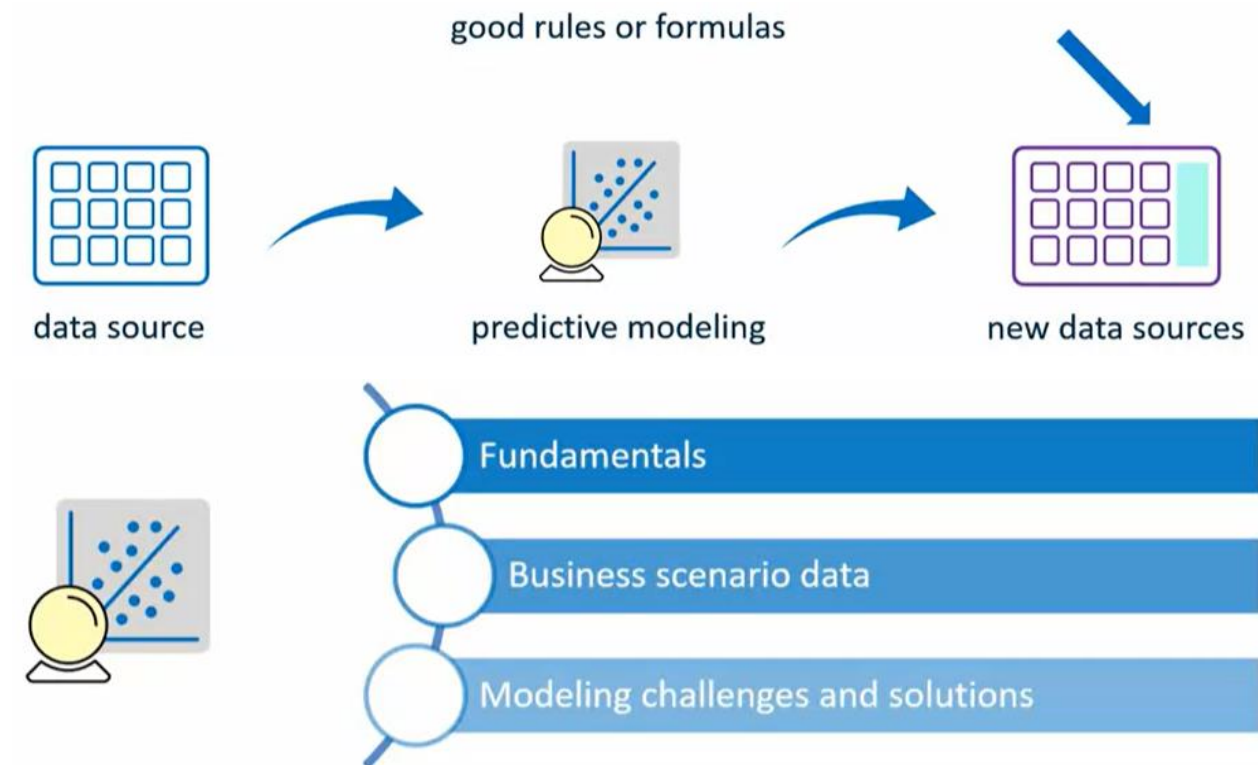


SBA Statistical Business Analyst with SAS

SBA3 Predictive Modeling with Logistic Regression

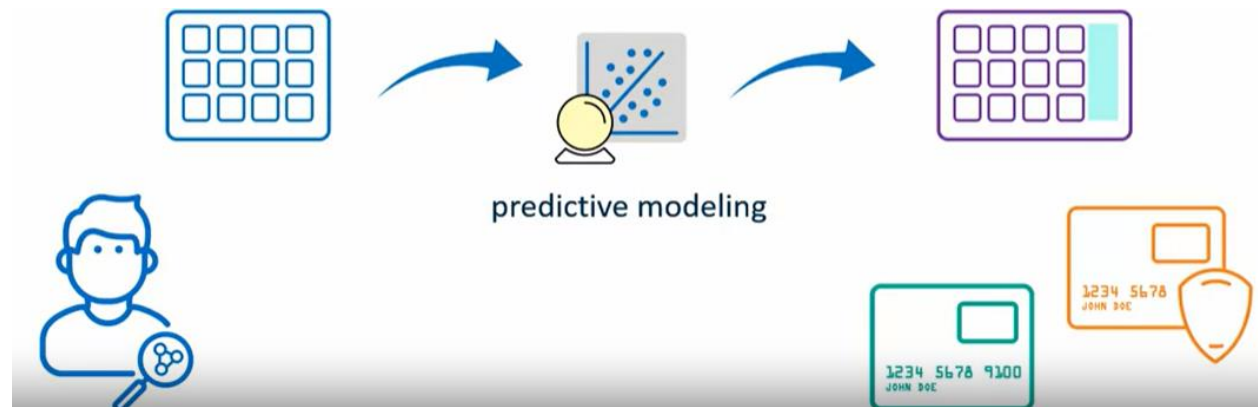
SBA301 Predictive Modeling Fundamentals and Predictive Modeling Challenges

Overview



Predictive Modeling Fundamentals

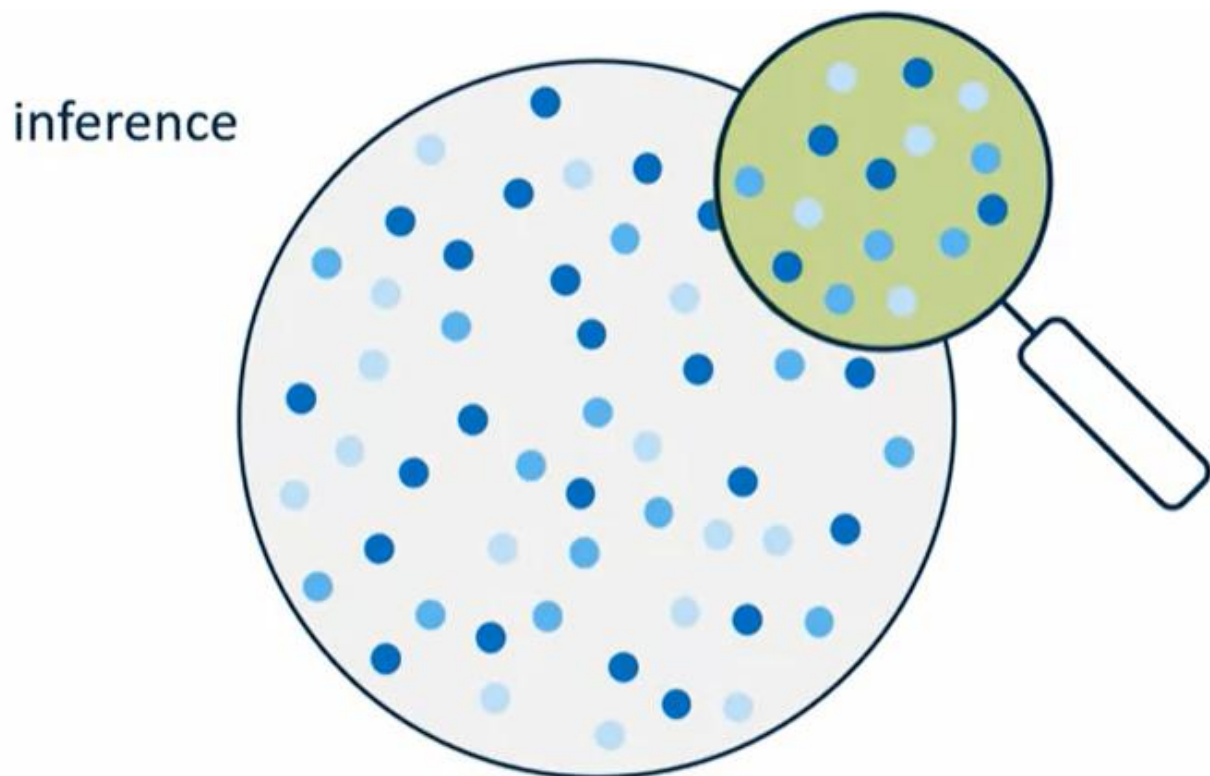
Introduction

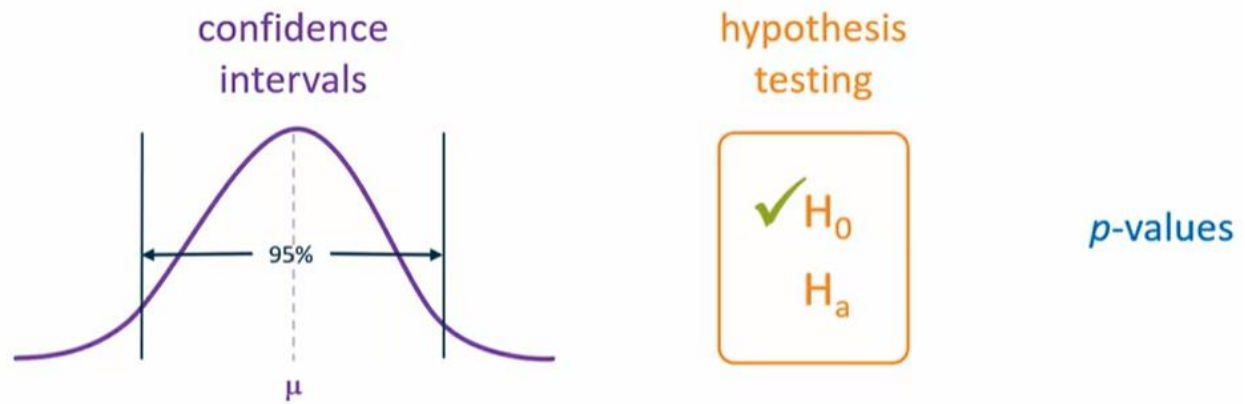


In this topic, you learn to do the following:

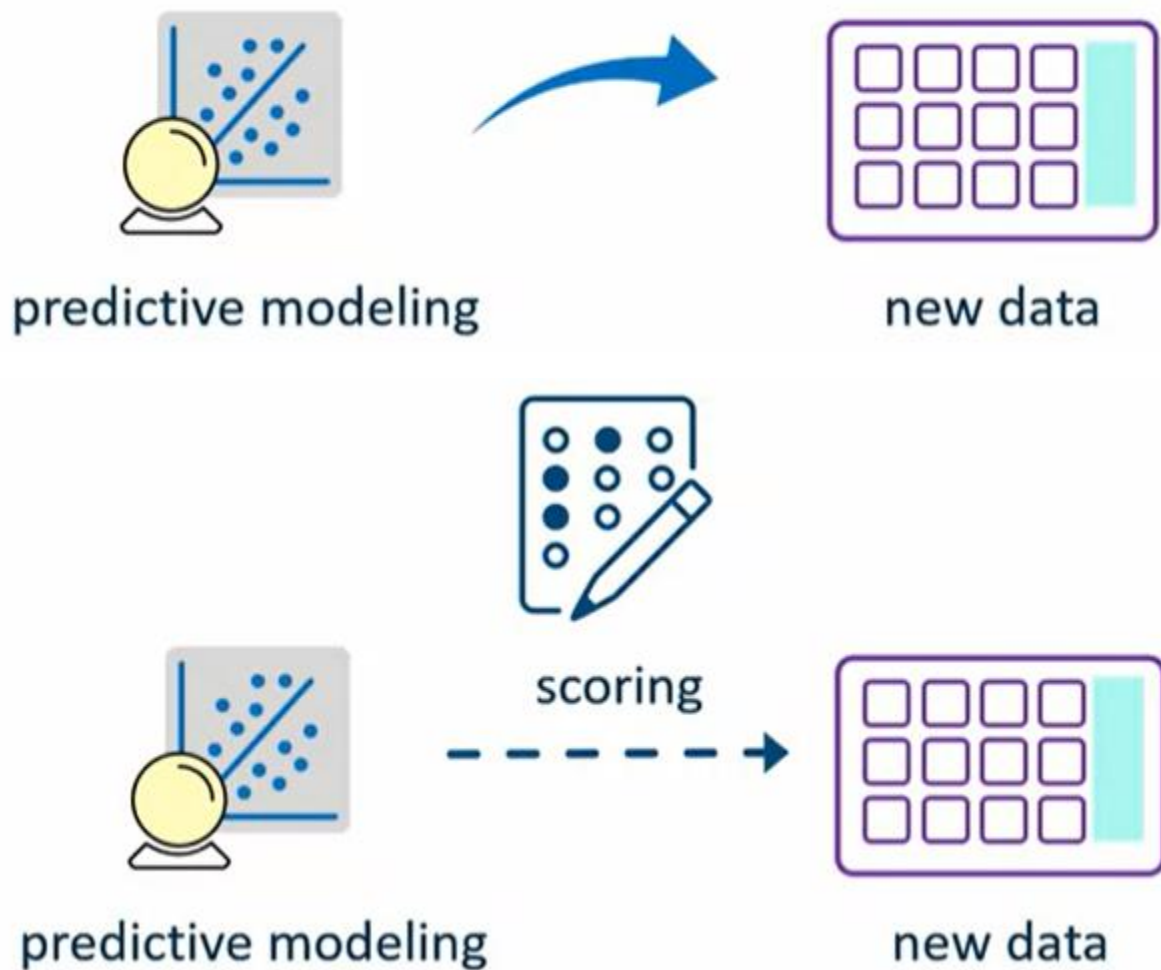
- identify the goals of predictive modeling
- define terminology of predictive modeling elements
- explain the basic steps of predictive modeling
- identify business applications of predictive modeling
- identify issues with the business scenario data

Goals of Predictive Modeling



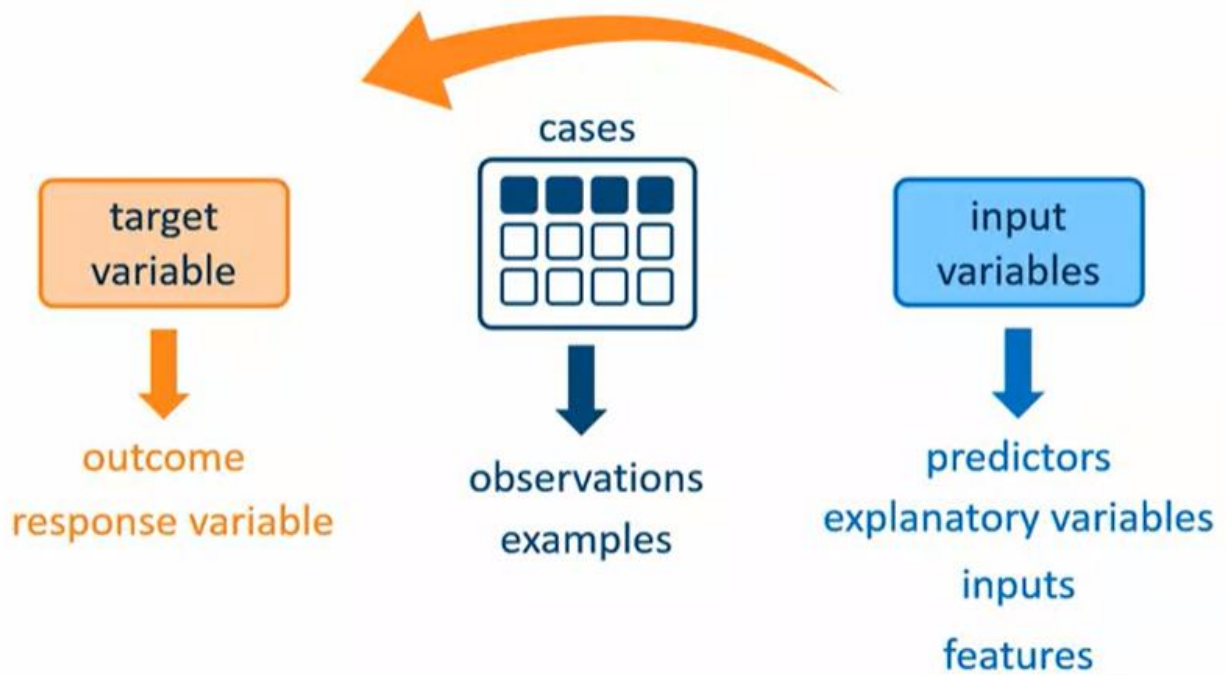


generalization





Terms for Elements in Predictive Modeling



Step 1

Build model on historical data.



predictive modeling

↓ ✓

	Y	X ₁	X ₂	...	X _k
1	1	■	■	...	■
2	1	■	■	...	■
3	0	■	■	...	■
⋮	⋮	⋮	⋮		⋮
n	0	■	■	...	■

Step 1

Supervised classification.



predictive modeling

discrete ↓ ✓

continuous ✗

	Y	X ₁	X ₂	...	X _k
1	1	■	■	...	■
2	1	■	■	...	■
3	0	■	■	...	■
⋮	⋮	⋮	⋮		⋮
n	0	■	■	...	■

Step 1

Supervised classification.



