2 Project stages

Three lenses of due diligence

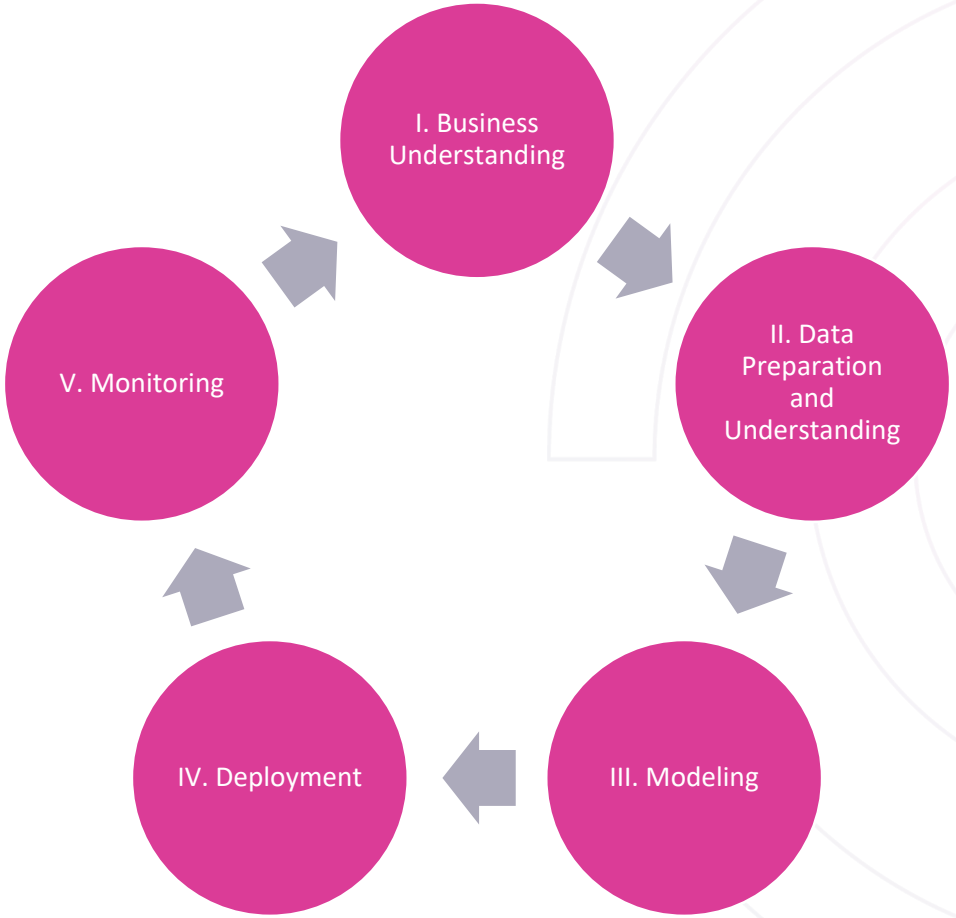
- Due diligence before the start of any project:

- Business viability: positive value creation
  - increase revenue, lower cost, boost efficiency or launch new business

- Technical feasibility: AI system can be built
  - meet desired performance, data availability, engineering timeline, etc.

- Human desirability: project is really wanted and ethically ok
  - AI developed/used by people

Data Science Life Cycle



I. Business Understanding

- Identify an opportunity to create value
  - Which part of your company process workflow can benefit from AI?

- How can AI help us?
  - Automate existing processes and facilitate human-machine collaboration
  - Improve existing algorithms to become more accurate or reliable
  - New business opportunities

- Pin down the project's goals
  - How is AI going to solve the problem?

