Introduction to Big Data Science

03rd Period

A Scenario of Business Analysis with Data Science Process

Contents

- A Scenario of Business Analysis With Data Science Process
- Provide a Scenario of Business Analysis
- Apply an Entire Data Science Process

Recap: CRISP-DM

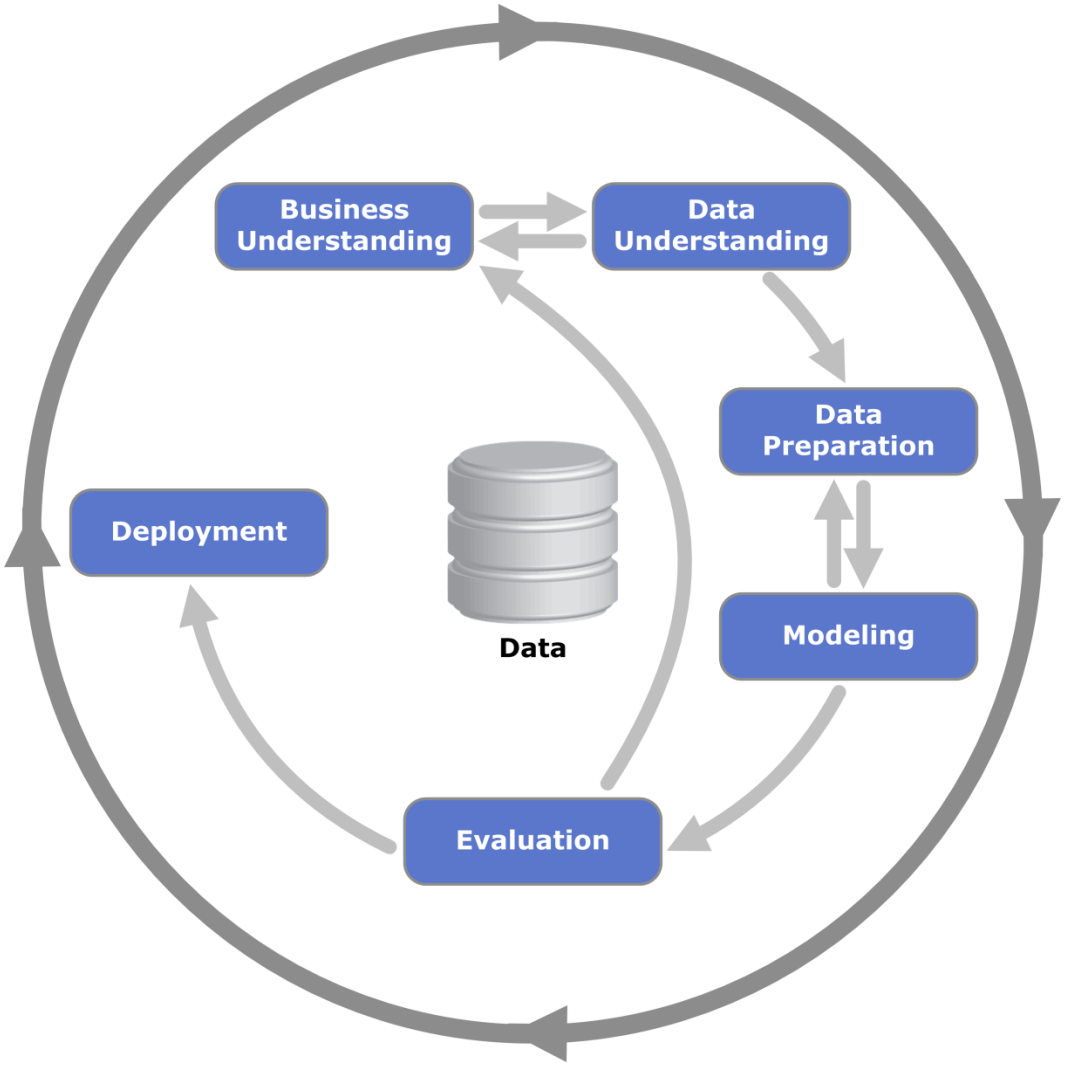
Cross-Industry Standard Process for Data Mining (CRISP-DM)

European Community funded effort to develop framework for data mining tasks

Goals:

- Encourage interoperable tools across entire data mining process
- Take the mystery/high-priced expertise out of simple data mining tasks

CRISP-DM overview



Note: Iteration is the rule rather than the exception.

CRISP-DM figure (wikipedia.org)

A Scenario

Example: Predicting Customer Churn

You just landed a great analytical job with MegaTelCo, one of the largest telecommunication firms in the United States.

Since the cell phone market is now saturated, the huge growth in the wireless market has tapered off. Communications companies are now engaged in battles to attract each other's customers while retaining their own.

They have allocated some marketing budget to spend customers. You have been called in to help figure out its marketing strategy.

Think carefully about what data you might use and how they would be used. Specifically, they need your help to choose a set of customers to receive their offer in order to best increase its profits.

(``Data Science for Business'' by Provost&Fawcett)

Figure 1: Fierce competition in mobile telecommunication industry (Japan).
Firms are trying to attract more customers..
The figure shows # of monthly increase in the carriers' contracts.