class label

binary

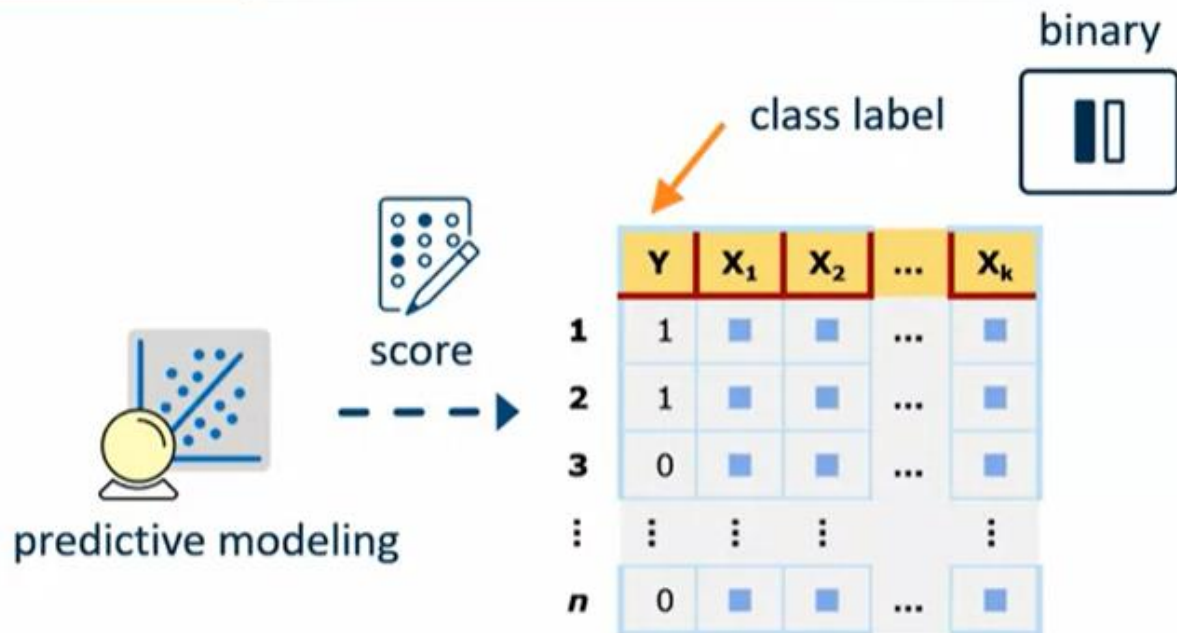
	Y	X ₁	X ₂	...	X _k
1	1	■	■	...	■
2	1	■	■	...	■
3	0	■	■	...	■
⋮	⋮	⋮	⋮	⋮	⋮
n	0	■	■	...	■

1 – response to an offer

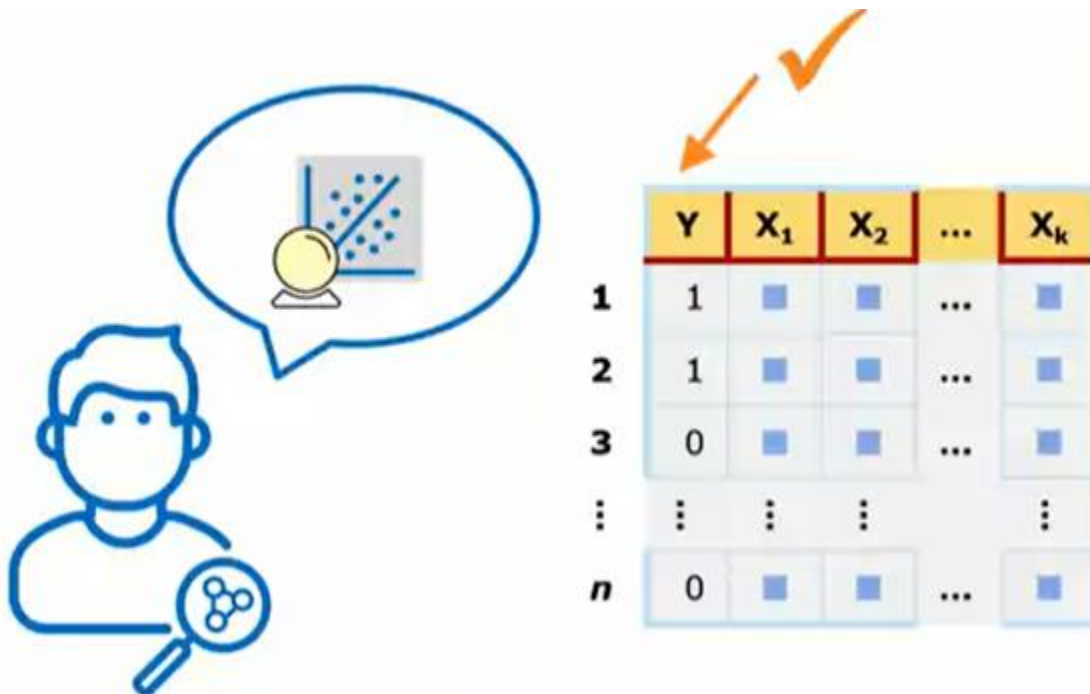
0 – no response to an offer

Step 1

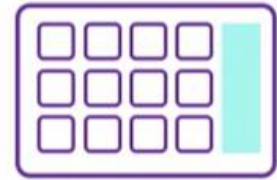
Supervised classification.



probability that a case belongs to a particular class



	Y	X ₁	X ₂	...	X _k
1	1	■	■	...	■
2	1	■	■	...	■
3	0	■	■	...	■
⋮	⋮	⋮	⋮		⋮
n	0	■	■	...	■



Step 1

Supervised classification.

Step 2

Generalization.

	Y	X ₁	X ₂	...	X _k
1	1	■	■	...	■
2	1	■	■	...	■
3	0	■	■	...	■
⋮	⋮	⋮	⋮		⋮
n	0	■	■	...	■



	Y	X ₁	X ₂	...	X _k
1	?	■	■	...	■
2	?	■	■	...	■
3	?	■	■	...	■
⋮	⋮	⋮	⋮		⋮
n	?	■	■	...	■



	Y	X ₁	X ₂	...	X _k
1	1	■	■	...	■
2	1	■	■	...	■
3	0	■	■	...	■
⋮	⋮	⋮	⋮		⋮
n	0	■	■	...	■

Step 1

Supervised classification.

- Prepare the inputs.

- handling missing values
- dealing with categorical inputs
- reducing redundancy
- performing variable screening



Step 1

Supervised classification.

- Prepare the inputs.
- Select the most predictive inputs and fit the models.

Step 2

Generalization.

- Assess the models.

Applications of Predictive Modeling



target marketing



	Y	X ₁	X ₂	...	X _k
1	1	■	■	...	■
2	1	■	■	...	■
3	0	■	■	...	■
⋮	⋮	⋮	⋮	⋮	⋮
n	0	■	■	...	■

improve sales
promotions and
product loyalty

target marketing



reponse
to past
promotion



previous purchase history
demographics

customers

	Y	X ₁	X ₂	...	X _k
1	1	■	■	...	■
2	1	■	■	...	■
3	0	■	■	...	■
⋮	⋮	⋮	⋮	⋮	⋮
n	0	■	■	...	■

target marketing



customers
who are
likely to
respond

	Y	X ₁	X ₂	...	X _k
1	1	■	■	...	■
2	1	■	■	...	■
3	0	■	■	...	■
⋮	⋮	⋮	⋮	⋮	⋮
n	0	■	■	...	■



	Y	X ₁	X ₂	...	X _k
1	?	■	■	...	■
2	?	■	■	...	■
3	?	■	■	...	■
⋮	⋮	⋮	⋮	⋮	⋮
n	?	■	■	...	■



customers
who are
likely to
respond



attrition prediction



	Y	X ₁	X ₂	...	X _k
1	1	■	■	...	■
2	1	■	■	...	■
3	0	■	■	...	■
⋮	⋮	⋮	⋮	⋮	⋮
n	0	■	■	...	■

customers at risk of churn

	Y	X ₁	X ₂	...	X _k
1	?	■	■	...	■
2	?	■	■	...	■
3	?	■	■	...	■
⋮	⋮	⋮	⋮	⋮	⋮
n	?	■	■	...	■

customers at risk of churn ✓

credit scoring



paid or defaulted



credit application
credit reports

past applicants

	Y	X ₁	X ₂	...	X _k
1	1	■	■	...	■
2	1	■	■	...	■
3	0	■	■	...	■
⋮	⋮	⋮	⋮	⋮	⋮
n	0	■	■	...	■

reduce defaults and serious delinquencies

fraud detection



fraudulent?



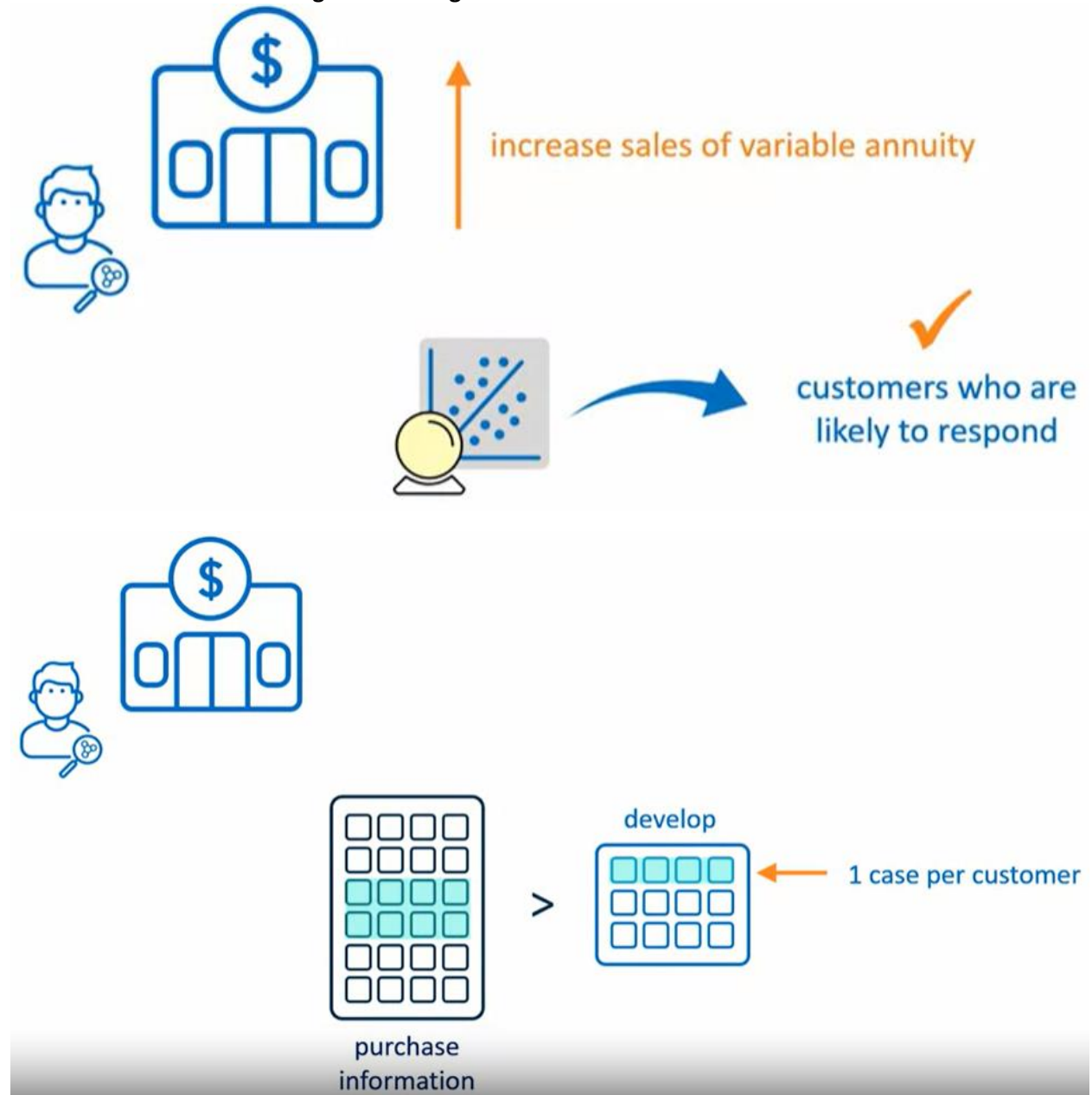
particulars/circumstances
of transactions

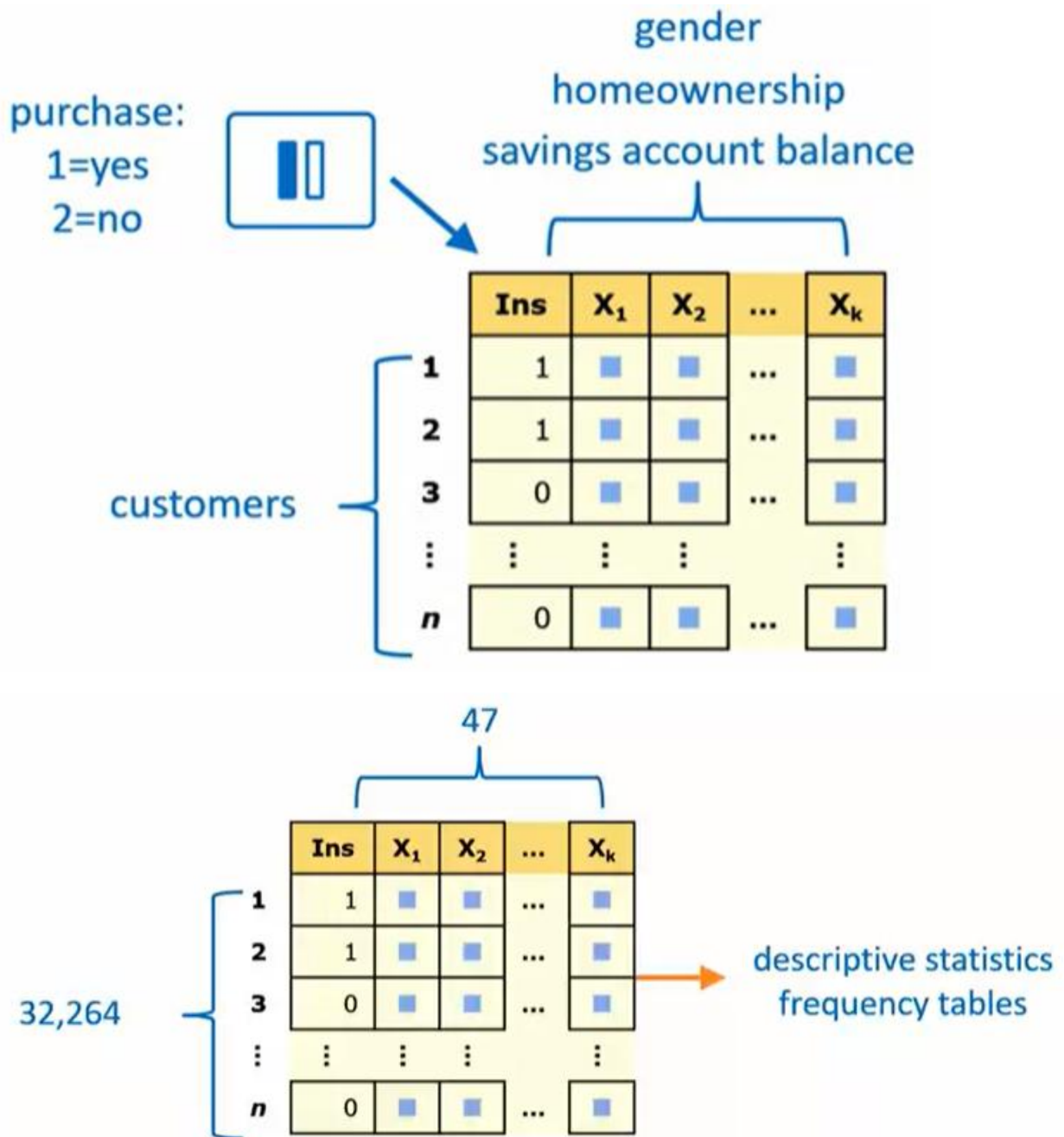
insurance
claims,
transactions

	Y	X ₁	X ₂	...	X _k
1	1	■	■	...	■
2	1	■	■	...	■
3	0	■	■	...	■
⋮	⋮	⋮	⋮	⋮	⋮
n	0	■	■	...	■

anticipate fraud on new transaction/claims

Demonstration Scenario: Target Marketing for a Bank





Demo: Examining the Code for Generating Descriptive Statistics and Frequency Tables

```
pmlr01d01.sas

/* Use a temporary data set for development. */
data work.develop;
    set pmlr.develop;
run;

/* Establish a macro variable for the list of numeric inputs */
%global inputs;
%let inputs=ACCTAGE DDA DDABAL DEP DEPAMT CASHBK
CHECKS DIRDEP NSF NSFAMT PHONE TELLER
SAV SAVBAL ATM ATMAMT POS POSAMT CD
CDBAL IRA IRABAL LOC LOCBAL INV
INVBAL ILS ILSBAL MM MMBAL MMCRED MTG
MTGBAL CC CCBAL CCPURC SDB INCOME
HMOWN LORES HMVAL AGE CRSCORE MOVED
INAREA;

/* Investigate the distribution of the numeric inputs */
proc means data=work.develop n nmiss mean min max;
    var &inputs;
run;

/* Investigate the distribution of the categorical inputs and
proc freq data=work.develop;
    tables ins branch res;
run;

/* ===== */
/* Lesson 1, Section 1: l1d1.sas
Demonstration: Examining the Code for Generating
Descriptive Statistics and Frequency Tables
[m641_1_i; derived from pmlr01d01.sas] */
/* ===== */
```

```
data work.develop;
    set pmlr.develop;
run;
```

```
%global inputs;  
%let inputs=ACCTAGE DDA DDABAL DEP DEPAMT CASHBK  
CHECKS DIRDEP NSF NSFAMT PHONE TELLER  
SAV SAVBAL ATM ATMAMT POS POSAMT CD  
CDBAL IRA IRABAL LOC LOCBAL INV  
INVBAL ILS ILSBAL MM MMBAL MMCRED MTG  
MTGBAL CC CCBAL CCPURC SDB INCOME  
HMOWN LORES HMVAL AGE CRSCORE MOVED  
INAREA;
```

```
proc means data=work.develop n nmiss mean min max;  
var &inputs;  
run;
```

```
proc freq data=work.develop;  
tables ins branch res;  
run;
```

