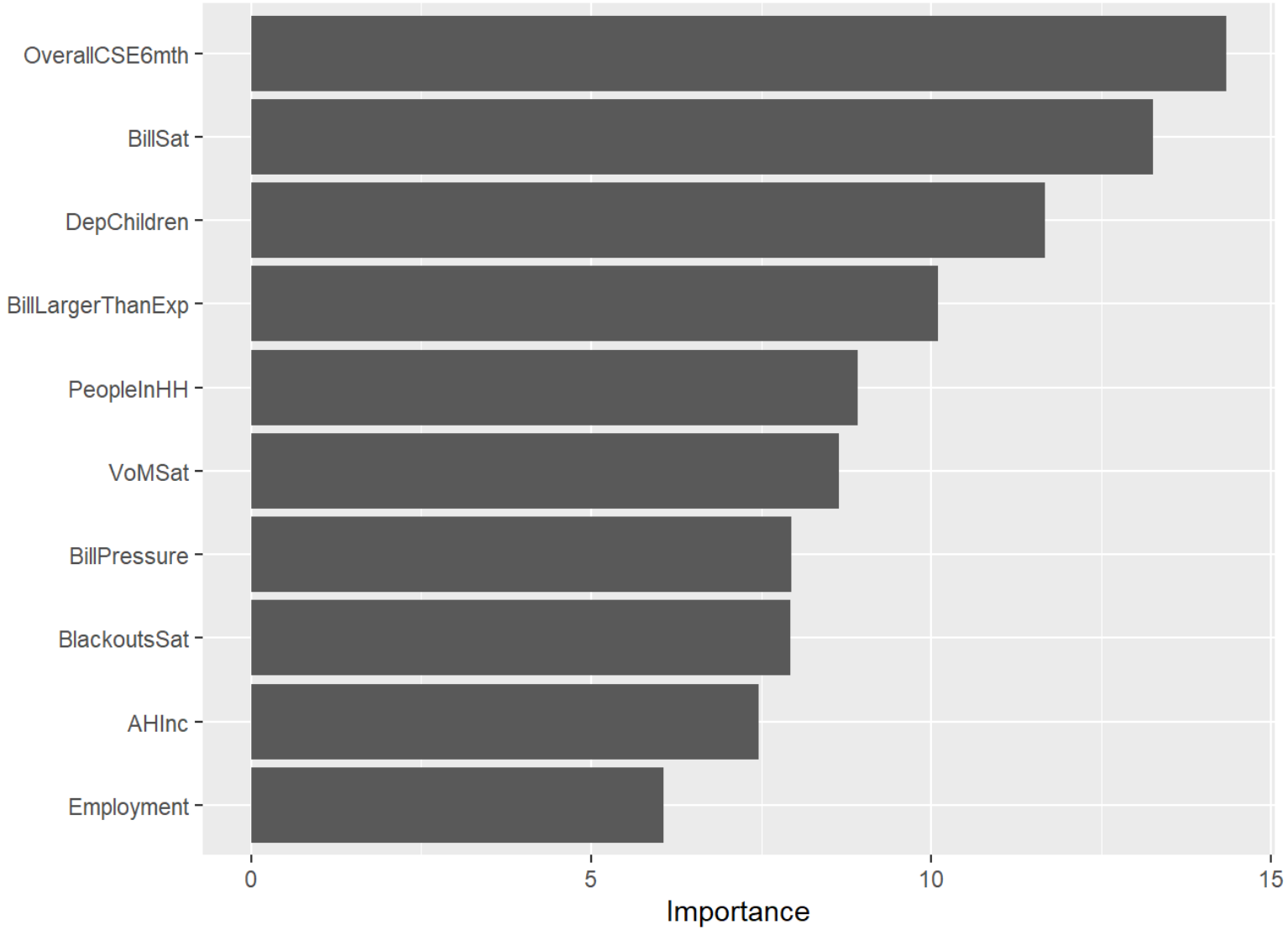
Bagging

- With tree base learner



Energy Consumer Sentiment Survey

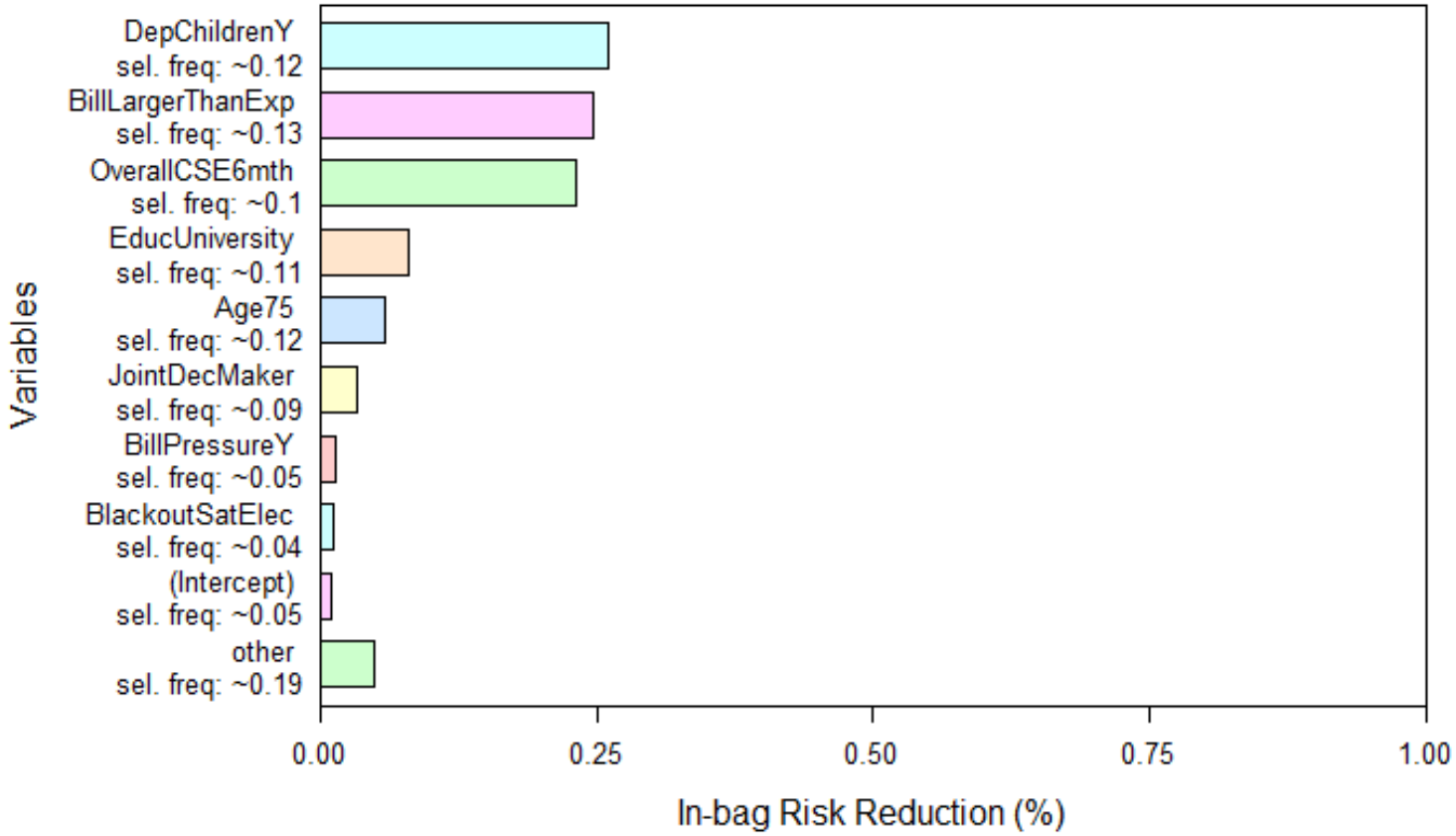
Random Forest



Energy Consumer Sentiment Survey

Boosting

- With logistic regression as base



Energy Consumer Sentiment Survey

Support vector machine

- See test set predictive assessment!

Energy Consumer Sentiment Survey

- Variable importance from training set – Top 5

	Logistic	Decision Tree	NeuralNet	K-NN	Bagging	Random forest	Boosting	SVM
1	BillLargerThan Exp	BillLargerThan Exp	HouseType		VoMSat	OverallCSE	DepChildren	
2	DepChildren	AHInc	AHInc		OverallCSE	BillSat	BillLargerThan Exp	
3	Age	VoMSat	Employment (PT)		BillSat	DepChildren	OverallCSE	
4	DecRole	Age	Age		BlackoutSat	BillLargerThan Exp	Education (UniDegree)	
5	OverallCSE	DepChildren	DecRole		PeopleInHH	PeopleInHH	Age75+	