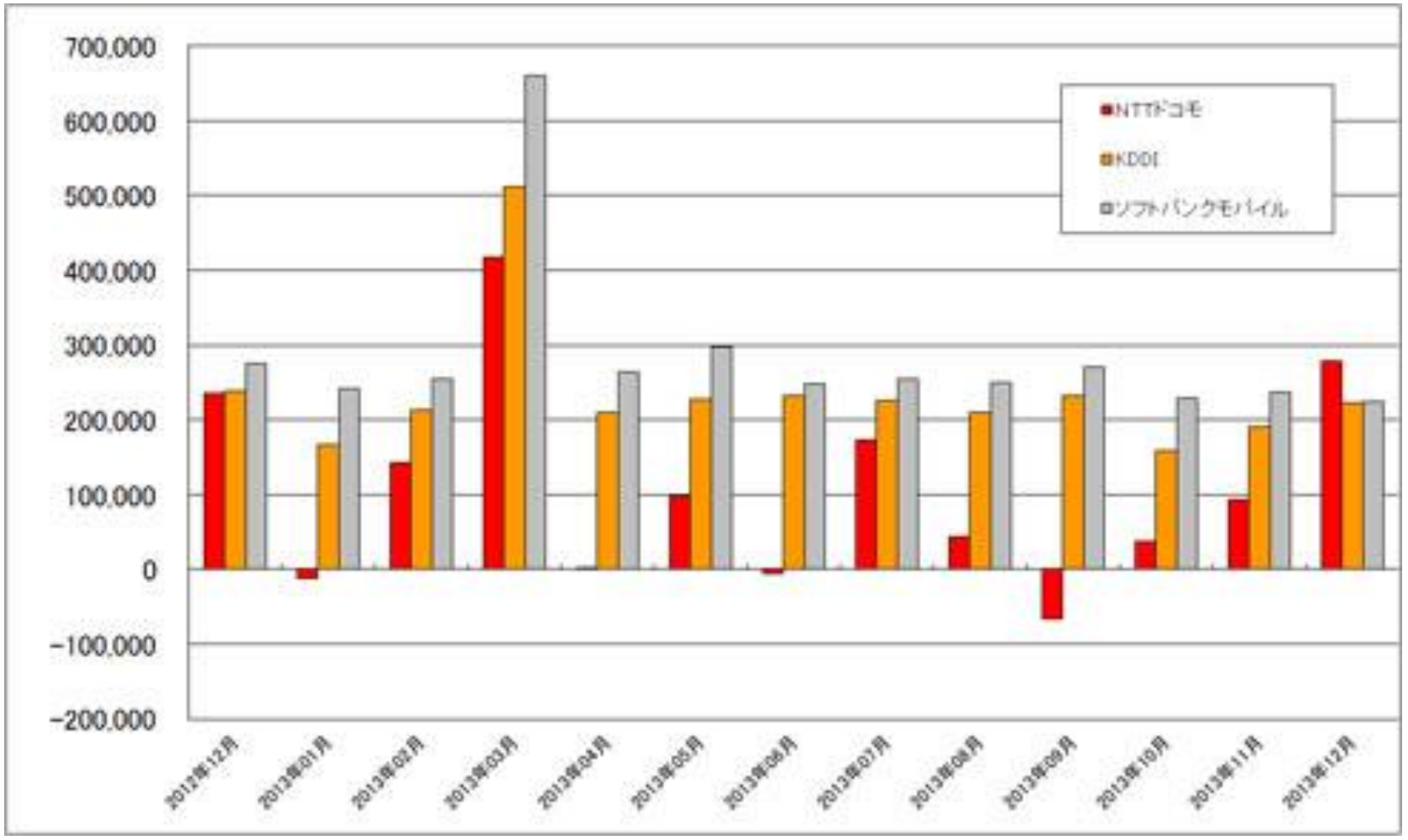
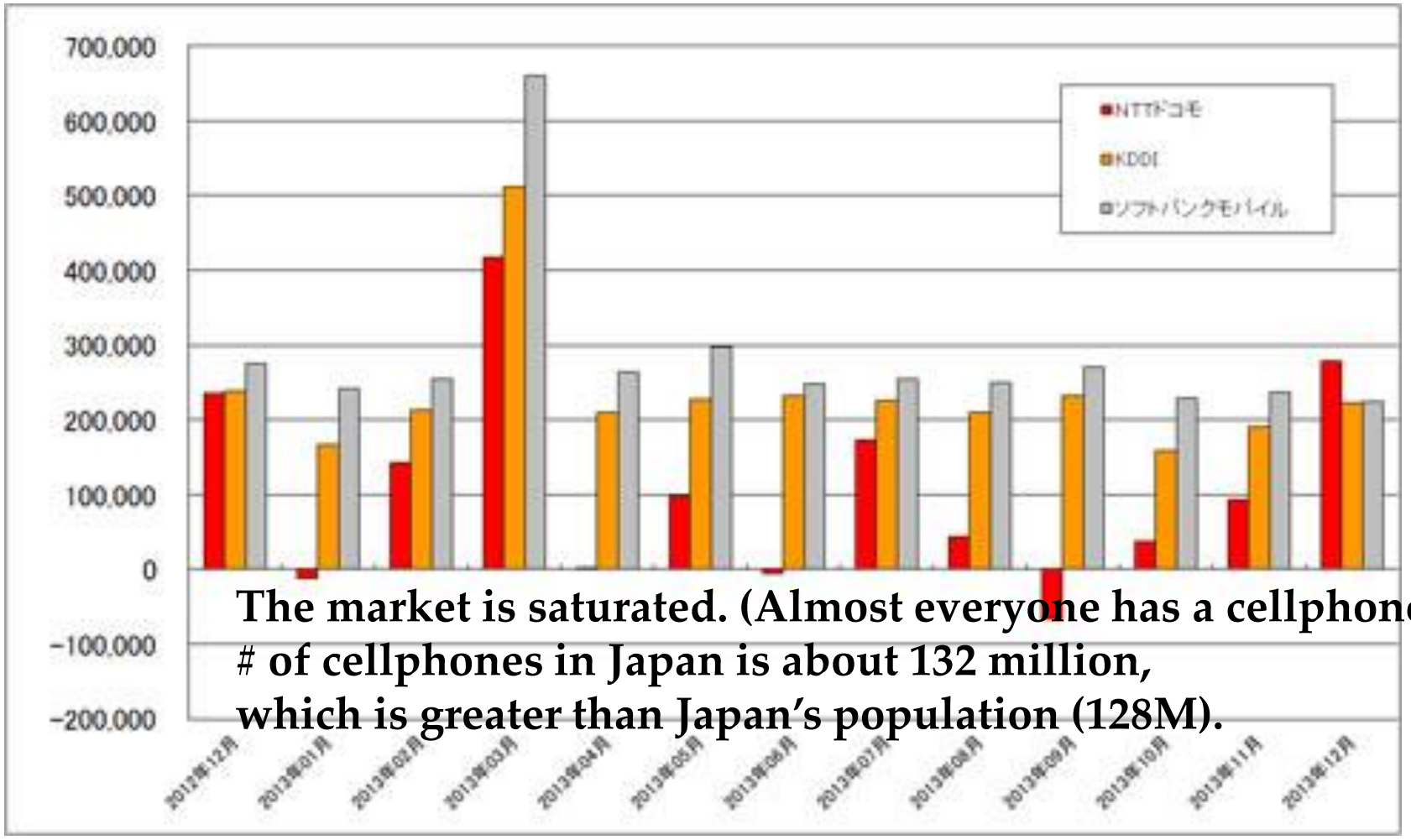


Figure 1: Fierce competition in mobile telecommunication industry (Japan).
Firms are trying to attract more customers..
The figure shows # of monthly increase in the carriers' contracts.



The market is saturated. (Almost everyone has a cellphone.)
# of cellphones in Japan is about 132 million,
which is greater than Japan's population (128M).

1. Business Understanding

- Determine business objectives
- Solve a specific problem
- Assess the current situation
- Convert the above into a data mining project
- Develop a project plan

1. Business Understanding

What is the best way to spend the marketing money?

Consider which group of people we should focus.

- Attract _new_ customers.
- Retain _existing_ customers.

1. Business Understanding

What is the best way to spend the marketing money?

Consider which group of people we should focus.

- Attract _new_ customers.

  Who should we target?

  - People who currently don't use any cell phone?
  - People who have some contract with another carrier?

- Retain _existing_ customers.

  Who should we target? (Who will likely leave/switch from us?)

  - Age?
  - Gender?
  - Income?
  - Depending on their plans?

1. Business Understanding

What is the best way to spend the marketing money?

Consider which group of people we should focus.

- Attract _new_ customers.
- Retain _existing_ customers.

After discussions with the marketing team, you found that …

1. Business Understanding

What is the best way to spend the marketing money?

Consider which group of people we should focus.

- Attract _new_ customers.
- Retain _existing_ customers.