The MEANS Procedure

Variable	Label	N	N Miss	Mean	Minimum	Maximum
AcctAge	Age of Oldest Account	30194	2070	5.9086772	0.3000000	61.5000000
DDA	Checking Account	32264	0	0.8156459	0	1.0000000
DDABal	Checking Balance	32264	0	2170.02	-774.8300000	278093.83
Dep	Checking Deposits	32264	0	2.1346082	0	28.0000000
DepAmt	Amount Deposited	32264	0	2232.76	0	484893.67
CashBk	Number Cash Back	32264	0	0.0159621	0	4.0000000
Checks	Number of Checks	32264	0	4.2599182	0	49.0000000
DirDep	Direct Deposit	32264	0	0.2955616	0	1.0000000
NSF	Number Insufficient Fund	32264	0	0.0870630	0	1.0000000
NSFAmt	Amount NSF	32264	0	2.2905464	0	666.8500000
Phone	Number Telephone Banking	28131	4133	0.4056024	0	30.0000000
Teller	Teller Visits	32264	0	1.3652678	0	27.0000000
Sav	Saving Account	32264	0	0.4668981	0	1.0000000
SavBal	Saving Balance	32264	0	3170.60	0	700026.94
ATM	ATM	32264	0	0.6099368	0	1.0000000
ATMAmt	ATM Withdrawal Amount	32264	0	1235.41	0	427731.26
POS	Number Point of Sale	28131	4133	1.0756816	0	54.0000000
POSAmt	Amount Point of Sale	28131	4133	48.9261782	0	3293.49
CD	Certificate of Deposit	32264	0	0.1258368	0	1.0000000
CDBal	CD Balance	32264	0	2530.71	0	1053900.00
IRA	Retirement Account	32264	0	0.0532792	0	1.0000000
IRABal	IRA Balance	32264	0	617.5704550	0	596497.60
LOC	Line of Credit	32264	0	0.0633833	0	1.0000000
LOCBal	Line of Credit Balance	32264	0	1175.22	-613.0000000	523147.24
Inv	Investment	28131	4133	0.0296826	0	1.0000000
InvBal	Investment Balance	28131	4133	1599.17	-2214.92	8323796.02
ILS	Installment Loan	32264	0	0.0495909	0	1.0000000
ILSBal	Loan Balance	32264	0	517.5692344	0	29162.79
MM	Money Market	32264	0	0.1148959	0	1.0000000
MMBal	Money Market Balance	32264	0	1875.76	0	120801.11
MMCred	Money Market Credits	32264	0	0.0563786	0	5.0000000
MTG	Mortgage	32264	0	0.0493429	0	1.0000000
MTGBal	Mortgage Balance	32264	0	8081.74	0	10887573.28
CC	Credit Card	28131	4133	0.4830969	0	1.0000000
CCBal	Credit Card Balance	28131	4133	9586.55	-2060.51	10641354.78
CCPurc	Credit Card Purchases	28131	4133	0.1541716	0	5.0000000
SDB	Safety Deposit Box	32264	0	0.1086660	0	1.0000000
Income	Income	26482	5782	40.5889283	0	233.0000000
HMOwn	Owns Home	26731	5533	0.5418802	0	1.0000000
LORes	Length of Residence	26482	5782	7.0056642	0.5000000	19.5000000

The FREQ Procedure

Ins	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	21089	65.36	21089	65.36
1	11175	34.64	32264	100.00

Branch of Bank				
Branch	Frequency	Percent	Cumulative Frequency	Cumulative Percent
B1	2819	8.74	2819	8.74
B10	273	0.85	3092	9.58
B11	247	0.77	3339	10.35
B12	549	1.70	3888	12.05
B13	535	1.66	4423	13.71
B14	1072	3.32	5495	17.03
B15	2235	6.93	7730	23.96
B16	1534	4.75	9264	28.71
B17	850	2.63	10114	31.35
B18	541	1.68	10655	33.02
B19	285	0.88	10940	33.91
B2	5345	16.57	16285	50.47
B3	2844	8.81	19129	59.29
B4	5633	17.46	24762	76.75
B5	2752	8.53	27514	85.28
B6	1438	4.46	28952	89.73
B7	1413	4.38	30365	94.11
B8	1341	4.16	31706	98.27
B9	558	1.73	32264	100.00

Area Classification				
Res	Frequency	Percent	Cumulative Frequency	Cumulative Percent
R	8077	25.03	8077	25.03
S	11506	35.66	19583	60.70
U	12681	39.30	32264	100.00

Practice: Exploring the Bank Data for the Target Marketing Project

Question 1

In this activity, you explore the **develop** data set to become familiar with the data and identify possible data issues. (This activity uses the code that is shown in the previous demonstration.)

Reminder: If you started a new SAS session, you must run **setup.sas** to define the **pmlr** library before you do this practice.

1. Open **create_work_develop.sas** from the **practices** folder. The DATA step creates the temporary data set **develop** from the **pmlr.develop** data set that is included in the data for this course.
2. Submit the code and then answer the questions that follow.

Note: All questions in this practice are free response and all answers are marked correct. Type your responses and compare them to the answers provided.

What type of variable is **Moved**?

Moved is a binary variable.

Question 2

How many variables have missing values?

A total of 15 variables have missing values. Missing values are an issue. You learn how to handle missing values later in the course.

Question 3

Look at the percentage of cases that have an **Ins** variable value of 1 versus those with a value of 0. What could you infer about the selection of cases for this data set?

The results of PROC FREQ show that 34.6% of the customers in the **develop** data set purchased the insurance product. You might think that this percentage seems artificially high. In fact, the target event (buying the insurance product) is rare — only 2% of the population. To build the **develop** data set, the bank included all cases that have an **Ins** variable value of 1 and a representative sample of cases that have an **Ins** variable value of 0. This oversampling of the events increases the efficiency of the analysis because you are using a smaller sample and therefore have fewer cases to process. However, this oversampling also biases the results. You learn more about oversampling events, and how to adjust the model for it, later in the course.

Question 4

How many bank branches are represented in the data? Do you think this is a useful number of levels for the analysis?

The **Branch of Bank** table (the frequency table for **Branch**) indicates that the customers represented in the data do their banking in 19 different branches. When you determine that a categorical input variable has too many levels to be useful, you can collapse the levels. You learn to do this later in the course.

Question 5

How many area classifications are represented in the data? Which area has the largest number of customers?

The **Area Classification** table indicates that **Res** has three levels: R (rural), S (suburban), and U (urban). The largest number of customers live in urban areas, followed by suburban areas, and then rural areas.

/* Practice: l1p1.sas step 1 */

```
data pmlr.pva(drop=control_number
              MONTHS_SINCE_LAST_PROM_RESP
              FILE_AVG_GIFT
              FILE_CARD_GIFT);
set pmlr.pva_raw_data;
STATUS_FL=RECENCY_STATUS_96NK in("F","L");
STATUS_ES=RECENCY_STATUS_96NK in("E","S");
home01=(HOME_OWNER="H");
nses1=(SES="1");
nses3=(SES="3");
nses4=(SES="4");
nses_=(SES="?");
nurbr=(URBANICITY="R");
nurbu=(URBANICITY="U");
```

```
nurbs=(URBANICITY="S");
nurbt=(URBANICITY="T");
nurb_=(URBANICITY="?");
```

```
run;
```

Table: PMLR.PVA
 View: Column names
 Filter: (none)

Columns
 Total rows: 19372
 Total columns: 58
 Rows 1-100

	TARGET_B	TARGET_D	MONTHS_SINCE_ORIGIN	DONOR_AGE	IN_HOUSE	URBA
1	0	.	101	87	0	?
2	1	10	137	79	0	R
3	0	.	113	75	0	S
4	0	.	92	.	0	U
5	0	.	101	74	0	R
6	0	.	101	63	0	U
7	0	.	89	71	0	R

```
proc contents data=pmlr.pva;
```

```
run;
```

The CONTENTS Procedure

Data Set Name	PMLR.PVA	Observations	19372
Member Type	DATA	Variables	58
Engine	V9	Indexes	0
Created	09/06/2021 23:14:06	Observation Length	432
Last Modified	09/06/2021 23:14:06	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Engine/Host Dependent Information	
Data Set Page Size	131072
Number of Data Set Pages	65
First Data Page	1
Max Obs per Page	303
Obs in First Data Page	281
Number of Data Set Repairs	0
Filename	/home/u58304328/EPMLR51/data/pva.sas7bdat
Release Created	9.0401M6
Host Created	Linux
Inode Number	10858277167
Access Permission	rw-r--r--
Owner Name	u58304328

File Size	8MB
File Size (bytes)	8650752

Alphabetic List of Variables and Attributes			
#	Variable	Type	Len
42	CARD_PROM_12	Num	8
8	CLUSTER_CODE	Char	2
4	DONOR_AGE	Num	8
10	DONOR_GENDER	Char	3
26	FREQUENCY_STATUS_97NK	Num	8
9	HOME_OWNER	Char	3
11	INCOME_GROUP	Num	8
5	IN_HOUSE	Num	8
41	LAST_GIFT_AMT	Num	8
37	LIFETIME_AVG_GIFT_AMT	Num	8
33	LIFETIME_CARD_PROM	Num	8
35	LIFETIME_GIFT_AMOUNT	Num	8
36	LIFETIME_GIFT_COUNT	Num	8
38	LIFETIME_GIFT_RANGE	Num	8
39	LIFETIME_MAX_GIFT_AMT	Num	8
40	LIFETIME_MIN_GIFT_AMT	Num	8
34	LIFETIME_PROM	Num	8
16	MEDIAN_HOME_VALUE	Num	8
17	MEDIAN_HOUSEHOLD_INCOME	Num	8
45	MONTHS_SINCE_FIRST_GIFT	Num	8

44	MONTHS_SINCE_LAST_GIFT	Num	8
3	MONTHS_SINCE_ORIGIN	Num	8
14	MOR_HIT_RATE	Num	8
43	NUMBER_PROM_12	Num	8
13	OVERLAY_SOURCE	Char	1
19	PCT_MALE_MILITARY	Num	8
20	PCT_MALE_VETERANS	Num	8
18	PCT_OWNER_OCCUPIED	Num	8
21	PCT_VIETNAM_VETERANS	Num	8
22	PCT_WWII_VETERANS	Num	8
23	PEP_STAR	Num	8
46	PER_CAPITA_INCOME	Num	8
12	PUBLISHED_PHONE	Num	8
25	REGENCY_STATUS_96NK	Char	5
30	RECENT_AVG_CARD_GIFT_AMT	Num	8
28	RECENT_AVG_GIFT_AMT	Num	8
32	RECENT_CARD_RESPONSE_COUNT	Num	8
29	RECENT_CARD_RESPONSE_PROP	Num	8
31	RECENT_RESPONSE_COUNT	Num	8
27	RECENT_RESPONSE_PROP	Num	8
24	RECENT_STAR_STATUS	Num	8
7	SES	Char	4
48	STATUS_ES	Num	8
47	STATUS_FL	Num	8
1	TARGET_B	Num	8
2	TARGET_D	Num	8

6	URBANICITY	Char	4
15	WEALTH_RATING	Num	8
49	home01	Num	8
50	nses1	Num	8
51	nses3	Num	8
52	nses4	Num	8
53	nses_	Num	8
58	nurb_	Num	8
54	nurbr	Num	8
56	nurbs	Num	8
57	nurbt	Num	8
55	nurbu	Num	8

```
proc means data=pmlr.pva mean nmiss max min;
```

```
  var _numeric_;
```

```
run;
```

The MEANS Procedure

Variable	Mean	N Miss	Maximum	Minimum
TARGET_B	0.2500000	0	1.0000000	0
TARGET_D	15.6243444	14529	200.0000000	1.0000000
MONTHS_SINCE_ORIGIN	73.4099732	0	137.0000000	5.0000000
DONOR_AGE	58.9190506	4795	87.0000000	0
IN_HOUSE	0.0731984	0	1.0000000	0
INCOME_GROUP	3.9075434	4392	7.0000000	1.0000000
PUBLISHED_PHONE	0.4977287	0	1.0000000	0
MOR_HIT_RATE	3.3616560	0	241.0000000	0
WEALTH_RATING	5.0053967	8810	9.0000000	0
MEDIAN_HOME_VALUE	1079.87	0	6000.00	0
MEDIAN_HOUSEHOLD_INCOME	341.9702147	0	1500.00	0
PCT_OWNER_OCCUPIED	69.6989986	0	99.0000000	0
PCT_MALE_MILITARY	1.0290109	0	97.0000000	0
PCT_MALE_VETERANS	30.5739211	0	99.0000000	0
PCT_VIETNAM_VETERANS	29.6032934	0	99.0000000	0
PCT_WWII_VETERANS	32.8524675	0	99.0000000	0
PEP_STAR	0.5044394	0	1.0000000	0
RECENT_STAR_STATUS	0.9311377	0	22.0000000	0
FREQUENCY_STATUS_97NK	1.9839975	0	4.0000000	1.0000000
RECENT_RESPONSE_PROP	0.1901275	0	1.0000000	0
RECENT_AVG_GIFT_AMT	15.3653959	0	260.0000000	0
RECENT_CARD_RESPONSE_PROP	0.2308077	0	1.0000000	0
RECENT_AVG_CARD_GIFT_AMT	11.6854703	0	300.0000000	0
RECENT_RESPONSE_COUNT	3.0431034	0	16.0000000	0
RECENT_CARD_RESPONSE_COUNT	1.7305389	0	9.0000000	0
LIFETIME_CARD_PROM	18.6680776	0	56.0000000	2.0000000
LIFETIME_PROM	47.5705141	0	194.0000000	5.0000000
LIFETIME_GIFT_AMOUNT	104.4257165	0	3775.00	15.0000000
LIFETIME_GIFT_COUNT	9.9797646	0	95.0000000	1.0000000
LIFETIME_AVG_GIFT_AMT	12.8583383	0	450.0000000	1.3600000
LIFETIME_GIFT_RANGE	11.5878758	0	997.0000000	0
LIFETIME_MAX_GIFT_AMT	19.2088081	0	1000.00	5.0000000
LIFETIME_MIN_GIFT_AMT	7.6209323	0	450.0000000	0
LAST_GIFT_AMT	16.5841988	0	450.0000000	0
CARD_PROM_12	5.3671278	0	17.0000000	0
NUMBER_PROM_12	12.9018687	0	64.0000000	2.0000000
MONTHS_SINCE_LAST_GIFT	18.1911522	0	27.0000000	4.0000000
MONTHS_SINCE_FIRST_GIFT	69.4820875	0	260.0000000	15.0000000
PER_CAPITA_INCOME	15857.33	0	174523.00	0
STATUS_FL	0.0833161	0	1.0000000	0
STATUS_ES	0.2399339	0	1.0000000	0
home01	0.5474912	0	1.0000000	0
nses1	0.3058022	0	1.0000000	0
nses3	0.1715362	0	1.0000000	0
nses4	0.0199773	0	1.0000000	0
nses_	0.0234359	0	1.0000000	0
nurbr	0.2067417	0	1.0000000	0
nurbu	0.1267809	0	1.0000000	0
nurbs	0.2318294	0	1.0000000	0
nurbt	0.2035928	0	1.0000000	0
nurb_	0.0234359	0	1.0000000	0

```
proc freq data=pmlr.pva nlevels;
  tables _character_;
run;
```

The FREQ Procedure

Number of Variable Levels	
Variable	Levels
URBANICITY	6
SES	5
CLUSTER_CODE	54
HOME_OWNER	2
DONOR_GENDER	4
OVERLAY_SOURCE	4
REGENCY_STATUS_96NK	6

URBANICITY	Frequency	Percent	Cumulative Frequency	Cumulative Percent
?	454	2.34	454	2.34
C	4022	20.76	4476	23.11
R	4005	20.67	8481	43.78
S	4491	23.18	12972	66.96
T	3944	20.36	16916	87.32
U	2456	12.68	19372	100.00

SES	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	5924	30.58	5924	30.58
2	9284	47.92	15208	78.51
3	3323	17.15	18531	95.66
4	387	2.00	18918	97.66
?	454	2.34	19372	100.00

