CRISP-DM

Overview

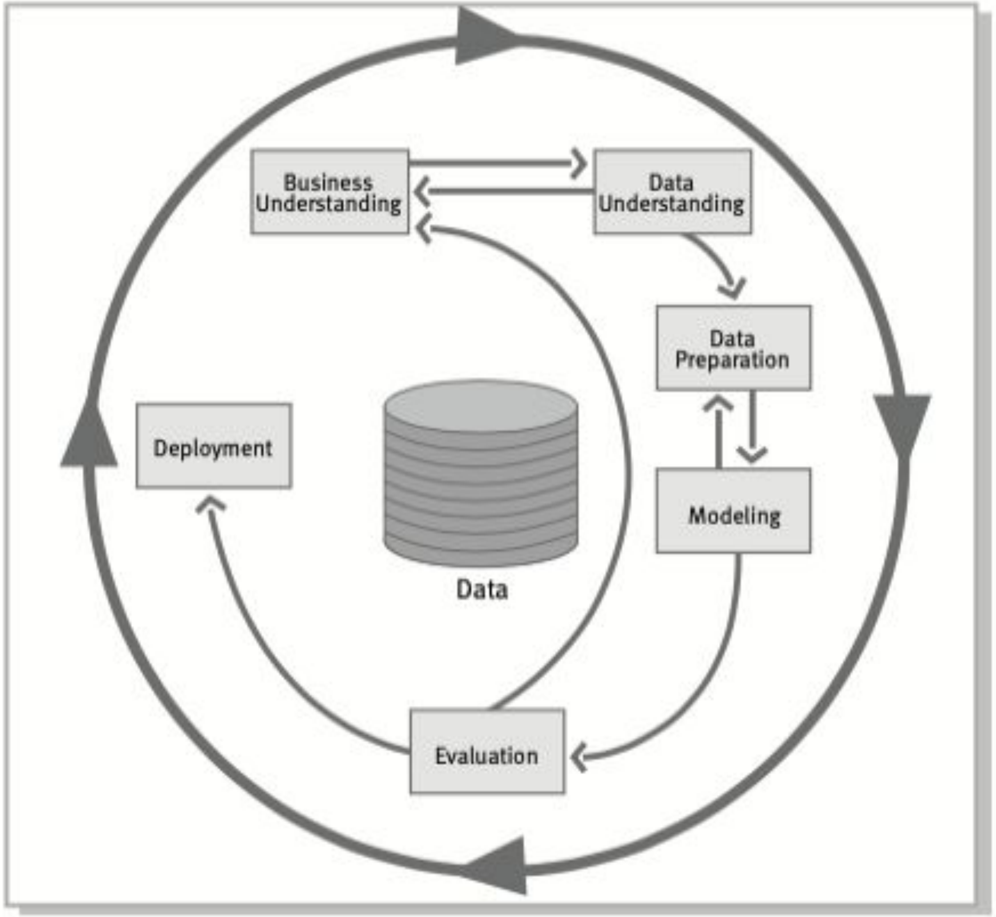


Figure 2: Phases of the CRISP-DM reference model

Source: CRISP-DM
step-by-step miniing guide,
SPSS

Overview