After discussions with the marketing team, you found that …

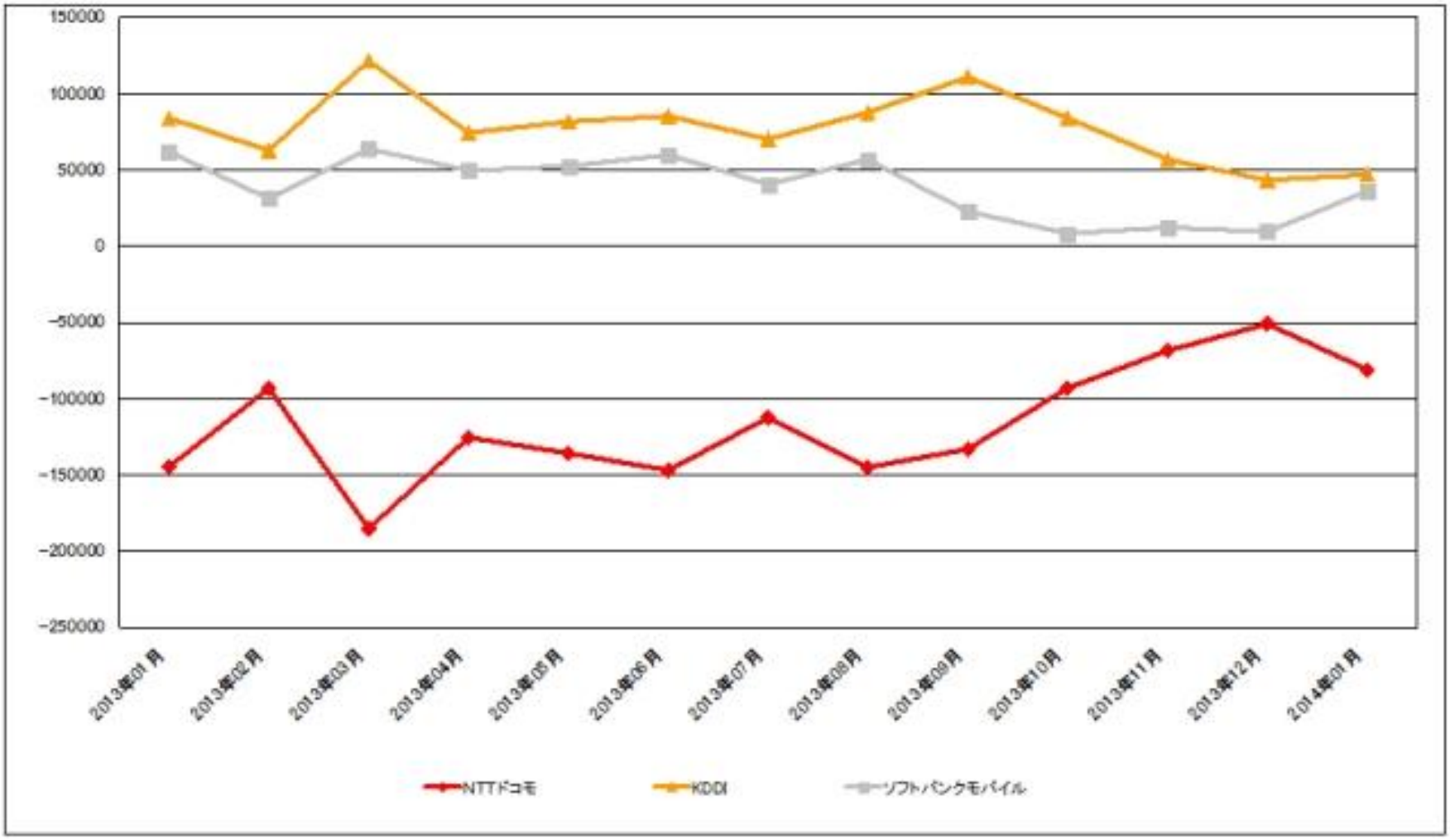
Attracting new customers is much more expensive than retaining existing ones, so a good deal of marketing budget is allocated to retain existing customers.

In fact, MegaTelCo is having a major problem with customer retention in their wireless business. In the mid-Atlantic region, 20% of cell phone customers leave when their contracts expire.

Terminology: Customers switching from one company to another is called ***churn***, and it is expensive all around: one company must spend on incentives to attract a customer while another company loses revenue when the customer departs.

Figure 2: Fierce competition in mobile telecommunication industry (Japan).

Churn is a severe problem. The figure shows increase/decrease in contracts via MNP (Mobile Number Portability). Docomo is losing 50,000 to 150,000 customers monthly..



- Transform the business problem into a data mining one

Marketing has already designed a special retention offer. Your task is to devise a precise, step-by-step plan for how the data science team should use MegaTelCo's vast data resources to decide which customers should be offered the special retention deal prior to the expiration of their contracts.

How should MegaTelCo choose a set of customers to receive their offer in order to best reduce churn for a particular incentive budget?

Your model should predict which customers will likely churn.

2. Data Understanding

- Initial Data Collection
- Data Description
- Data Exploration
- Data Quality Verification
- Data Selection
- Related data can come from many sources

2. Data Understanding

We have a historical data set of 20,000 customers. At the point of collecting the data, each customer either had stayed with the company or had left (churned). Each customer is described by the variables listed below.

Variable	Explanation
COLLEGE	Is the customer college educated?
INCOME	Annual income
OVERAGE	Average overcharges per month
LEFTOVER	Average number of leftover minutes per month
HOUSE	Estimated value of dwelling (from census tract)
HANDSET_PRICE	Cost of phone
LONG_CALLS_PER_MONTH	Average number of long calls (15 mins or over) per month
AVERAGE_CALL_DURATION	Average duration of a call
REPORTED_SATISFACTION	Reported level of satisfaction
REPORTED_USAGE_LEVEL	Self-reported usage level
LEAVE (Target variable)	Did the customer stay or leave (churn)?

Data looks like …

ID	COLLEGE	INCOME	OVERAGE	LEFTOVER	HOUSE	HANDSET_PRICE	LONG_CALLS_PER_MONTH	AVERAGE_CALL_DURATION	REPORTED_SATISFACTION	REPORTED_USAGE_LEVEL	LEAVE (Target variable)
0001	1	50000	30	0	1M	150	3	15			0
0002	1	60000	10	20	.5M	50	.8	5			1
0003	0	30000	15	5		75	1.2	6			1
0004	0		7	0	0	30	.5	7			0
:	:	:	:	:	:	:	:	:	:	:	:

Data looks like …

ID	COLLEGE	INCOME	OVERAGE	LEFTOVER	HOUSE	HANDSET_PRICE	LONG_CALLS_PER_MONTH	AVERAGE_CALL_DURATION	REPORTED_SATISFACTION	REPORTED_USAGE_LEVEL	LEAVE (Target variable)
0001	1	50000	30	0	1M	150	3	15			0
0002	1	60000	10	20	.5M	50	.8	5			1
0003	0	30000	15	5		75	1.2	6			1
0004	0		7	0	0	30	.5	7			0
:	:	:	:	:	:	:	:	:	:	:	:

Available data to predict churn.

This is what we want to predict.

3. Data Preparation

- Clean selected data for better quality
  - Treat missing values
  - Identify or remove outliers
  - Resolve redundancy caused by data integration
  - Correct inconsistent data
- Transform data
  - Convert different measurements of data into a unified numerical scale by using simple mathematical formulations

For further reference: “Introduction to Data Mining” by Tan, Steinbach and Kumar.

ID	COLLEGE	INCOME	OVERAGE	LEFTOVER	HOUSE	HANDSET_PRICE	LONG_CALLS_PER_MONTH	AVERAGE_CALL_DURATION	REPORTED_SATISFACTION	REPORTED_USAGE_LEVEL	LEAVE (Target variable)
0001	1	50000	30	0	1M	150	3	15			0
0002	1	60000	10	20	.5M	50	.8	5			1
0003	0	30000	15	5		75	1.2	6			1
0004	0		7	0	0	30	.5	7			0
:	:	:	:	:	:	:	:	:	:	:	:

Example: Treat missing values

- Eliminate data objects with missing values.
- Ignore attributes with missing values.
- Fill in missing values.

ID	COLLEGE	INCOME	OVERAGE	LEFTOVER	HOUSE	HANDSET_PRICE	LONG_CALLS_PER_MONTH	AVERAGE_CALL_DURATION	REPORTED_SATISFACTION	REPORTED_USAGE_LEVEL	LEAVE (Target variable)
0001	1	50000	30	0	1M	150	3	15			0
0002	1	60000	10	20	.5M	50	.8	5			1
0003	0	30000	15	5		75	1.2	6			1
0004	0		7	0	0	30	.5	7			0
:	:	:	:	:	:	:	:	:	:	:	:

Example: Treat missing values

- (E) Eliminate data objects with missing values.
- (I) Ignore attributes with missing values.
- (F) Fill in missing values.

Consider "REPORTED_USAGE_LEVEL" field.

(E) Most customers don't provide "REPORTED_USAGE_LEVEL" info.

If we do (E), we eliminate almost all records. → Not appropriate

(I) Most customers don't provide "REPORTED_USAGE_LEVEL" info.

If we do (I), almost all records will remain. → Looks good.