II. Data Preparation

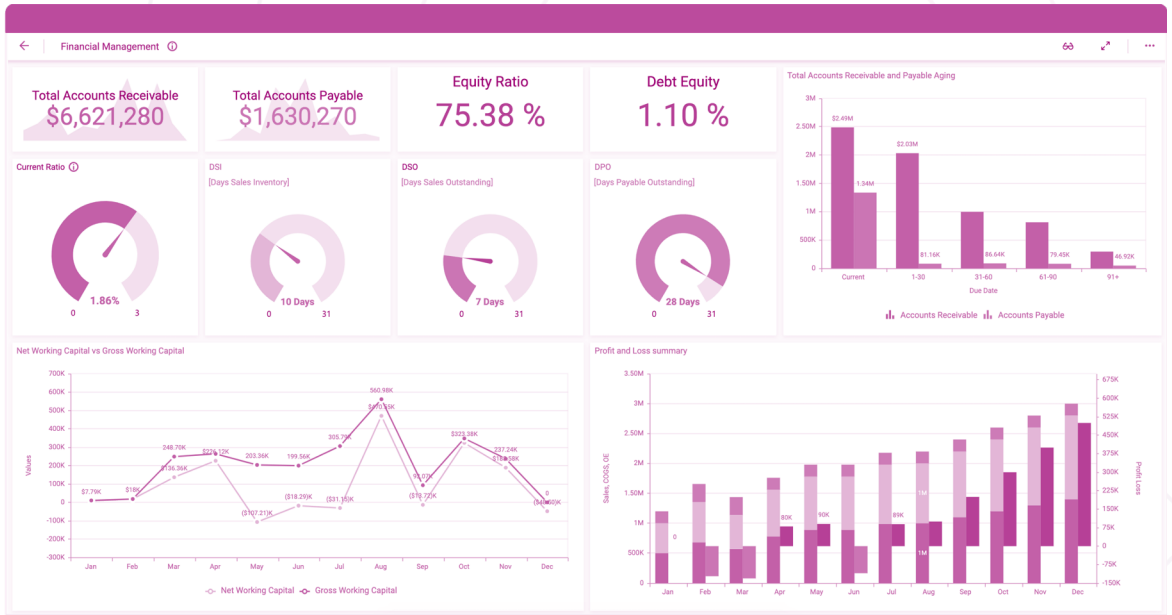
- Data collection
  - Internal & external data sources to acquire relevant and comprehensive data
  - Focus on data quality

- Data processing
  - Cleaning to deal with missing and inconsistent data
  - Preparation for modeling phase
  - Can be very time consuming

- Need a structured way to deal with data and centralize data flow

II. Data Understanding

- Summarize the main characteristics of the data set
- Represent dataset visually in a dashboard
- Understanding the patterns and bias in the data
- Gain insight into the data
- Assesses quality of the data



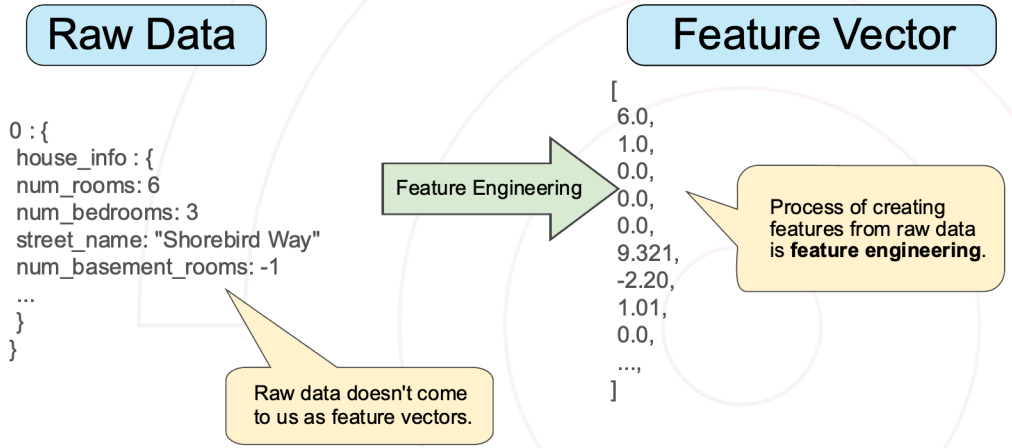
Key questions regarding data

- What kind of data do I have available?
- Where can I find it in my organization?
- Who owns the data and am I allowed to use it?
- What is the format/quality of the data?
- Can I trust my data?
- Possible to enrich own data with external data?

- Is this the right data to solve my business problem at hand?

