Customer Analytics 2020: Review & Practice session

Professor Vincent Nijs

Rady School of Management @ UCSD

Customer Analytics

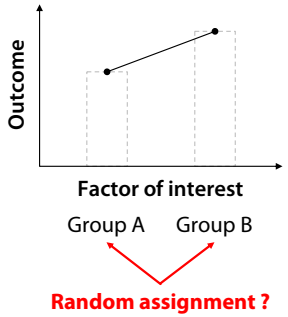
Final exam: What you can expect (online proctored)

- Causality check-lists
- Customer Lifetime Value calculations (CLV)
- RFM (Independent , Sequential)
- Manipulate data (e.g., transform variables, 'bin' a continuous variable)
- Exploratory Data Analysis (EDA)
- Linear and Logistic regression (interpret coefficients / odds-ratios)
- Evaluate relative importance of explanatory variables
- Estimate and interpret interactions
- Use training and test samples
- Predict using RFM, Linear/Logistic regression, NN, Decision Trees, Random Forests, GBM, etc.
- Use lift, gains, and profit charts and evaluate overfitting
- Determine profits and return on marketing expenditures for actions
- Understand benefits and limitations of partial factorials
- Estimate logistic regression on experimental data
- Calculate Difference-in-Differences (DiD)
- Bias-Variance tradeoff
- Tuning ML models using Cross-Validation
- …

The causality checklist

CHECK FOR PROBABILISTIC EQUIVALENCE

Were units randomly assigned to groups?



If no, then differences may not have a causal interpretation

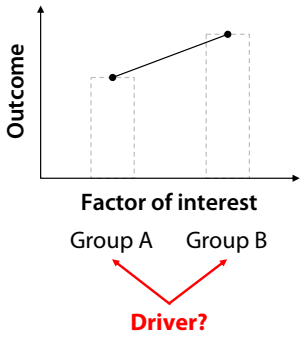
If yes, then the analysis passes the causality checklist

IDENTIFY GROUP DRIVERS

Initial evaluation
What drivers influenced assignment of units to groups?

Digging deeper

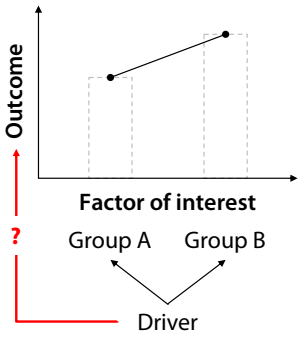
- Did firm influence group assignment? If so, based on what drivers?
- Did units self select into groups? If so, based on what drivers?
- Are groups separated by time? If so, what outcome related drivers vary over time?



Note: Consider all possible drivers

CHECK FOR CONFOUNDS

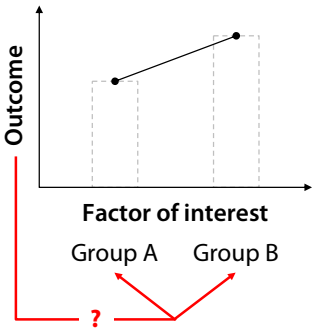
Could a driver have a direct effect on the outcome?



If yes, then the causality check fails because driver is a confound

CHECK FOR REVERSE CAUSALITY

Could group outcomes have direct impact on the factor of interest?



If yes, then the causality check fails due to reverse causality

If no, then the analysis passes the causality checklist

If an analysis passes the causality check list we can conclude that differences in the outcome variable across groups are **caused** by differences in the factor of interest

TASK 1: Apply the checklist to evaluate causal statements

A company sells many snowmobiles in Canada but very few in Mexico. The company also advertises extensively in Canada but does not advertise at all in Mexico

Causal claim: Advertising works

Factor of interest:

Groups:

Outcome:

Group assignment:

Drivers of group assignment:

Confound:

Reverse causality:

TASK 1: Apply the checklist to evaluate causal statements

Snowmobile sales are below expectations in January and a dealership in Toronto plans to run a promotion in February. During the promotional period an unexpected snow-storm hits the Toronto area. Sales of snowmobiles in February are 10% higher than expected

Causal claim: The promotion caused a 10% increase in sales

Factor of interest:

Groups:

Outcome:

Group assignment:

Drivers of group assignment:

Confound:

Reverse causality:

TASK 1: Apply the checklist to evaluate causal statements

Door dash is a logistics software startup. Affiliated drivers deliver restaurant food to customers. A restaurant decides to put a link to Door Dash on their website, starting in January. The number of orders for take-out in January are 5% lower than in December

Causal claim: Adding the link to the Door Dash site caused a decrease in sales

Factor of interest:

Groups:

Outcome:

Group assignment:

Drivers of group assignment:

Confound:

Reverse causality:

TASK 1: Apply the checklist to evaluate causal statements

A manufacturer of kitchen knives has improved the quality of their product each year. The company also increased prices each year to cover the costs of these quality improvements. A regression of price on demand (i.e., demand = a + b x price) gives a coefficient for price very close to 0 that is not statistically significant

Causal claim: Customers are not sensitive to price changes so the manufacturer can continue to increase prices, even if quality is not improved

Factor of interest:

Groups:

Outcome:

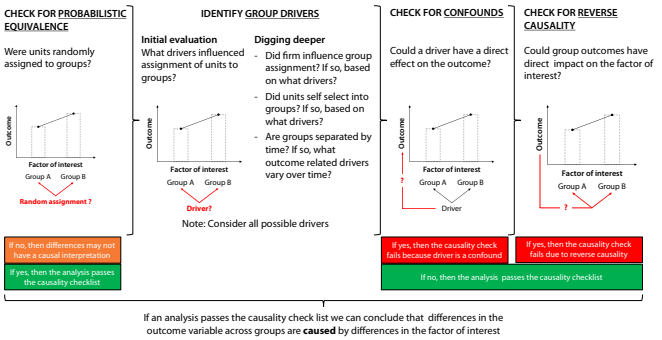
Group assignment:

Drivers of group assignment:

Confound:

Getting from prediction to prescription requires careful deliberation

PREDICTION → PRESCRIPTION CHECKLIST



Is the data available for prediction relevant for the desired prescription?

Yes / **No**

Gather additional data or experiment

Is it reasonable to attach a causal interpretation to the estimated effect of the prescription?

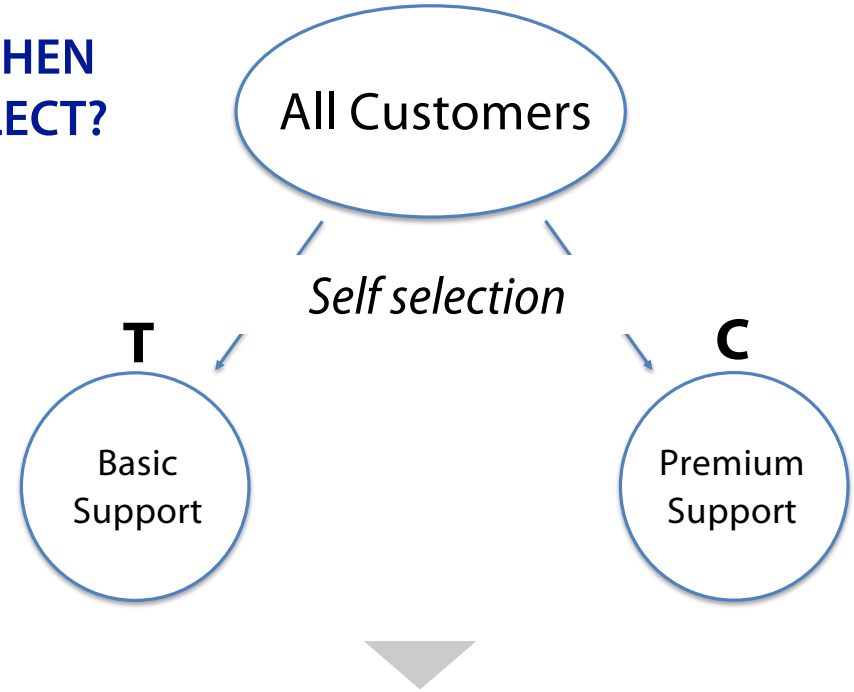
Yes / **No**

- Based on Exhibit 1?
- Based on Testing?

Experiment or fix the analysis

How to “fix” an analysis with Matching

LEARN FROM DATA WHEN CUSTOMERS SELF-SELECT?



amazon web services

When we can’t run an experiment, we may still be able to create *“observationally equivalent”* groups using matching

Customer support plan and billing data

ID	Premium	Billings
2618643	No	$ 3,160
3199888	No	$ 14,380
2844779	No	$ 22,700
3118111	No	$ 26,280
2775543	No	$ 15,120
3325986	No	$ 6,320
3382231	No	$ 9,840
3680449	No	$ 10,800
1833421	No	$ 14,040
2409691	No	$ 8,400
1800623	Yes	$ 27,480
1946297	Yes	$ 14,800
2955486	Yes	$ 27,140
3338292	Yes	$ 9,860
3838594	Yes	$ 6,800
2051366	Yes	$ 15,760
2702835	Yes	$ 33,360
2913122	Yes	$ 67,220
3136996	Yes	$ 39,140
2093118	Yes	$ 11,240

Average billings: $13,104

Average billings: $25,280

Does premium support increase customer billings by (approx.) $12,000?

	Billings	Type	Tenure	Revenue (M)
$	13,104	50%	25.9	12.5
$	25,280	30%	40	20.4
$	12,176			

Customer support plan and billing data

ID	Premium	Billings	Type	Tenure	Revenue (M)
2618643	No	$ 3,160	Tech	8	3
3199888	No	$ 14,380	Other	19	14
2844779	No	$ 22,700	Other	35	22
3118111	No	$ 26,280	Other	114	24
2775543	No	$ 15,120	Tech	6	15
3325986	No	$ 6,320	Other	16	6
3382231	No	$ 9,840	Tech	12	9
3680449	No	$ 10,800	Tech	27	10
1833421	No	$ 14,040	Other	2	14
2409691	No	$ 8,400	Tech	20	8
1800623	Yes	$ 27,480	Other	114	21
1946297	Yes	$ 14,800	Other	20	12
2955486	Yes	$ 27,140	Other	37	22
3338292	Yes	$ 9,860	Tech	13	8
3838594	Yes	$ 6,800	Other	40	5
2051366	Yes	$ 15,760	Tech	8	13
2702835	Yes	$ 33,360	Other	108	26
2913122	Yes	$ 67,220	Other	1	56
3136996	Yes	$ 39,140	Other	37	32
2093118	Yes	$ 11,240	Tech	22	9

ID	Premium	Billings	Type	Tenure	Revenue (M)
2618643	No	$ 3,160	Tech	8	3
3199888	No	$ 14,380	Other	19	14
2844779	No	$ 22,700	Other	35	22
3118111	No	$ 26,280	Other	114	24
2775543	No	$ 15,120	Tech	6	15
3325986	No	$ 6,320	Other	16	6
3382231	No	$ 9,840	Tech	12	9
3680449	No	$ 10,800	Tech	27	10
1833421	No	$ 14,040	Other	2	14
2409691	No	$ 8,400	Tech	20	8
1800623	Yes	$ 27,480	Other	114	21
1946297	Yes	$ 14,800	Other	20	12
2955486	Yes	$ 27,140	Other	37	22
3338292	Yes	$ 9,860	Tech	13	8
3838594	Yes	$ 6,800	Other	40	5
2051366	Yes	$ 15,760	Tech	8	13
2702835	Yes	$ 33,360	Other	108	26
2913122	Yes	$ 67,220	Other	1	56
3136996	Yes	$ 39,140	Other	37	32
2093118	Yes	$ 11,240	Tech	22	9

Matching has constructed two observationally equivalent samples for analysis

ID	Premium	Billings	Type	Tenure	Revenue (M)
3199888	No	$ 14,380	Other	19	14
2844779	No	$ 22,700	Other	35	22
3118111	No	$ 26,280	Other	114	24
2775543	No	$ 15,120	Tech	6	15
3382231	No	$ 9,840	Tech	12	9
2409691	No	$ 8,400	Tech	20	8
1946297	Yes	$ 14,800	Other	20	12
2955486	Yes	$ 27,140	Other	37	22
3338292	Yes	$ 9,860	Tech	13	8
2051366	Yes	$ 15,760	Tech	8	13
2702835	Yes	$ 33,360	Other	108	26
2093118	Yes	$ 11,240	Tech	22	9

Average billings: $16,120 (was $13,104)

Difference after matching: $2,500 (was $12,000)

Average billings: $18,693 (was $25,280)

Billings	Type	Tenure	Revenue (M)
$ 16,120	30%	34.3	15.3
$ 18,693	30%	34.7	15.0
$ 2,573			