Energy Consumer Sentiment Survey

- Test set accuracy metrics (@ cutoff probability of 0.5)

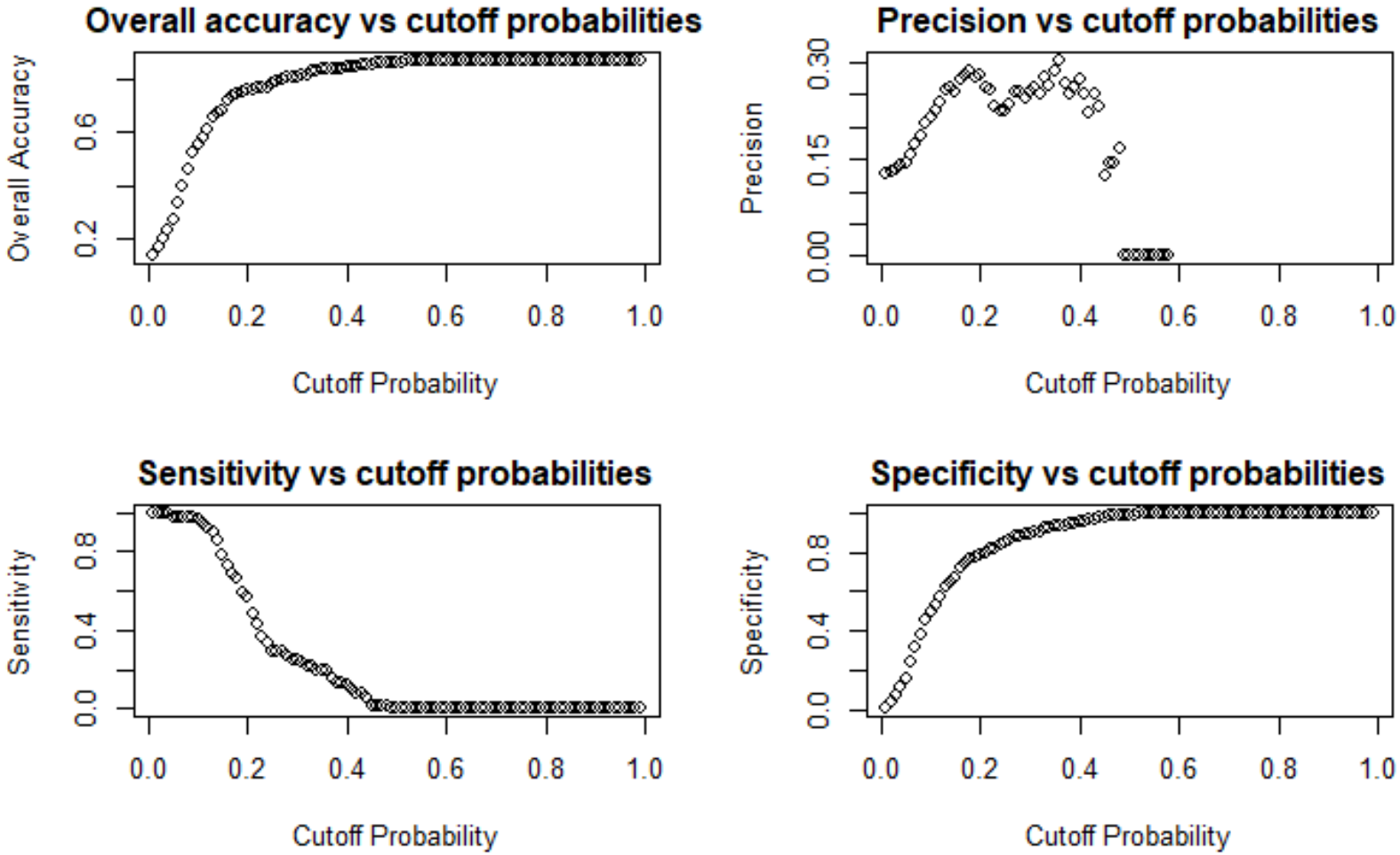
	Overall	Precision	Sensitivity	Specificity	AUC	Gini
Naive	0.874	NaN	0.000	1.000	0.500	0.000
Logistic	0.864	**0.000**	**0.000**	0.989	**0.781**	**0.562**
DecisionTree	0.851	0.263	0.098	0.960	**0.648**	**0.295**
NeuralNet	**0.817**	0.220	0.176	**0.909**	0.731	0.463
k-NN	0.874	**NaN**	**0.000**	**1.000**	0.690	0.380
Bagging	0.854	0.318	0.137	0.958	0.682	0.363
RandomForest	**0.876**	**1.000**	0.020	**1.000**	0.727	0.454
Boosting	0.874	**NaN**	**0.000**	**1.000**	0.771	0.543
SVM	0.874	**NaN**	**0.000**	**1.000**	0.662	0.324

Energy Consumer Sentiment Survey

Are predictions most accurate at cutoff = 0.5?