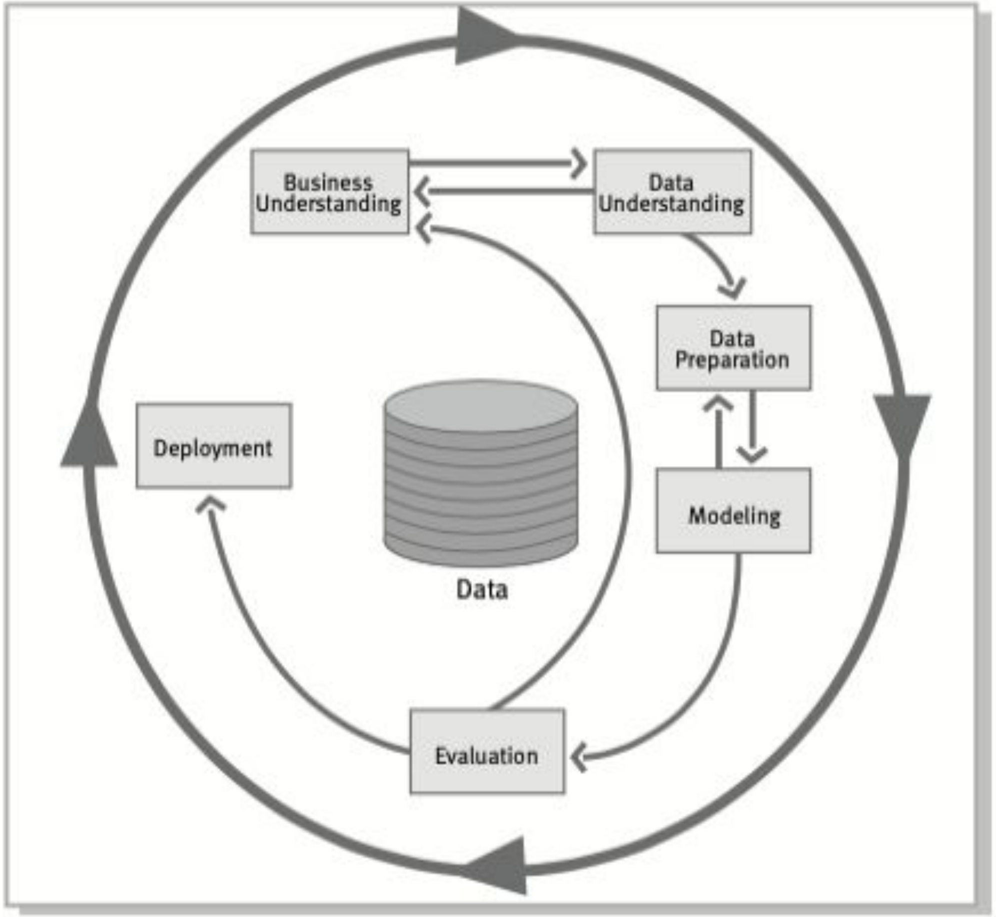


Figure 2: Phases of the CRISP-DM reference model

Overview

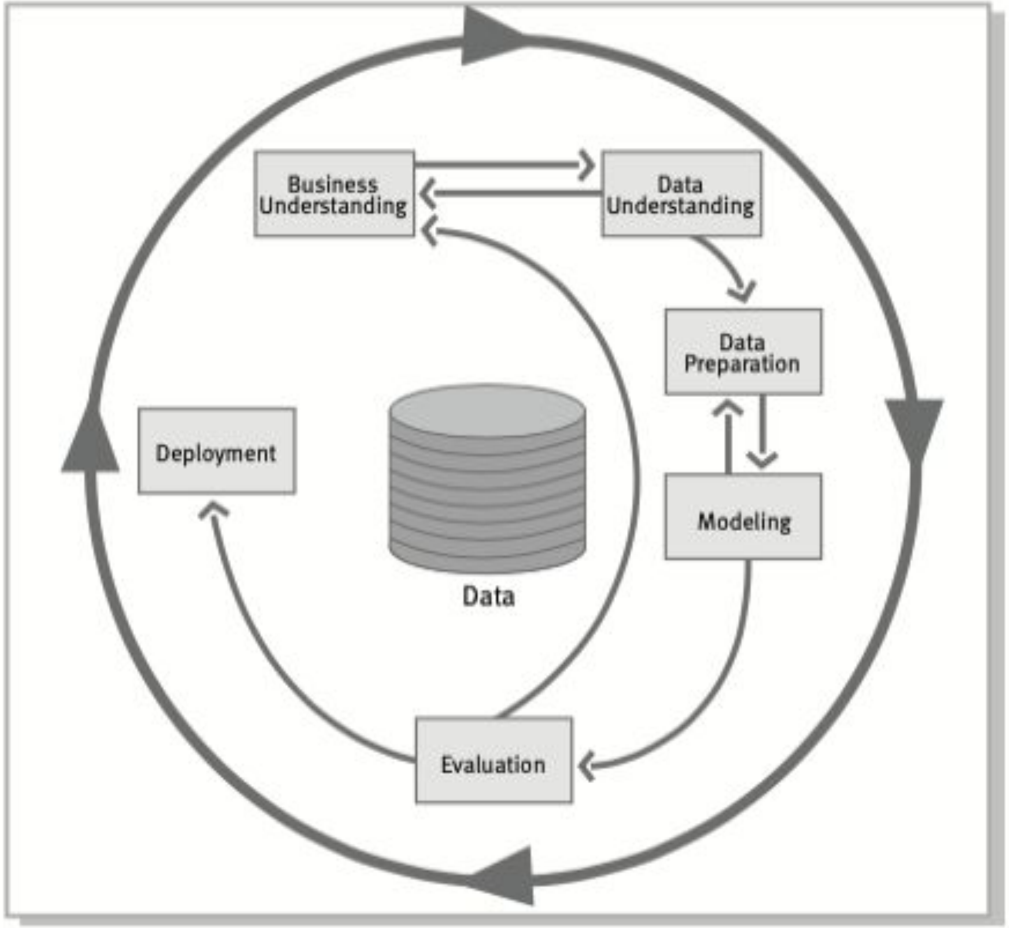
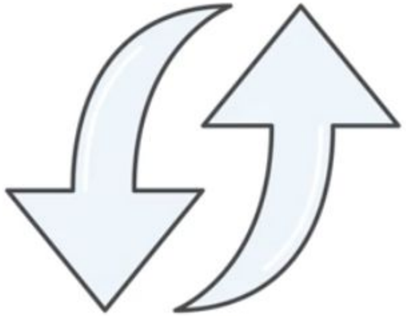