ID	COLLEGE	INCOME	OVERAGE	LEFTOVER	HOUSE	HANDSET_PRICE	LONG_CALLS_PER_MONTH	AVERAGE_CALL_DURATION	REPORTED_SATISFACTION	REPORTED_USAGE_LEVEL	LEAVE (Target variable)
0001	1	50000	30	0	1M	150	3	15			0
0002	1	60000	10	20	.5M	50	.8	5			1
0003	0	30000	15	5		75	1.2	6			1
0004	0		7	0	0	30	.5	7			0
:	:	:	:	:	:	:	:	:	:	:	:

Example: Treat missing values

- (E) Eliminate data objects with missing values.
- (I) Ignore attributes with missing values.
- (F) Fill in missing values.

Consider "INCOME" field.

(E) Some customers don't provide "INCOME" info.

If we do (E), we eliminate some fraction of records. → Okay.

(I) Many customers provide "INCOME" info.

If we do (I), we eliminate many. → Not appropriate.

(F) We may be able to reasonably estimate income from other data, e.g. COLLEGE.

Fill in a blank with average income of college graduate/non-graduate. → Okay.

4. Modeling

- Data Treatment
  - Training set
  - Test set
- Data Mining Techniques
  - Regression (prediction)
  - Association
  - Classification
  - Clustering

4. Modeling

We have clean data and want to predict churn.

Which variables we should use to predict churn?

Restate: Which of the variables would be best to segment these people into groups, "churn" and "non-churn"?

Information gain

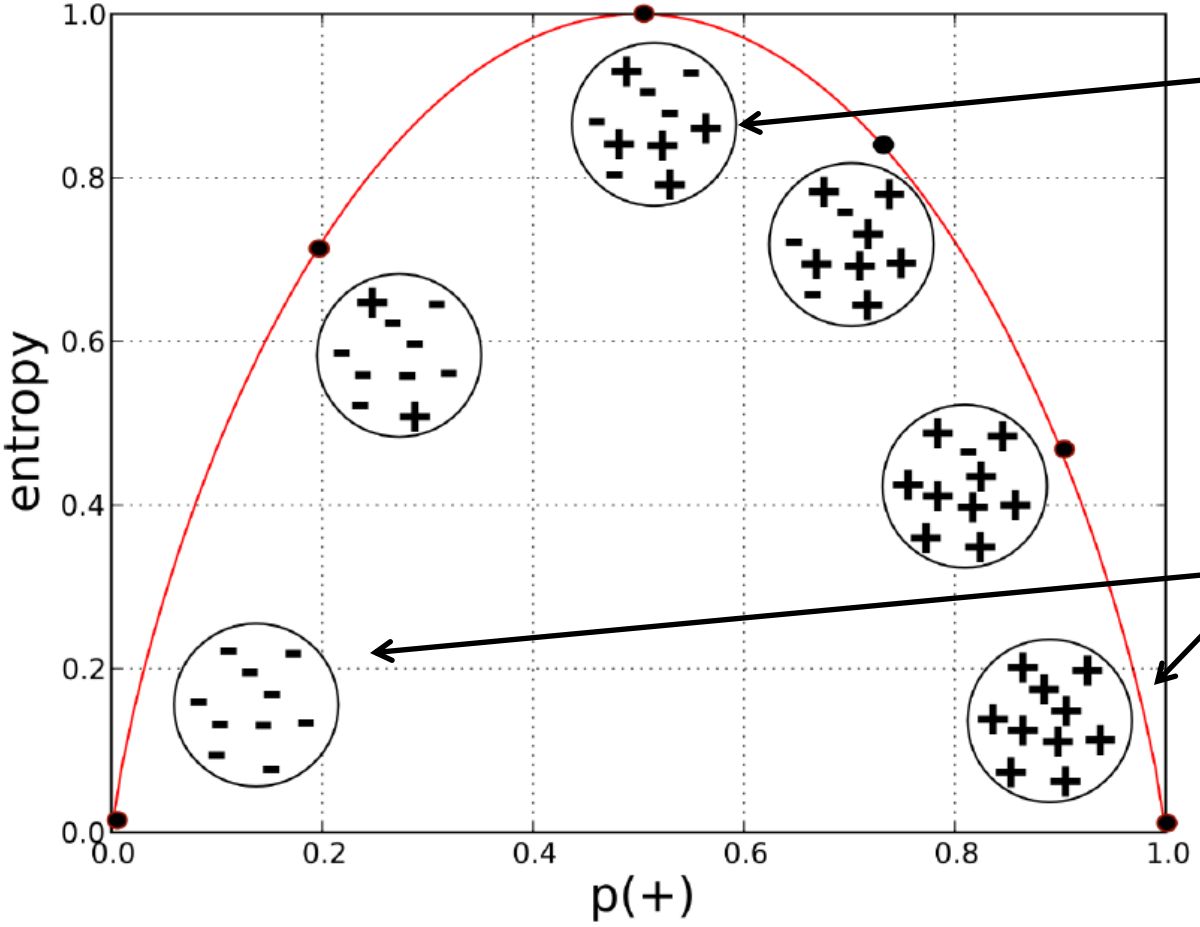
Information gain will tell us how informative an attribute is.

Entropy

$$entropy = -p1 \log p1 - p2 \log p2 \cdots$$

Each pX is the probability of property X within the set.

(pX=1 implies that all observations have property X.)



p(+) is probability that a customer churned.

The model is more predictive if it achieves lower entropy.

FSelector pkg in R

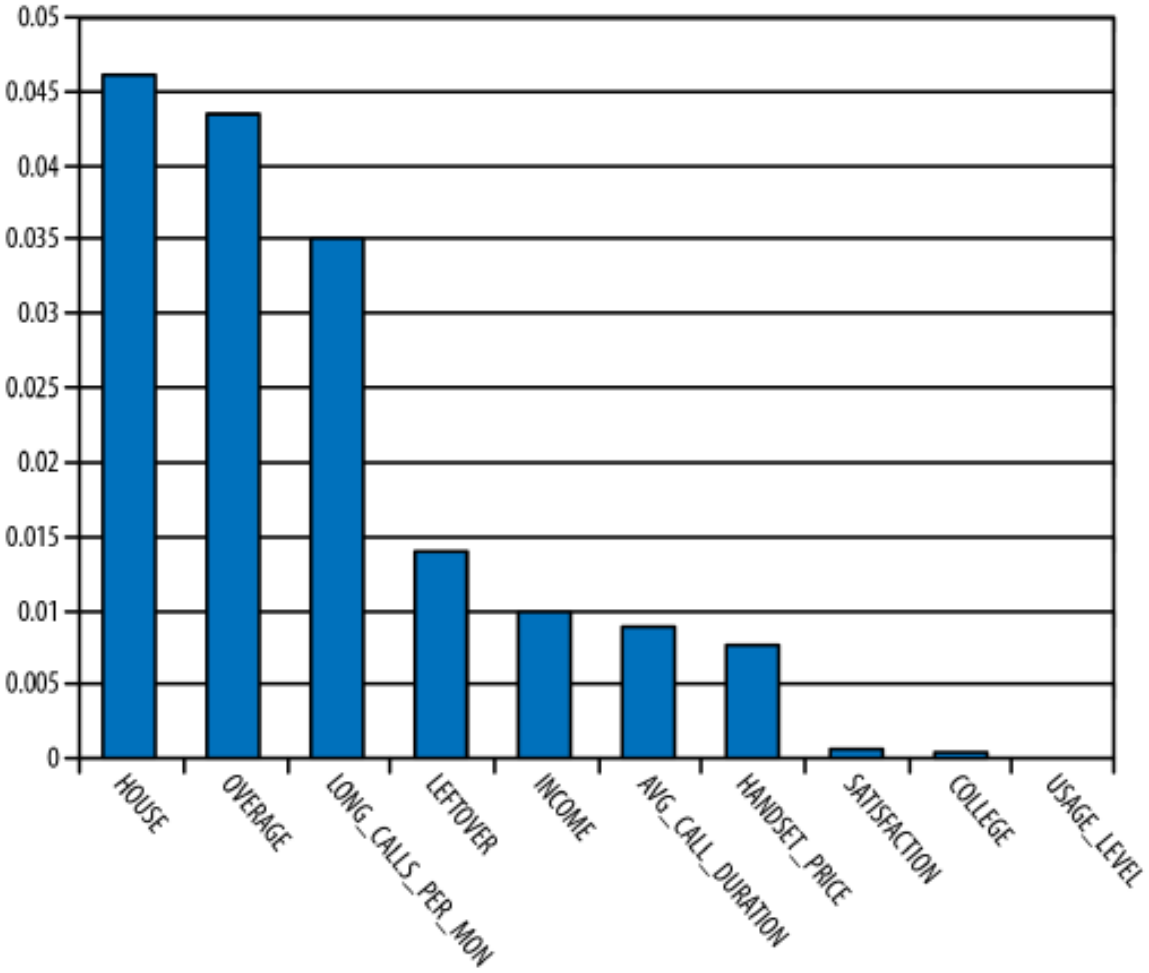
You don't need to be worried about computation…

Any statistical software will do for you.