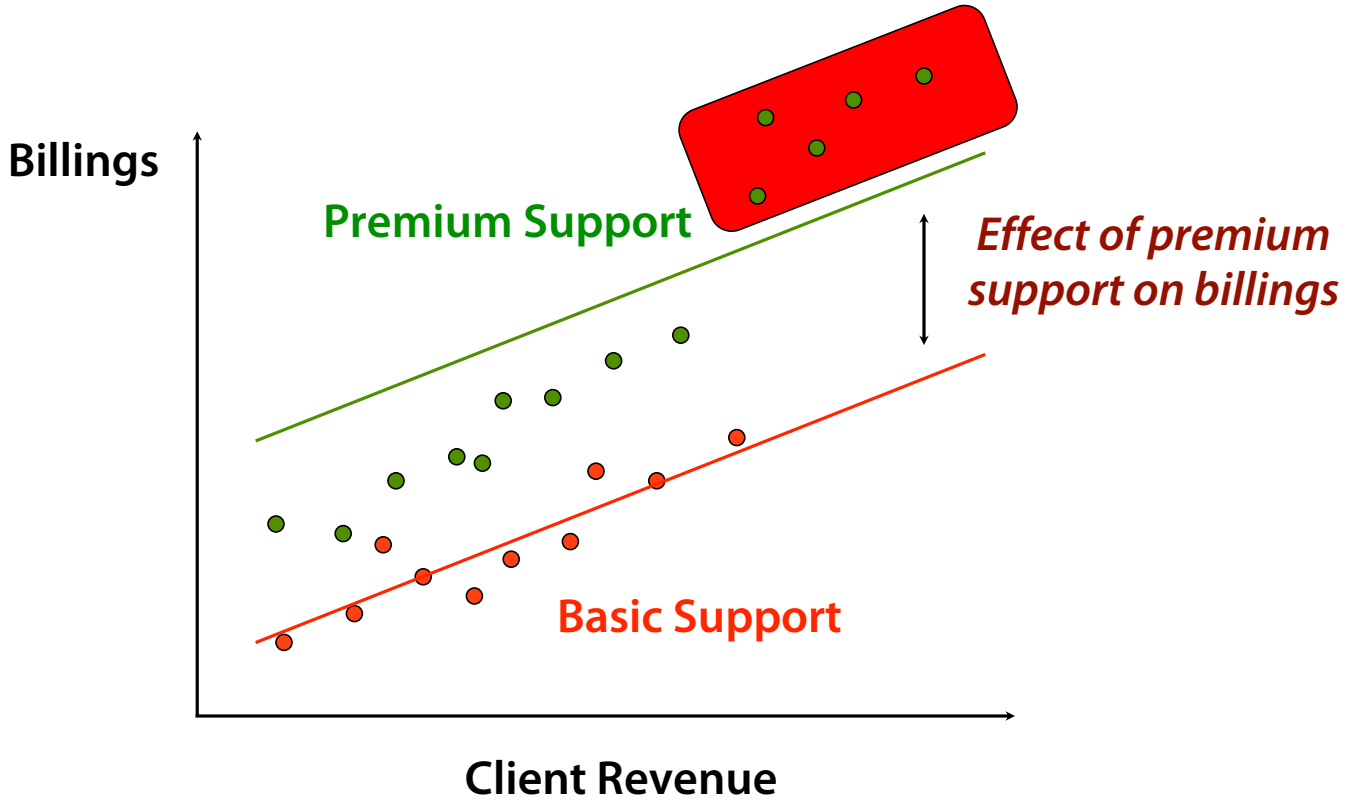
https://cran.r-project.org/web/packages/optmatch/vignettes/fullmatch-vignette.pdf

https://rstudio-pubs-static.s3.amazonaws.com/284461_5fabe52157594320921fc9e4d539ebc2.html

https://florianwilhelm.info/2017/04/causal_inference_propensity_score/

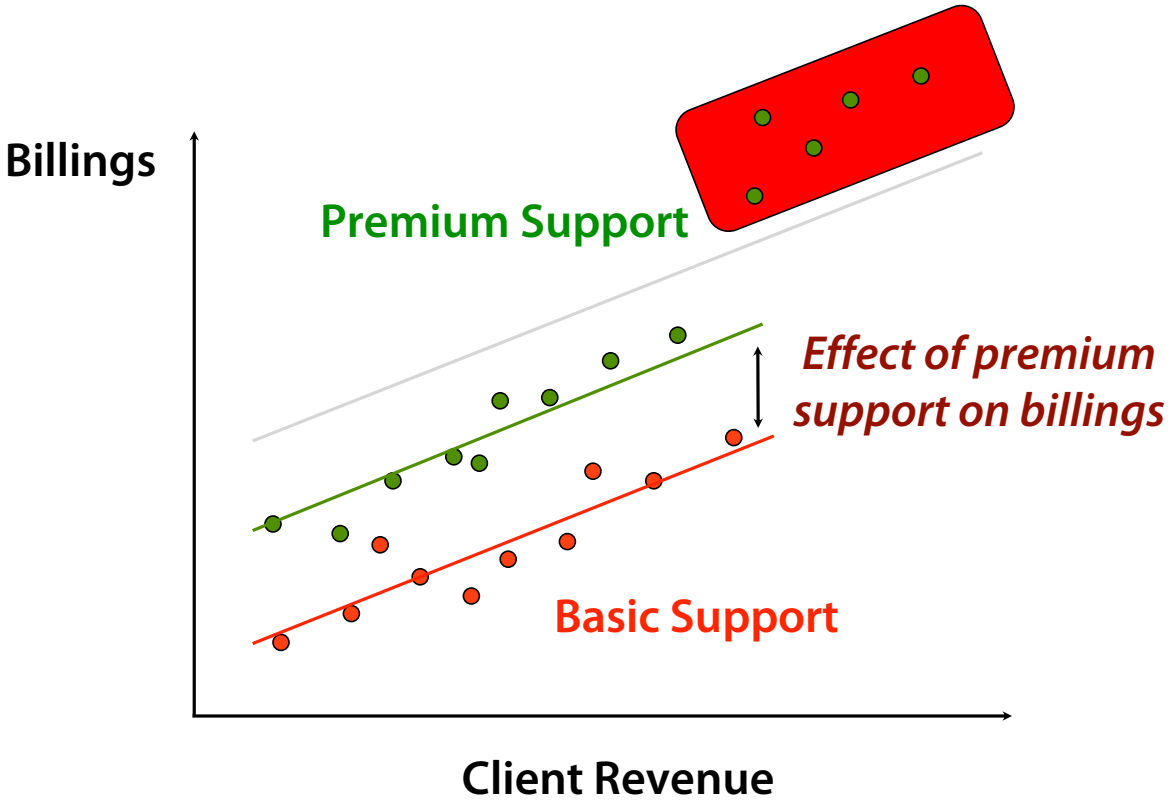
Regression includes data from fundamentally incomparable clients

COMPARING MATCHING AND REGRESSION



which we have information on both basic and premium support customers

COMPARING MATCHING AND REGRESSION



Matching, by design, ensures that groups are comparable on observables

TASK 2: Fix an analysis with Difference in Differences (Diff-in-Diff)

SALES COMPENSATION EXAMPLE IN THE HOME ALARM INDUSTRY:

	Region 1	Region 2
Period 1	*Control Group 1* **Average LTV $350**	*Control Group 2* **Average LTV $390**
Period 2	*Control Group 3* **Average LTV $320**	*Target Group* **Average LTV $400**

- What is the effect of the initiative, net of the seasonal and geographic effect?
- Estimate the effect of the new salesforce compensation plan using **data/did-sales-force.pkl**
  - Create a "pivot" table of the mean CLV
  - Determine the effect of the new plan using a linear regression
- Write out the full "interaction table" for the regression model (example on next slide)

Is the Friday/Saturday effect different during Comic-Con?

	coefficient	std.error	t.value	p.value	
(Intercept)	2.158	0.014	148.974	0.000	***
time	0.000	0.000	-3.380	0.001	***
fs|yes	0.445	0.016	28.065	0.000	***
ccon|yes	0.324	0.022	14.725	0.000	***
fs|yes:ccon|yes	0.180	0.039	4.549	0.000	***

	Intercept	fs|yes	ccon|yes	fs|yes * ccon|yes
FS, Comic-Con	1	1	1	1
Not FS, Comic-Con	1	0	1	0
FS, Not Comic-Con	1	1	0	0
Not FS, Not Comic-Con	1	0	0	0

"INTERACTION TABLE"

	8pm	**10pm**	**12am**
time	0	120	240
FS, Comic-Con	$3.107	$3.068	$3.029
Not FS, Comic-Con	$2.482	$2.444	$2.405
FS, Not Comic-Con	$2.603	$2.564	$2.525
Not FS, Not Comic-Con	$2.158	$2.120	$2.081
FS effect, Comic-Con	$0.624	$0.624	$0.624
FS effect, Not Comic-Con	$0.445	$0.445	$0.445

See also **week5/interactions.xlsx** on Canvas

Notice the similarity with calculations of effects in experiments

SALES PER 10,000 EXPOSURES BY EXPERIMENTAL CONDITION

Shipping threshold	Discount: 15%	Discount: 20%
$300	**500**	**560**
$200	**580**	

- Decreasing the **shipping threshold** increases sales by **80**
- Increasing the **tools discount** increases sales by **60**

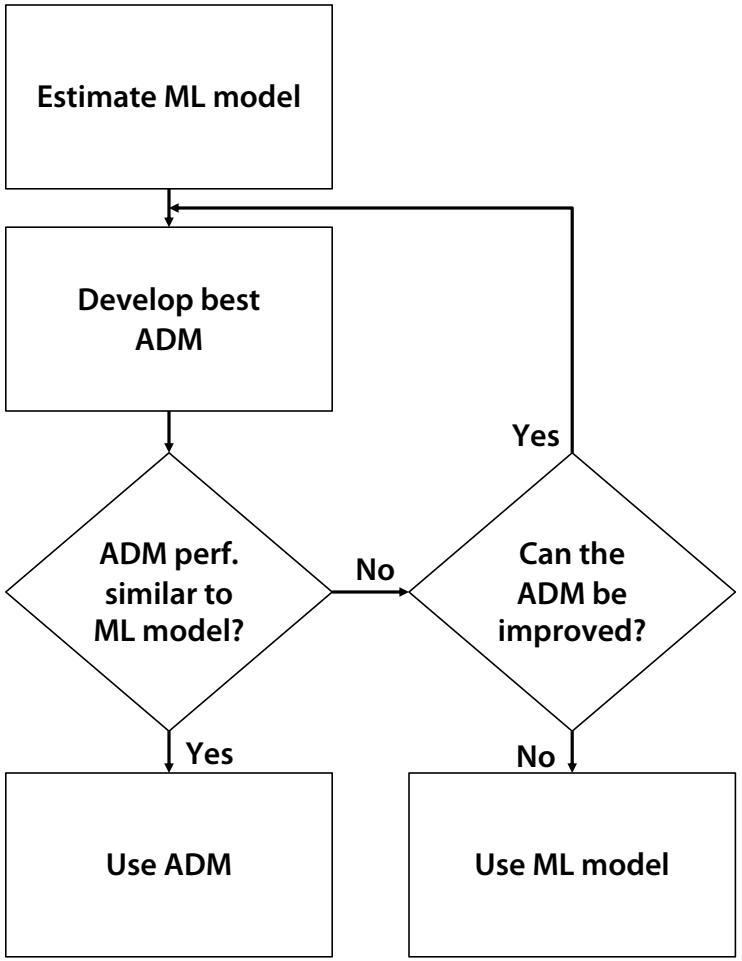
Notice the similarity with calculations of effects in experiments

SALES PER 10,000 EXPOSURES BY EXPERIMENTAL CONDITION

Shipping threshold	Discount: 15%	Discount: 20%
$300	**500**	**560**
$200	**580**	**680**

- Decreasing the **shipping threshold** increases sales by **80**
- Increasing the **tools discount** increases sales by **60**
- Decreasing the **shipping threshold AND** increasing the **tools discount** increases sales by an additional **40** units

Machine Learning (ML) models can be used in combination with Analyst Driven Models (ADM)

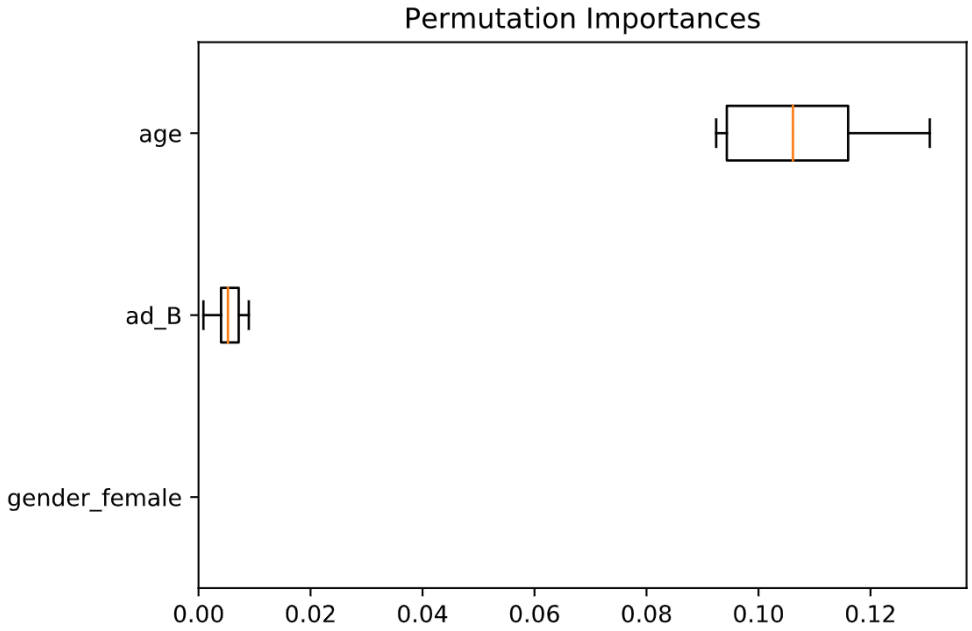


Core idea:

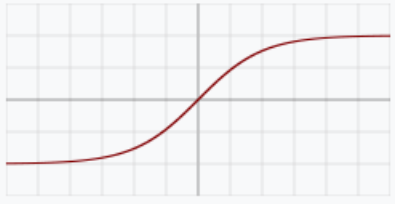
- Use predictions from ML model as performance benchmark
- Use ADM for interpretation

TASK 4: What predicts ad click-through? (see python/facebook.py)

- Reproduce the plot on the right using a NN (1)
- How does the plot change as we add another node to the hidden layer, i.e., NN(2)? Why does it change?



