Data & Things

(Spring 25)

Monday February 24

Lecture 11: Improving and selecting machine learning models

Jens Ulrik Hansen

Roskilde Universitet

Outline of this lecture

- Hand-in for the exam
- The case of today
- The data science process revisited
- The business problem and the data
- Pre-processing techniques
- The model training process
- Model evaluation
- Hyper-parameter tuning

Course practicalities

- **The exam**
  - Exam Dates:
    - **Hand-in of written product: Sunday March 16 at 10.00am on eksamen.ruc.dk**
    - **Oral exam: Thursday March 20 and Friday March 21**
  - Format: 20 min individual oral exam based on a written product

Course practicalities

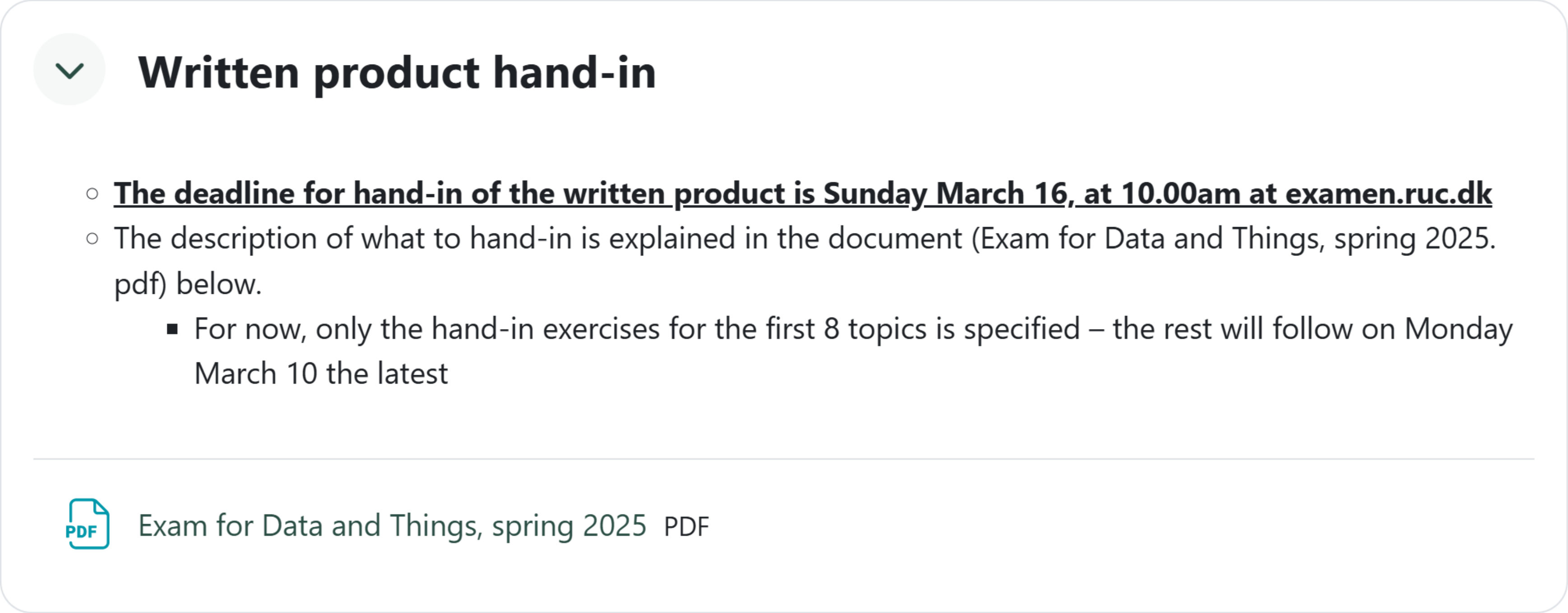
- **<u>The oral exam</u>**
  - Dates: March 20&21
  - There will 20 minutes for each student, including time for assessment and feedback
  - At the beginning of the exam, the student draws a random number. Each number will correspond to an exam topic (14 topics – see exam document on moodl).
  - The student then present on the topic (3-5 min) including potential selected exercises handed in as part of the written product.
  - This is followed by questions about the presentation and the exam question from the examiners (5-10 min).
  - At the end, the examiners might relate their questions to some of other exam topics.
  - For each of the exam topics, the student is expected to know the central concepts, methods, theories, and problems discussed in class and be able to explain and exemplify them. Moreover, for those exam topics where there are selected exercises handed in as part of the written product, the student is expected to be able to explain how they solved the exercises and be able to explain their entire code.