CLUSTER_CODE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	454	2.34	454	2.34
01	239	1.23	693	3.58
02	380	1.96	1073	5.54
03	300	1.55	1373	7.09
04	113	0.58	1486	7.67
05	199	1.03	1685	8.70
06	123	0.63	1808	9.33
07	184	0.95	1992	10.28
08	378	1.95	2370	12.23
09	153	0.79	2523	13.02
10	387	2.00	2910	15.02
11	484	2.50	3394	17.52
12	631	3.26	4025	20.78
13	579	2.99	4604	23.77
14	454	2.34	5058	26.11

34	284	1.47	11707	60.43
35	727	3.75	12434	64.19
36	716	3.70	13150	67.88
37	204	1.05	13354	68.93
38	240	1.24	13594	70.17
39	512	2.64	14106	72.82
40	830	4.28	14936	77.10
41	431	2.22	15367	79.33
42	284	1.47	15651	80.79
43	468	2.42	16119	83.21
44	383	1.98	16502	85.18
45	482	2.49	16984	87.67
46	369	1.90	17353	89.58
47	185	0.95	17538	90.53
48	180	0.93	17718	91.46
49	675	3.48	18393	94.95
50	156	0.81	18549	95.75
51	460	2.37	19009	98.13
52	60	0.31	19069	98.44
53	303	1.56	19372	100.00

HOME_OWNER	Frequency	Percent	Cumulative Frequency	Cumulative Percent
H	10606	54.75	10606	54.75
U	8766	45.25	19372	100.00

DONOR_GENDER	Frequency	Percent	Cumulative Frequency	Cumulative Percent
A	1	0.01	1	0.01
F	10401	53.69	10402	53.70
M	7953	41.05	18355	94.75
U	1017	5.25	19372	100.00

OVERLAY_SOURCE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
B	8732	45.08	8732	45.08
M	1480	7.64	10212	52.72
N	4392	22.67	14604	75.39
P	4768	24.61	19372	100.00

REGENCY_STATUS_96NK	Frequency	Percent	Cumulative Frequency	Cumulative Percent
A	11918	61.52	11918	61.52
E	427	2.20	12345	63.73
F	1521	7.85	13866	71.58
L	93	0.48	13959	72.06
N	1192	6.15	15151	78.21
S	4221	21.79	19372	100.00

Practice: Exploring the Veterans' Organization Data Used in the Practices

Question 1

A national veterans' organization wants to better target its solicitations for donation. By soliciting only the most likely donors, the organization can spend less money on solicitation efforts and more money on charitable concerns. Throughout the practices in this course, you'll apply your predictive modeling skills to this project.

The original sample of the veterans' organization's entire solicitations database, **pmlr.pva_raw_data**, needs further modification before it can be used for modeling. Open the PDF below to read a description and list of variables in **pva_raw_data**.

PVA_RAW_DATA Description.pdf PDF File

In this practice, you create and explore the **pmlr.pva** data set.

Reminder: If you started a new SAS session, you must run **setup.sas** to define the **pmlr** library before you do this practice.

Step 1: Open **l1p01.sas** in your SAS software. Notice that the DATA step code drops several of the variables from **pmlr.pva_raw_data** and creates dummy variables for four existing variables. The dummy variables are used in a later practice.

Submit the code and check the log to verify that the **pmlr.pva** data set was created.

How many observations and variables are in **pmlr.pva**?

Note: This question is free response and all answers are marked correct. Type your response and compare it to the answers provided.
The **pmlr.pva** data set has 19372 observations and 58 variables.

Question 2

Step 2: To examine the contents of **pmlr.pva**, write a PROC CONTENTS step, submit it, and review the results.

How many character variables are in the data set? **Note:** Type a numeric value for your answer.
The **pmlr.pva** data set has 7 character variables.

For the solution code, open **l1p1_s.sas** from the **practices/solutions** folder and see Step 2.

Question 3

Step 3: Write a PROC MEANS step that generates the following descriptive statistics for the numeric variables in the **pmlr.pva** data set: mean, number of missing values, maximum value, and minimum value. To specify only the numeric variables in the input data set, use the special SAS name list **_NUMERIC_** in the VAR statement.

Submit the code for this step, view the results, and answer the following questions:

- Is the proportion of events in the sample equal to the proportion of events in the population? (Hint: To find the proportion of events in the population, read the description of **pmlr.pva_raw_data**.)
- What is the average number of months since the last gift to the organization?
- How many numeric variables have missing values?

Note: This is a free response question and all attempts are marked correct. Type your responses and compare them to the answers provided.

- The proportion of events in the sample and the population are different. The proportion of events in the sample is 0.25, the mean of **Target_B**. The proportion of events in the population was reported to be 0.05.
- The average number of months since the last gift is the mean of the variable **Months_Since_Last_Gift**, which is 18.2.
- Four numeric variables have missing values: **Target_D**, **Donor_Age**, **Income_Group**, and **Wealth_Rating**.

For the solution code, open **llp1_s.sas** from the **practices/solutions** folder and see Step 3.

Question 4

Step 4: Write a PROC FREQ step that generates frequency tables for the character variables in the **pmlr.pva** data set. To specify only the character variables in the input data set, add the special SAS name list **_CHARACTER_** to the TABLES statement. To display the Number of Variable Levels table for each variable specified in the TABLES statement, include the NLEVELS option in the PROC FREQ statement.

Submit the code for this step, view the results, and answer the following questions:

- Which character variable has the highest number of levels?
- How many dummy variables would need to be created for the character variable that has the highest number of levels?

Note: This is a free response question and all attempts are marked correct. Type your response and compare your answer to the answer provided.

- **Cluster_Code** has the highest number of levels (54).
- To represent **Cluster_Code**, you would need to create 53 ($54 - 1$) dummy variables.

For the solution code, open **llp1_s.sas** from the **practices/solutions** folder and see Step 4.

Predictive Modeling Challenges

Introduction

In this topic, you learn to do the following:

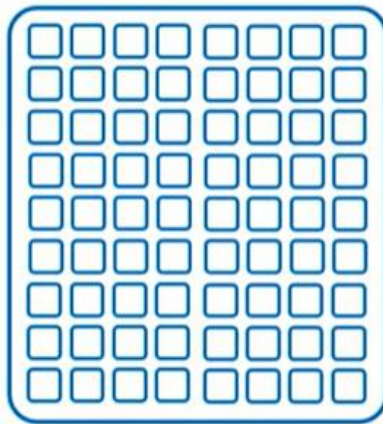
- describe challenges that predictive modelers commonly encounter
- identify solutions to some of these challenges
- define honest assessment
- split the data

Data Challenges



Data Challenges
observational data
mixed measurement scales
high dimensionality
rare target events

interval ordinal
nominal



AcctType



design
variables

Value	Label	D1	D2
1	Checking	1	0
2	Savings	0	1
3	Other	0	0

Data Challenges

observational data

mixed measurement scales

high dimensionality

rare target events

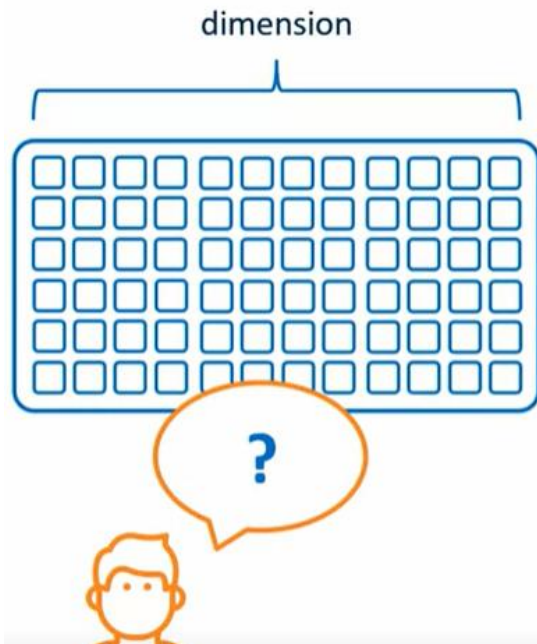
Data Challenges

observational data

mixed measurement scales

high dimensionality

rare target events



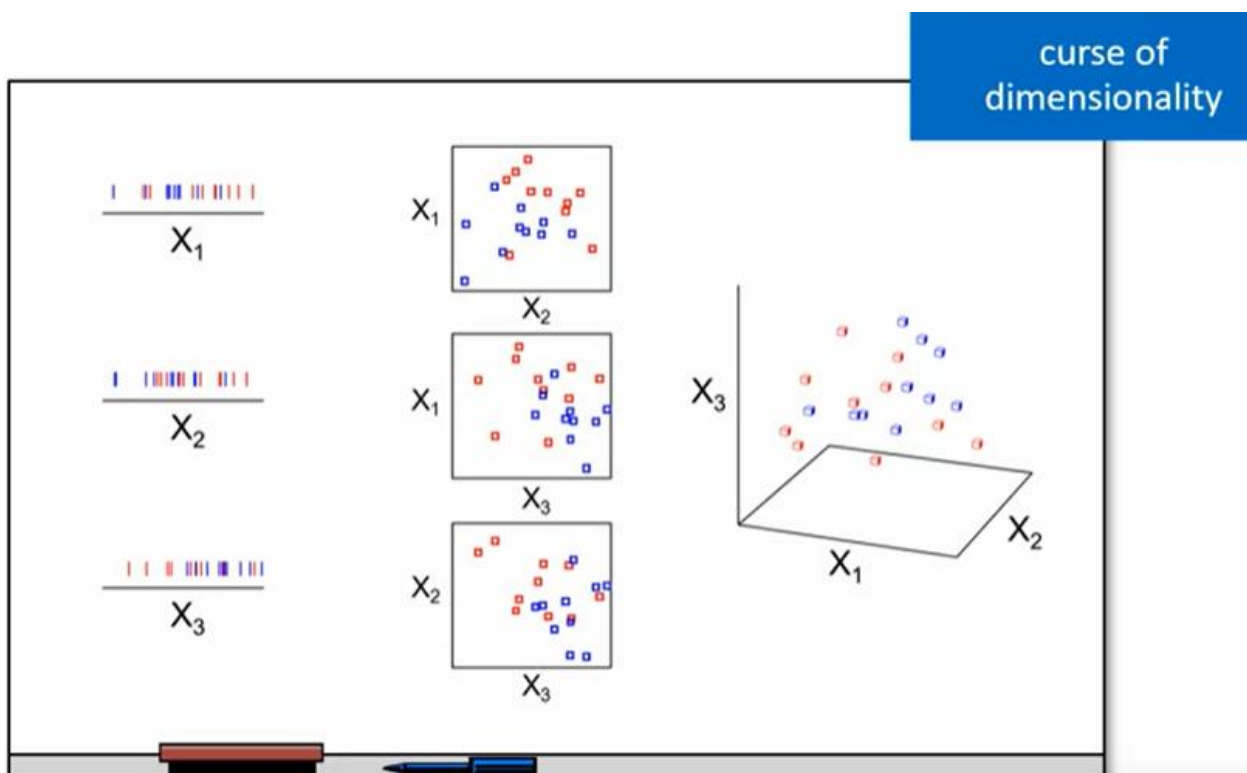
Data Challenges

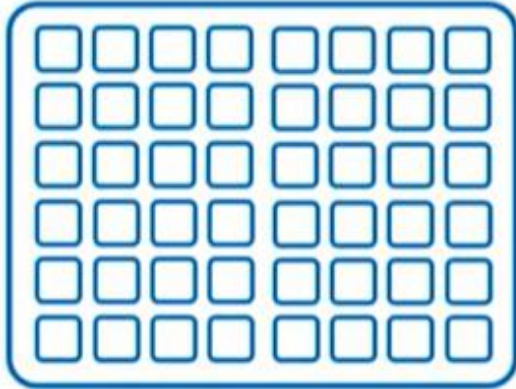
observational data

mixed measurement scales

high dimensionality

rare target events





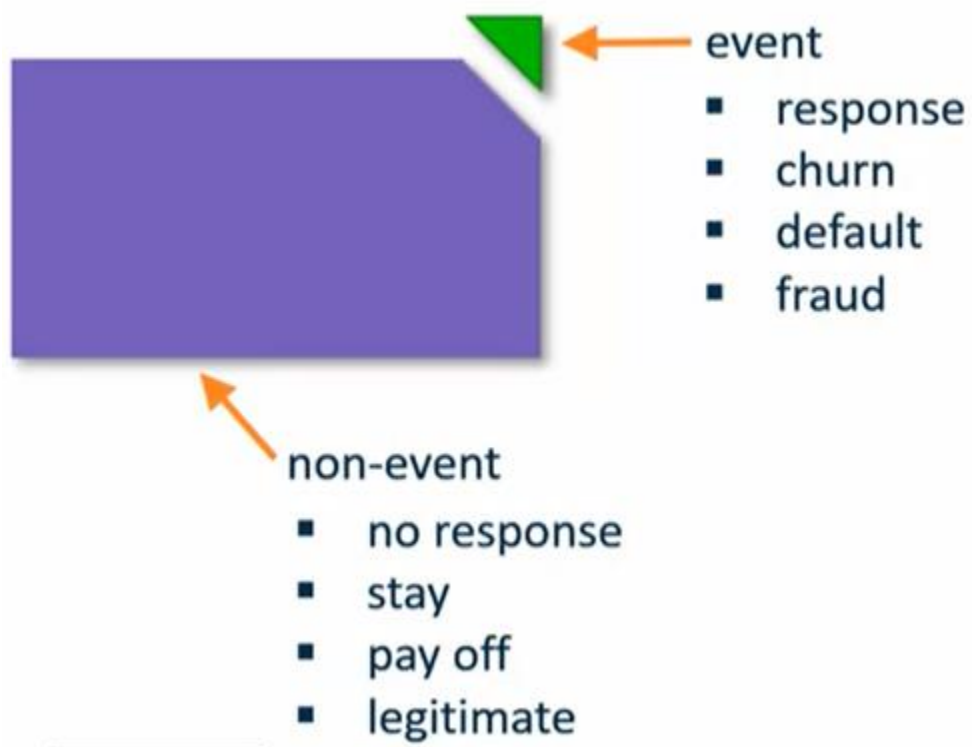
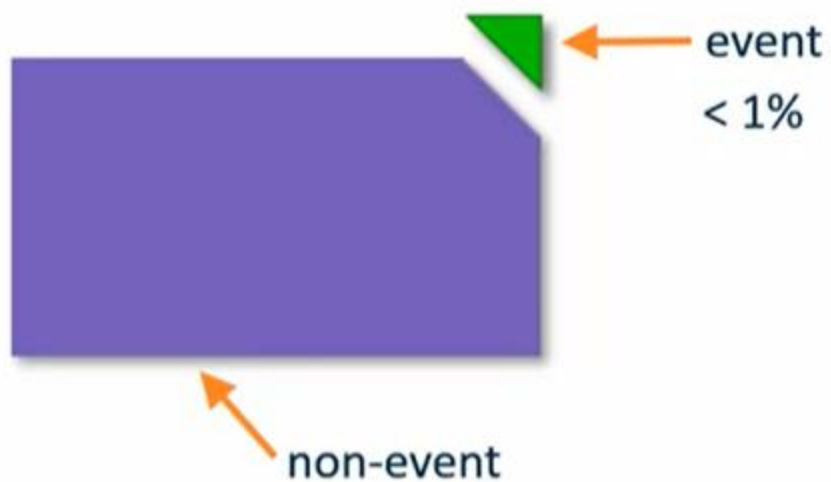
Data Challenges