Figure 2: Phases of the CRISP-DM reference model

Phase 1: Business Understanding

Understanding business requirements

Analyzing supporting information



Converting to a Data Mining problem

Preparing a preliminary plan

Phase 1: Business Understanding

Understanding Business Requirements

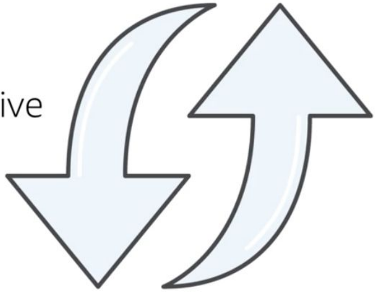
- Understand business requirements
- Form a business question
- Highlight project's critical features

Analysing supporting information

- List required resources and assumptions
- Analyze associated risks
- Plan for contingencies
- Compare costs and benefits

Converting to a data mining Problem

- Review machine learning question
- Create technical data mining objective
- Define the criteria for successful outcome of the project



Preparing a preliminary plan

- Number and duration of stages
- Dependencies
- Risks
- Goals
- Evaluation methods
- Tools and techniques

Business Understanding – Case
- Titanic dataset challenge

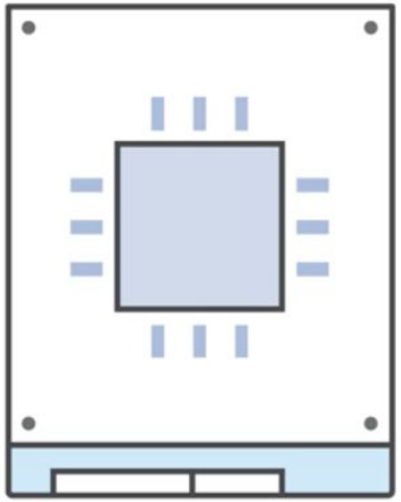
❑ Business question

- What factors was associated with a person survive in the Titanic disaster?

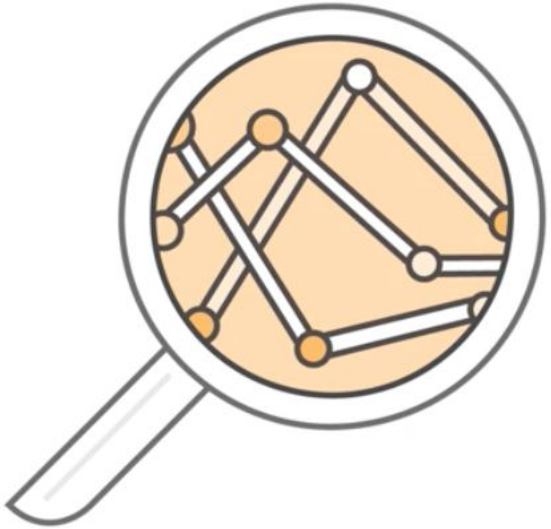
❑ Features:

1. How to analyze the data?
   1. Language: Python,
   2. Data preparation: Pandas
   3. Data visualisation: matplotlib; seaborn
   4. Data modeling & evaluation: sklearn
2. What were the passager types (Ages, Gender, Class, etc)
3. What factors helped survive the sink?