Course practicalities

- **The written product**
  - Hand-in date: Sunday, March 16 at 10.00am
  - The written product will consist of answers to selected exercises - these will be selected from those that have already been done in class. Only some of the exam topics will have such selected exercises (most of them will!).
  - The list of selected exercises will be made public on the last day of class (Monday March 10).
  - The handed in answers to the selected exercises should be in the format of a single Jupyter Notebook or a zip file containing a notebook for each of the selected exercises.
  - The students can do the selected exercises in self-made groups and either hand in as a group on Eksamen.ruc.dk, or hand in the notebook(s) individually.

Hand-in for the exam

- The hand-in exercises for the first 8 (out of 14) topics is now on Moodle… - go read it now…



Written product hand-in

- ○ **<u>The deadline for hand-in of the written product is Sunday March 16, at 10.00am at examen.ruc.dk</u>**
- ○ The description of what to hand-in is explained in the document (Exam for Data and Things, spring 2025. pdf) below.
    - ▪ For now, only the hand-in exercises for the first 8 topics is specified – the rest will follow on Monday March 10 the latest

PDF Exam for Data and Things, spring 2025 PDF

Outline of this lecture

- Hand-in for the exam
- The case of today
- The data science process revisited
- The business problem and the data
- Pre-processing techniques
- The model training process
- Model evaluation
- Hyper-parameter tuning

A Churn case

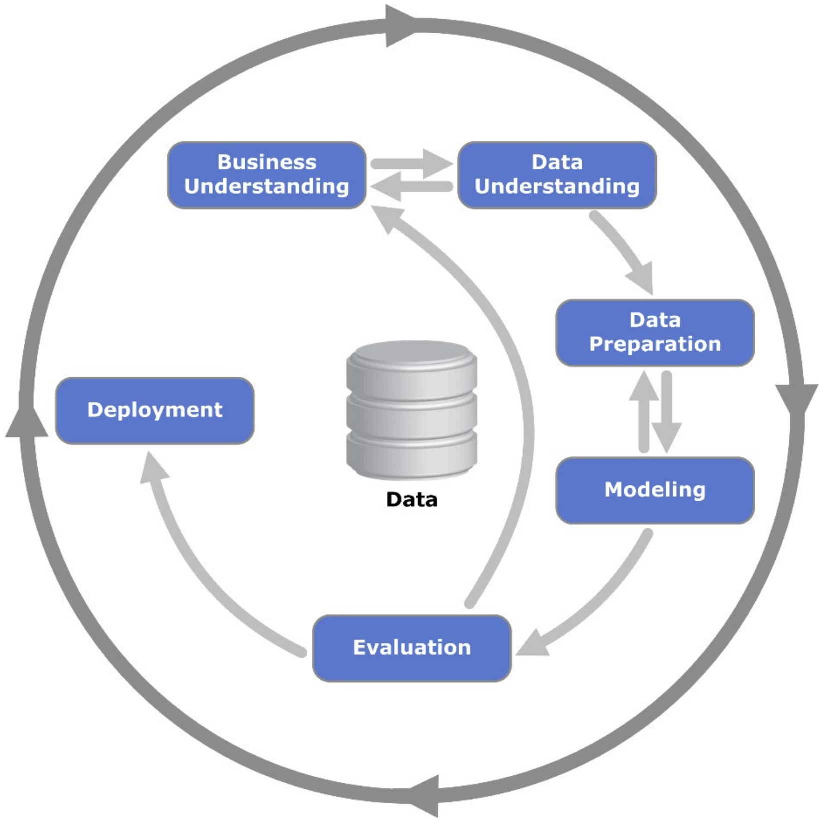
- Today we will all work on a Churn case that was originally posed as a challenge by Maven Analytics, see:
  - https://www.mavenanalytics.io/blog/maven-churn-challenge.
- The dataset can be downloaded Maven Analytics' website, from Kaggle, or from moodle. The dataset contains 3 files:
  - The ***telecom_customer_churn.csv*** file contains information on all 7,043 customers from a Telecommunications company in California in Q2 2022. Each record represents one customer, and contains details about their demographics, location, tenure, subscription services, status for the quarter (joined, stayed, or churned), and more!* The
  - ***telecom_zipcode_population.csv*** file contains complimentary information on the estimated populations for the California zip codes in the Customer Churn table* The
  - telecom_data_dictionary.csv* file contains metadata about the other two files in the sense of a dictionary of the variables in the two other files.
- The goal according to the Maven Churn Challenge is to ***"help the company improve retention by identifying high value customers and churn risks, and have been asked to present your findings to the CMO in the form of a single page report or dashboard."*** *(*Maven Churn Challenge, 2022).