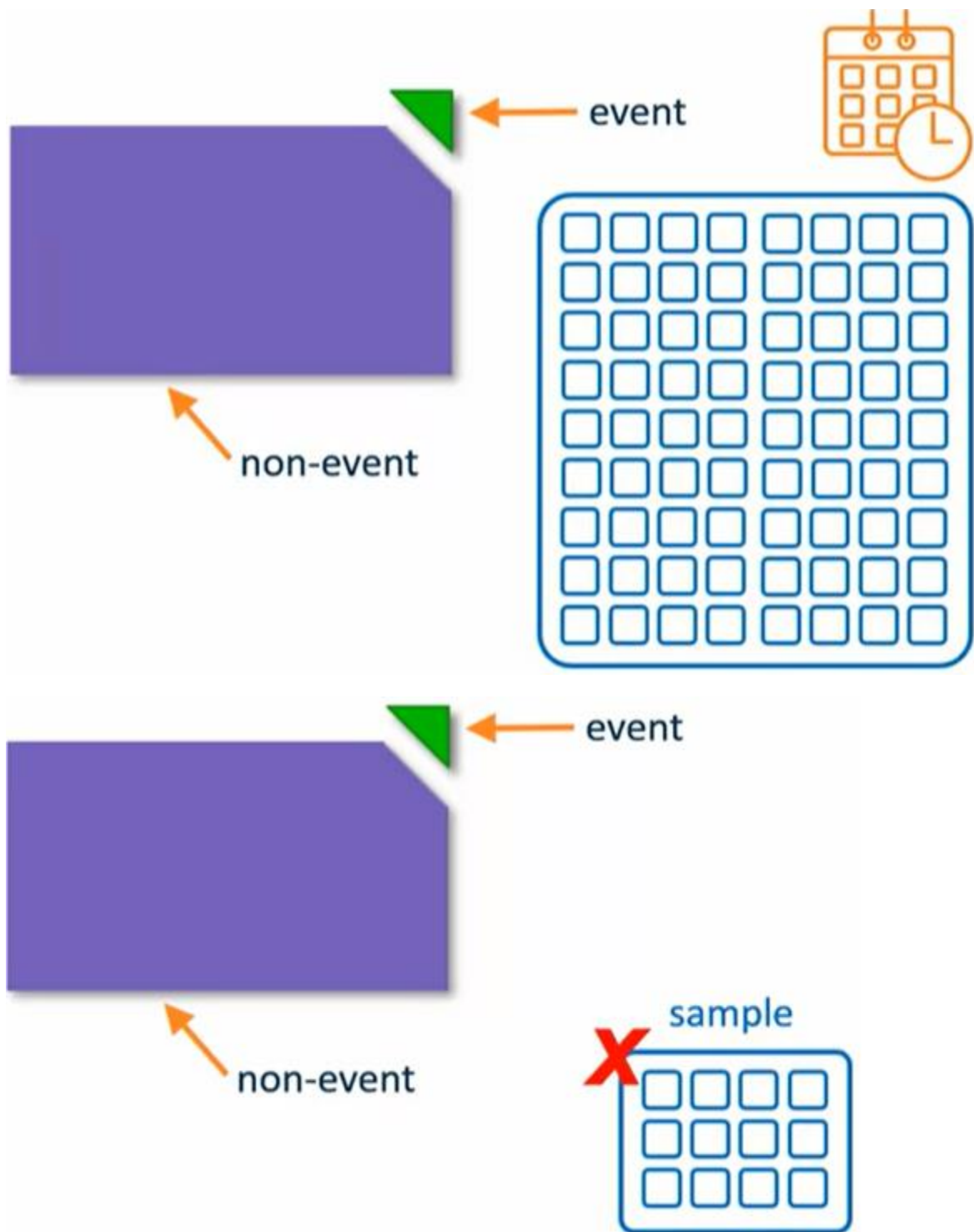
observational data

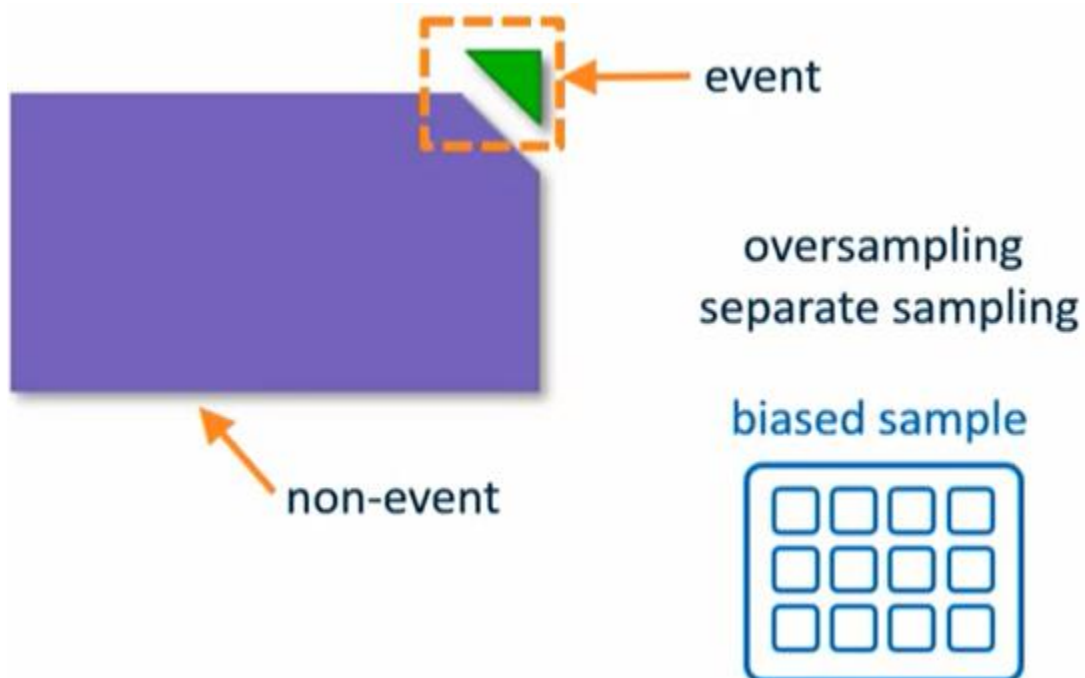
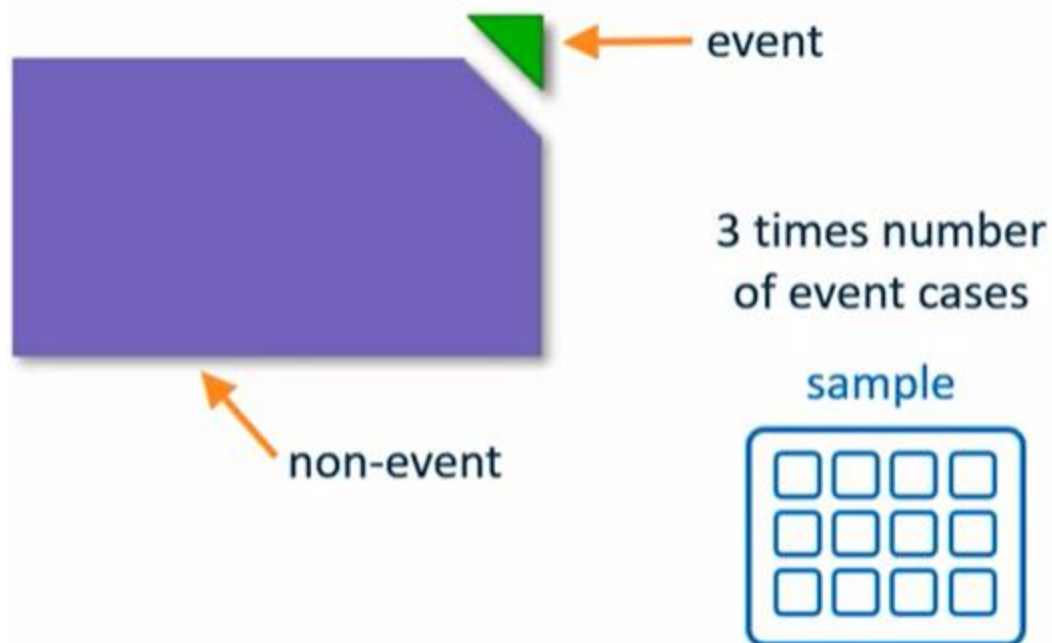
mixed measurement scales

high dimensionality

rare target events









oversampling

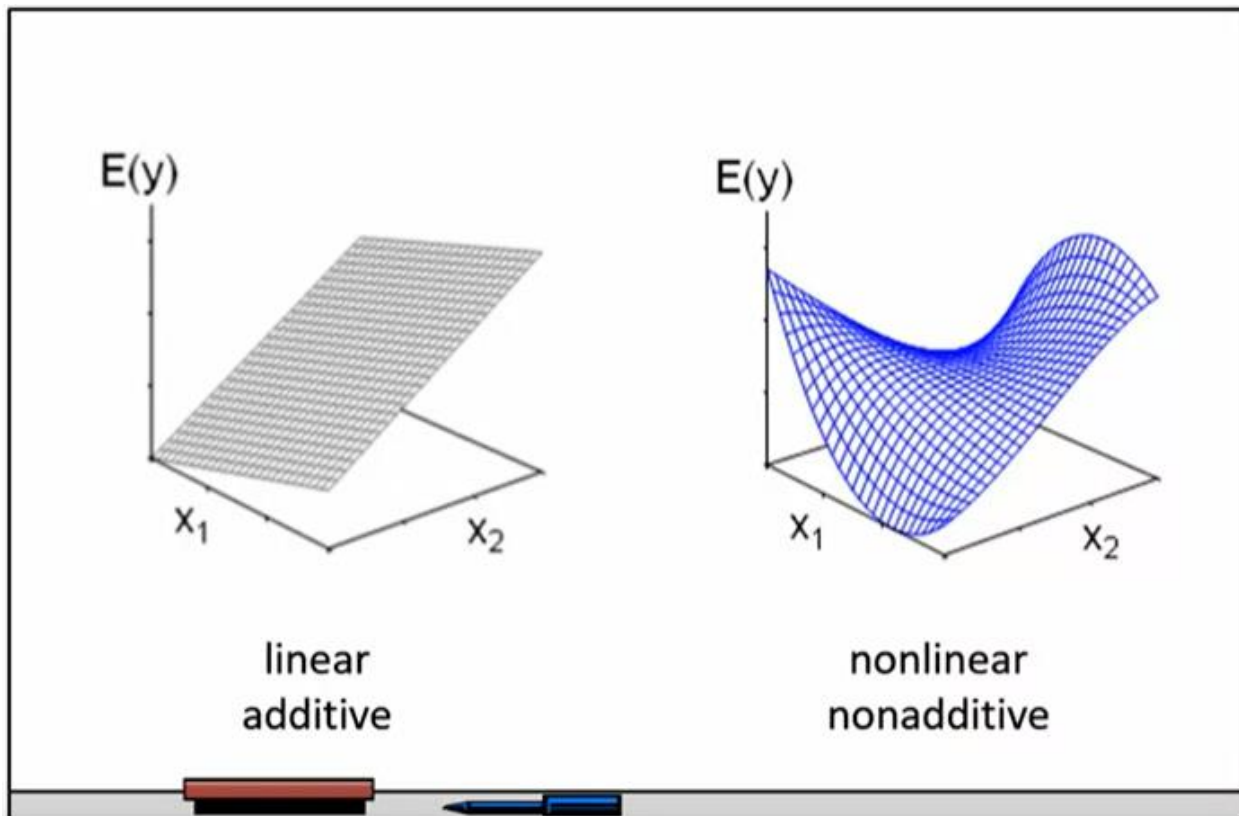


Analytical Challenges

Analytical Challenges

nonlinearities and interactions

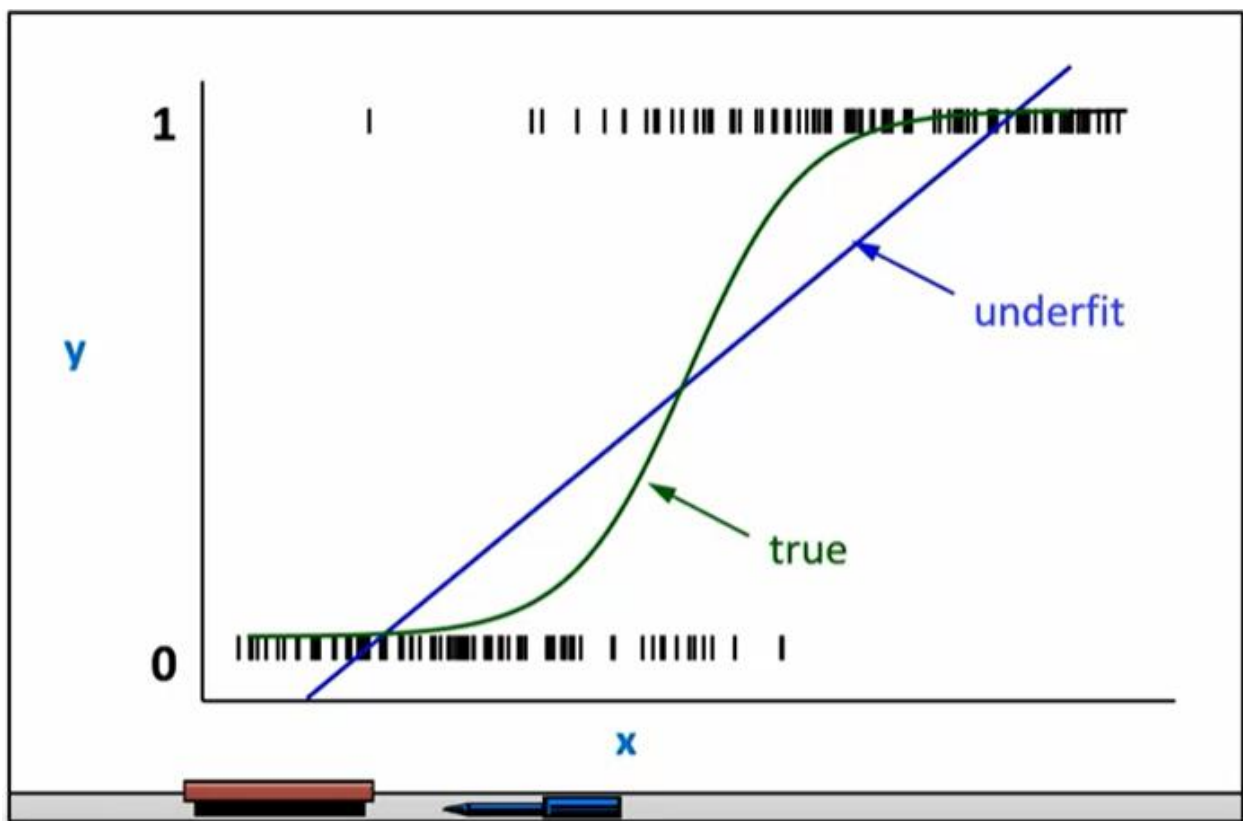
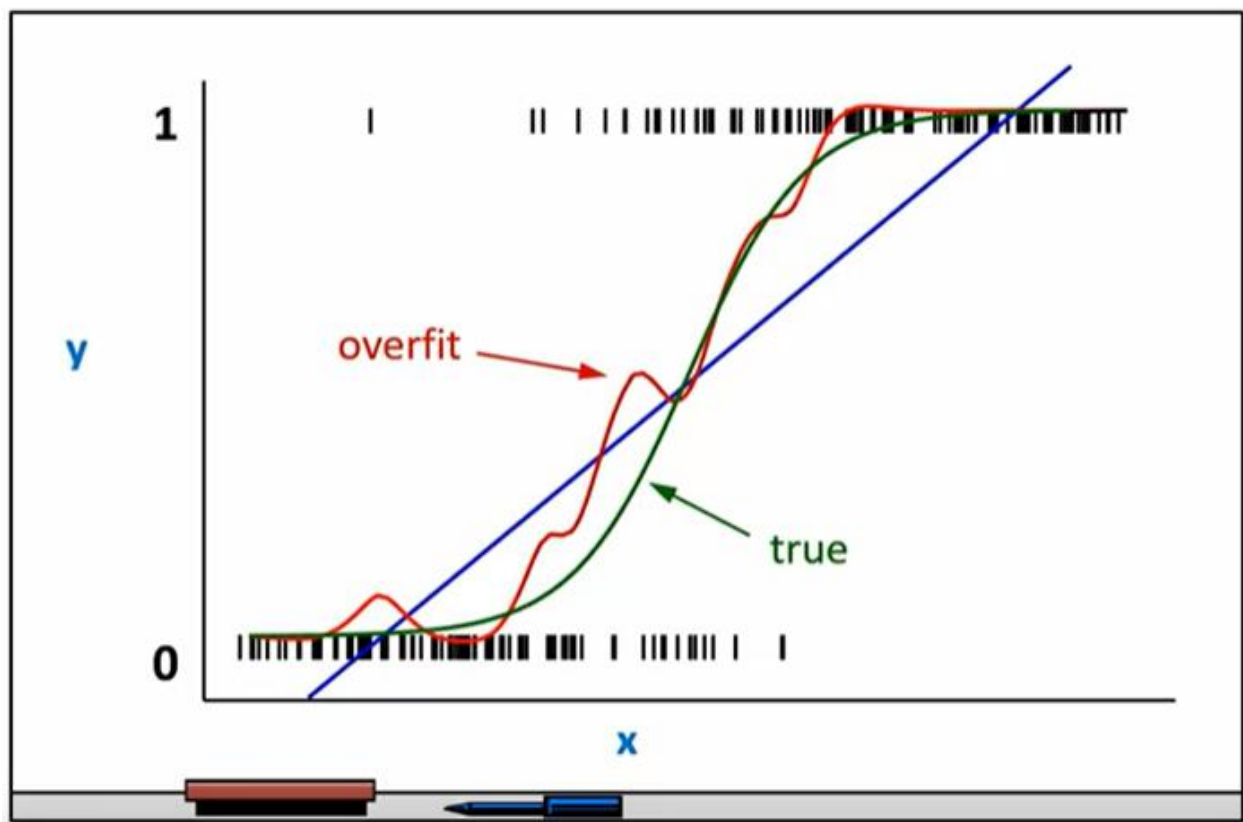
model selection