TanH



$$f(x) = \tanh(x) = \frac{(e^{x} - e^{-x})}{(e^{x} + e^{-x})}$$

TASK 4: What is the effect of ad A vs B on (fe)male prospects?

USE DATA/FACEBOOK.PKL TO REPLICATE

CONSTRUCT THE "INTERACTION TABLE"

	OR	OR%	coefficient	std.error	z.value	p.value	
(Intercept)			-4.377	0.230	-19.008	< .001	***
age	1.050	5.0%	0.049	0.005	9.091	< .001	***
gender|male	0.360	-64.0%	-1.023	0.147	-6.968	< .001	***
ad|B	0.309	-69.1%	-1.175	0.144	-8.145	< .001	***
gender|male:ad|B	7.806	680.6%	2.055	0.205	10.019	< .001	***

Interpretation:

- **Men**: the odds of click-through are _____ by ____ for Ad B compared to Ad A
- **Women**: the odds of click-through are _____ by ____ for Ad B compared to Ad A

Why do we need to multiply odds ratios?

see interactions/interactions.xlsx

male, ad|A as the base

	OR	coefficient	std.error	z.value	p.value	
(Intercept)	0.000	-5.399	0.251	-21.504	0.000	***
age	1.050	0.049	0.005	9.091	0.000	***
gender|female	2.780	1.023	0.147	6.968	0.000	***
ad|B	2.410	0.880	0.146	6.033	0.000	***
gender|female:ad|B	0.128	-2.055	0.205	-10.019	0.000	***

	female	ad|B	female * ad|B
male ad|A	0	0	0
male ad|B	0	1	0
female ad|A	1	0	0
female ad|B	1	1	1

ad|B
gender|female
gender|female * ad|B * gender|female:ad|B

	Compare to male ad|A	Compare to male ad|B	Compare to female ad|A	Compare to female ad|B
male ad|A	1.000	0.415	0.360	1.165
male ad|B	2.410	1.000	0.867	2.808
female ad|A	2.780	1.154	1.000	3.239
female ad|B	0.858	0.356	0.309	1.000

Coefficients	Odds Ratios
a	e^a
b	e^b
a + b	$e^{a+b} = e^a e^b$
a - b	$e^{a-b} = e^a / e^b$

TASK 5: Calculate CLV (use data/clv.xlsx)

		Years				
	Start of CLV Calc.	1	2	3	4	5
Revenues	$0	$400	$400			
Product/Service Costs	$0	$80	$80			
Marketing Costs	$0	$0	$0			
Customer Profit	$0	$320	$320			
Prob. of being active at end of period	100.00%	100.00%	59.00%	34.81%		
Profit expected on average	$0	$320.00	$188.80			
Present Value of Exp. Profits	$0	$320				

- Discount rate is 10% annually
- What is the churn rate? What about the retention rate?
- What assumption are we making about the timing of churn (Optimistic or Pessimistic)?
- What assumption are we making about the timing of payment (Optimistic or Pessimistic)?

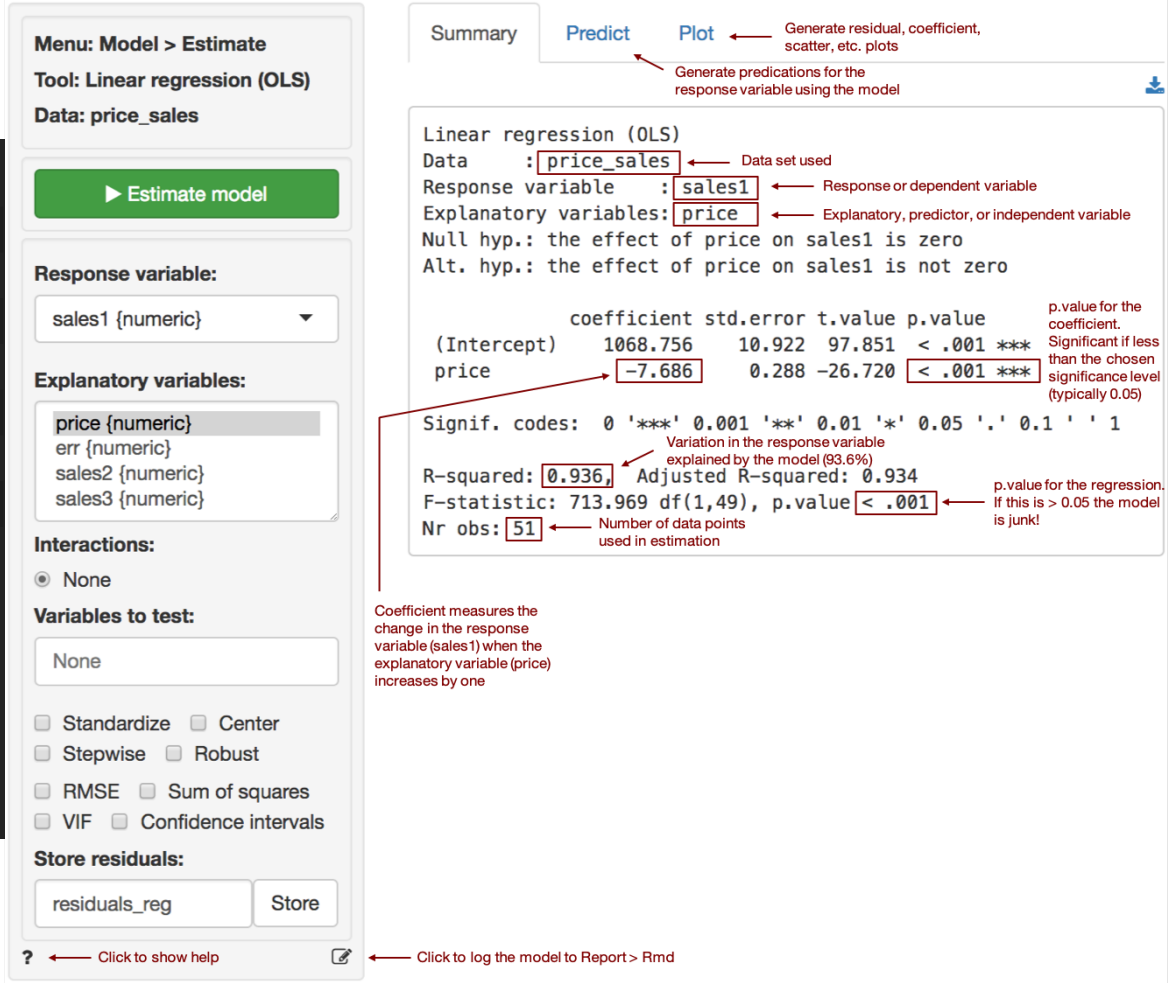
Linear regression (see python/linear-regression.py)

OLS Regression Results

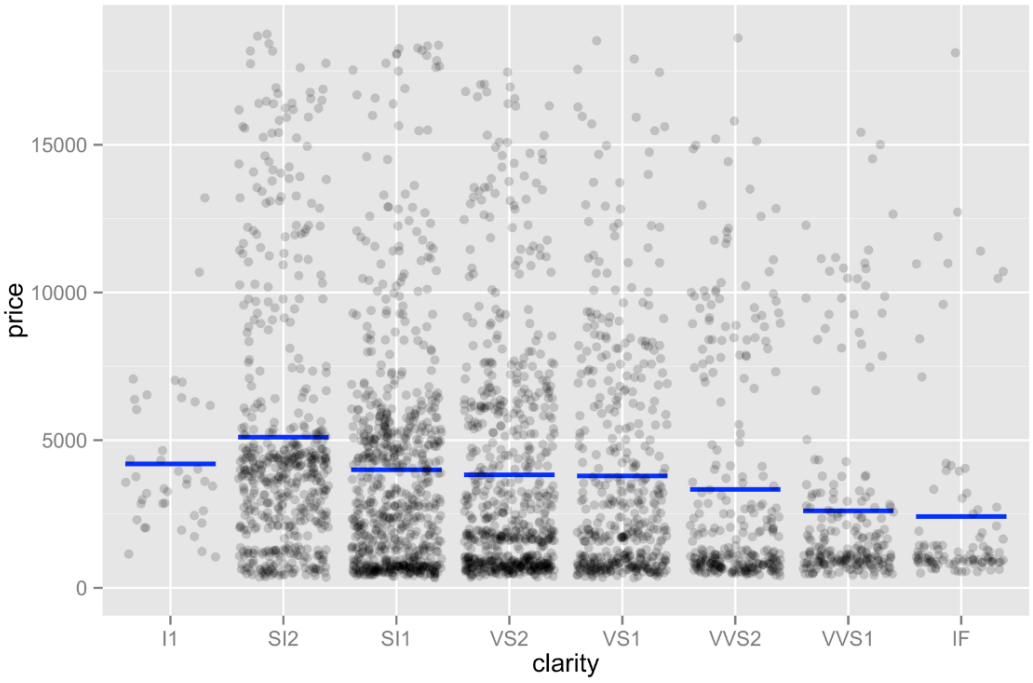
Dep. Variable:	sales1	R-squared:	0.936
Model:	OLS	Adj. R-squared:	0.934
Method:	Least Squares	F-statistic:	714.0
Date:	Mon, 16 Mar 2020	Prob (F-statistic):	7.19e-31
Time:	00:33:16	Log-Likelihood:	-245.21
No. Observations:	51	AIC:	494.4
Df Residuals:	49	BIC:	498.3
Df Model:	1		
Covariance Type:	nonrobust		

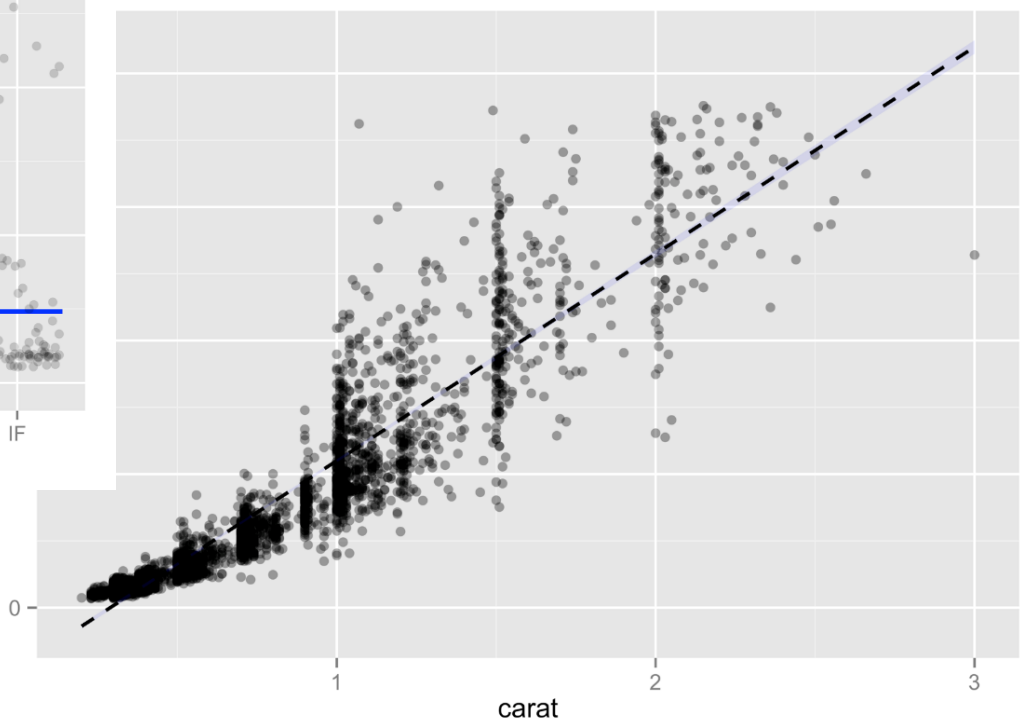
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1068.7560	10.922	97.851	0.000	1046.807	1090.705
price	-7.6863	0.288	-26.720	0.000	-8.264	-7.108