Energy Consumer Sentiment Survey

- Let us take the example of the logistic regression…
- How do the various accuracy metrics change if we alter the cutoff?



Energy Consumer Sentiment Survey

Observations:

- Overall accuracy plateau at approx. 87%
- Lower cutoff gives better precision (max at cutoff = 0.36)
  - … but this comes at a cost of lower specificity
- Sensitivity is greatest at the lowest cutoff considered
  - … but specificity declines drastically in this region!

ALWAYS A TRADE OFF!

Energy Consumer Sentiment Survey

Out-of-time robustness assessment

- How do the models perform in predicting the sample that are collected at different time points?
- This will show you how robust the models are in predicting into the “unknown”
- Can only be done if you have data available

Energy Consumer Sentiment Survey

Out-of-time robustness assessment

- We investigate the out-of-time robustness of three candidate models:

    1. Neural network
    2. Bagging
    3. Random forest

- Out-of-time sample: June 2020

Energy Consumer Sentiment Survey

- Test set accuracy metrics (@ cutoff probability of 0.5)

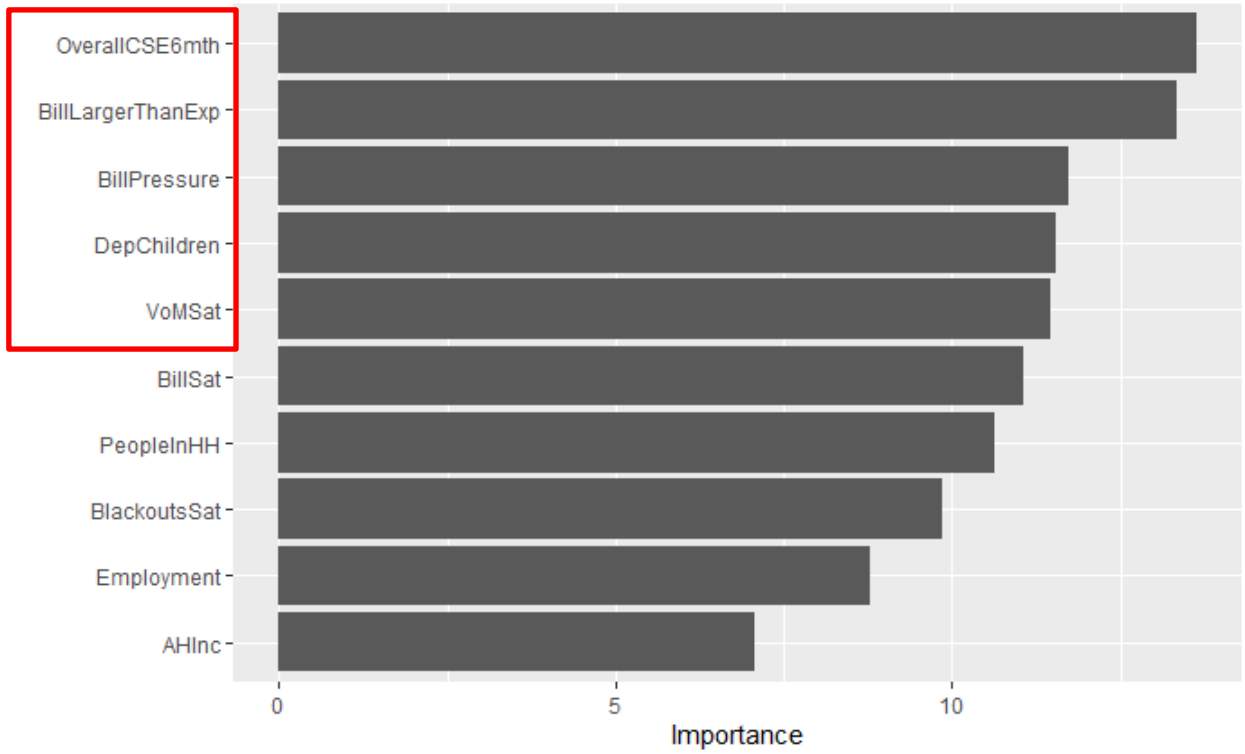
December 19	Overall	Precision	Sensitivity	Specificity	AUC	Gini
NeuralNet	**0.817**	0.220	**0.176**	**0.909**	**0.731**	**0.463**
Bagging	0.854	0.318	0.137	0.958	**0.682**	**0.363**
RandomForest	**0.876**	**1.000**	**0.020**	**1.000**	0.727	0.454