Example with R

```
> library(Fselector)
> data(iris)
>
> weights <- information.gain(Species~., iris)
>
> print(weights)
              attr_importance
Sepal.Length       0.6522837
Sepal.Width        0.3855963
Petal.Length       1.3565450
Petal.Width        1.3784027
```

More informative attributes

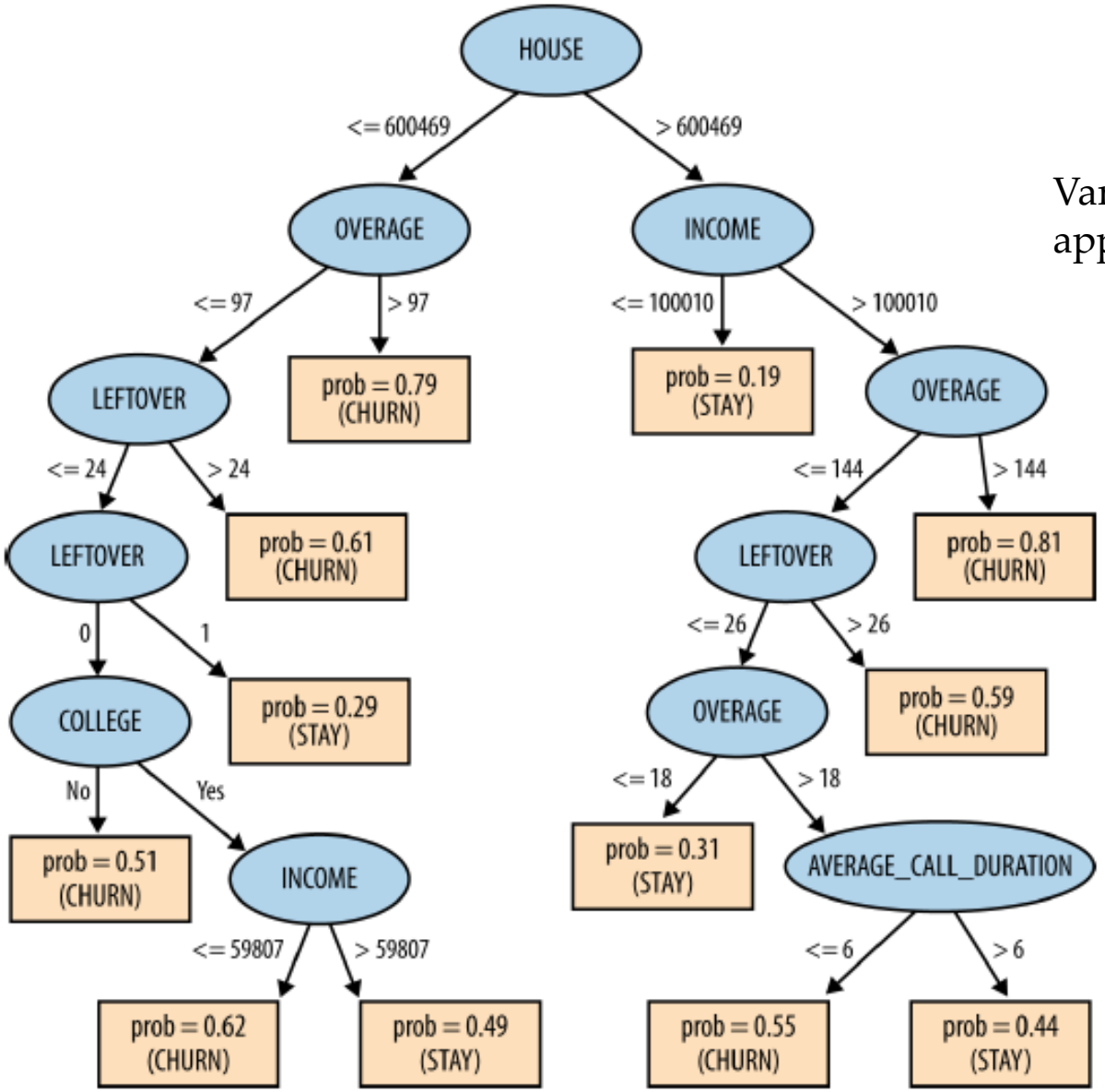


Rank	Info. gain	Attribute name
1	0.0461	HOUSE
2	0.0436	OVERAGE
3	0.0350	LONG_CALLS_PER_MON
4	0.0136	LEFTOVER
5	0.0101	INCOME
6	0.0089	AVG_CALL_DURATION
7	0.0076	HANDSET_PRICE
8	0.0003	SATISFACTION
9	0.000	COLLEGE
10	0.000	USAGE_LEVEL

Information gain for variables in churn example.

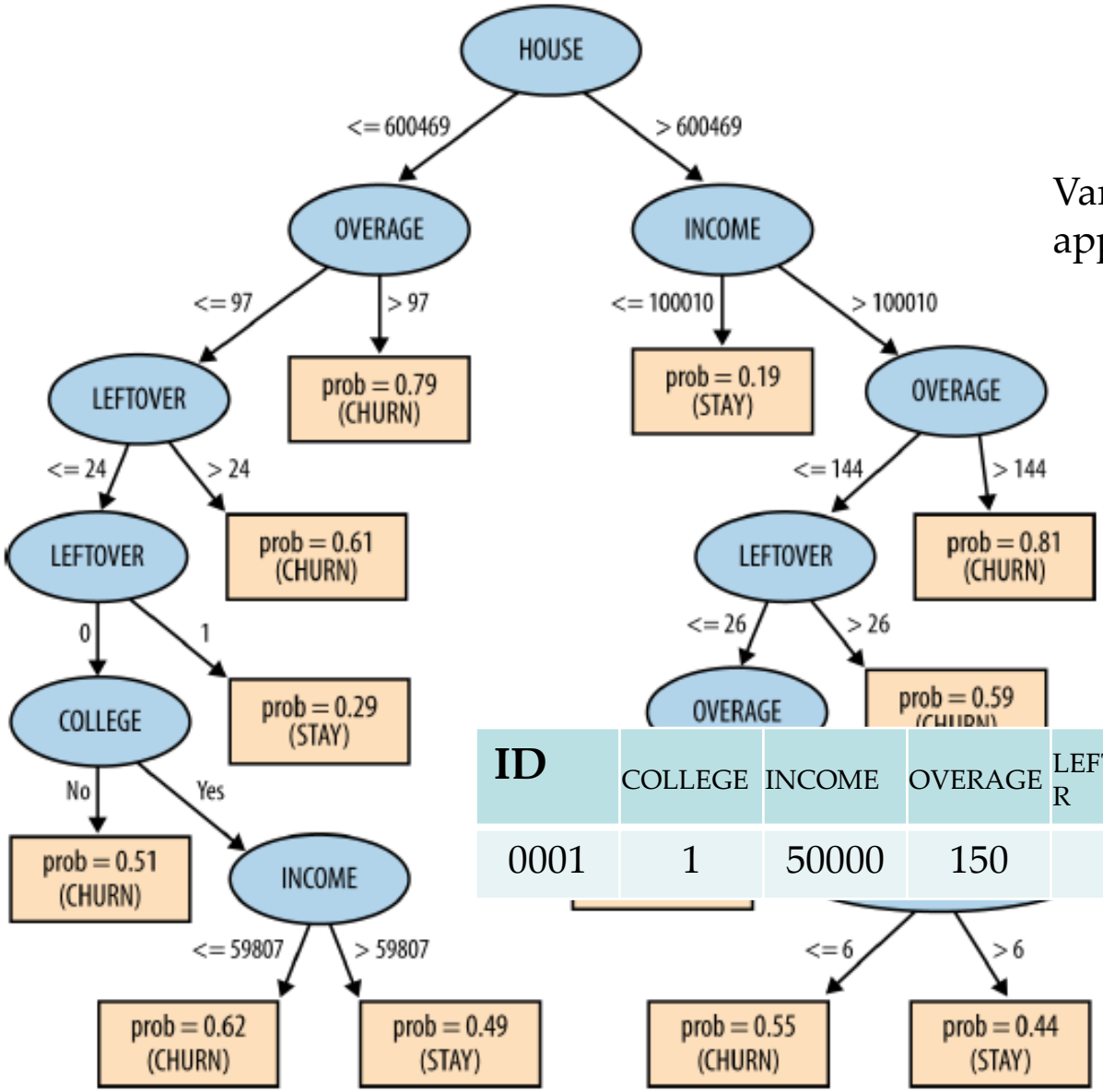
It would be good to include into the model variables which have relatively high IG.