Phase 2: Data Understanding



Data collection

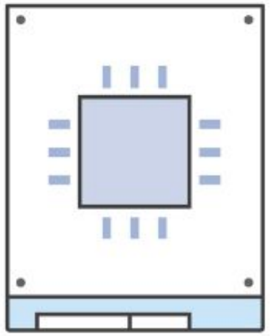


Data properties

Quality

Phase 2: Data Understanding

Data collection



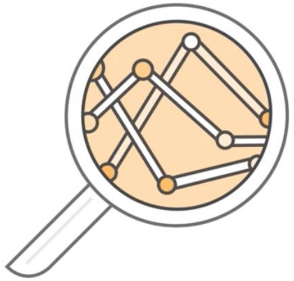
- Detail various sources and steps to extract data
- Analyze data for additional requirements
- Consider other data sources

Quality

- Verifying attributes
- Identifying missing data
- Reveal inconsistencies
- Report solution

Data properties

- Describe the data, amount of data used, and metadata properties
- Find key features and relationships in the data
- Use tools and techniques to explore data properties



Data Understanding - Case

- Titanic dataset challenge

- Study the data dictionary (whenever available)
- Inspect the dataset

```python
import pandas as pd
import requests
import io

dataset = 'https://gist.githubusercontent.com/michhar/2dfd2de0d4f8727f873422c5d959fff5/raw/fa71405126017e6a37bea592440b4bee94bf7b9e/titanic.csv'

csv_raw = requests.get(dataset).content
csv_io =  io.StringIO(csv.decode('utf8'))

df = pd.read_csv(csv_io)
df.head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Data Dictionary

test.csv

- Data Source - [http://www.kaggle.com/c/titanic-gettingStarted/data]
- Data Information - Test data for Kaggle Titanic introductory comp.

train.csv

- Data Source - [http://www.kaggle.com/c/titanic-gettingStarted/data]
- Data Information - Training data for Kaggle Titanic introductory comp.

Data Variables (Test / Train)

Describes the variables in the test / train .csv files. This data dictionary and subsequent info was obtained from Kaggle.

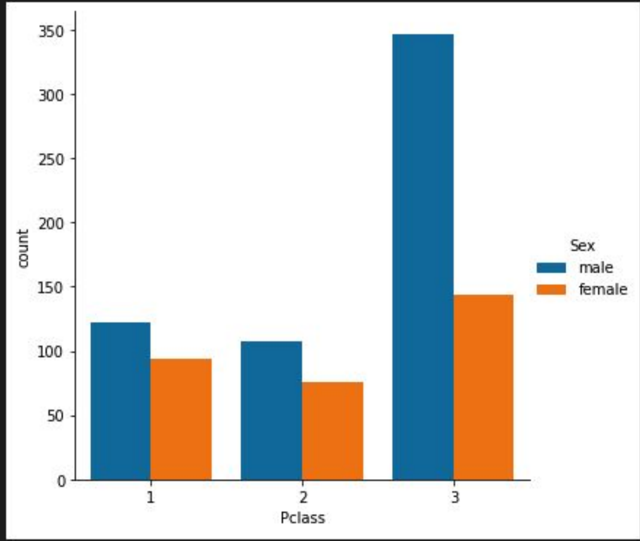
Variable	Description	Details
survival	Survival	0 = No; 1 = Yes
pclass	Passenger Class	1 = 1st; 2 = 2nd; 3 = 3rd
name	First and Last Name	
sex	Sex	
age	Age	
sibsp	Number of Siblings/Spouses Aboard	
parch	Number of Parents/Children Aboard	
ticket	Ticket Number	
fare	Passenger Fare	
cabin	Cabin	
embarked	Port of Embarkation	C = Cherbourg; Q = Queenstown; S = Southampton

Data Understanding - Case
- Titanic dataset challenge

- Identify missing data
- Detect key associations
- Analyse empty cells

```
sns.catplot('Pclass', data=df,kind='count', hue='Sex')
```

<seaborn.axisgrid.FacetGrid at 0x7fb2343b20d0>