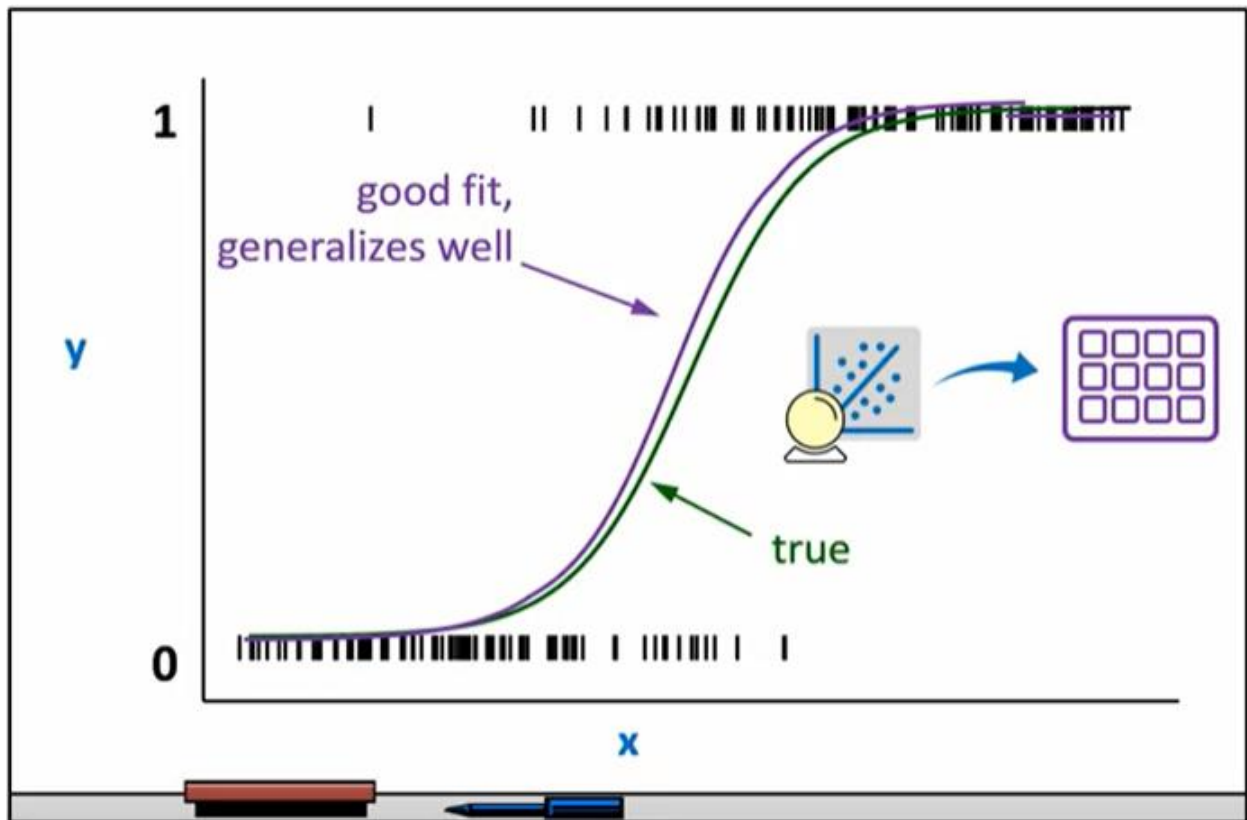


Analytical Challenges

nonlinearities and interactions

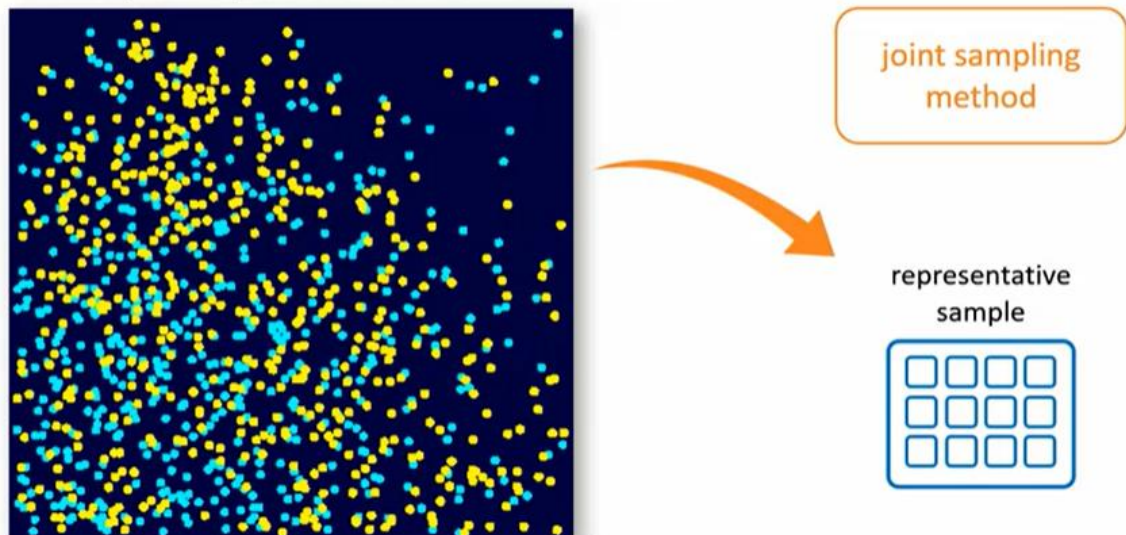
model selection



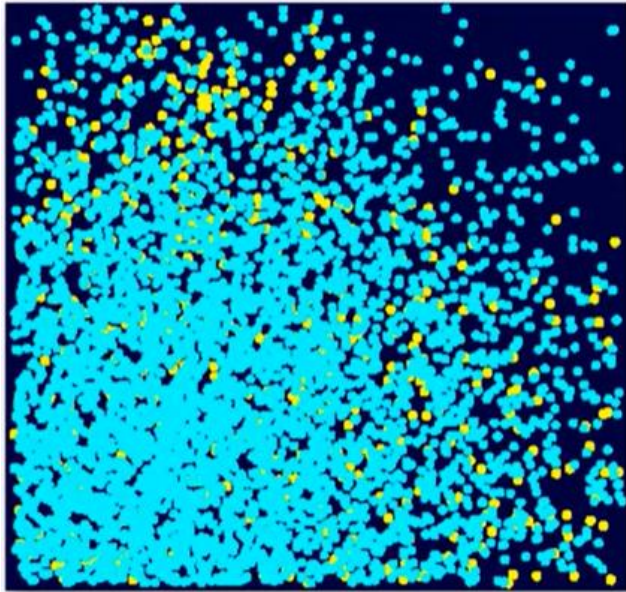


Separate Sampling

Data Set with Equal Proportion of Events to Non-Events



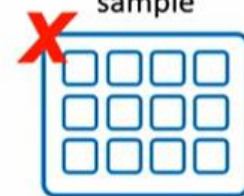
Data Set with Rare Event



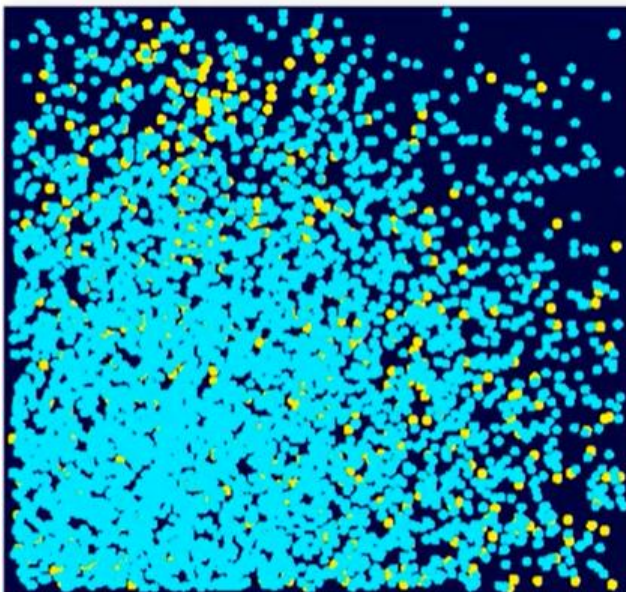
X joint sampling method



representative sample



Data Set with Rare Event



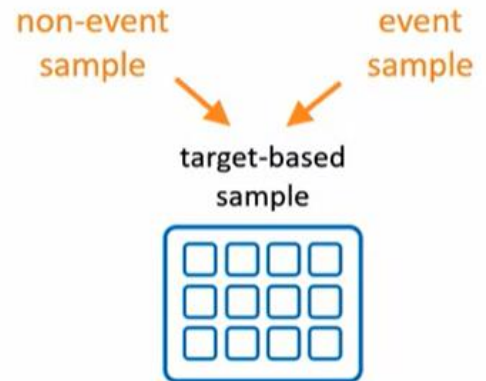
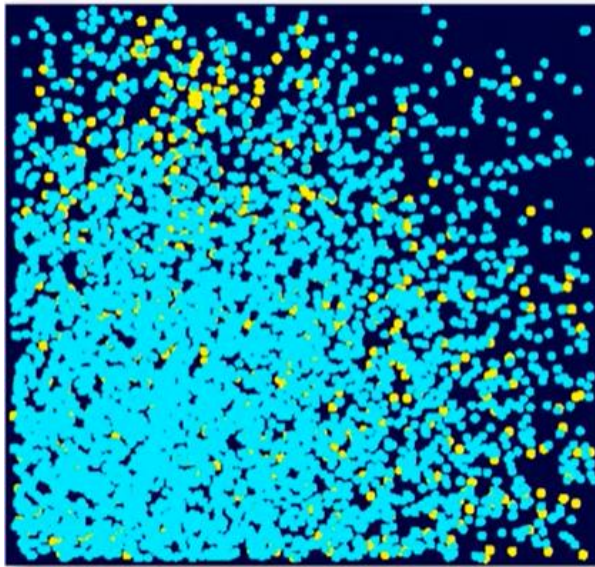
separate sampling



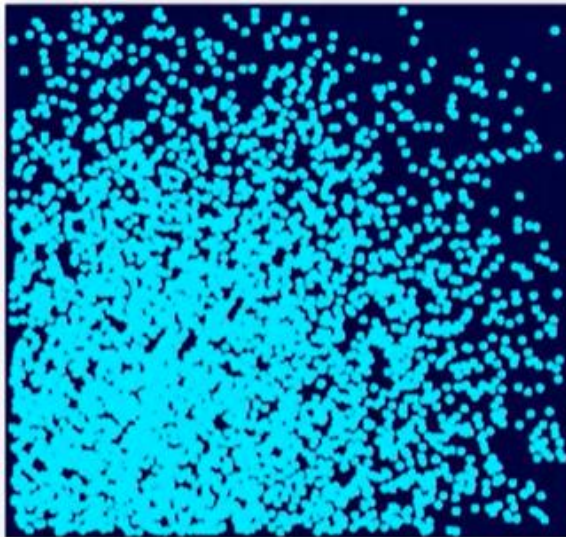
target-based sample



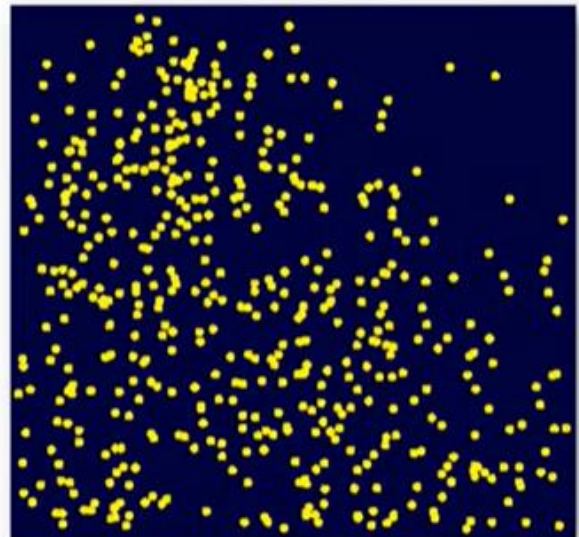
Data Set with Rare Event



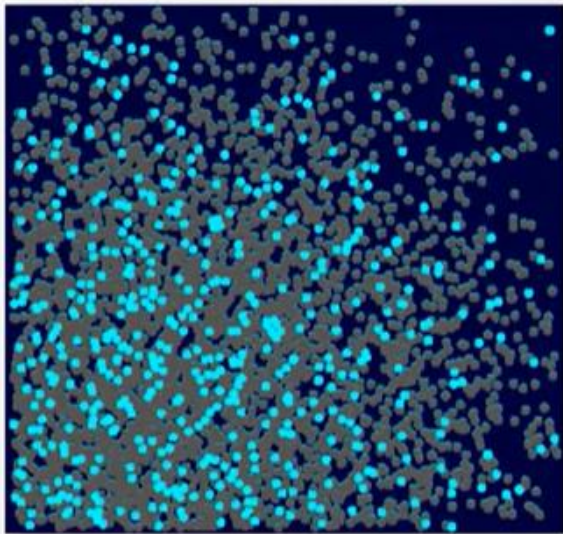
Secondary Outcome (Non-Event)



Primary Outcome (Event)

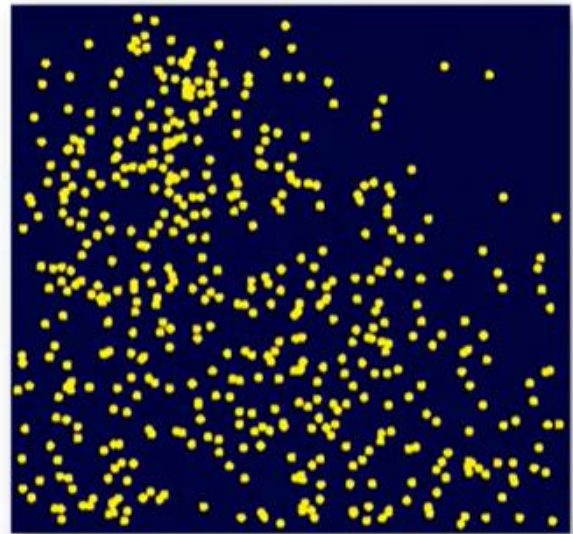


Secondary Outcome (Non-Event)



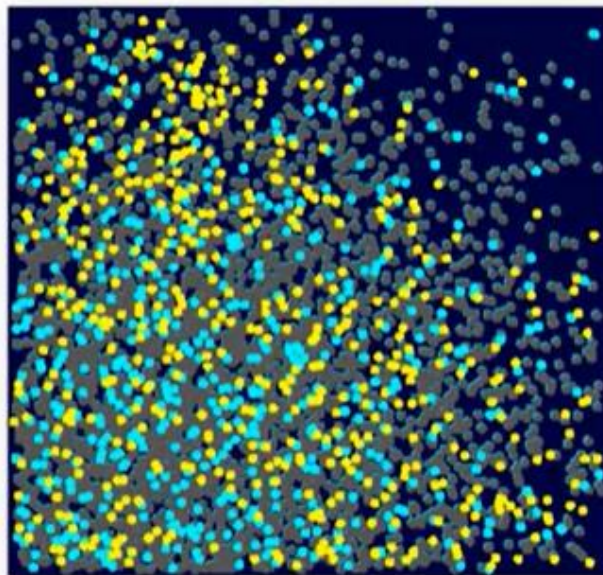
Select some cases.

Primary Outcome (Event)

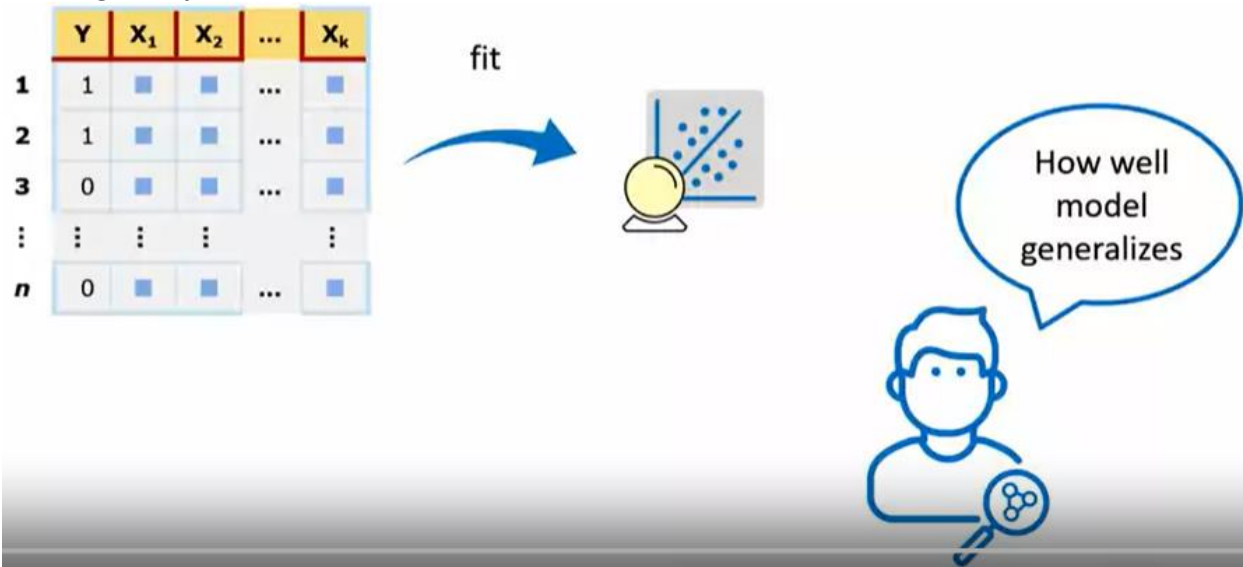


Select all cases.