Omnibus:	1.982	Durbin-Watson:	2.110
Prob(Omnibus):	0.371	Jarque-Bera (JB):	1.140



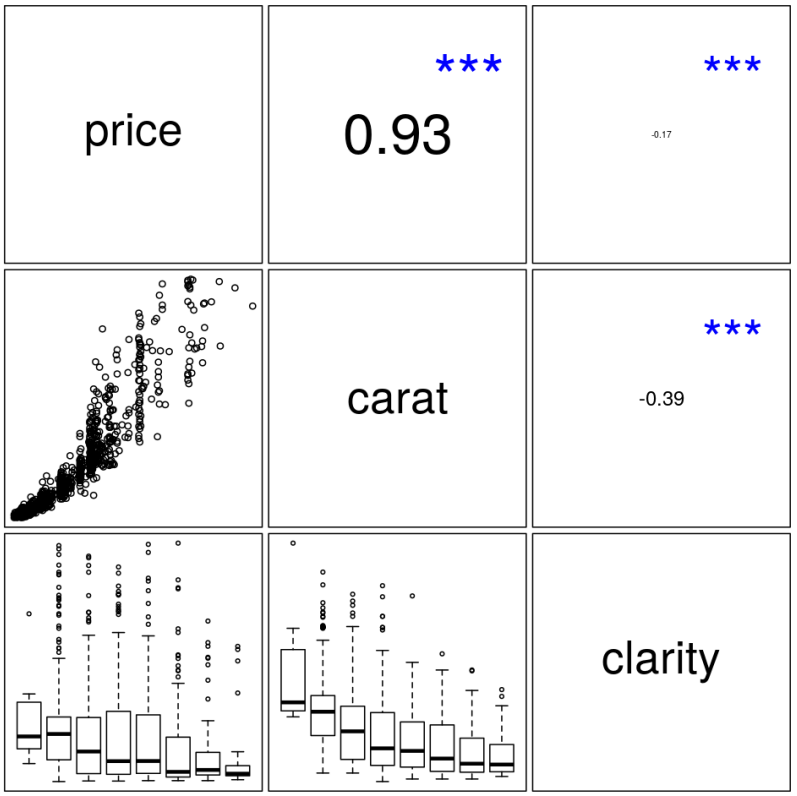
Omitted Variable Bias (OVB)





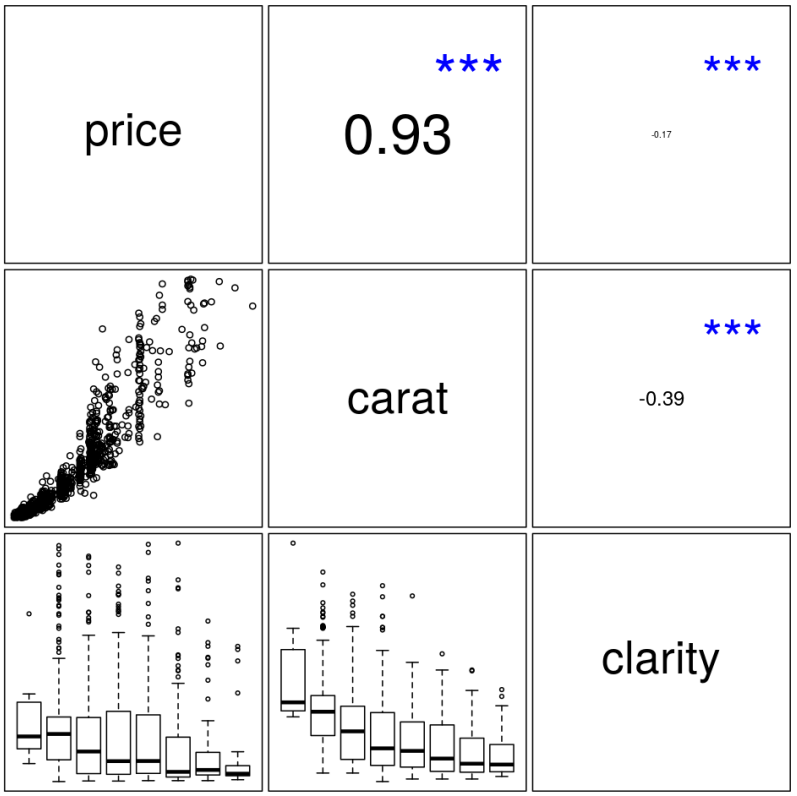
Omitted Variable Bias (OVB) and Multi-collinearity (MC)

```
Linear regression (OLS)
Data     : diamonds
Response variable    : price
Explanatory variables: clarity
Null hyp.: the effect of clarity on price is zero
Alt. hyp.: the effect of clarity on price is not zero

              coefficient std.error t.value p.value
 (Intercept)     4194.775   616.530   6.804  < .001 ***
 clarity|SI2      905.414   639.415   1.416   0.157
 clarity|SI1     -196.198   633.401  -0.310   0.757
 clarity|VS2     -371.808   634.911  -0.586   0.558
 clarity|VS1     -405.594   643.823  -0.630   0.529
 clarity|VVS2    -856.955   658.518  -1.301   0.193
 clarity|VVS1   -1586.315   669.318  -2.370   0.018 *
 clarity|IF     -1783.078   730.540  -2.441   0.015 *

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

R-squared: 0.031,  Adjusted R-squared: 0.029
F-statistic: 13.759 df(7,2992), p.value < .001
Nr obs: 3,000
```



Omitted Variable Bias (OVB) and Multi-collinearity (MC)

```
Linear regression (OLS)
Data     : diamonds
Response variable    : price
Explanatory variables: carat, clarity
Null hyp.: the effect of x on price is zero
Alt. hyp.: the effect of x on price is not zero

              coefficient std.error t.value p.value
 (Intercept)    -6780.993   204.952 -33.086  < .001 ***
 carat           8438.030    51.101 165.125  < .001 ***
 clarity|SI2     2790.760   201.395  13.857  < .001 ***
 clarity|SI1     3608.531   200.508  17.997  < .001 ***
 clarity|VS2     4249.906   201.607  21.080  < .001 ***
 clarity|VS1     4461.956   204.592  21.809  < .001 ***
 clarity|VVS2    5109.476   210.207  24.307  < .001 ***
 clarity|VVS1    5027.669   214.251  23.466  < .001 ***
 clarity|IF      5265.170   233.658  22.534  < .001 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

R-squared: 0.904,  Adjusted R-squared: 0.904
F-statistic: 3530.024 df(8,2991), p.value < .001
Nr obs: 3,000
```



TASK 6: Click Ball Point Pens

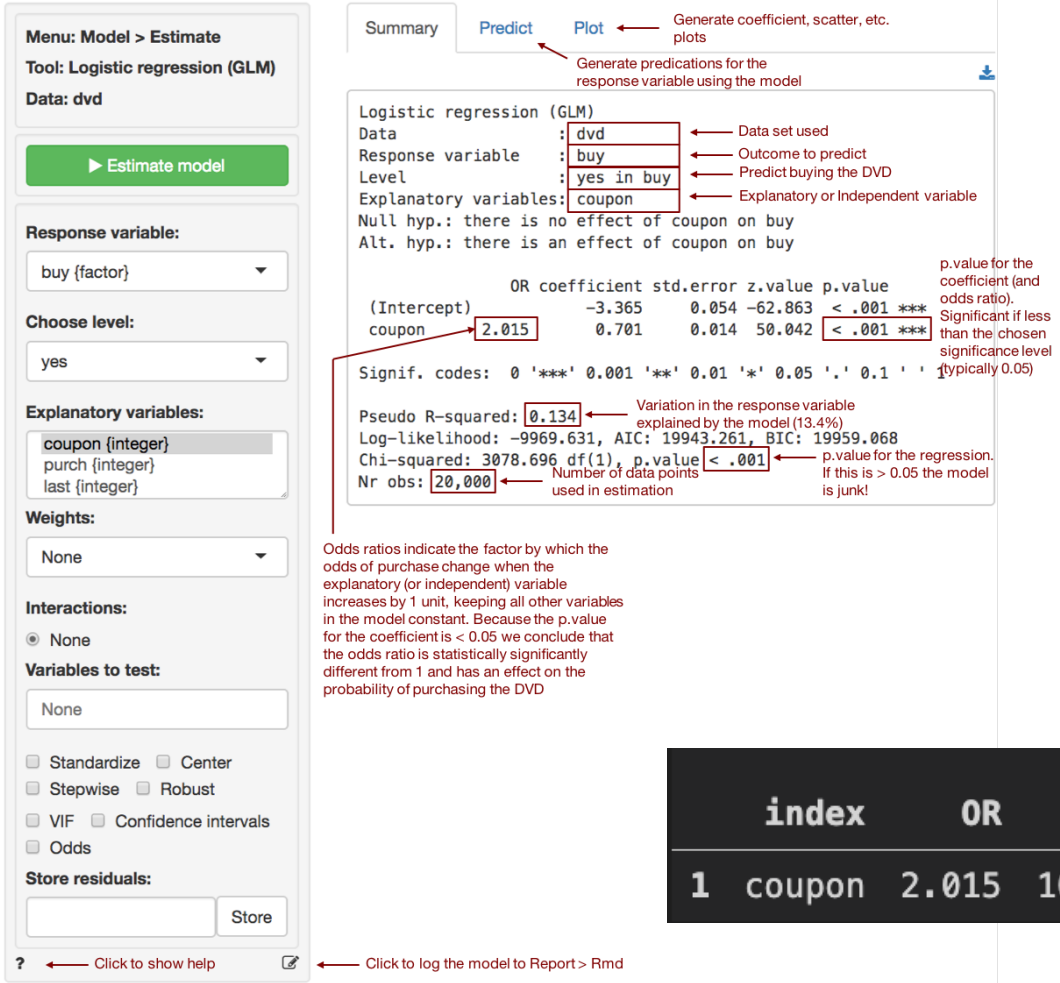
- **Company:** A national manufacturer of ball point pens.

- **Managerial problem:**
  - What is the value of an advertising spot?
  - How much should we pay sales reps?

- **Data:** Sales data for 40 markets/territories along with measures of marketing effort

- Review using **data/click.pkl**

Logistic regression (see python/python-regression.py)



	index	OR	OR%	2.5%	97.5%
1	coupon	2.015	101.518%	1.961	2.071

Variable importance

Menu: Model > Estimate
Tool: Logistic regression (GLM)
Data: dvd

Estimate model

Response variable:
buy {factor}

Choose level:
yes

Explanatory variables:
coupon {integer}
purch {integer}
last {integer}

Weights:
None

Interactions:
None 2-way 3-way

Variables to test:
None

Standardize Center

Summary Predict Plot

```
Logistic regression (GLM)
Data                 : dvd
Response variable    : buy
Level                : yes in buy
Explanatory variables: coupon, purch, last
Null hyp.: there is no effect of x on buy
Alt. hyp.: there is an effect of x on buy

               OR    OR% coefficient std.error z.value p.value
 (Intercept)                 -3.038     0.063 -48.136  < .001 ***
 coupon      2.169 116.9%     0.774     0.015  51.240  < .001 ***
 purch       1.095   9.5%     0.091     0.005  17.879  < .001 ***
 last        0.933  -6.7%    -0.069     0.002 -35.388  < .001 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Pseudo R-squared: 0.208
Log-likelihood: -9110.529, AIC: 18229.058, BIC: 18260.672
Chi-squared: 4796.899 df(3), p.value < .001
Nr obs: 20,000
```

	index	OR	OR%	2.5%	97.5%
1	coupon	2.169	116.866%	2.105	2.234
2	purch	1.095	9.539%	1.085	1.106
3	last	0.933	-6.678%	0.930	0.937