```
def empty(x): return x.isnull().sum()
df.agg(['dtype','count','nunique',empty])
```

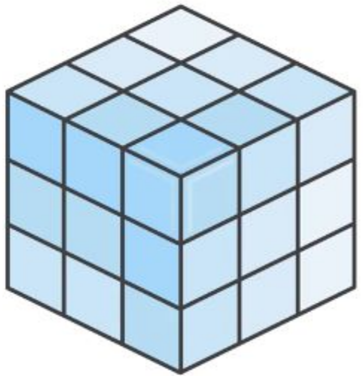
	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Title
dtype	int64	int64	int64	object	object	float64	int64	int64	object	float64	object	object	object
count	891	891	891	891	891	714	891	891	891	891	204	889	891
nunique	891	2	3	891	2	88	7	7	681	248	147	3	5
empty	0	0	0	0	0	177	0	0	0	0	687	2	0

Phase 3: Data preparation

Final dataset selection

Analyze constraints:

- Total size
- Included and excluded columns
- Record selection
- Data type



Data preparation

1. Cleaning
2. Transforming
3. Merging
4. Formatting

Phase 3: Data preparation

Cleaning

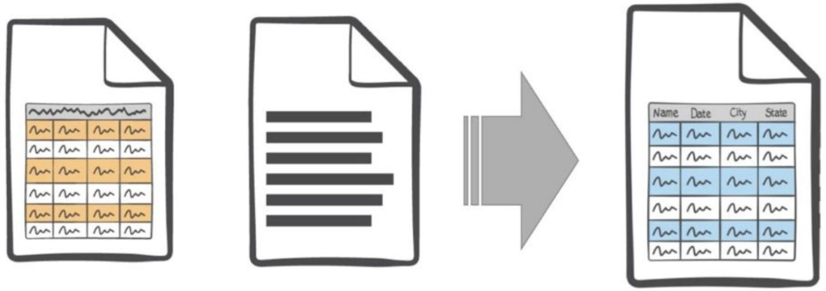
- How is the missing data handled?
  - Dropping rows with missing values
  - Adding a default value or a mean value for missing data
  - Using statistical methods to calculate the value (e.g., regression)
- Clean attributes with corrupt data or variable noise.

Transformation

- Derive additional attributes from the original attributes
- Normalization
- Attribute transformation

YES	1
YES	1
NO	0
YES	1
NO	0

Merging

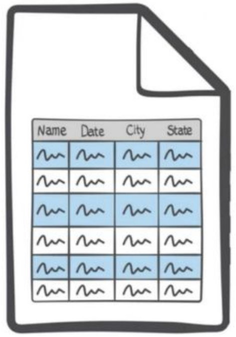


Recommended: Revisit the Data Understanding phase afterward.

Formatting

Format for modeling tool needs:

- Rearrange attributes
- Randomly shuffle data
- Remove constraints of the modeling tool (e.g. removing Unicode characters)



Data preparation - Case

- Select columns and variables:
    - Do we drop rows? Which?
- Drop columns: which factors?
- Treat empty values: when? How?

```python
def empty(x): return x.isnull().sum()
df.agg(['dtype','count','nunique',empty])
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Title
dtype	int64	int64	int64	object	object	float64	int64	int64	object	float64	object	object	object
count	891	891	891	891	891	714	891	891	891	891	204	889	891
nunique	891	2	3	891	2	88	7	7	681	248	147	3	5
empty	0	0	0	0	0	177	0	0	0	0	687	2	0