June 20	Overall	Precision	Sensitivity	Specificity	AUC	Gini
NeuralNet	**0.815**	0.265	0.217	**0.907**	**0.640**	**0.281**
Bagging	0.858	0.361	0.082	0.977	0.658	0.316
RandomForest	**0.867**	**0.571**	**0.015**	**0.998**	**0.713**	**0.427**

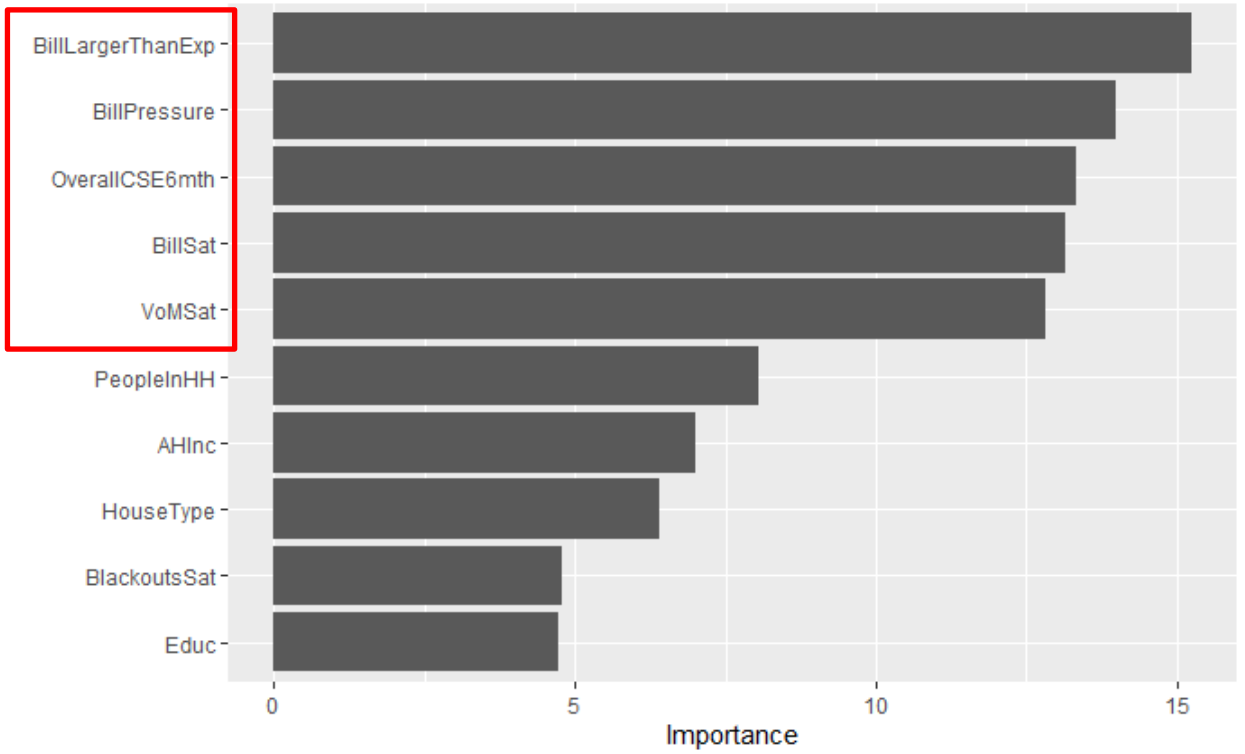
Energy Consumer Sentiment Survey

December 2019 vs June 2020 – What has changed?

- Variable importance from random forest



December 19



June 20

Predicting Classification

Some concluding remarks

- Regression based model – translation function to compute probabilities
  - Generalized linear models (GLMs)
- Machine learning methods generally applicable to classification
- Predictive assessment – different considerations to numeric outputs
  - Matching/non-matching
  - Consequences of non-matching needs to be clear

Predictive Analytics – An evolving process!

Understanding the context & usage

Accessibility to users