Menu: Model > Estimate
Tool: Logistic regression (GLM)
Data: dvd

Estimate model

Response variable:
buy {factor}

Choose level:
yes

Explanatory variables:
coupon {integer}
purch {integer}
last {integer}

Weights:
None

Interactions:
None 2-

Variables to tes
None

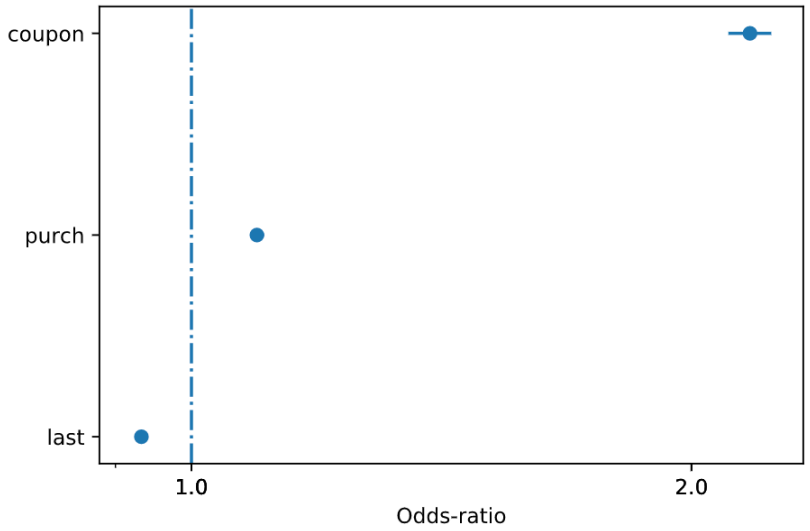
Standardize

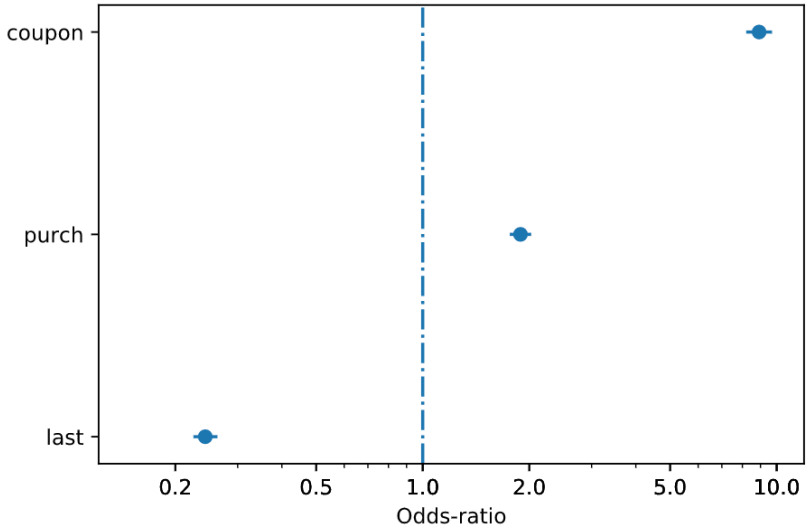
Summary Predict Plot

```
Logistic regression (GLM)
Data                 : dvd
Response variable    : buy
Level                : yes in buy
Explanatory variables: coupon, purch, last
Null hyp.: there is no effect of x on buy
Alt. hyp.: there is an effect of x on buy
**Standardized odds-ratios and coefficients shown (2 X SD)**

               OR    OR% coefficient std.error z.value p.value
 (Intercept)                 -1.400     0.021 -65.424  < .001 ***
 coupon      8.921 792.1%     2.188     0.043  51.240  < .001 ***
 purch       1.889  88.9%     0.636     0.036  17.879  < .001 ***
 last        0.243 -75.7%    -1.415     0.040 -35.388  < .001 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Pseudo R-squared: 0.208
Log-likelihood: -9110.529, AIC: 18229.058, BIC: 18260.672
Chi-squared: 4796.899 df(3), p.value < .001
Nr obs: 20,000
```

	index	OR	OR%	2.5%	97.5%	Importance
1	coupon	8.921	792.117%	8.205	9.700	8.921000
3	last	0.243	-75.717%	0.225	0.263	4.115226
2	purch	1.889	88.879%	1.762	2.025	1.889000

Variable importance





label	OR	OR%	coefficient	std.error	z.value	p.value	sig_star	dummy	mean	sd	min	max	importance	OR_normal	OR%_normal
(Intercept)	0.000	0.000	-1.400	0.021	-65.424	0.000	***	0	1	0	1	1	0	0	0
coupon	8.921	7.921	2.188	0.043	51.240	0.000	***	0	3.009	1.414	1	5	8.921	2.169	1.169
purch	1.889	0.889	0.636	0.036	17.879	0.000	***	0	3.896	3.490	1	12	1.889	1.095	0.095
last	0.243	-0.757	-1.415	0.040	-35.388	0.000	***	0	15.137	10.240	1	35	4.118	0.933	-0.067

2.169^(2 * 1.414) = 8.921
Coupon OR
(standardized)

⇔

8.921^(1 / (2 * 1.414)) = 2.169

exp(2.188 / (2 * 1.414)) = 2.169

1/0.243 = 4.118

"negative" OR to Importance

TASK 7: Evaluate model performance (see python/slow-auc.py)

CONVERT A PROBABILITY TO A BINARY OUTCOME USING BREAKEVEN AS THE THRESHOLD

		Predicted	
		Pos.	**Neg.**
Actual	**Pos.**	TP	FN
	Neg.	FP	TN

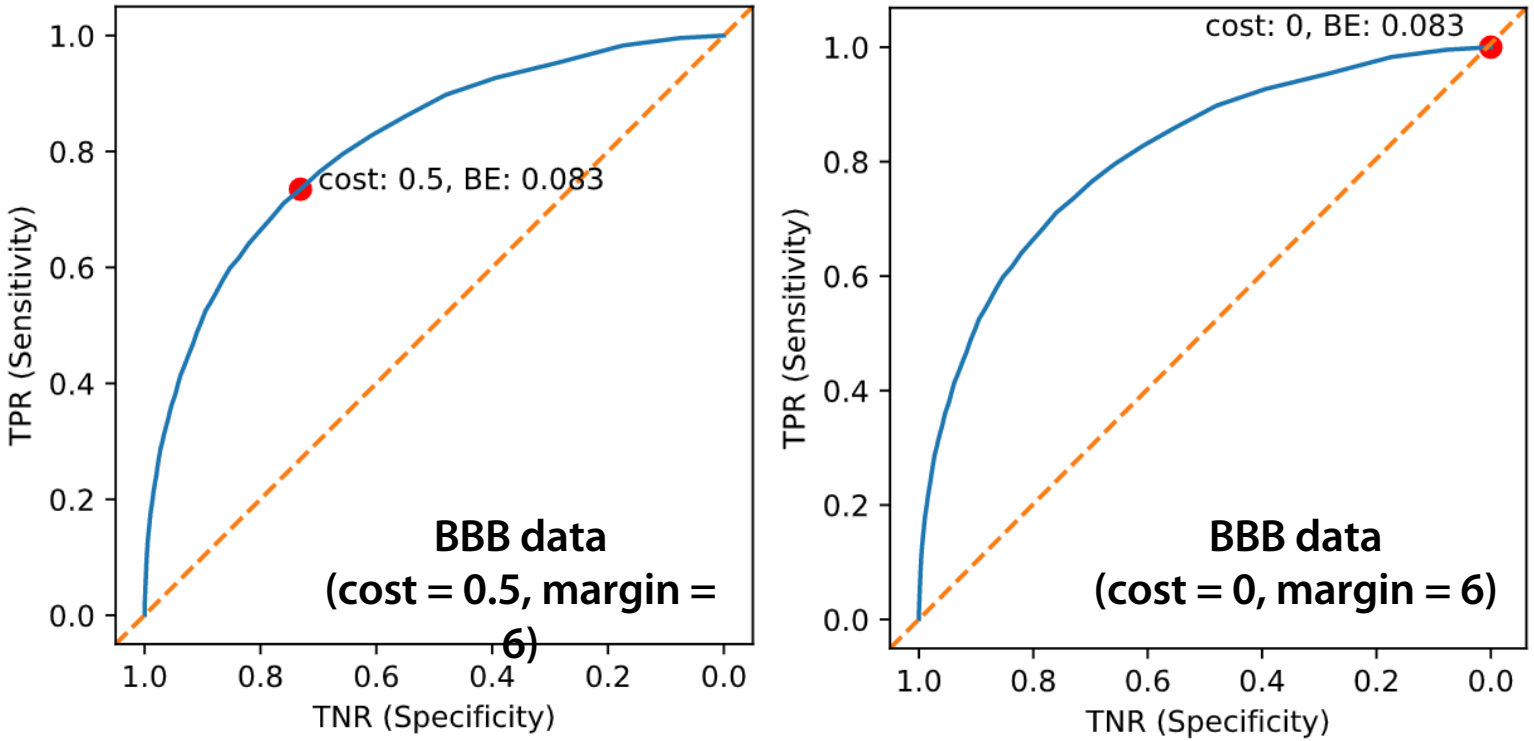
		Predicted	
		Pos.	**Neg.**
Actual	**Pos.**	655	176
	Neg.	10,871	16,176

- **TP**: True positive (predicted pos, actual pos)
- **TN**: True negative (predicted neg, actual neg)
- **FP**: False positive (predicted pos, actual neg)
- **FN**: False negative (predicted neg, actual pos)

Additional performance metrics used in practice

- **Accuracy**: Proportion of all outcomes that was correctly predicted as either positive or negative, i.e., (TP + TN) / (TP + TN + FP + FN)
- **Kappa**: Corrects the accuracy measure for the probability of generating a correct prediction purely by chance
- **True positive rate (TPR)**: Proportion of **actual positive outcomes** in the data that received a **positive prediction** (i.e., TP / (TP + FN)). Also known as **sensitivity** or **recall**
- **True negative rate (TNR)**: Proportion of **actual negative outcomes** in the data that received a **negative prediction** (i.e., TN / (TN + FP)). Also known as **specificity**
- **AUC:** Area Under the (ROC) Curve. The ROC curve plots the FPR against the TPR for all possible classification thresholds. AUC is the area under this curve. The maximum AUC value is 1 and the minimum value is 0.5

AUC is a measure of model performance at all possible thresholds



- True positive rate (TPR): Proportion of actual positive outcomes in the data that received a positive prediction (i.e., TP / (TP + FN)). Also known as sensitivity or recall
- True negative rate (TNR): Proportion of actual negative outcomes in the data

Probabilistic interpretation of AUC

AUC is the probability that Pred(X) > Pred(Y) where X is a randomly selected buyer and Y is a randomly selected non-buyer

```
(
    np.random.choice(pred_did_buy, nr) >
    np.random.choice(pred_did_not_buy, nr)
).mean()
```

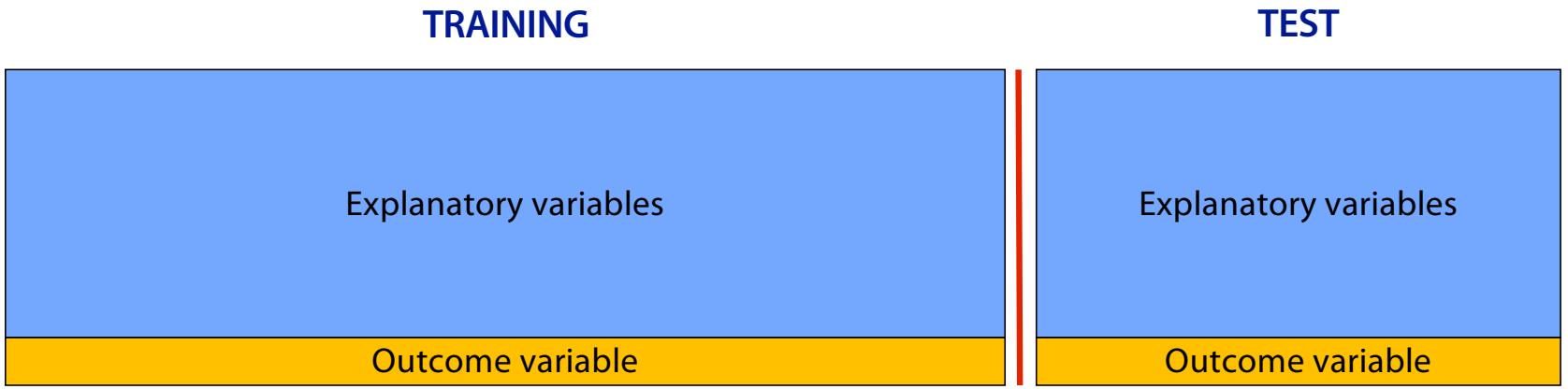
What does an AUC = 1 imply about “pred_did_buy” vs “pred_did_not_buy”?

What does an AUC = 0 imply about “pred_did_buy” vs “pred_did_not_buy”?

What does an AUC = 0.5 imply about “pred_did_buy” vs “pred_did_not_buy”?

https://www.alexejgossmann.com/auc/

TASK 8: How to “tune” the hyper parameters to avoid overfitting?