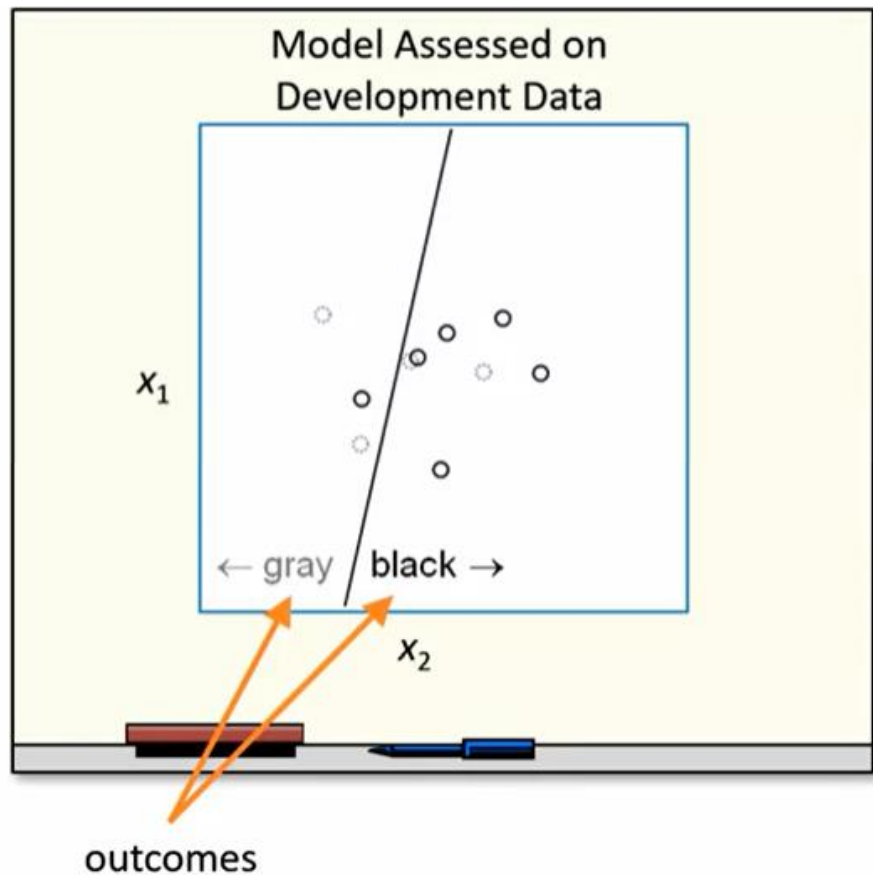
Modeling Sample

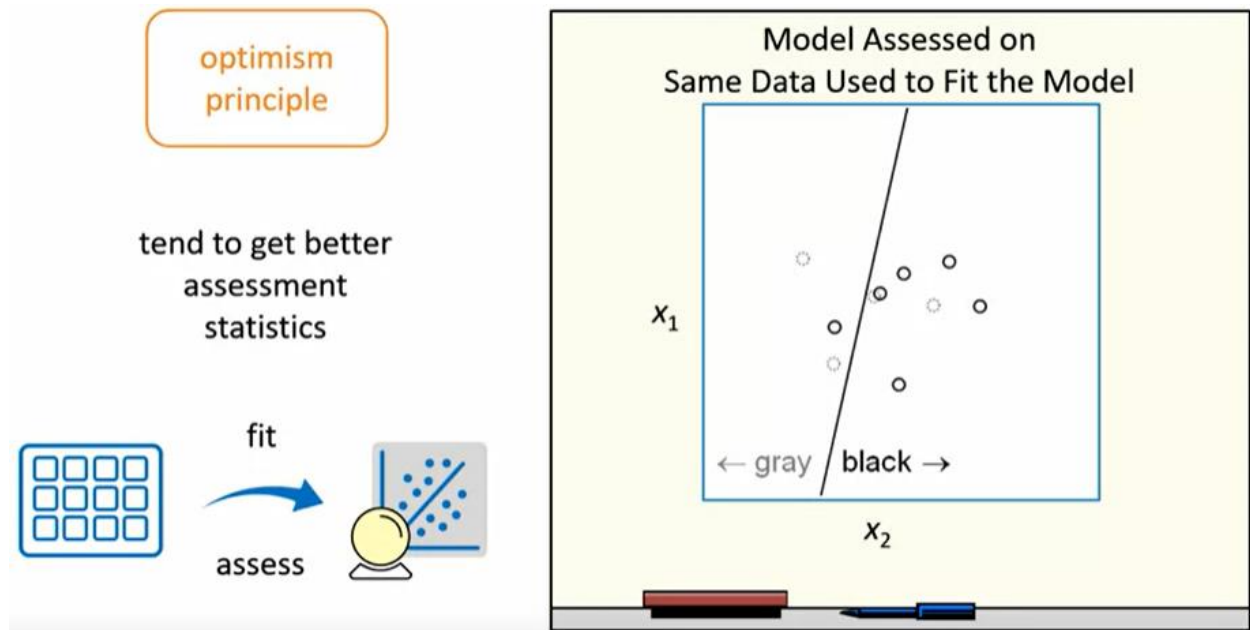
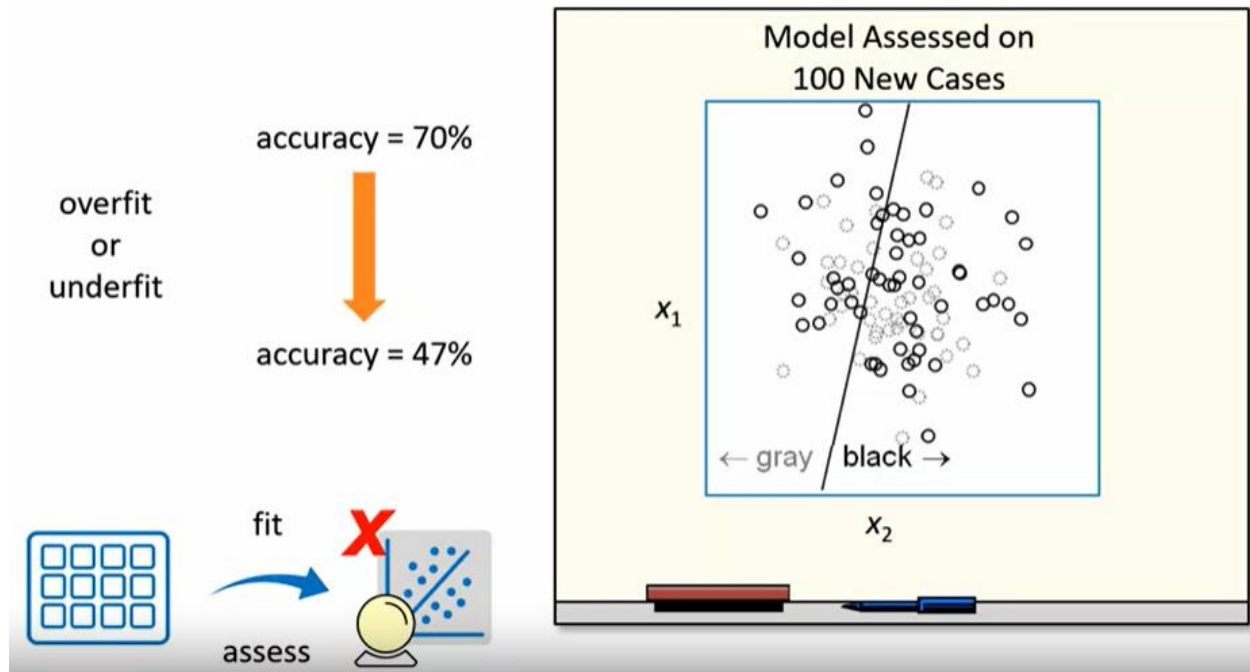


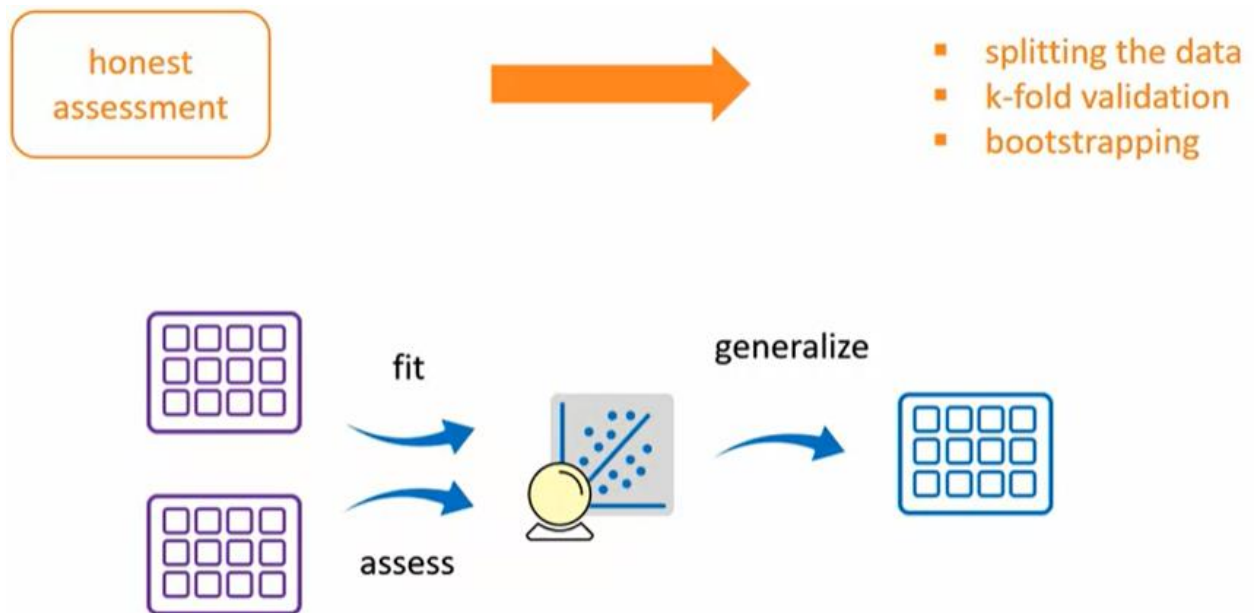
Avoiding the Optimism Bias: Honest Assessment



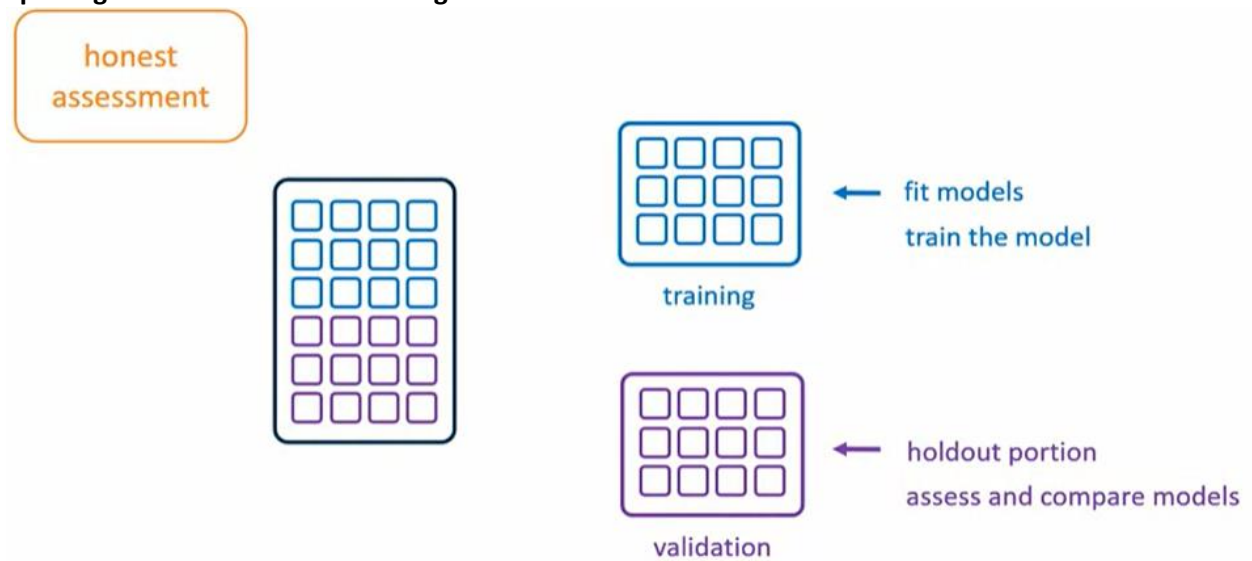
10-case data
set with two
variables

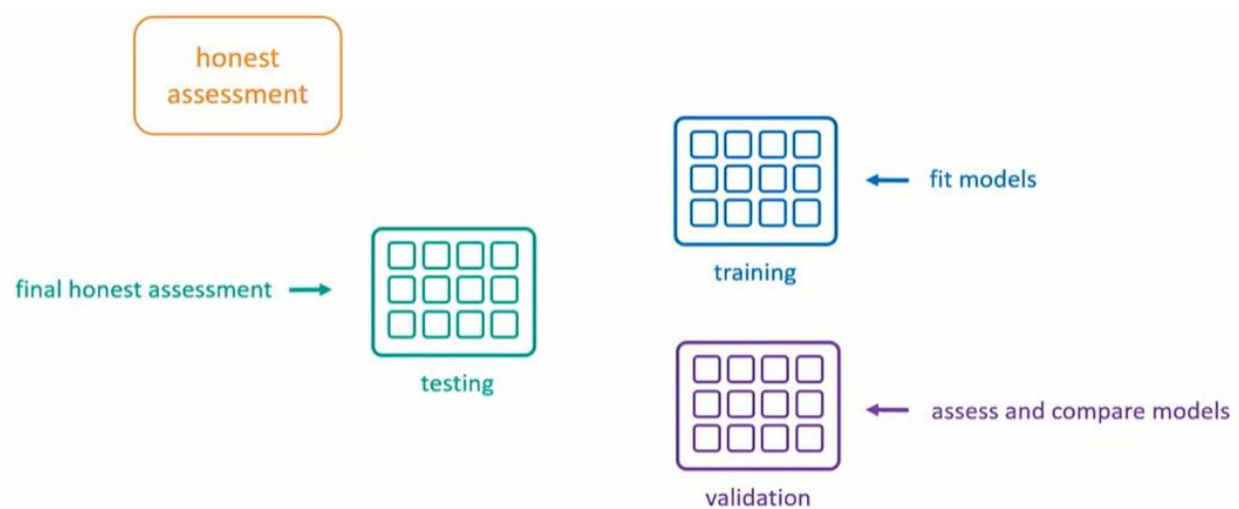








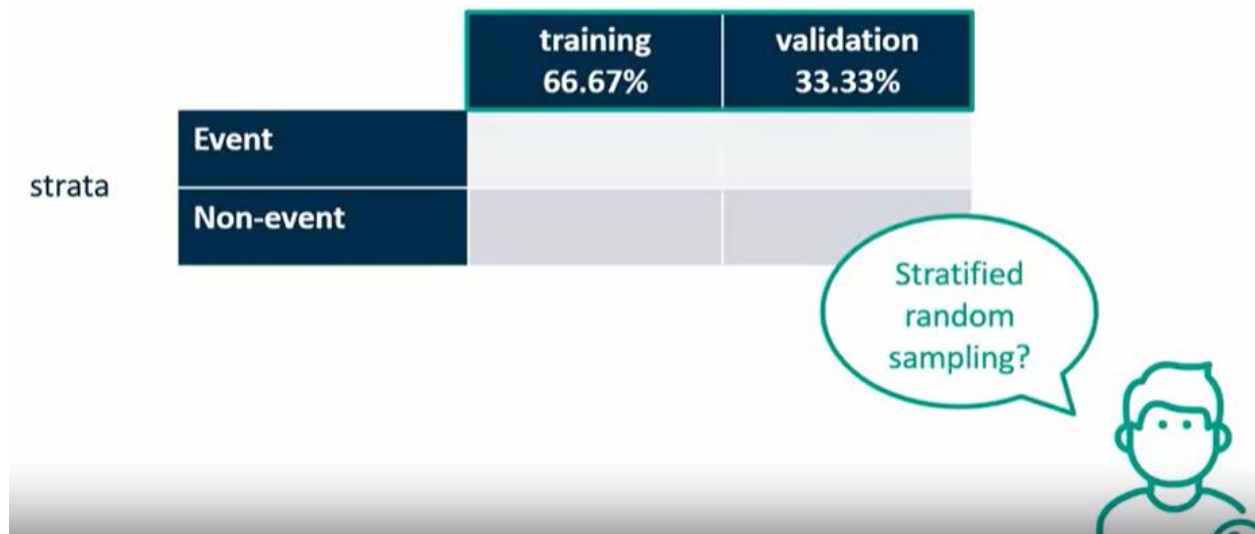
Splitting the Data for Model training and assessment





	training	validation
 <p>Percentage of cases?</p>		$\frac{1}{4}$ to $\frac{1}{2}$ of the data

	training 66.67%	validation 33.33%
 <p>Random sampling? X</p>		

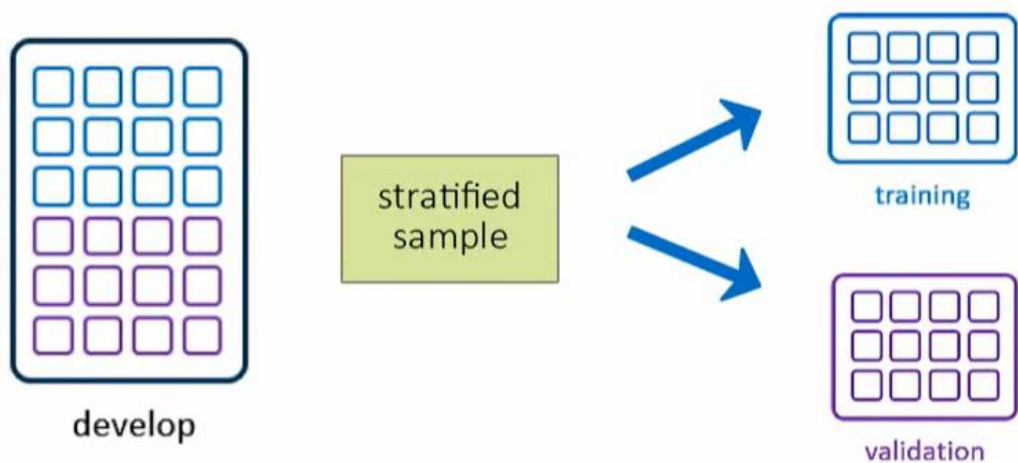


develop



	training 66.67%	validation 33.33%	
Event	7,451 (35%)	3,724 (35%)	11,175 (35%)
Non-event	14,061 (65%)	7,028 (65%)	21,089 (65%)
	21,512 (100%)	10,752 (100%)	32,264 (100%)