Modeling Example: choice of tree classification



Variables which have higher IG appear first. But not always..

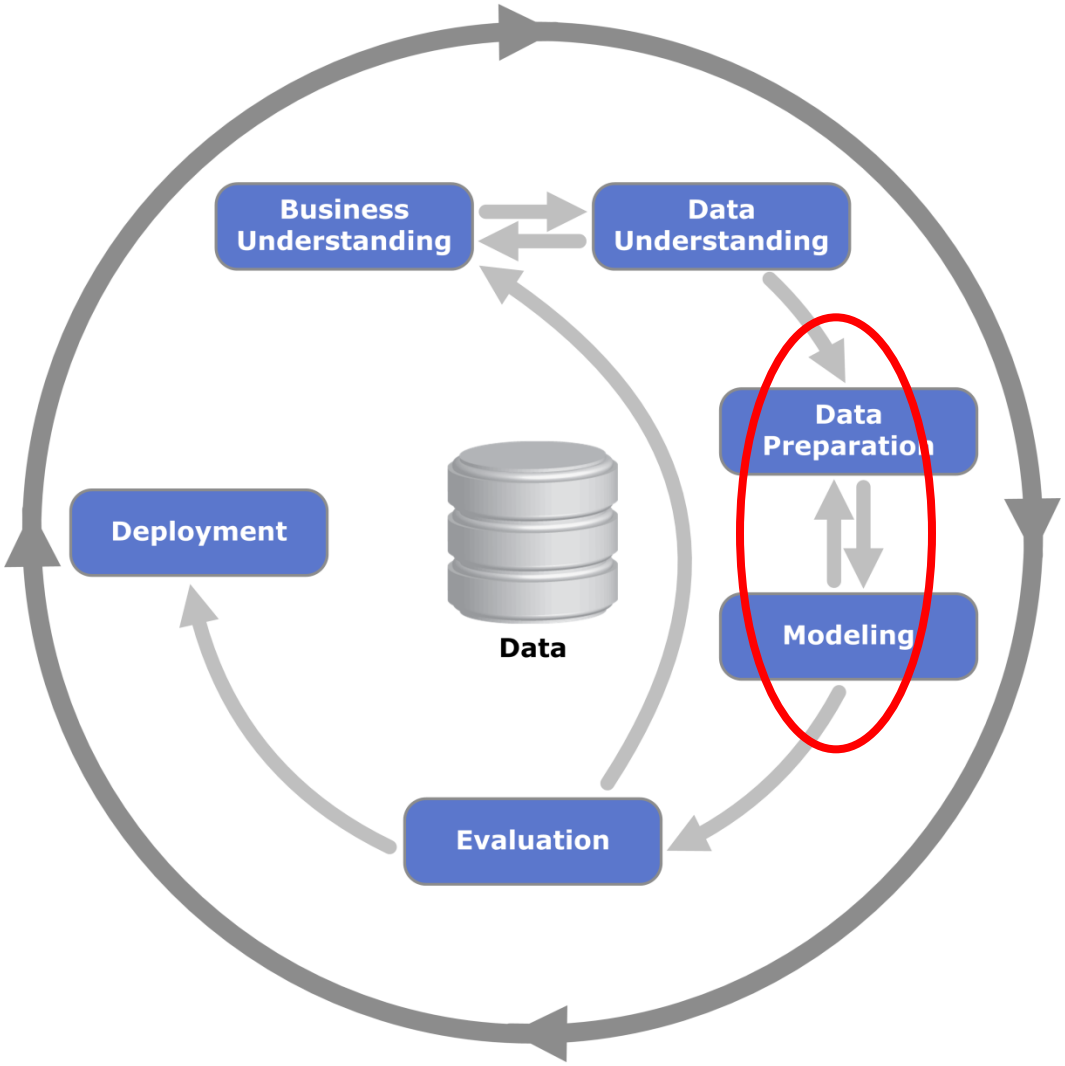
Modeling Example: choice of tree classification



Variables which have higher IG appear first. But not always..

ID	COLLEGE	INCOME	OVERAGE	LEFTOVER	HOUSE	…..	LEAVE (Target variable)
0001	1	50000	150	0	1M	….	0

CRISP-DM overview



Note:
After Modeling stage, we may need to go back to Data Preparation.

Your model may require a specific form of data as input.

CRISP-DM figure (wikipedia.org)

5. Evaluation

- Does model meet business objectives?
- Any business objectives not addressed?
- Does model make sense?
- Is model actionable?
- It should be possible to make business decisions after this step.
- All important objectives should be achieved.

5. Evaluation

To evaluate the model, one of the most important fundamental notions:

Avoid Overfitting.

Overfitting is the tendency of data mining procedures to tailor models to the training data, at the expense of generalization.

An extreme example:

Consider the following *table* model.

We stored the feature vector for each customer who has churned in a database table. Let's call that Tc. In use, when the model is presented with a customer to determine the likelihood of churning, it takes the customer's feature vector, looks her up in Tc, and reports "100% likelihood of churning" if she is in Tc and "0% likelihood of churning" if she is not in Tc.

When the tech team applies our model to the historical dataset, the model predicts perfectly. The model 100% fits with the historical dataset.

ID	COLLEGE	INCOME	OVERAGE	LEFTOVER	HOUSE	HANDSET_PRICE	LONG_CALLS_PER_MONTH	AVERAGE_CALL_DURATION	REPORTED_SATISFACTION	REPORTED_USAGE_LEVEL	LEAVE (Target variable)
0001	1	50000	30	0	1M	150	3	15			0
0002	1	60000	10	20	.5M	50	.8	5			1
0003	0	30000	15	5		75	1.2	6			1
0004	0		7	0	0	30	.5	7			0
:	:	:	:	:	:	:	:	:	:	:	:

Use ID to predict LEAVE.
Does it work?

table model (illustration):

Your prediction system stores customer IDs who left MegaTelCo.

If you input a customer ID and it matches some stored ID, then your system predicts churn.

If you apply the existing data to your prediction system, you will obtain 100% accuracy.

What is the problem of the *table* model?

Consider how we'll use the model in practice.

