GBUS 738 Data Mining

Introduction to CRISP-DM

David Svancer – George Mason University School of Business

GEORGE MASON UNIVERSITY
School of Business

Data Mining Steps

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

The CRISP-DM methodology was designed specifically for data mining but is used for most data science/analytical projects. The steps include:

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

Image: https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

Data Mining Steps

CRISP-DM – Business Understanding

The **Business Understanding** step is focused on understanding project requirements from a business standpoint. It includes the following steps:

Identify Business Objective

Identifying the business objective

a) What problem is the business trying to solve?

- For example, customers are spending to much time on a company's website searching for content
- This is leading to customers leaving the website without purchasing products
- An objective may be to find a way to display relevant content to the customer based on their website activity, leading to increased sales

Data Mining Steps

CRISP-DM – Business Understanding

The **Business Understanding** step is focused on understanding project requirements from a business standpoint.

Determining the analytical goals

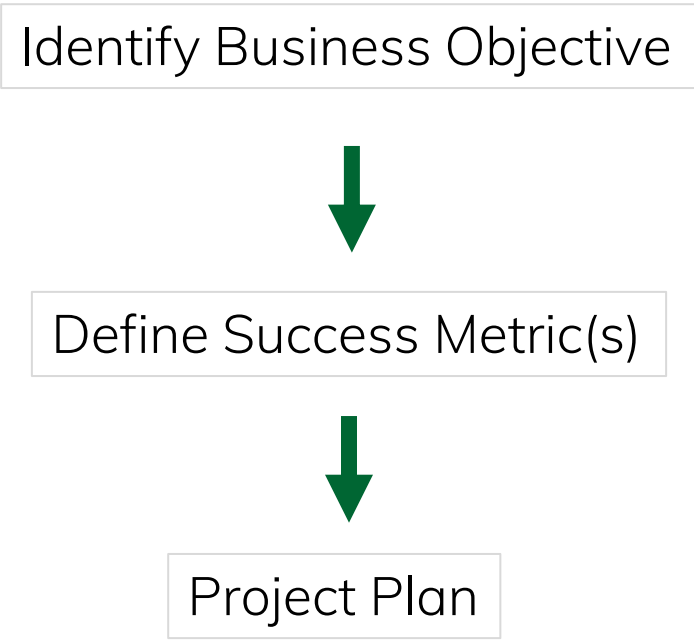
a) What does success look like?

- Decreased average time to product purchase?

b) Which data mining algorithms can help the business meet its objectives? Depends on the goal:

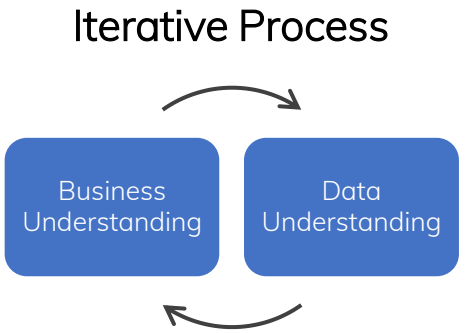
- **Inference** – determining what factors drive customers to purchase products and generalizing this to a population
- **Prediction** – being able to predict relevant content with high accuracy
- Some combination of both

Producing a project plan



Data Mining Steps

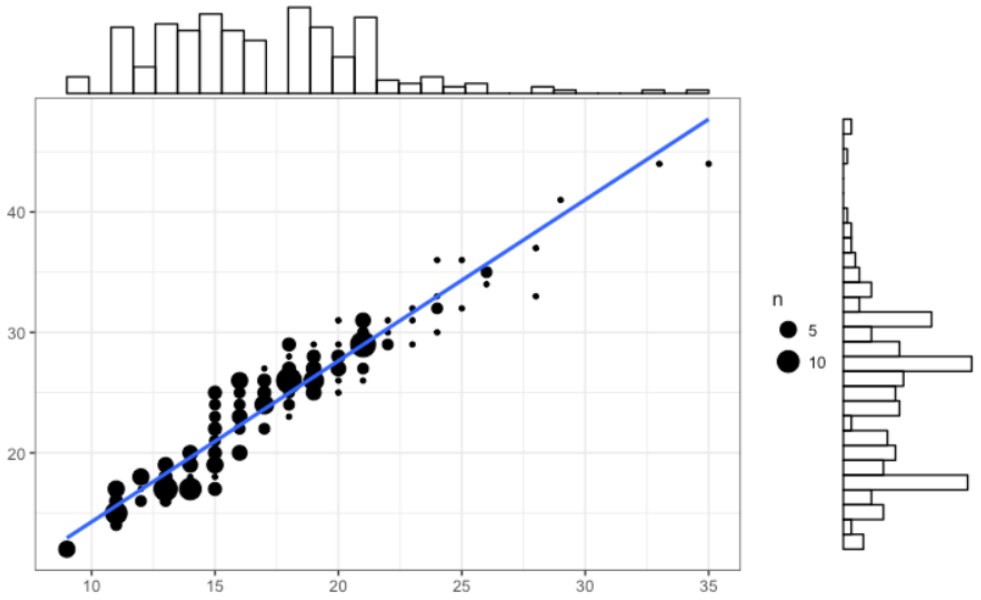
CRISP-DM – Data Understanding



The **Data Understanding** step is focused on collecting, describing, and exploring data that may be useful in solving the business problem. It also includes verifying the quality of the data