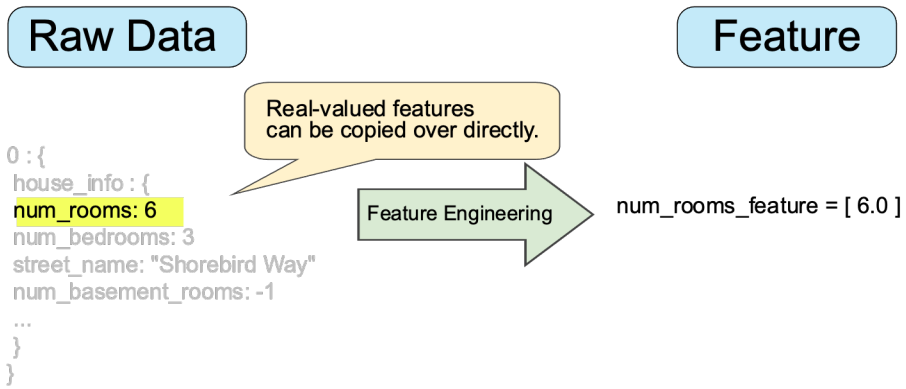
III. Modeling – Feature Engineering

- Transform raw data into usable features:
  - Google – ML Crash Course

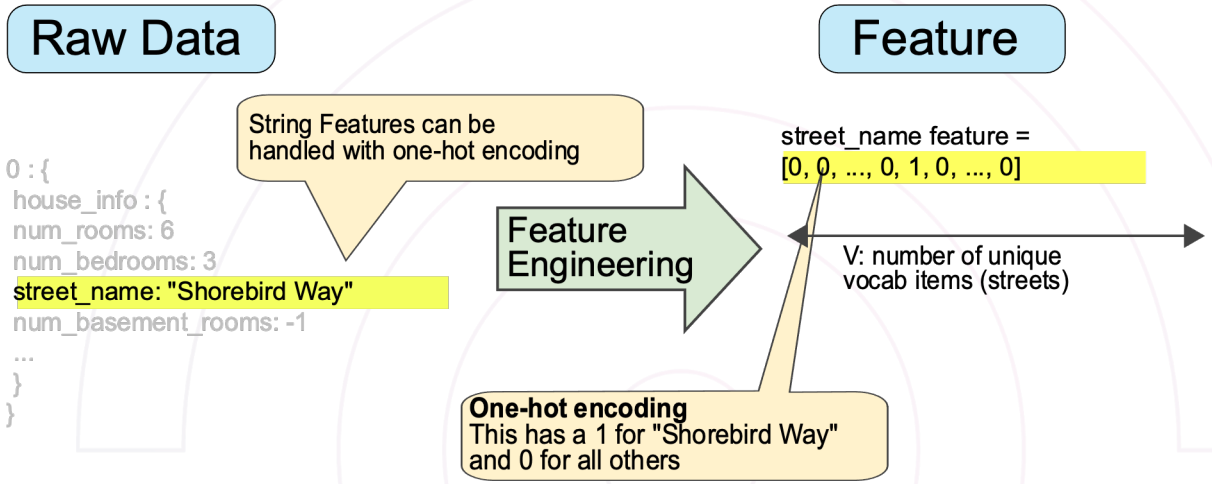


- Feature construction
  - creating new features from the ones that you already have
- Feature selection
  - Remove irrelevant features that add more noise than information

