DSA/ISE 5103 Intelligent Data Analytics

CRISP-DM and Project Understanding

Charles Nicholson, Ph.D.
cnicholson@ou.edu

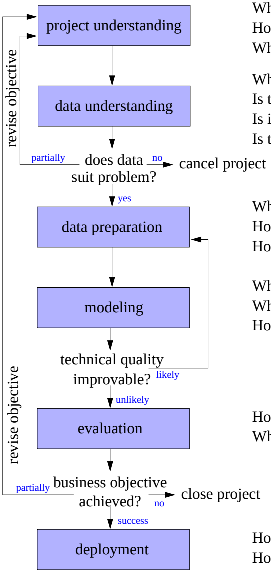
University of Oklahoma
Gallogly College of Engineering
School of Industrial and Systems Engineering

Outline

1. **Process: CRISP-DM**
2. Project Understanding

CRISP-DM

Cross Industry Standard Process for Data Mining



What exactly is the problem, the expected benefit?
How would a solution look like?
What is known about the domain?

What data do we have available?
Is the data relevant to the problem?
Is it valid? Does it reflect our expectations?
Is the data quality, quantity, recency sufficient?

Which data should we concentrate on?
How is the data best transformed for modeling?
How may we increase the data quality?

What kind of model architecture suits the problem best?
What is the best technique/method to get the model?
How good does the model perform technically?

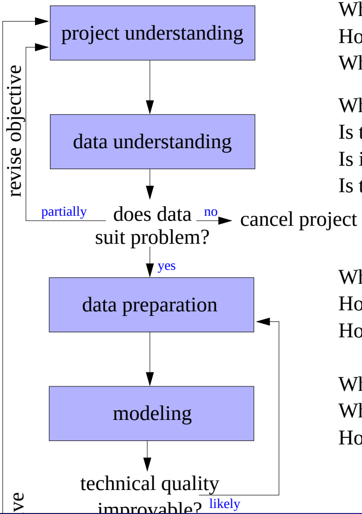
How good is the model in terms of project requirements?
What have we learned from the project?

How is the model best deployed?
How do we know that the model is still valid?

Outline

1. Process: CRISP-DM
2. **Project Understanding**

project understanding



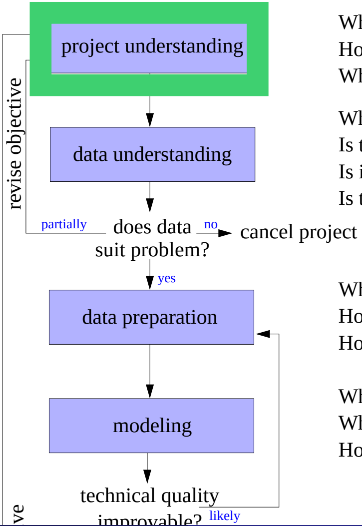
What exactly is the problem, the expected benefit?
How would a solution look like?
What is known about the domain?

What data do we have available?
Is the data relevant to the problem?
Is it valid? Does it reflect our expectations?
Is the data quality, quantity, recency sufficient?

Which data should we concentrate on?
How is the data best transformed for modeling?
How may we increase the data quality?

What kind of model architecture suits the problem best?
What is the best technique/method to get the model?
How good does the model perform technically?

project understanding



What exactly is the problem, the expected benefit?
How would a solution look like?
What is known about the domain?

What data do we have available?
Is the data relevant to the problem?
Is it valid? Does it reflect our expectations?
Is the data quality, quantity, recency sufficient?

Which data should we concentrate on?
How is the data best transformed for modeling?
How may we increase the data quality?

What kind of model architecture suits the problem best?
What is the best technique/method to get the model?
How good does the model perform technically?

project understanding

The 80-20 Rule!

- Average time spent for project and data understanding within the CRISP-DM model: 20%
- Importance for success: 80%

tasks in project understanding

tasks in project understanding

1. Determine business objectives

Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rüdiger Wirth (2000); *CRISP-DM 1.0 Step-by-step data mining guides.*

tasks in project understanding

1. Determine business objectives
2. Assess situation

Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rüdiger Wirth (2000); *CRISP-DM 1.0 Step-by-step data mining guides.*

tasks in project understanding

1. Determine business objectives
2. Assess situation
3. Determine data mining goals

Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rüdiger Wirth (2000); *CRISP-DM 1.0 Step-by-step data mining guides.*

tasks in project understanding

1. Determine business objectives
2. Assess situation
3. Determine data mining goals
4. Produce project plan

Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rüdiger Wirth (2000); *CRISP-DM 1.0 Step-by-step data mining guides.*

determine business objectives

Tasks:

- Understand problem from a *business perspective*
- Identify critical factors, competing objectives and constraints

Output:

- Background description
- Primary objectives
- Success criteria

determine business objectives

Tasks:

- Understand problem from a *business perspective*
- Identify critical factors, competing objectives and constraints

Output:

- Background description
- Primary objectives
- Success criteria

determine business objectives

Tasks:

- Understand problem from a *business perspective*
- Identify critical factors, competing objectives and constraints

Output:

- Background description
- Primary objectives
- Success criteria

determine business objectives

Tasks:

- Understand problem from a *business perspective*
- Identify critical factors, competing objectives and constraints

Output:

- Background description
- Primary objectives
- Success criteria

determine business objectives

Tasks:

- Understand problem from a *business perspective*
- Identify critical factors, competing objectives and constraints

Output:

- Background description
- Primary objectives
- Success criteria

understanding the problem

understanding the problem

- What exactly is the problem?

understanding the problem

- What exactly is the problem?
- What would a potential solution look like?

understanding the problem

- What exactly is the problem?
- What would a potential solution look like?
  - What would you do with a solution?

understanding the problem

- What exactly is the problem?
- What would a potential solution look like?
  - What would you do with a solution?
  - What could you do with a solution?

case study: Automobile Insurance Fraud

In spite of having a fraud investigation team that investigates up to 30% of all claims made, an automobile insurance company is still losing too much money due to fraudulent claims.

case study: Automobile Insurance Fraud

In spite of having a fraud investigation team that investigates up to 30% of all claims made, an automobile insurance company is still losing too much money due to fraudulent claims.

What predictive analytics solutions could be proposed to help address this business problem?

case study: Automobile Insurance Fraud

Claim prediction

case study: Automobile Insurance Fraud

Claim prediction

- Model: predict whether an insurance claim is fraudulent or not

case study: Automobile Insurance Fraud

Claim prediction

- Model: predict whether an insurance claim is fraudulent or not
- Use: assign likely fraudulent claims to investigation team

case study: Automobile Insurance Fraud

Claim prediction

- Model: predict whether an insurance claim is fraudulent or not
- Use: assign likely fraudulent claims to investigation team
- Benefit: investigation team's time allocated better, reducing $ lost to fraud

case study: Automobile Insurance Fraud

Claim prediction

- Model: predict whether an insurance claim is fraudulent or not
- Use: assign likely fraudulent claims to investigation team
- Benefit: investigation team's time allocated better, reducing $ lost to fraud

case study: Automobile Insurance Fraud

Claim prediction

- Model: predict whether an insurance claim is fraudulent or not
- Use: assign likely fraudulent claims to investigation team
- Benefit: investigation team's time allocated better, reducing $ lost to fraud

Member prediction

case study: Automobile Insurance Fraud

Claim prediction

- Model: predict whether an insurance claim is fraudulent or not
- Use: assign likely fraudulent claims to investigation team
- Benefit: investigation team's time allocated better, reducing $ lost to fraud

Member prediction

- Model: predict if a member will commit fraud in near future

case study: Automobile Insurance Fraud

Claim prediction

- Model: predict whether an insurance claim is fraudulent or not
- Use: assign likely fraudulent claims to investigation team
- Benefit: investigation team's time allocated better, reducing $ lost to fraud

Member prediction

- Model: predict if a member will commit fraud in near future
- Use: risk mitigation action, e.g. contact members with warning

case study: Automobile Insurance Fraud

Claim prediction

- Model: predict whether an insurance claim is fraudulent or not
- Use: assign likely fraudulent claims to investigation team
- Benefit: investigation team's time allocated better, reducing $ lost to fraud

Member prediction

- Model: predict if a member will commit fraud in near future
- Use: risk mitigation action, e.g. contact members with warning
- Benefit: potentially stop fraud before it happens – save significant $

case study: Automobile Insurance Fraud

Application prediction

case study: Automobile Insurance Fraud

Application prediction

- Model: predict, at time of application, if member will commit fraud ever

case study: Automobile Insurance Fraud

Application prediction

- Model: predict, at time of application, if member will commit fraud ever
- Use: reject applications with high likelihood of fraud

case study: Automobile Insurance Fraud

Application prediction

- Model: predict, at time of application, if member will commit fraud ever
- Use: reject applications with high likelihood of fraud
- Benefit: potentially stop fraud before it happens – save significant $

case study: Automobile Insurance Fraud

Application prediction

- Model: predict, at time of application, if member will commit fraud ever
- Use: reject applications with high likelihood of fraud
- Benefit: potentially stop fraud before it happens – save significant $

case study: Automobile Insurance Fraud

Application prediction

- Model: predict, at time of application, if member will commit fraud ever
- Use: reject applications with high likelihood of fraud
- Benefit: potentially stop fraud before it happens – save significant $

Payment prediction

case study: Automobile Insurance Fraud

Application prediction

- Model: predict, at time of application, if member will commit fraud ever
- Use: reject applications with high likelihood of fraud
- Benefit: potentially stop fraud before it happens – save significant $