In this stage, you are getting familiar with the data through exploratory data analysis (EDA)

- Exploration of variables in the data set with summary statistics, histograms, boxplots
- Looking for associations among variables
- Data visualization to study complex relationships
- Transforming variables to create new features/characteristics that might improve the performance of machine learning algorithms



School of Business

Data Mining Steps

CRISP-DM – Data Preparation

The **Data Preparation** includes all steps in the process of converting raw data into **a numeric data matrix** for machine learning applications

This is sometimes referred to as *ETL* (Extract-Transform-Load). To accomplish this task, analysts might perform some or all the following tasks:

1. Import and clean the raw data
   - Clean unstructured text fields, format numeric variables, transform numeric variables
2. Constructing the final data set
   - Writing the code, whether in SQL or any other software language, that produces the final data set from the raw data
3. Integrating the data into a production environment

Raw Data

Purchased	Age	Metro Area	Amount	LinkedIn URL
No	29	D.C.	$0	Linkedin.com/hillary
Yes	67	Baltimore	$50	Linkedin.com/matt
Yes	67	D.C.	$30	Linkedin.com/mike
No	37	Boston	$0	*NULL*
No	41	Pittsburgh	$0	Linkedin.com/sara

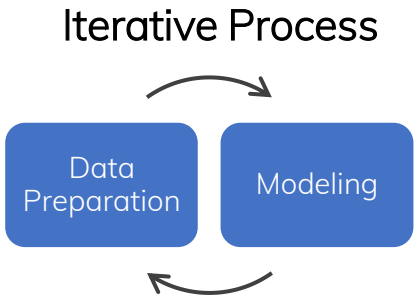
Outcome

Purchased	Age	Metro_Boston	Metro_DC	Amount	Education_MS
0	29	0	1	0.0	0
1	67	0	0	50.0	1
1	67	0	1	30.0	0
0	37	1	0	0.0	0
0	41	0	0	0.0	1

Predictors

Data Mining Steps

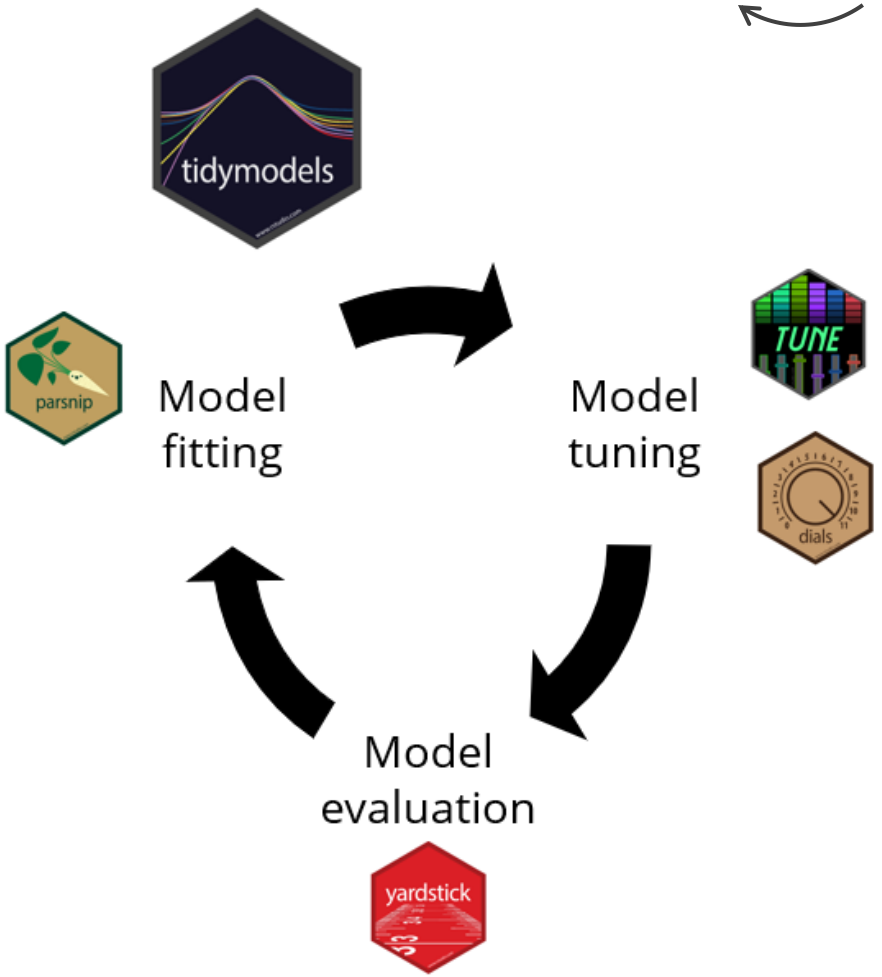
CRISP-DM – Modeling



Modeling

Develop, train, and assess machine learning algorithms to determine which work best for solving the business problem