When a previously unseen customer's contract is about to expire, we'll want to apply the model. Of course, this customer was not part of the historical dataset, so the lookup will fail since there will be no exact match, and the model will predict "0% likelihood of churning" for this customer. In fact, the model will predict this for every customer (not in the training data).

The model looked perfect, but it is completely useless in practice!

The table model `overfits` to the training data and looses generalization.

Generalization is the property of a model or modeling process, whereby the model applies to data that were not used to build the model. In this example, the model does not generalize at all beyond the data that were used to build it.

The issue is more complex than it looks because the answer is _not_ to use a data mining procedure that doesn't overfit. All of procedures do overfit.

How do we recognize overfitting?

A simple analytic tool: Fitting graph

It shows the accuracy of a model as a function of complexity.

First, we need to "hold out" some data.

"hold out" data: has the value of the target variable, but will not be used to build the model.

The issue is more complex than it looks because the answer is _not_ to use a data mining procedure that doesn't overfit. All of procedures do overfit.

How do we recognize overfitting?

A simple analytic tool: Fitting graph

It shows the accuracy of a model as a function of complexity.

First, we need to "hold out" some data.

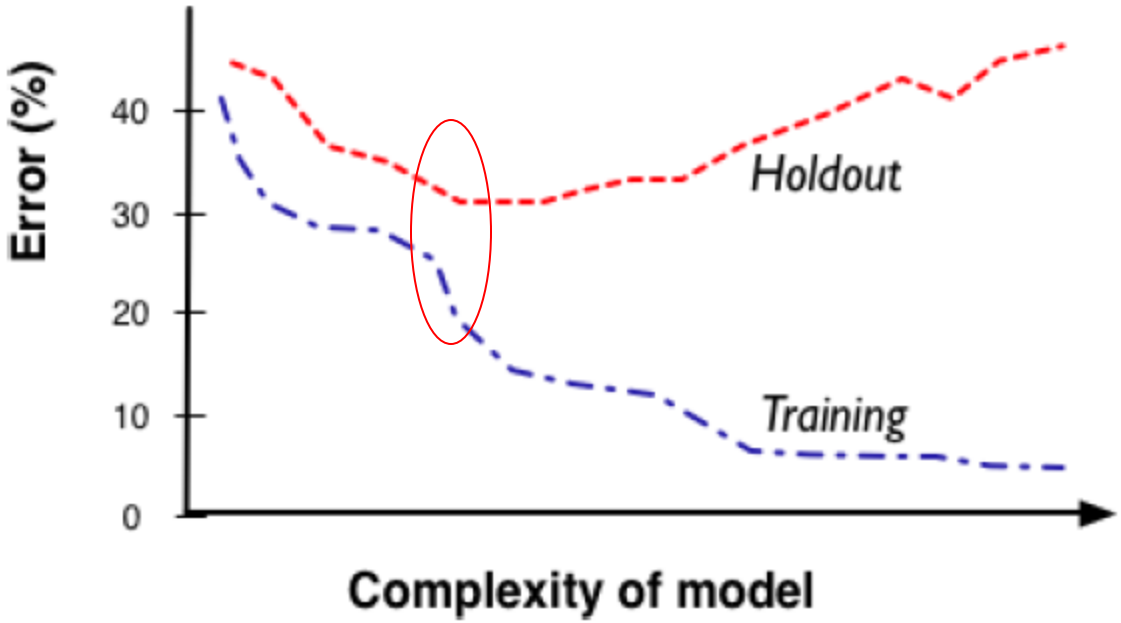
"hold out" data: has the value of the target variable, but will not be used to build the model.

Example:

You have data of 100,000 customers. Then,

50,000 customers: You apply a DM technique to build a model.

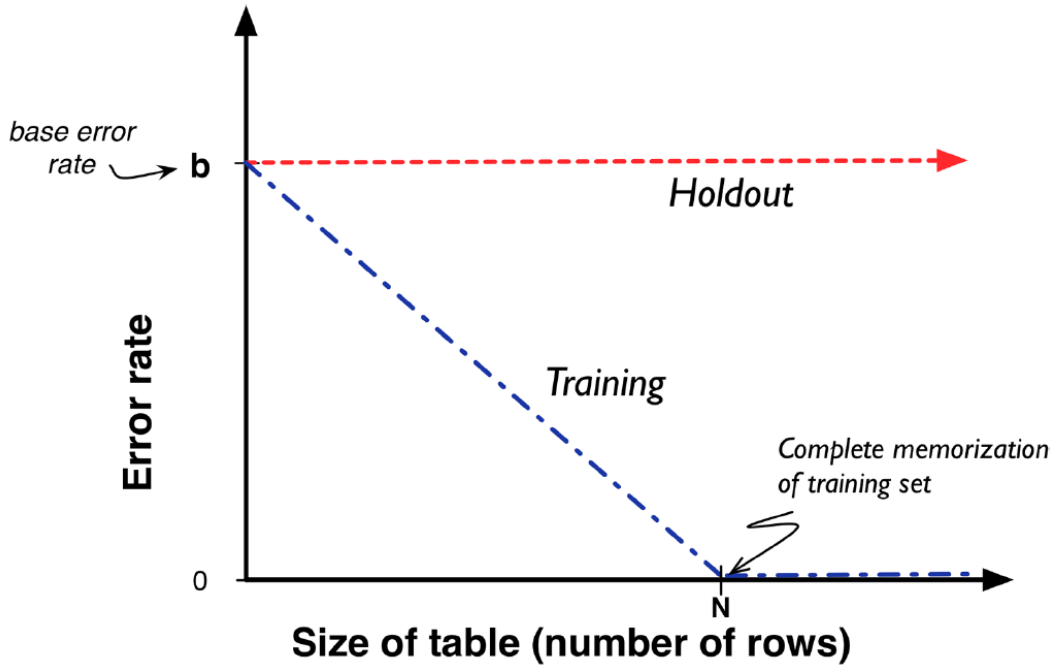
50,000 : You "hold out" for evaluation of your model.



Fitting graph
Complexity: the number of variables we use in the model.

Each point on a curve represents an accuracy estimation of a model with a specified complexity (as indicated on the horizontal axis).

When the model is not allowed to be complex enough, it is not very accurate. As the models get too complex, they look very accurate on the training data, but in fact are overfitting—the training accuracy diverges from the holdout (generalization) accuracy.



Fitting graph (the *table* model)
Complexity: Size of table

What would **b** be?

Since the table model always predicts no churn for every new case.

It will get every no churn case right and every churn case wrong.

Thus the error rate will be the percentage of churn cases in the population.

This is known as the base rate, and a classifier that always selects the majority class is called a base rate classifier.

6. Deployment

- Ongoing monitoring and maintenance
  - Evaluate performance against success criteria
  - Market reaction & competitor changes