Outline of this lecture

- Hand-in for the exam
- The case of today
- The data science process revisited
- The business problem and the data
- Pre-processing techniques
- The model training process
- Model evaluation
- Hyper-parameter tuning

The data science process revisited

- **The Data Science process**
  - *CRISP-DM*: Cross-industry standard process for data mining
  - An iterative process, where we can return to any step at any time!



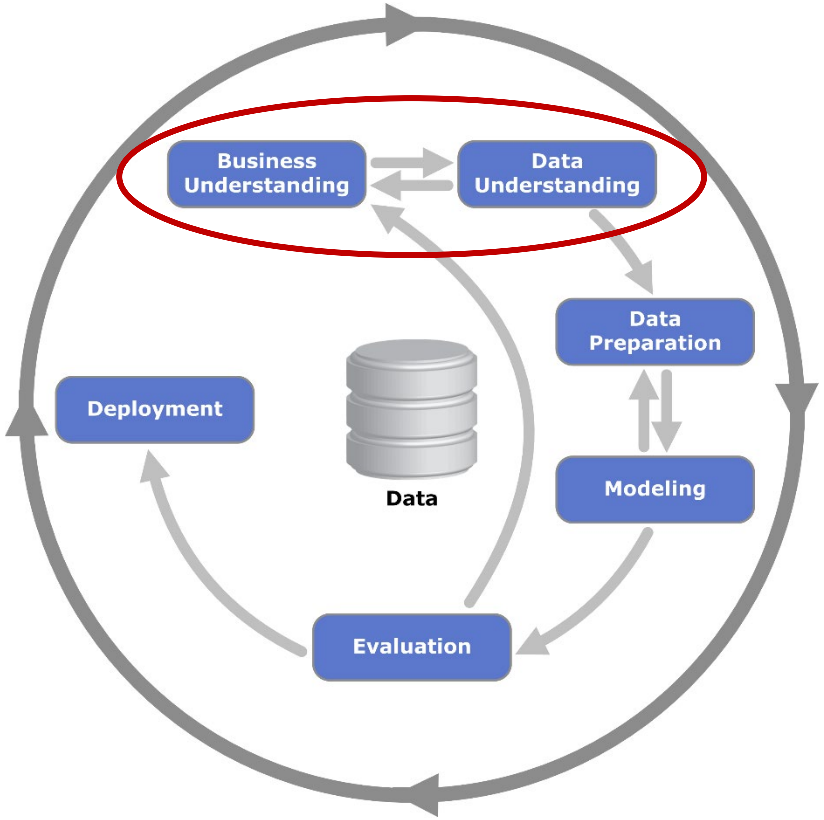
https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

Outline of this lecture

- Hand-in for the exam
- The case of today
- The data science process revisited
- The business problem and the data
- Pre-processing techniques
- The model training process
- Model evaluation
- Hyper-parameter tuning