- All candidate machine learning algorithms are trained to optimize the business objective
- Having a clear business objective that is ***quantifiable*** is critical before proceeding with the Modeling stage



Data Mining Steps

CRISP-DM – Evaluation

The **Evaluation** step is concerned with determining whether the machine learning algorithm meets the business objective on new data

A learning algorithm should be tested on new, previously unseen data

- The algorithm was trained to optimize the business objective on the training data
    - Therefore, our ***training error*** is an ***overly optimistic*** estimate of what our future error will be on new data
- In this step, we find out if the algorithm generalizes well to new incoming data and produces acceptable results
- From our earlier example
    - Does our final model lead to increased product purchases on new data?

GEORGE MASON UNIVERSITY
School of Business

Data Mining Steps

CRISP-DM – Deployment

In the **Deployment** phase, the final model is deployed throughout the business

- This may be as simple as producing monthly dashboards with model results to drive business decisions
- It may be the complex process of deploying a machine learning algorithm in a production environment, such as Microsoft Azure or AWS, where the results of the algorithm must be stored and passed between several business applications with customer interactions in real-time

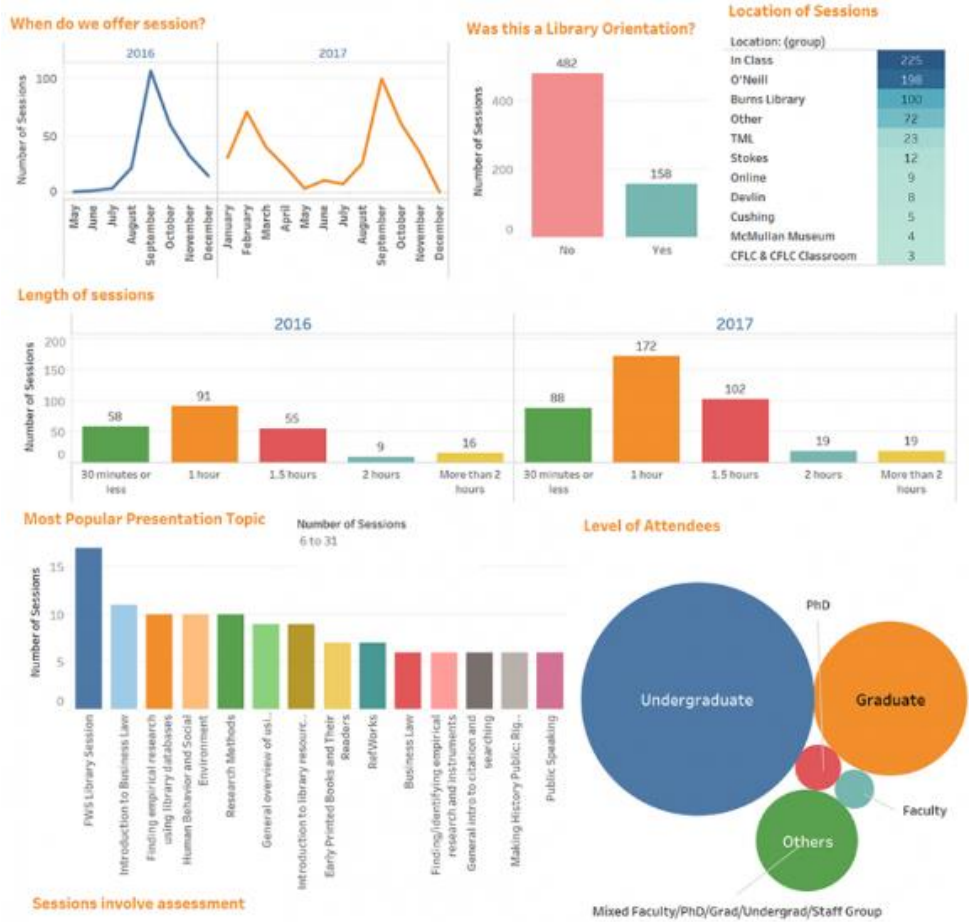


Image: Mauricio Vacas, https://www.kdnuggets.com/2017/04/models-from-lab-factory.html

Model Maintenance

Monitoring and Data Governance

Once the model is deployed into production

- Routines are developed to re-train the model as new data is available
  - Ensures new customer activity is continuously captured by the machine learning algorithm
- Governance
  - Strict rules must be put in place to ensure the raw data is not unknowingly altered, resulting in a crash of the machine learning system
  - Legal governance of model output and results
    - GDPR and customer privacy compliance
    - Monitor for potential algorithmic bias (Gender, Age, Ethnicity) that may put the company at legal risk