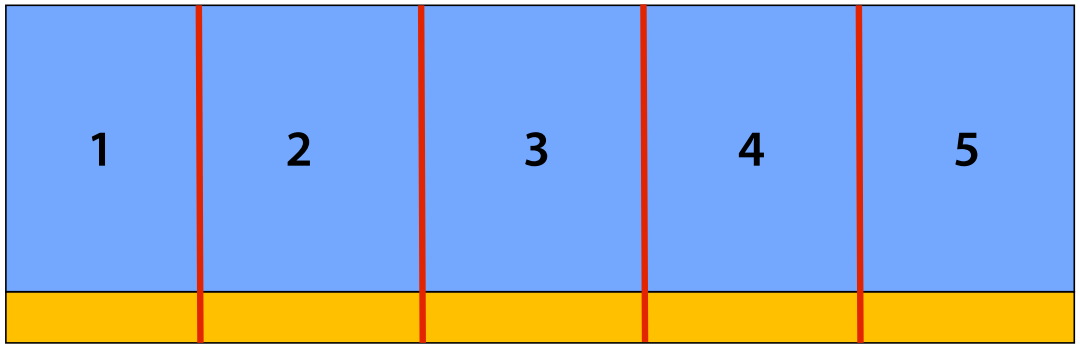


SIZE 1 2 3 4 5

DECAY 0 0.1 0.2 0.3 0.4 0.5

	Size				
Decay	1	2	3	4	5
0	(1 , 0)	(2 , 0)	(3 , 0)	(4 , 0)	(5 , 0)
0.1	(1, 0.1)	(2, 0.1)	(3, 0.1)	(4, 0.1)	(5, 0.1)
0.2	(1, 0.2)	(2, 0.2)	(3, 0.2)	(4, 0.2)	(5, 0.2)
0.3	(1, 0.3)	(2, 0.3)	(3, 0.3)	(4, 0.3)	(5, 0.3)
0.4	(1, 0.4)	(2, 0.4)	(3, 0.4)	(4, 0.4)	(5, 0.4)
0.5	(1, 0.5)	(2, 0.5)	(3, 0.5)	(4, 0.5)	(5, 0.5)

K-fold cross validation to “tune” hyper parameters



TRAIN	VALIDATE
1-4	5
2-5	1
3-1	2
4-2	3
5-3	4

HYPER PARAMETER GRID

	Size				
Decay	1	2	3	4	5
0	(1 , 0)	(2 , 0)	(3 , 0)	(4 , 0)	(5 , 0)
0.1	(1, 0.1)	(2, 0.1)	(3, 0.1)	(4, 0.1)	(5, 0.1)
0.2	(1, 0.2)	(2, 0.2)	(3, 0.2)	(4, 0.2)	(5, 0.2)
0.3	(1, 0.3)	(2, 0.3)	(3, 0.3)	(4, 0.3)	(5, 0.3)
0.4	(1, 0.4)	(2, 0.4)	(3, 0.4)	(4, 0.4)	(5, 0.4)
0.5	(1, 0.5)	(2, 0.5)	(3, 0.5)	(4, 0.5)	(5, 0.5)

The model associated with each cell in the “grid” is evaluated 5 times in a training-validation pair. The average performance metric for each grid cell is then used to determine the best hyper parameters to use.

TASK 8: K-fold cross validation to “tune” hyper parameters for NN (classification) – see python/bbb_sklearn.py

```
result <- nn(
  dvd,
  rvar = "buy",
  evar = c("coupon", "purch", "last"),
  lev = "yes",
  size = 1,
  decay = 0.5,
  seed = 1234
)
summary(result, prn = TRUE)
cv.nn(result, size = 1:5, decay = seq(0, 0.5, 0.1), fun = auc)
```

	auc (mean)	std	min	max	decay	size
1	0.8033182	0.011697243	0.7895879	0.8178669	0.2	2
2	0.8033004	0.004589567	0.7971142	0.8068806	0.1	2
3	0.8032534	0.007469075	0.7932854	0.8105454	0.1	1
4	0.8032138	0.008531422	0.7932959	0.8130288	0.2	3
5	0.8031805	0.010234796	0.7951378	0.8208853	0.2	1
6	0.8031724	0.005997718	0.7946074	0.8093790	0.1	3
7	0.8031589	0.009481509	0.7914550	0.8149171	0.5	1
8	0.8031366	0.009842612	0.7946294	0.8174317	0.5	5
9	0.8031187	0.005555916	0.7939298	0.8077333	0.5	4
10	0.8031108	0.007351791	0.7957296	0.8118604	0.0	1

TASK 9: Experimental design and partial factorials (see R/bizware.nb.html – python version to be added soon)

price	message	promotion	response
USD150	speed	trial	0.14
USD150	power	gift	0.40
USD160	power	trial	0.09
USD160	speed	gift	0.13
USD170	power	trial	0.06
USD170	speed	gift	0.10
USD180	speed	trial	0.01
USD180	power	gift	0.07

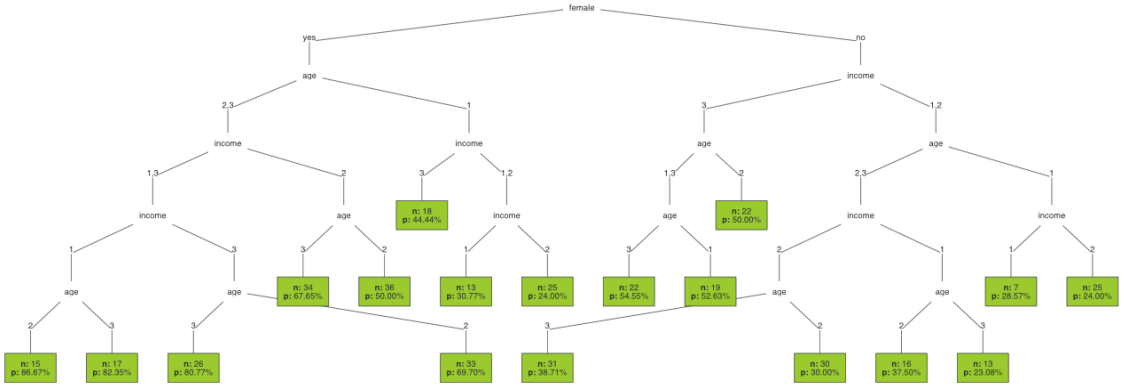
source: Boost your Marketing ROI with Experimental Design (HBR)
Authors: Eric Almquist and Gordon Wyner

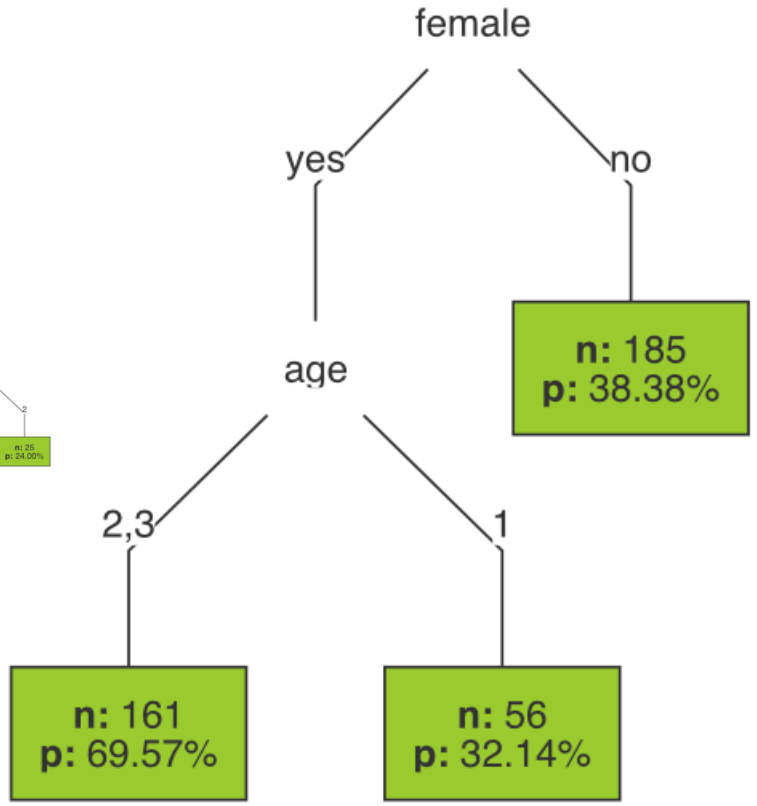
Assume the sample size for each cell was 2,000

- Generate a partial factorial design using information about factors and levels shown in the response table (**use radiant**)
- Did you get the same design? Why (not)?
- Estimate a logistic regression based on the response table shown and predict response for all profiles
- Use **data/bizware.xlsx**
- What are the 2 top offers?
- What are the 2 worst offers?

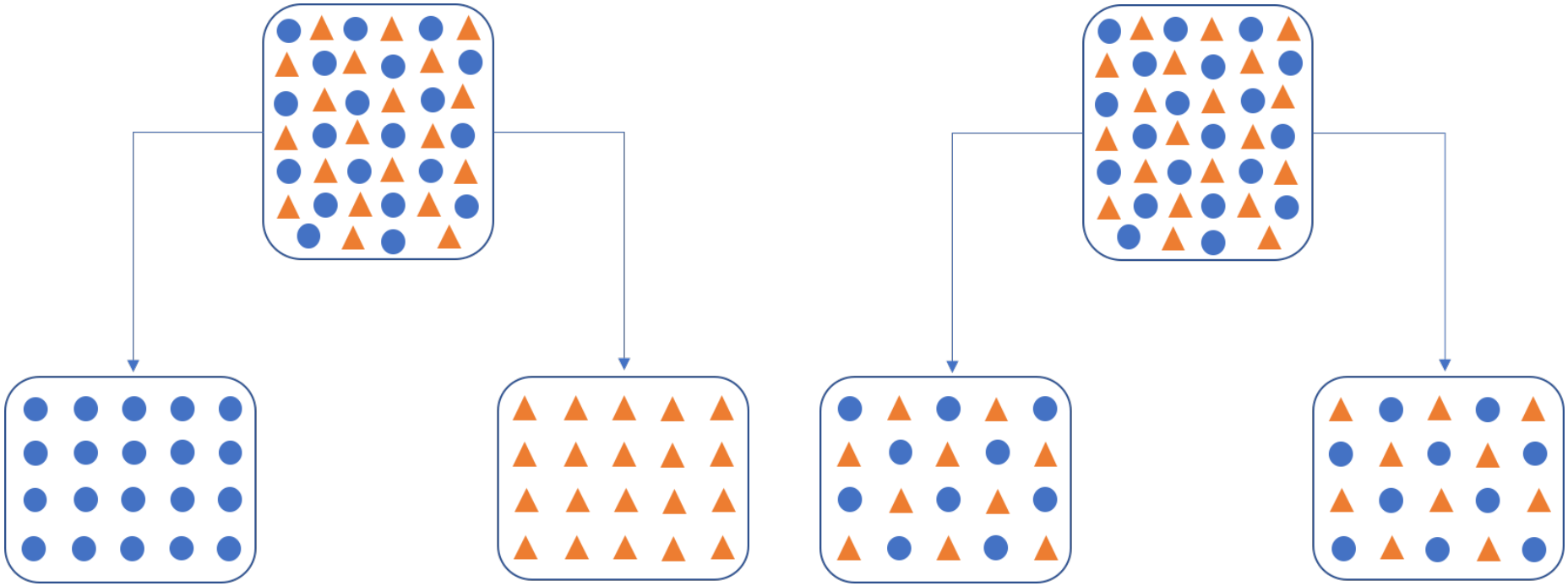
What is the decision tree splitting process?

USE DATA/CART_DEMO50.RDS TO REPLICATE





Decision trees: Best possible split vs worst possible split



TASK 10: Calculate "node impurity" when there are two classes (CART)

$$I(A) = p_1 \times (1 - p_1) + p_2 \times (1 - p_2)$$

$$\Delta I = N(A)I(A) - N(A_L)I(A_L) - N(A_R)I(A_R)$$

$I(A)$ is the level of *impurity* in the node we want to split

$N(A)$ is the number of observations in the node we want to split

$I(A_L)$ and $I(A_R)$ represent the level of *impurity* in the node's children after the split

$N(A_L)$ and $N(A_L)$ are the number of observations in the node's children after the split

Creating a decision tree starts at the root (node)

Female variable:

- Cross tab "female" and "response"
- Calculate the reduction in impurity from the split

```
Pivot table
Data        : cart_demo50
Categorical : response female

 female yes  no Total
    yes 130  87   217
     no  71 114   185
  Total 201 201   402
```

female == “yes” vs female == “no”

9.26

- Root: 402 x (201/402 x (1-201/402) + 201/402 x (1 - 201/402))

- Impurity reduction: 201 – 104.24 – 87.5 = 9.26

Now evaluate all possible splits of the root node using age

Age variable:

- Cross tab "age" and "response"
- Calculate the reduction in impurity for each split

```
Pivot table
Data        : cart_demo50
Categorical : response age

   age yes  no Total
     1  36  71   107
     2  80  72   152
     3  85  58   143
 Total 201 201   402
```

- **age == 1 vs age == 2 | age == 3**
- age == 2 vs age == 1 | age == 3
- age == 3 vs age == 1 | age == 2