The business problem and the data

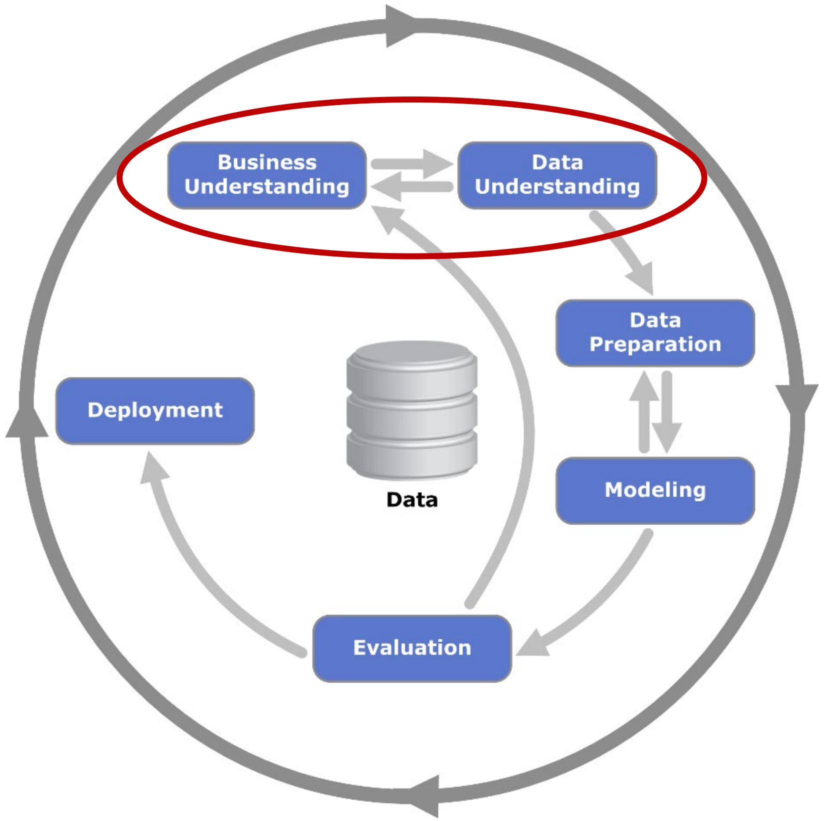
- The goal according to the Maven Churn Challenge is to ***"help the company improve retention by identifying high value customers and churn risks, and have been asked to present your findings to the CMO in the form of a single page report or dashboard."*** *(*Maven Churn Challenge, 2022).
- Does this match the dataset (***telecom_customer_churn.csv***) given?
- Go do Task 1 and 2 in the notebook "Churn case.ipynb"



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

The business problem and the data

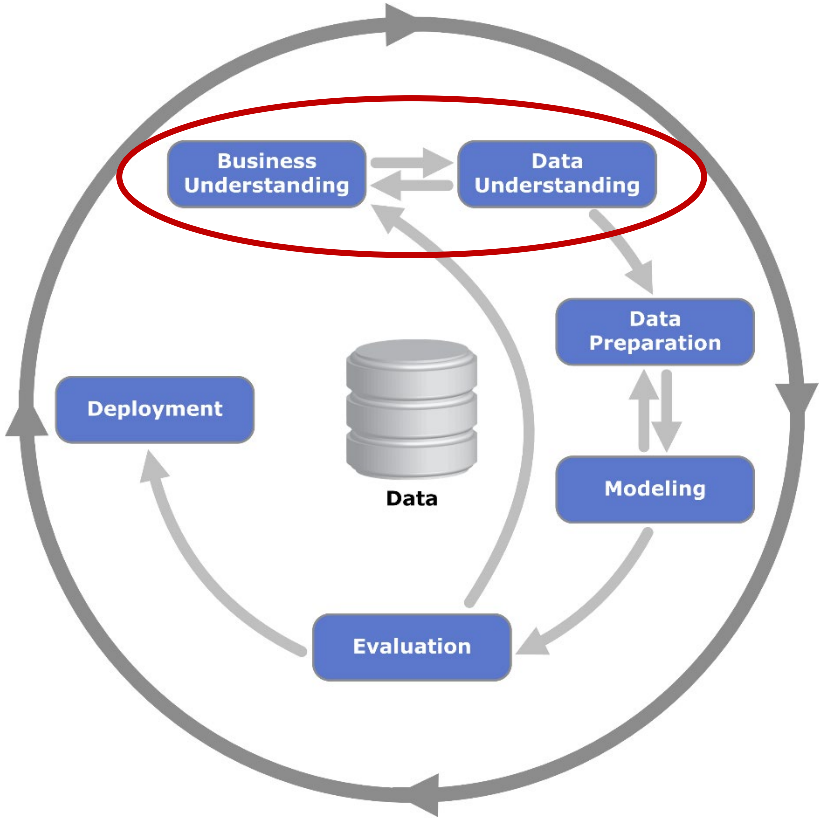
- The goal according to the Maven Churn Challenge is to ***"help the company improve retention by identifying high value customers and churn risks, and have been asked to present your findings to the CMO in the form of a single page report or dashboard."*** *(*Maven Churn Challenge, 2022).
- **It seems like we have data that can help us identify high value customers**
- **It seems like we have data that would allow us to train a supervised classification model to predict whether a costumer churn**
- Go do Task 3 and 4 in the notebook “Churn case.ipynb”