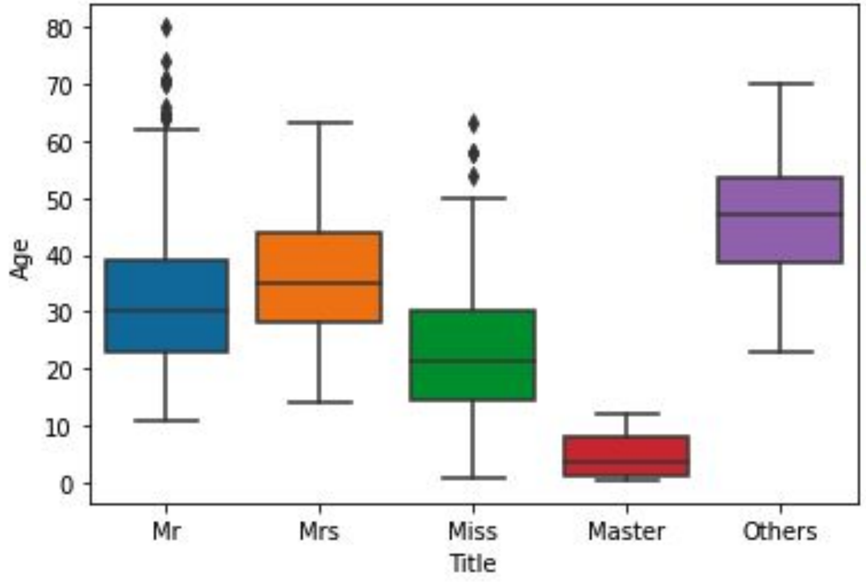
Data preparation - Case

Ex: filling Age with median of title

- Verify that the age does not vary much per title
- Some titles with small occurrence: Others

```
AgeMedian_by_titles = df.groupby('Title')['Age'].median()
AgeMedian_by_titles
```

```
Title
Master     3.5
Miss      21.5
Mr        30.0
Mrs       35.0
Others    47.0
Name: Age, dtype: float64
```



Data preparation - Case

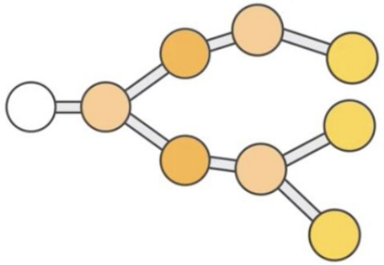
Ex:

- Getting dummies: embarked
- Boolean (sex): turn into 0/1

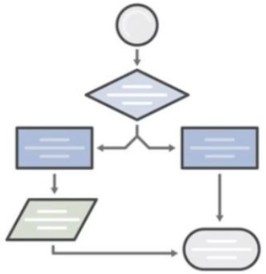
```
df['Sex'] = df.Sex.map({'male': 1, 'female':0})

df_final = pd.get_dummies(df[[ 'Survived','Pclass',
    'Sex', 'Age','SibSp', 'Parch', 'Fare', 'Embarked']]).fillna(0)
df_final
```

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked_C	Embarked_Q	Embarked_S
0	0	3	0.0	22.0	1	0	7.2500	0	0	1
1	1	1	0.0	38.0	1	0	71.2833	1	0	0
2	1	3	0.0	26.0	0	0	7.9250	0	0	1
3	1	1	0.0	35.0	1	0	53.1000	0	0	1
4	0	3	0.0	35.0	0	0	8.0500	0	0	1

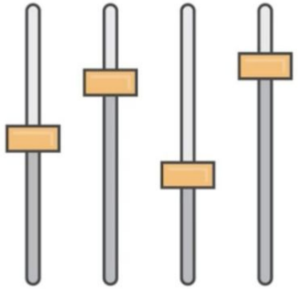
Phase 4: Modeling



Model selection
and creation

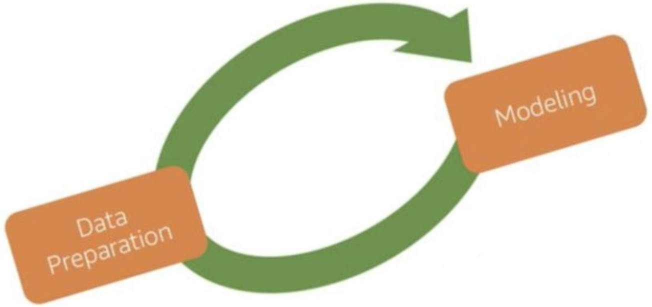


Model testing
plan



Parameter
tuning/testing

Works together with the Data Preparation phase

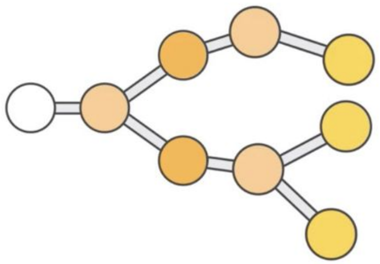


Phase 4: Modeling

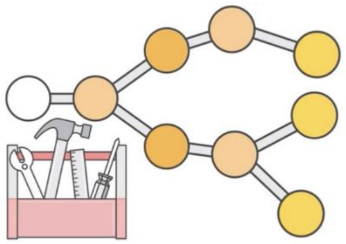
Model selection

Identify:

- Modeling technique
- Constraints of modeling technique and tool
- Ways in which constraints tie back to Data Preparation phase



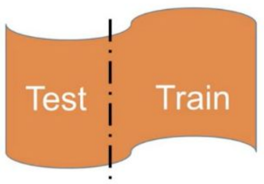
Building the Model



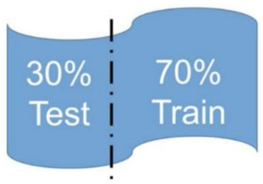
- Train the model
- Tweak the model for better performance
- Build multiple models with different parameter settings
- Describe the trained models and report on the findings

Generating a Model Testing Plan

Before training your model:



Test model effectiveness



Dataset split

Model evaluation criterion

Modeling - Case

Model selection:

- Is Linear regression the best model?
- Which other models could we use?
  - decision tree
- Which columns should we select?

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix
from sklearn.metrics import precision_recall_fscore_support
```

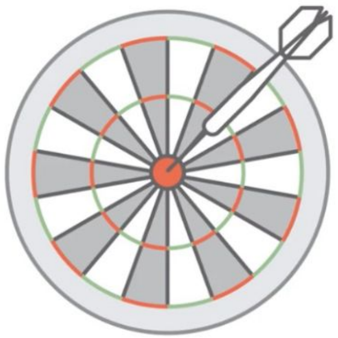
```
X = df_final.drop('Survived', axis = 1)
y = df_final['Survived']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.33, random_state = 0)

log_reg = LogisticRegression()
log_reg.fit(X_train, y_train)
y_pred = log_reg.predict(X_test)
```

Phase 5: Model evaluation

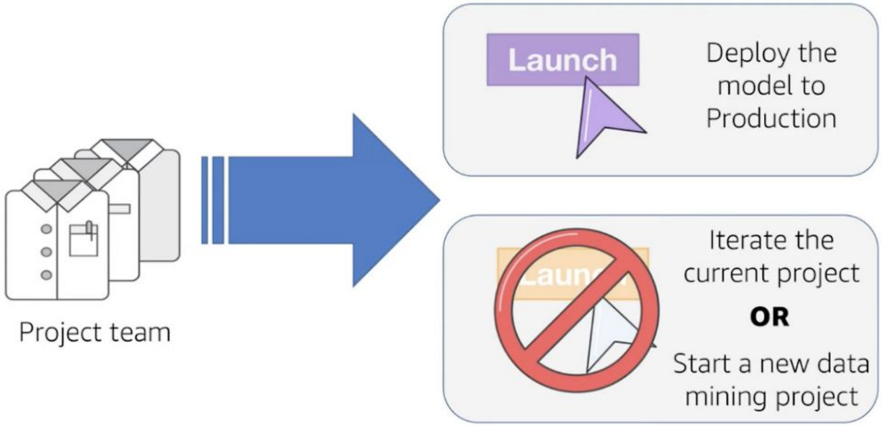
Evaluate the Model

Evaluation depends on:

- Accuracy of the model
- Model generalization on unseen/unknown data
- Evaluation of the model using the business success criteria



Determine the Next Steps



Review the Project

Assess the steps taken in each phase

- Was any important criteria overlooked?

Perform quality assurance checks

- Model performance using the determined data
- Is the data available for future training?

Phase 5: Model evaluation

Results:

```
metrics = precision_recall_fscore_support(
    y_pred, y_test, average = 'macro')
metrics = list(metrics)
print(confusion_matrix(y_pred, y_test),
    '''

    precision: {0},
    recall: {1},
    f_score: {2}'''.format(*metrics[:3]))
```

```
[[160  64]
 [ 24  47]]

  precision: 0.6464943204073639,
  recall: 0.688128772635815,
  f_score: 0.6503986209868562
```