7.80
0.34
3.96

Finally, evaluate all possible splits of the root node using income

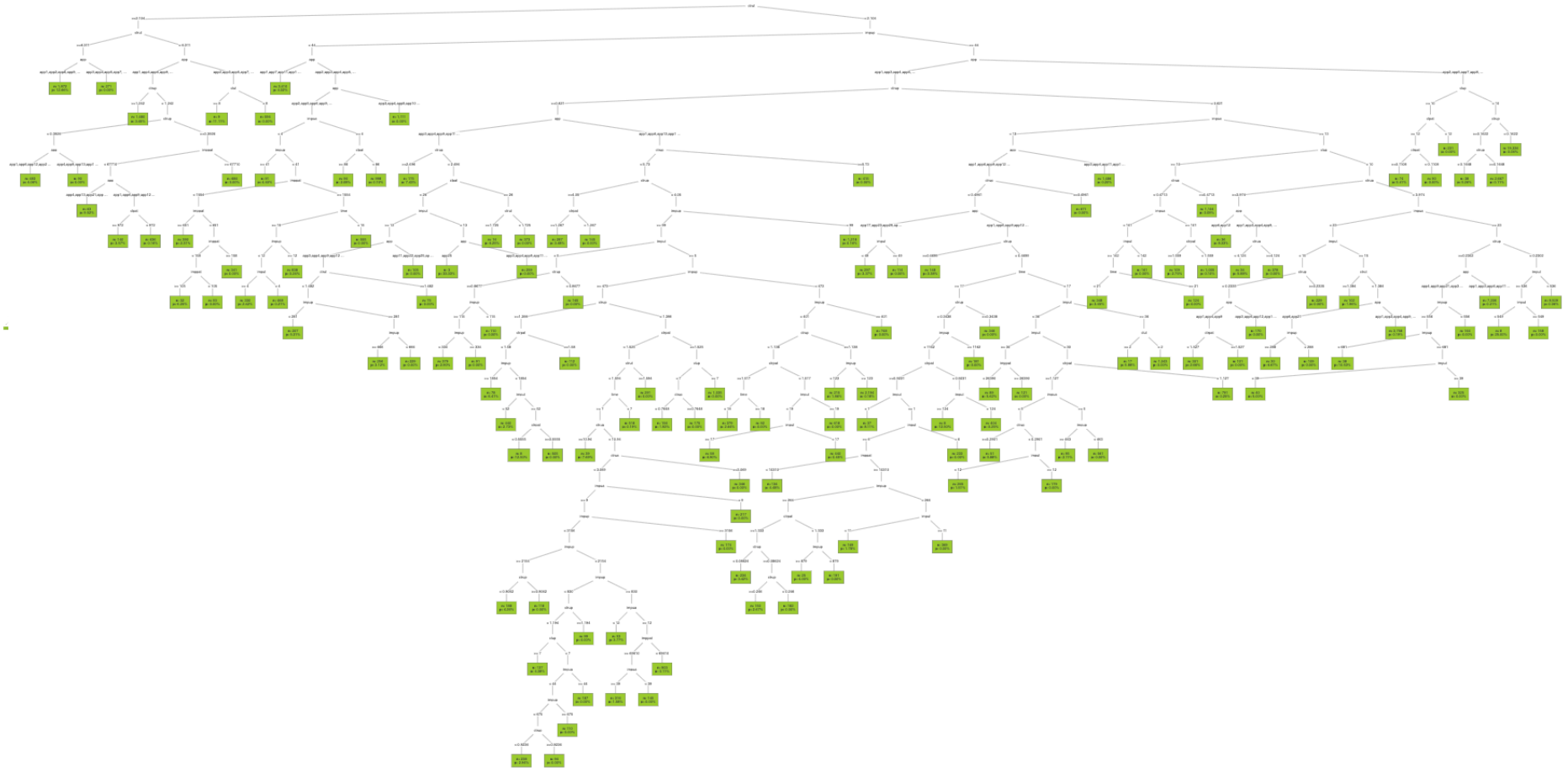
Income variable:

- Cross tab "income" and "response"
- Calculate reduction in impurity for each split

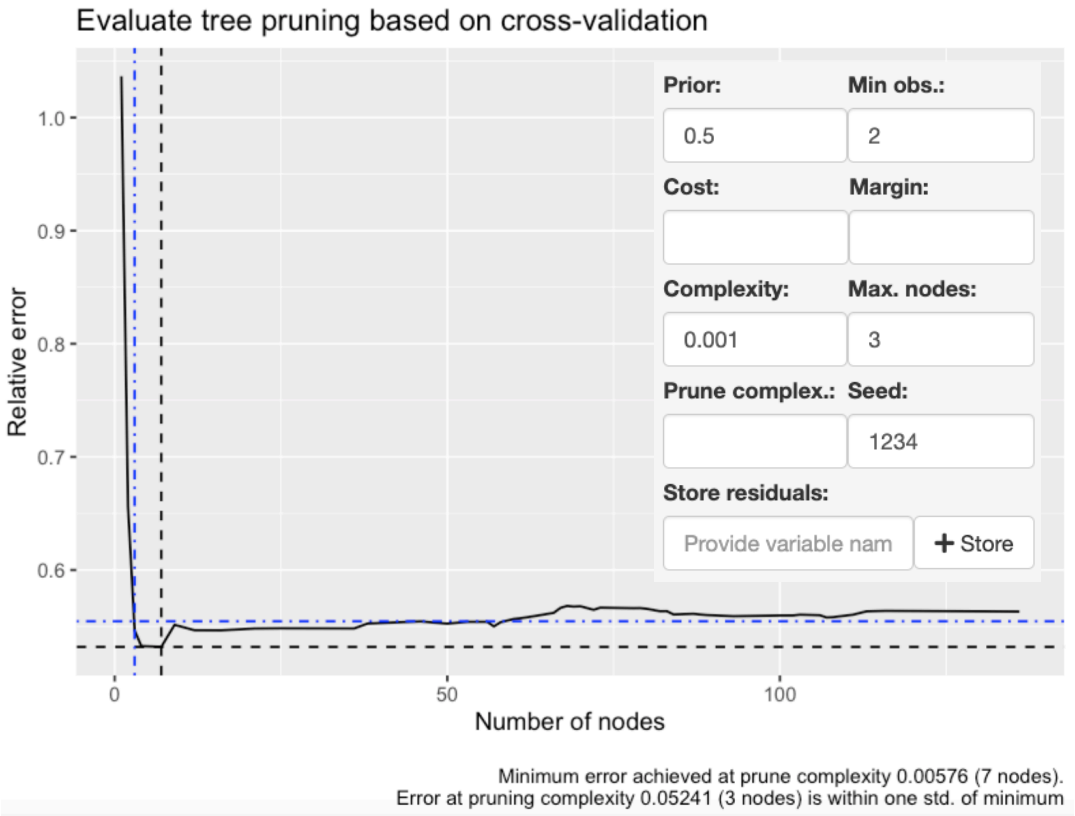
```
Pivot table
Data        : cart_demo50
Categorical : response income

 income yes  no Total
      1  42  39    81
      2  74 107   181
      3  85  55   140
  Total 201 201   402
```

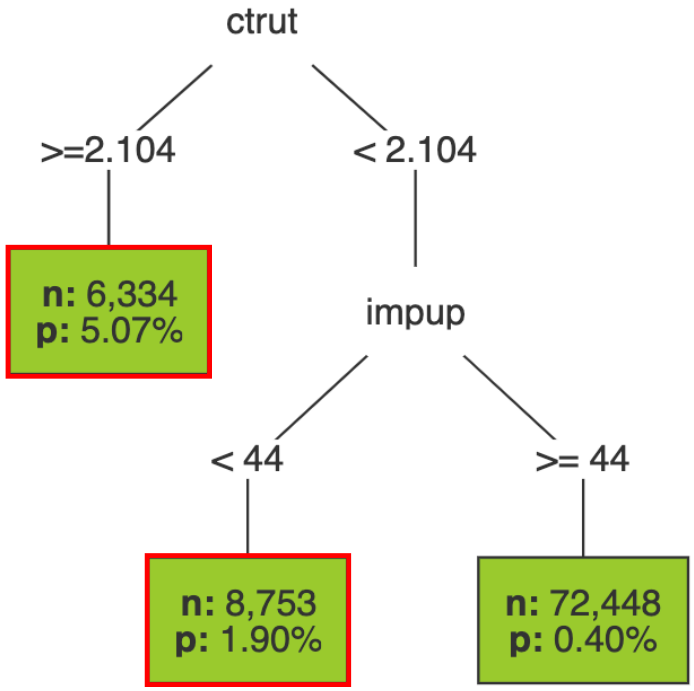
- income == 1 vs income == 2 | income == 3 — 0.01
- **income == 2 vs income == 1 | income == 3** — **5.47**
- income == 3 vs income == 1 | income == 2 — 4.93

Step 1: Construct a big tree on the training sample (training == 1)

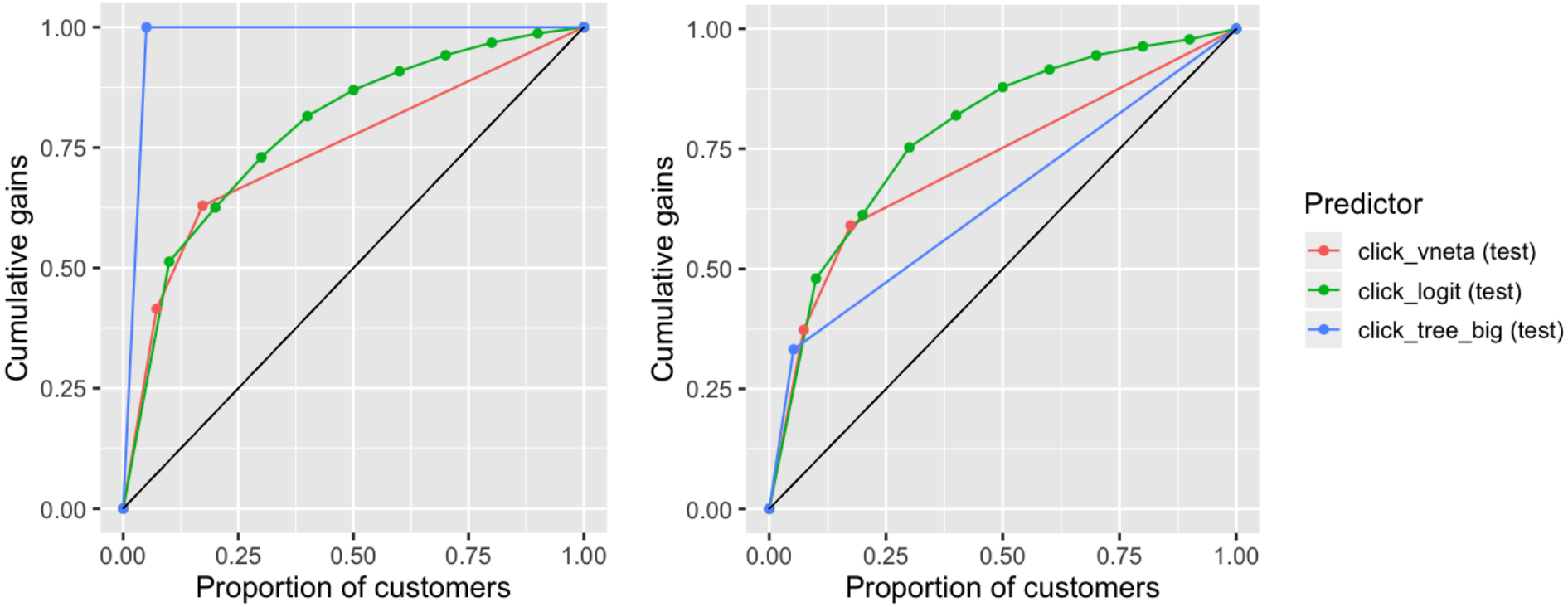


Step 2: Prune the tree based on cross-validation

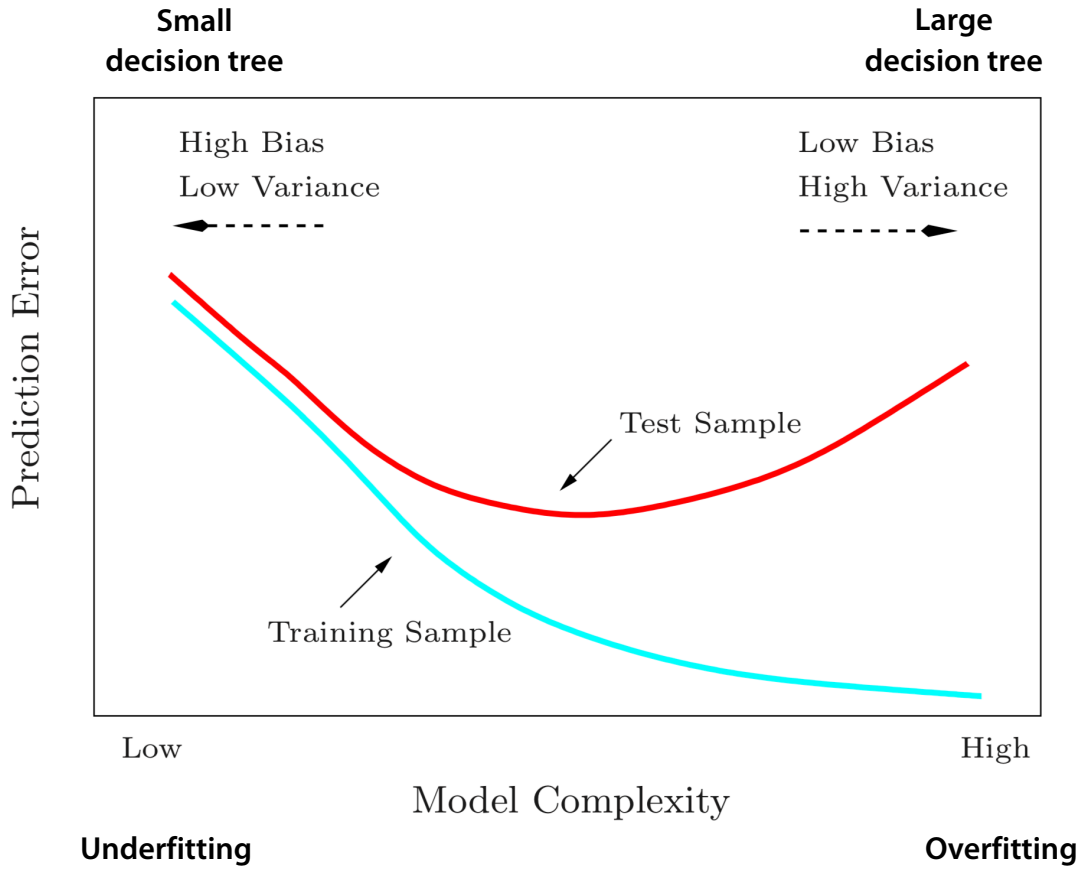




The un-pruned decision tree over-fits the training data massively!