Demo Splitting the Data



```
/* ===== */
```

```
/* Lesson 1, Section 2: l1d2.sas
```


Demonstration: Splitting the Data

```
/* ===== */

/* Sort the data by the target in preparation for stratified sampling. */

proc sort data=work.develop out=work.develop_sort;
    by ins;
run;

/* The SURVEYSELECT procedure will perform stratified sampling
on any variable in the STRATA statement. The OUTALL option
specifies that you want a flag appended to the file to
indicate selected records, not simply a file comprised
of the selected records. */

proc surveyselect noprint data=work.develop_sort
    samprate=.6667 stratumseed=restore
    out=work.develop_sample
    seed=44444 outall;
    strata ins;
run;

/* Verify stratification. */

proc freq data=work.develop_sample;
    tables ins*selected;
run;
```

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of Ins by Selected			
	Ins	Selected(Selection Indicator)		
		0	1	Total
	0	7028	14061	21089
		21.78	43.58	65.36
		33.33	66.67	
		65.36	65.36	
	1	3724	7451	11175
		11.54	23.09	34.64
		33.32	66.68	
		34.64	34.64	
	Total	10752	21512	32264
		33.33	66.67	100.00

/* Create training and validation data sets. */

data work.train(drop=selected SelectionProb SamplingWeight)

work.valid(drop=selected SelectionProb SamplingWeight);

set work.develop_sample;

if selected then output work.train;

else output work.valid;

run;

NOTE: There were 32264 observations read from the data set WORK.DEVELOP_SAMPLE.

NOTE: The data set WORK.TRAIN has 21512 observations and 48 variables.

NOTE: The data set WORK.VALID has 10752 observations and 48 variables.

Question 1.03

What would happen if you split the data by taking a simple random sample in PROC SURVEYSELECT?

Assume that, as in the previous demonstration, you split the data into two data sets (a training data set and a validation data set) and specify a sampling rate of 0.6667.

The proportion of the events in the training data set would probably be different from the proportion of events in the validation data set.

Unlike a stratified random sample, a simple random sample does not guarantee an equal percentage of events in the training and validation data sets. However, because the sampling rate is the same as in the demonstration

(0.6667), the training data set (SELECTED=1) will contain 66.67 percent of the observations regardless of the sampling method.

```
/* Run this code before doing practice l1p2 */
```

```
/* ===== */
```

```
/* Lesson 1, Practice 1
```

```
Practice: Exploring the Veterans' Organization Data
```

```
Used in the Practices */
```

```
/* ===== */
```

```
data pmlr.pva(drop=control_number  
MONTHS_SINCE_LAST_PROM_RESP  
FILE_AVG_GIFT  
FILE_CARD_GIFT);  
  
set pmlr.pva_raw_data;  
  
STATUS_FL=RECENCY_STATUS_96NK in("F","L");  
  
STATUS_ES=RECENCY_STATUS_96NK in("E","S");  
  
home01=(HOME_OWNER="H");  
  
nses1=(SES="1");  
  
nses3=(SES="3");  
  
nses4=(SES="4");  
  
nses_=(SES="?");  
  
nurbr=(URBANICITY="R");  
  
nurbu=(URBANICITY="U");  
  
nurbs=(URBANICITY="S");  
  
nurbt=(URBANICITY="T");  
  
nurb_=(URBANICITY="?");  
  
run;
```

```
proc contents data=pmlr.pva;
```

```
run;
```

```
proc means data=pmlr.pva mean nmiss max min;
```

```
var _numeric_;
```

```
run;
```

```
proc freq data=pmlr.pva nlevels;
```

```
tables _character_;
```

```
run;
```

The CONTENTS Procedure

Data Set Name	PMLR.PVA	Observations	19372
Member Type	DATA	Variables	58
Engine	V9	Indexes	0
Created	09/11/2021 22:36:15	Observation Length	432
Last Modified	09/11/2021 22:36:15	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Engine/Host Dependent Information	
Data Set Page Size	131072
Number of Data Set Pages	65
First Data Page	1
Max Obs per Page	303
Obs in First Data Page	281
Number of Data Set Repairs	0
Filename	/home/u58304328/EPMLR51/data/pva.sas7bdat
Release Created	9.0401M6
Host Created	Linux
Inode Number	10858262263
Access Permission	rw-r--r--
Owner Name	u58304328
File Size	8MB
File Size (bytes)	8650752

Alphabetic List of Variables and Attributes			
#	Variable	Type	Len
42	CARD_PROM_12	Num	8
8	CLUSTER_CODE	Char	2
4	DONOR_AGE	Num	8
10	DONOR_GENDER	Char	3
26	FREQUENCY_STATUS_97NK	Num	8
9	HOME_OWNER	Char	3
11	INCOME_GROUP	Num	8
5	IN_HOUSE	Num	8
41	LAST_GIFT_AMT	Num	8
37	LIFETIME_AVG_GIFT_AMT	Num	8
33	LIFETIME_CARD_PROM	Num	8
35	LIFETIME_GIFT_AMOUNT	Num	8
36	LIFETIME_GIFT_COUNT	Num	8
38	LIFETIME_GIFT_RANGE	Num	8
39	LIFETIME_MAX_GIFT_AMT	Num	8
40	LIFETIME_MIN_GIFT_AMT	Num	8
34	LIFETIME_PROM	Num	8
16	MEDIAN_HOME_VALUE	Num	8
17	MEDIAN_HOUSEHOLD_INCOME	Num	8
45	MONTHS_SINCE_FIRST_GIFT	Num	8
44	MONTHS_SINCE_LAST_GIFT	Num	8
3	MONTHS_SINCE_ORIGIN	Num	8
14	MOR_HIT_RATE	Num	8
43	NUMBER_PROM_12	Num	8
13	OVERLAY_SOURCE	Char	1
19	PCT_MALE_MILITARY	Num	8

The MEANS Procedure

Variable	Mean	N Miss	Maximum	Minimum
TARGET_B	0.2500000	0	1.0000000	0
TARGET_D	15.6243444	14529	200.0000000	1.0000000
MONTHS_SINCE_ORIGIN	73.4099732	0	137.0000000	5.0000000
DONOR_AGE	58.9190506	4795	87.0000000	0
IN_HOUSE	0.0731984	0	1.0000000	0
INCOME_GROUP	3.9075434	4392	7.0000000	1.0000000
PUBLISHED_PHONE	0.4977287	0	1.0000000	0
MOR_HIT_RATE	3.3616560	0	241.0000000	0
WEALTH_RATING	5.0053967	8810	9.0000000	0
MEDIAN_HOME_VALUE	1079.87	0	6000.00	0
MEDIAN_HOUSEHOLD_INCOME	341.9702147	0	1500.00	0
PCT_OWNER_OCCUPIED	69.6989986	0	99.0000000	0
PCT_MALE_MILITARY	1.0290109	0	97.0000000	0
PCT_MALE_VETERANS	30.5739211	0	99.0000000	0
PCT_VIETNAM_VETERANS	29.6032934	0	99.0000000	0
PCT_WWII_VETERANS	32.8524675	0	99.0000000	0
PEP_STAR	0.5044394	0	1.0000000	0
RECENT_STAR_STATUS	0.9311377	0	22.0000000	0
FREQUENCY_STATUS_97NK	1.9839975	0	4.0000000	1.0000000
RECENT_RESPONSE_PROP	0.1901275	0	1.0000000	0
RECENT_AVG_GIFT_AMT	15.3653959	0	260.0000000	0
RECENT_CARD_RESPONSE_PROP	0.2308077	0	1.0000000	0
RECENT_AVG_CARD_GIFT_AMT	11.6854703	0	300.0000000	0
RECENT_RESPONSE_COUNT	3.0431034	0	16.0000000	0
RECENT_CARD_RESPONSE_COUNT	1.7305389	0	9.0000000	0
LIFETIME_CARD_PROM	18.6680776	0	56.0000000	2.0000000
LIFETIME_PROM	47.5705141	0	194.0000000	5.0000000
LIFETIME_GIFT_AMOUNT	104.4257165	0	3775.00	15.0000000
LIFETIME_GIFT_COUNT	9.9797646	0	95.0000000	1.0000000
LIFETIME_AVG_GIFT_AMT	12.8583383	0	450.0000000	1.3600000
LIFETIME_GIFT_RANGE	11.5878758	0	997.0000000	0
LIFETIME_MAX_GIFT_AMT	19.2088081	0	1000.00	5.0000000
LIFETIME_MIN_GIFT_AMT	7.6209323	0	450.0000000	0
LAST_GIFT_AMT	16.5841988	0	450.0000000	0
CARD_PROM_12	5.3671278	0	17.0000000	0
NUMBER_PROM_12	12.9018687	0	64.0000000	2.0000000
MONTHS_SINCE_LAST_GIFT	18.1911522	0	27.0000000	4.0000000
MONTHS_SINCE_FIRST_GIFT	69.4820875	0	260.0000000	15.0000000
PER_CAPITA_INCOME	15857.33	0	174523.00	0
STATUS_FL	0.0833161	0	1.0000000	0

The FREQ Procedure

Number of Variable Levels	
Variable	Levels
URBANICITY	6
SES	5
CLUSTER_CODE	54
HOME_OWNER	2
DONOR_GENDER	4
OVERLAY_SOURCE	4
RECENCY_STATUS_96NK	6

URBANICITY	Frequency	Percent	Cumulative Frequency	Cumulative Percent
?	454	2.34	454	2.34
C	4022	20.76	4476	23.11
R	4005	20.67	8481	43.78
S	4491	23.18	12972	66.96
T	3944	20.36	16916	87.32
U	2456	12.68	19372	100.00

SES	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	5924	30.58	5924	30.58
2	9284	47.92	15208	78.51
3	3323	17.15	18531	95.66
4	387	2.00	18918	97.66
?	454	2.34	19372	100.00

CLUSTER_CODE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	454	2.34	454	2.34
01	239	1.23	693	3.58
02	380	1.96	1073	5.54
03	300	1.55	1373	7.09
04	113	0.58	1486	7.67
05	199	1.03	1685	8.70
06	123	0.63	1808	9.33
07	184	0.95	1992	10.28
08	378	1.95	2370	12.23
09	153	0.79	2523	13.02
10	387	2.00	2910	15.02
11	484	2.50	3394	17.52
12	631	3.26	4025	20.78
13	579	2.99	4604	23.77
14	454	2.34	5058	26.11
15	223	1.15	5281	27.26
16	384	1.98	5665	29.24
17	349	1.80	6014	31.04
18	619	3.20	6633	34.24
19	98	0.51	6731	34.75
20	317	1.64	7048	36.38
21	353	1.82	7401	38.20
22	251	1.30	7652	39.50
23	293	1.51	7945	41.01
24	795	4.10	8740	45.12
25	273	1.41	9013	46.53

26	202	1.04	9215	47.57
27	666	3.44	9881	51.01
28	343	1.77	10224	52.78
29	170	0.88	10394	53.65
30	519	2.68	10913	56.33
31	249	1.29	11162	57.62
32	152	0.78	11314	58.40
33	109	0.56	11423	58.97
34	284	1.47	11707	60.43
35	727	3.75	12434	64.19
36	716	3.70	13150	67.88
37	204	1.05	13354	68.93
38	240	1.24	13594	70.17
39	512	2.64	14106	72.82
40	830	4.28	14936	77.10
41	431	2.22	15367	79.33
42	284	1.47	15651	80.79
43	468	2.42	16119	83.21
44	383	1.98	16502	85.18
45	482	2.49	16984	87.67
46	369	1.90	17353	89.58
47	185	0.95	17538	90.53
48	180	0.93	17718	91.46
49	675	3.48	18393	94.95
50	156	0.81	18549	95.75
51	460	2.37	19009	98.13
52	60	0.31	19069	98.44
53	303	1.56	19372	100.00

HOME_OWNER	Frequency	Percent	Cumulative Frequency	Cumulative Percent
H	10606	54.75	10606	54.75
U	8766	45.25	19372	100.00

DONOR_GENDER	Frequency	Percent	Cumulative Frequency	Cumulative Percent
A	1	0.01	1	0.01
F	10401	53.69	10402	53.70
M	7953	41.05	18355	94.75
U	1017	5.25	19372	100.00

OVERLAY_SOURCE	Frequency	Percent	Cumulative Frequency	Cumulative Percent
B	8732	45.08	8732	45.08
M	1480	7.64	10212	52.72
N	4392	22.67	14604	75.39
P	4768	24.61	19372	100.00

REGENCY_STATUS_96NK	Frequency	Percent	Cumulative Frequency	Cumulative Percent
A	11918	61.52	11918	61.52
E	427	2.20	12345	63.73
F	1521	7.85	13866	71.58
L	93	0.48	13959	72.06
N	1192	6.15	15151	78.21
S	4221	21.79	19372	100.00

```
/* Solution for l1p2 */
```

```
/* step 2*/
```

```
proc sort data=pmlr.pva out=work.pva_sort;
```

```
  by target_b;
```

```
run;
```

```
proc surveyselect noprint data=work.pva_sort
```

```
  samprate=0.5 out=pva_sample seed=27513
```

```
  outall stratumseed=restore;
```

```
  strata target_b;
```

```
run;
```

```
data pmlr.pva_train(drop=selected SelectionProb SamplingWeight)
```

```
  pmlr.pva_valid(drop=selected SelectionProb SamplingWeight);
```

```
  set work.pva_sample;
```

```
  if selected then output pmlr.pva_train;
```

```
  else output pmlr.pva_valid;
```

```
run;
```

NOTE: There were 19372 observations read from the data set WORK.PVA_SAMPLE.

NOTE: The data set PMLR.PVA_TRAIN has 9687 observations and 58 variables.

NOTE: The data set PMLR.PVA_VALID has 9685 observations and 58 variables.

Practice: Splitting the Data

Question 1

For the veterans' organization project, split the **pmlr.pva** data set into training and validation data sets.

Reminder: If you started a new SAS session, you must run **setup.sas** to define the **pmlr** library before you do this practice.

Step 1: Open **l1p02_runFirst.sas** from the **practices** folder and run the code. You can add to this program or open a new editor to continue the practice.

Step 2: Use PROC SURVEYSELECT to split the **pmlr.pva** data set into two data sets: a training data set named **pmlr.pva_train** and a validation data set named **pmlr.pva_valid**. Use 50% of the data for each data set role. Stratify on the target variable, use a seed of 27513, and use the STRATUMSEED=RESTORE option. Submit the code and check the log.

How many observations are in **pmlr.pva_train**?
9687

Correct

The log indicates that the **pmlr.pva_train** data set has 9687 observations and the **pmlr.pva_valid** data set has 9685 observations. The training and validation data sets do not have the same number of observations because the original data set (**pmlr.pva**) has an odd number of events and non-events. The process of stratified random sampling resulted in one extra event and non-event in the training data set.

For the solution code, open **l1p2_s.sas** from the **practices/solutions** folder and see Step 2.

Question 2

How many observations are in **pmlr.pva_valid**?
9685

Correct

The log indicates that the **pmlr.pva_train** data set has 9687 observations and the **pmlr.pva_valid** data set has 9685 observations. The training and validation data sets do not have the same number of observations because the original data set (**pmlr.pva**) has an odd number of events and non-events. The process of stratified random sampling resulted in one extra event and non-event in the training data set.

For the solution code, open **l1p2_s.sas** from the **practices/solutions** folder and see Step 2.