```
metrics = precision_recall_fscore_support(
    y_pred, y_test, average = 'macro')
metrics = list(metrics)
print(confusion_matrix(y_pred, y_test),
    '''

    precision: {0},
    recall: {1},
    f_score: {2}'''.format(*metrics[:3]))
```

```
[[156  58]
 [ 28  53]]

  precision: 0.6626517822169996,
  recall: 0.6916464751355718,
  f_score: 0.6680014656616415
```

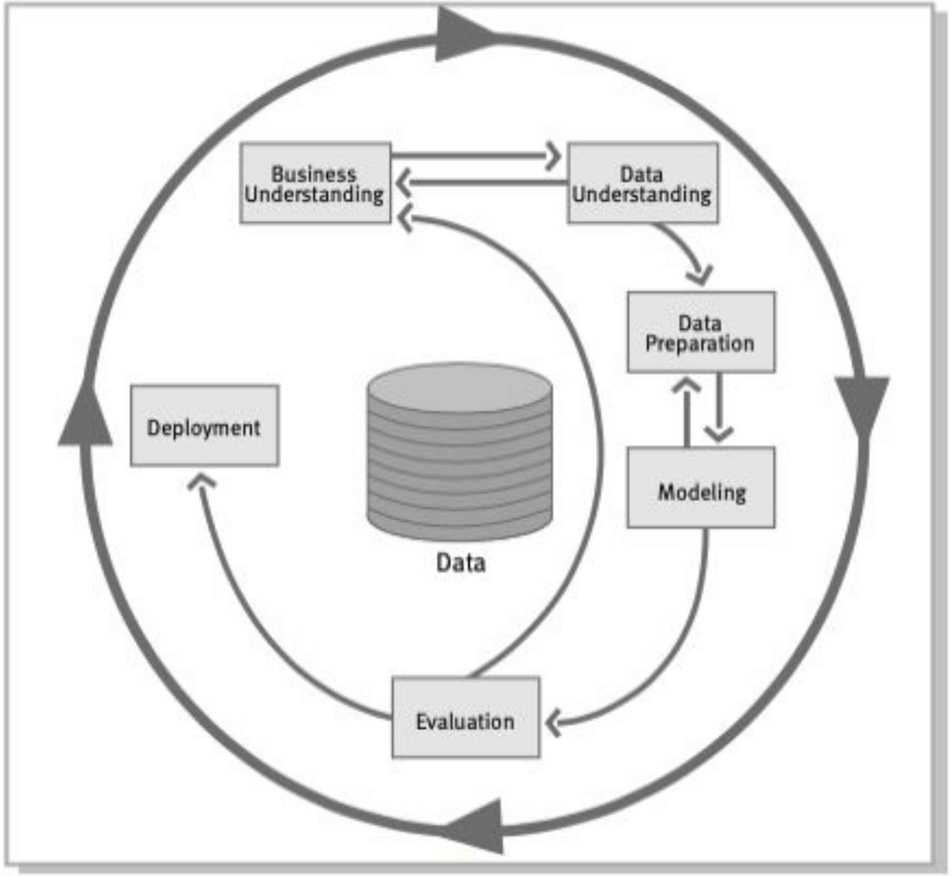


Figure 2: Phases of the CRISP-DM reference model

Phase 6: Deployment

Workflow

Phase 6: Deployment

Runtime

IBM Cloud Functions Apache OpenWhisk

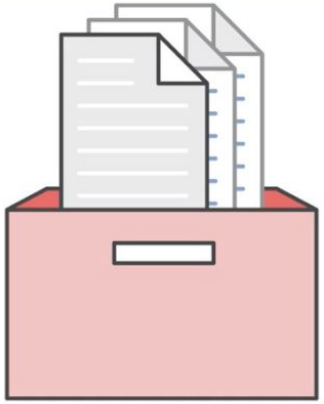
CLOUDFOUNDRY kubernetes

Management and Monitoring

mlflow

Kubeflow

Project review



Assess the outcomes of the project:

- Summarize results and write thorough documentation
  - Common pitfalls
  - Choosing the right ML solution
- Generalize the whole process to make it useful for the next iteration

Final report

- Highlight processes used in the project
- Analyze if all the goals for the project were met
- Detail the findings of the project
- Identify and explain the model used and reason behind using the model
- Identify the customer groups to target using this model

references

- https://medium.com/swlh/my-perspective-on-eda-for-the-titanic-dataset-38a4ecc94500

- https://www.aws.training/Details/eLearning?id=27200