III. Modeling- Feature Engineering



Quantitative or numerical features

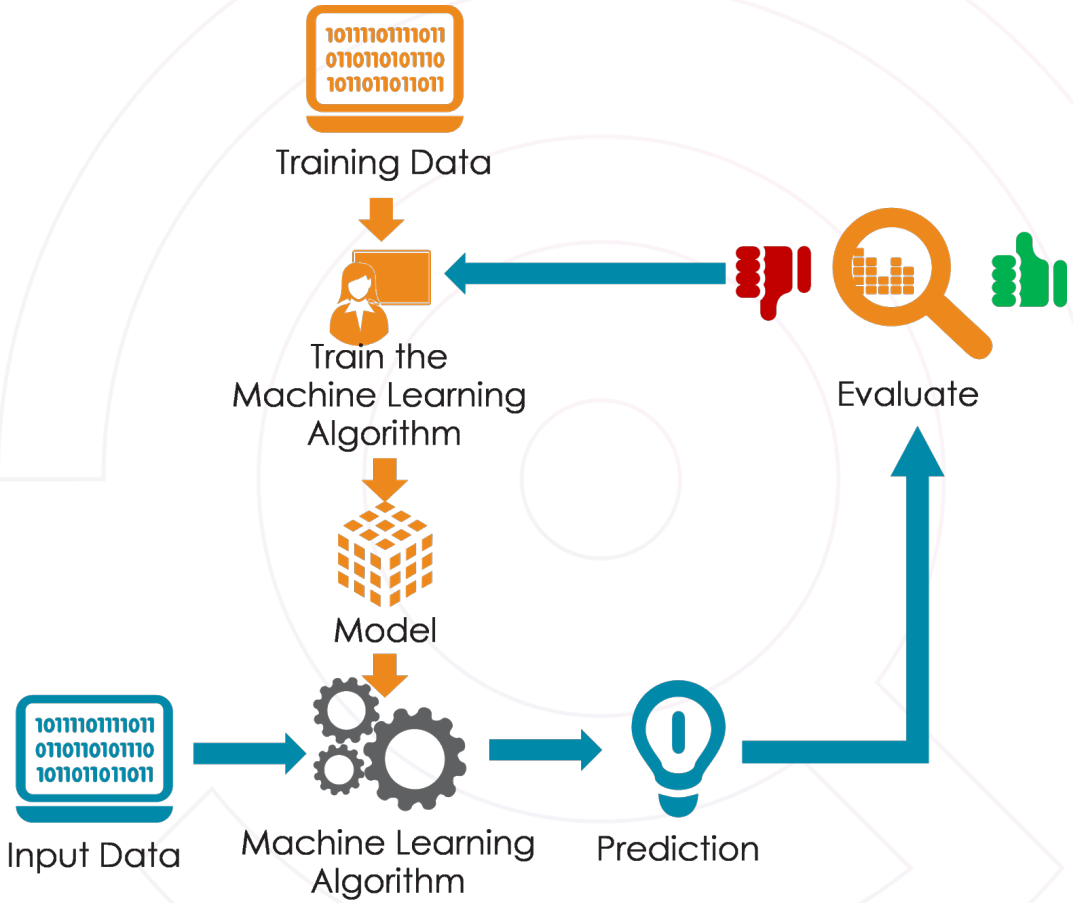


Qualitative or categorical features

Street	Feature vector
Charleston Road	[1,0,0,0]
North Shoreline Boulevard	[0,1,0,0]
Shorebird Way	[0,0,1,0]
Rengstorff Avenue	[0,0,0,1]

III. Modeling – Machine Learning

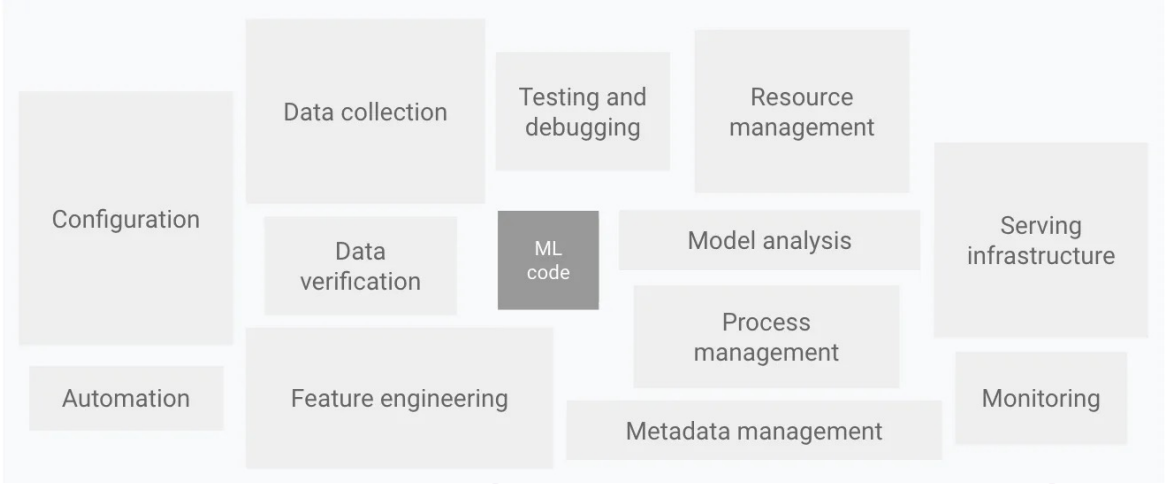
- Train various algorithms to develop models
- Evaluate model performance on new unseen data samples
- Ensure that the outcomes make sense and are significant
- Typically iterative process
  - Review model
  - Review data



IV. Deployment

- Integration of the ML model into an existing production environment
  - ML code only constitutes a tiny part of the full production architecture

- Done by MLOps engineer



Google Cloud - MLOps

- Proof of Concept (PoC) to production gap
  - Many researched ML solutions don’t see daylight

V. Monitoring

- The real world is constantly changing
  - Data drift: the distribution of input features changes (e.g., houses become smaller over time because of space scarceness)
  - Concept drift: the mapping from features to target changes (e.g., popularity for small houses makes these more expensive)

- Ensure that algorithms keep doing a good job once deployed
  - Constantly evaluate their performance with regard to a baseline
  - Identify degrading solutions early on

- Brainstorm statistics/metrics to track over time
  - Visualize in a dashboard
  - Set thresholds for alarms
  - Iterative process: adjust metrics + thresholds over time