“Ensembles” of trees address key weakness of single decision trees



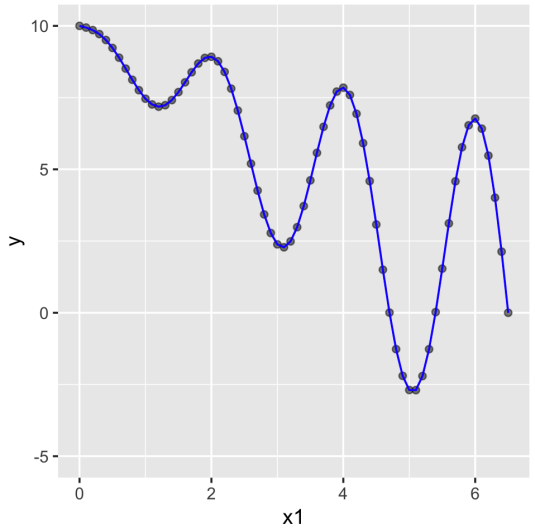
- Random Forests combine many large (overfit) decision tree to reduce variance
- Boosted Decision tree combine may small (underfit) decision trees to reduce bias
- Graph source: **The Elements of Statistical Learning**

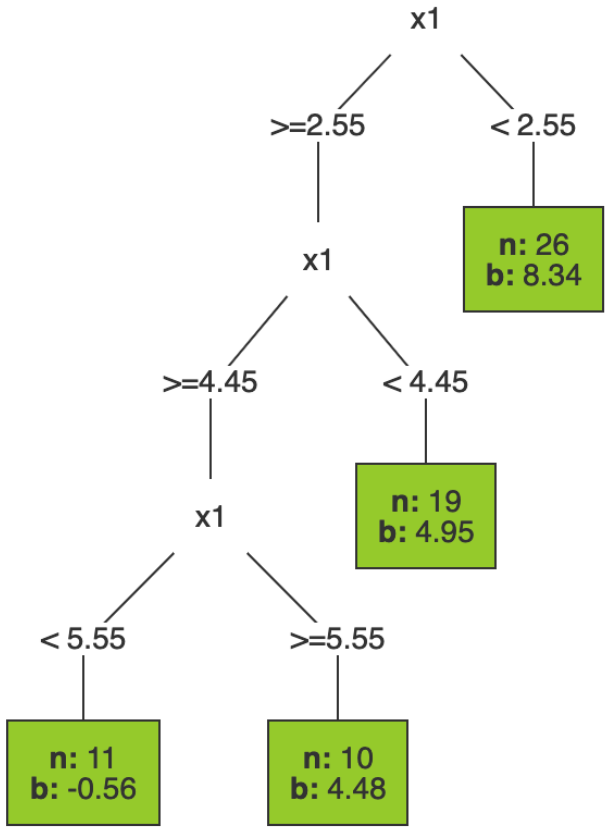
How does a random forests work?

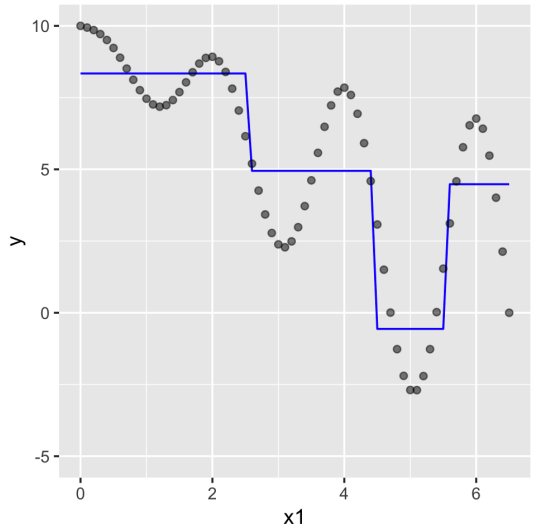
RANDOM FOREST IDEA

- Algorithm uses randomness to address overfitting for decision trees (Breiman and Cutler)
- Can be used with different decision tree algorithms (e.g., CART)
- Key idea is to create many decision trees, each of based on a
  - randomly chosen subsample of the data
  - randomly chosen subset of the explanatory variables at each node
- Very accurate predictor that can handle large numbers of explanatory variables [**WHY?**]

Boosted Decision Trees are even more popular (e.g., XGBoost)

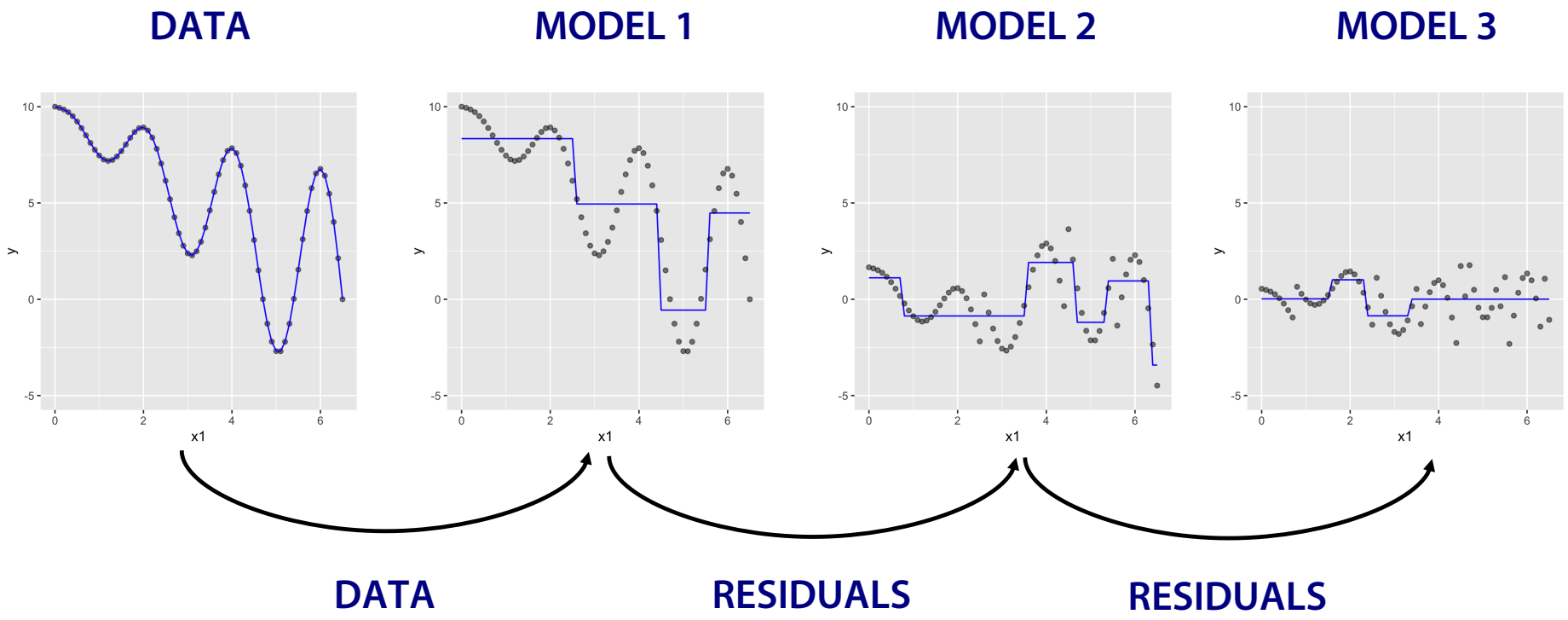




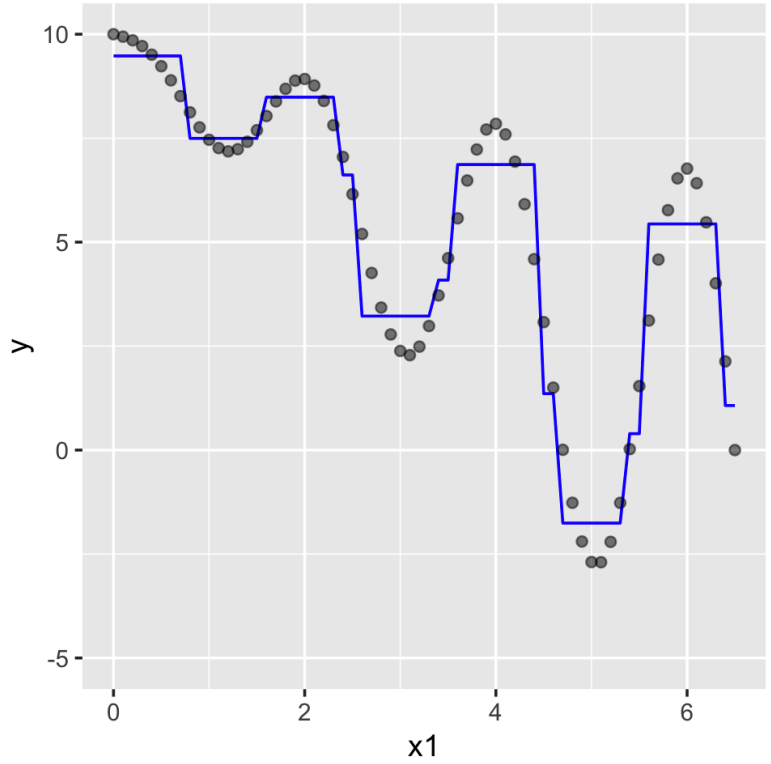


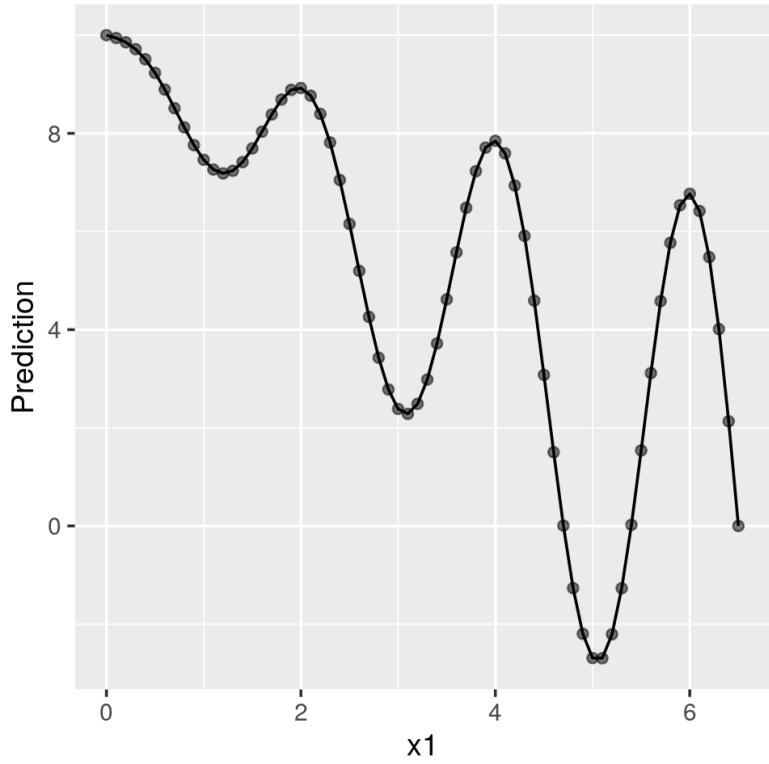
TASK 11: Reproduce a simple boosted regression tree (see python/boosting_regression.ipynb)

Boosted Decision Trees combine “weak learners” applied to residuals from previous model(s)



To generate a final prediction we can sum the 3 tree predictions





- Note: The above prediction uses a “learning rate” of 1. In practice, we would use a much smaller number (e.g., 0.01) and build (many) more trees

TASK 12: Perform RFM analysis

BOOKBINDERS ANALYSIS (USE DATA/BBB.PKL)

```
recency   = last
frequency = purch
monetary  = total
```

- Construct sequential RFM index
- Calculate break-even response rate
- Create bar charts of **rec_sq**, **freq_sq**, and **mon_sq** vs the response rate
- Create a bar-chart the **rfm_sq** index vs the response rate
- Construct a **mailto_sq** variable based **on the training data**
- Estimate a logistic regression that has the same "response curve" as RFM (i.e., use RFM index as an explanatory variable)

- Extra:
  - Calculate profitability of sequential RFM, projecting to the remaining 500,000 customers in the database based on predictions in the test set
  - Assume cost per contact is $1 and margin per sale is $9 (excl. contact cost)