Payment prediction

- Model: predict the appropriate amount of money that should be paid

case study: Automobile Insurance Fraud

Application prediction

- Model: predict, at time of application, if member will commit fraud ever
- Use: reject applications with high likelihood of fraud
- Benefit: potentially stop fraud before it happens – save significant $

Payment prediction

- Model: predict the appropriate amount of money that should be paid
- Use: predicted amount used to offer settlement to members

case study: Automobile Insurance Fraud

Application prediction

- Model: predict, at time of application, if member will commit fraud ever
- Use: reject applications with high likelihood of fraud
- Benefit: potentially stop fraud before it happens – save significant $

Payment prediction

- Model: predict the appropriate amount of money that should be paid
- Use: predicted amount used to offer settlement to members
- Benefit: limits the effort/expense of undergoing claim investigations

case study: Automobile Insurance Fraud

Each also has it downsides...

- **Claim** prediction: what if investigative team does not agree, e.g., the workload is too little, too much, too different
- **Member** prediction: if you issue a warning to a good customer, you may lose that customer
- **Application** prediction: how much future revenue are you giving up from good customers?
- **Payment** prediction: settlements might be too high

case study: Automobile Insurance Fraud

Each also has it downsides...

- **Claim** prediction: what if investigative team does not agree, e.g., the workload is too little, too much, too different
- **Member** prediction: if you issue a warning to a good customer, you may lose that customer
- **Application** prediction: how much future revenue are you giving up from good customers?
- **Payment** prediction: settlements might be too high

case study: Automobile Insurance Fraud

Each also has it downsides...

- **Claim** prediction: what if investigative team does not agree, e.g., the workload is too little, too much, too different
- **Member** prediction: if you issue a warning to a good customer, you may lose that customer
- **Application** prediction: how much future revenue are you giving up from good customers?
- **Payment** prediction: settlements might be too high

case study: Automobile Insurance Fraud

Each also has it downsides...

- **Claim** prediction: what if investigative team does not agree, e.g., the workload is too little, too much, too different
- **Member** prediction: if you issue a warning to a good customer, you may lose that customer
- **Application** prediction: how much future revenue are you giving up from good customers?
- **Payment** prediction: settlements might be too high

objectives and success

- The objective of the project should be clearly defined.

objectives and success

- The objective of the project should be clearly defined.
- Criteria to measure the success of the project should be defined.

objectives and success

Poor examples...

objectives and success

Poor examples...

- Increase sales

objectives and success

Poor examples...

- Increase sales
- Model loyal customers to increase sales

objectives and success

Poor examples...

- Increase sales
- Model loyal customers to increase sales
- “I read something in Forbes magazine – Walmart is doing something where they found that purchases of fishing tackle in late October is indicative of very profitable sales of decorative lamps in the Spring.” *Do that for us.*

objectives and success

Good example...

objective:	increase revenues (per customer) in direct mailing campaigns by personalized offer and individual customer selection
deliverable:	application that automatically selects a pre-specified number of customers from the database to whom the mailing shall be sent; runtime max: half-day
success criteria:	improve purchase rate by 5% or total revenues by 5%, measured within 4 weeks after mailing.

assess situation

assess situation

Task:

Detailed fact-finding

assess situation

Task:

Detailed fact-finding

Outputs

- Inventory of resources
- Requirements, assumptions, and constraints
- Terminology
- Costs and benefits

assess situation

Task:

Detailed fact-finding

Outputs

- Inventory of resources
- Requirements, assumptions, and constraints
- Terminology
- Costs and benefits

assess situation

Task:

Detailed fact-finding

Outputs

- Inventory of resources
- Requirements, assumptions, and constraints
- Terminology
- Costs and benefits

assess situation

Task:

Detailed fact-finding

Outputs

- Inventory of resources
- Requirements, assumptions, and constraints
- Terminology
- Costs and benefits

determine data mining goals

Task:

Map the problem to a data analysis task

determine data mining goals

Task:

Map the problem to a data analysis task **Output:**

- Data mining goals
- Data mining success criteria

determine data mining goals

Task:
Map the problem to a data analysis task **Output:**

- Data mining goals
- Data mining success criteria

produce project plan

Task:

Describe and document plan

produce project plan

Task:

Describe and document plan

Output:

- Project plan

  This is a *dynamic* document!

- Initial assessment of tools and techniques

produce project plan

Task:

Describe and document plan

Output:

- Project plan

  This is a *dynamic* document!

- Initial assessment of tools and techniques

produce project plan

Task:

Describe and document plan

Output:

- Project plan

  This is a *dynamic* document!

- Initial assessment of tools and techniques

project understanding

"A problem well stated is a problem half solved."

– Charles Kettering