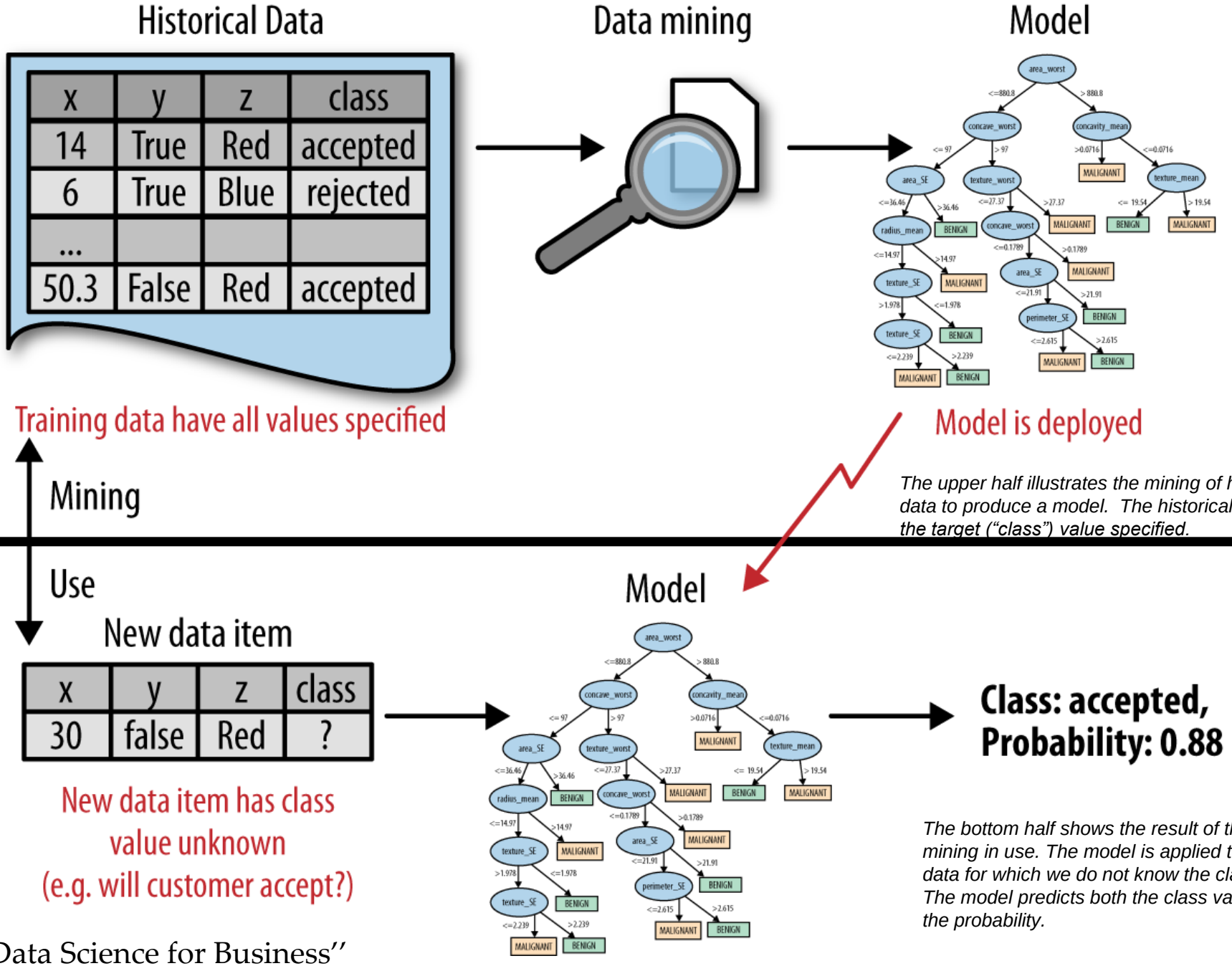
6. Deployment

A deployment scenario

We want to use the model to predict which of our customers will leave. Specifically, assume that data mining has created a class probability estimation model M.

Given each existing customer, described using a set of characteristics, M takes these characteristics as input and produces a score or probability estimate of attrition. This is the use of the results of data mining. The data mining produces the model M from some other, often historical, data.

The next figure illustrates these two phases. Data mining produces the probability estimation model, as shown in the top half of the figure. In the use phase (bottom half), the model is applied to a new, unseen case and it generates a probability estimate for it.



The upper half illustrates the mining of historical data to produce a model. The historical data have the target ("class") value specified.

The bottom half shows the result of the data mining in use. The model is applied to new data for which we do not know the class value. The model predicts both the class value and the probability.

Pitfalls

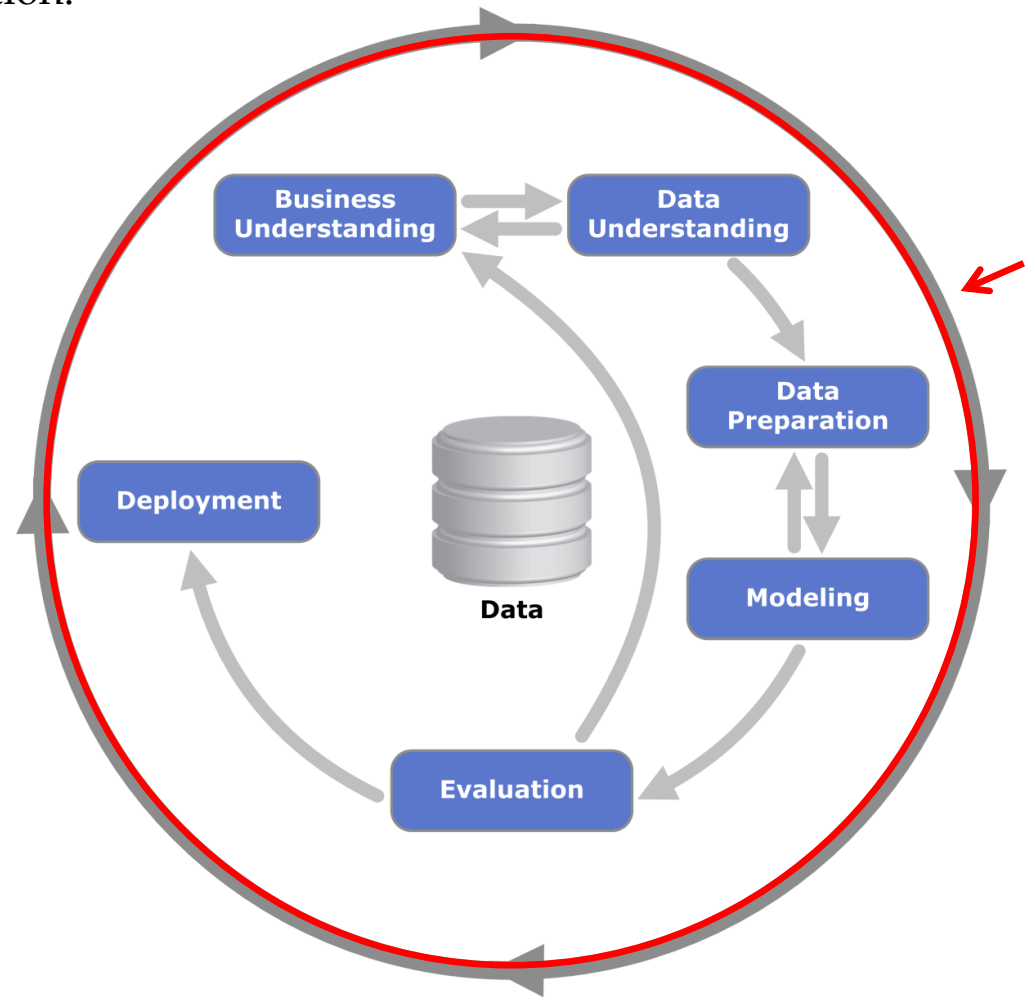
Deploying a model into a production system typically requires that the model be recoded for the production environment, usually for greater speed or compatibility with an existing system.

There are risks with "over the wall" transfers from data science to development. Remember that: "A deployed model is not what the data scientists design, it's what the engineers build."

It is advisable to have members of the development team involved early on in the data science project. They can begin as advisors, providing critical insight to the data science team.

The process never ends…

Regardless of whether deployment is successful, the process often returns to the Business Understanding phase. The process of mining data produces a great deal of insight into the business problem and the difficulties of its solution. A second iteration can yield an improved solution.



Note: Iteration is the rule rather than the exception.

Hint:

- Replace the missing value with the field mean/mode.
  A standard choice:
  Use mean for numerical values.
  Use mode for categorical values.

- Use Z-Score to identify outliers.
  How far an observation is from the field mean value.
  Z-Score of X = (X – mean(X))/SD(X)
  mean(X): the field mean
  SD(X): standard deviation of the field values.
  e.g. potential outliers if the absolute value of Z-Score exceeds 3.

- Modeling - Evaluation
  Recall one of questions to ask in Evaluation:
  Does model make sense?

  Explain how we should test the "odd number" model.