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

The business problem and the data

- **The definition of a high value customers**
  - The number of services a customer has?
  - To what extent the customer uses the services?
  - To which extent the customer referred other customers to the company?
  - How long the customer has been with the telecom company?
  - How many offers the customer has accepted?
  - How much revenue the customer has/is generated(ing)? – On average or in its lifetime (CLV – Customer Lifetime Values)
- **Relevant feature variables in predicting churn**
- We want variables that:
  - Could affect the target
  - Do not have to many categorical values
  - Contains substantial variance
  - Variables whose information is already captured by other variables
  - To start with, we might not want to include to complicated variables.



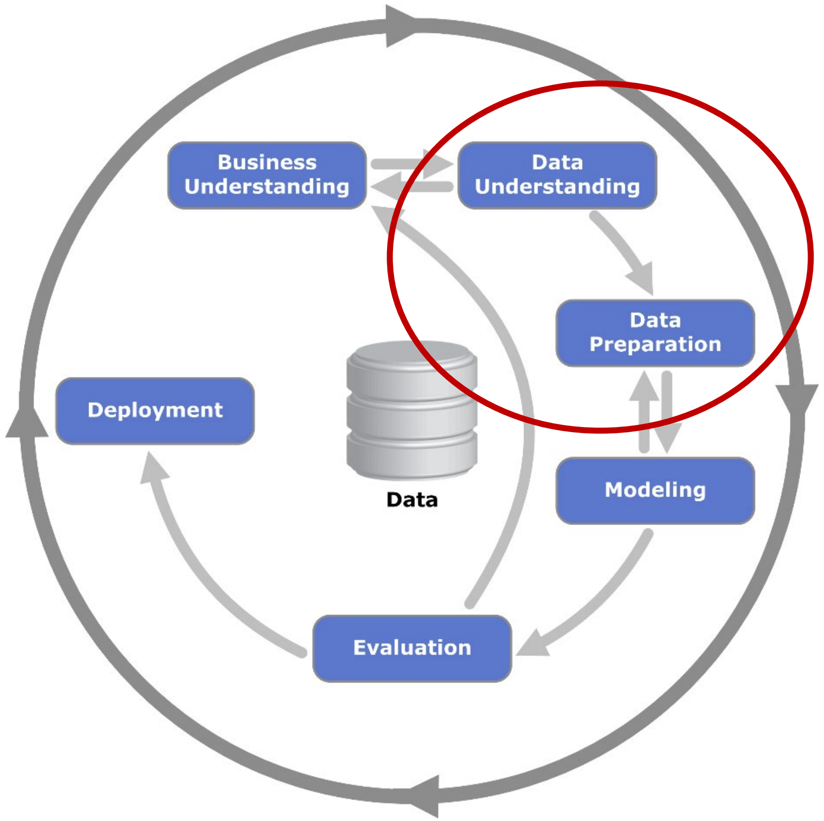
https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

Outline of this lecture

- Hand-in for the exam
- The case of today
- The data science process revisited
- The business problem and the data
- Pre-processing techniques
- The model training process
- Model evaluation
- Hyper-parameter tuning

Pre-processing techniques

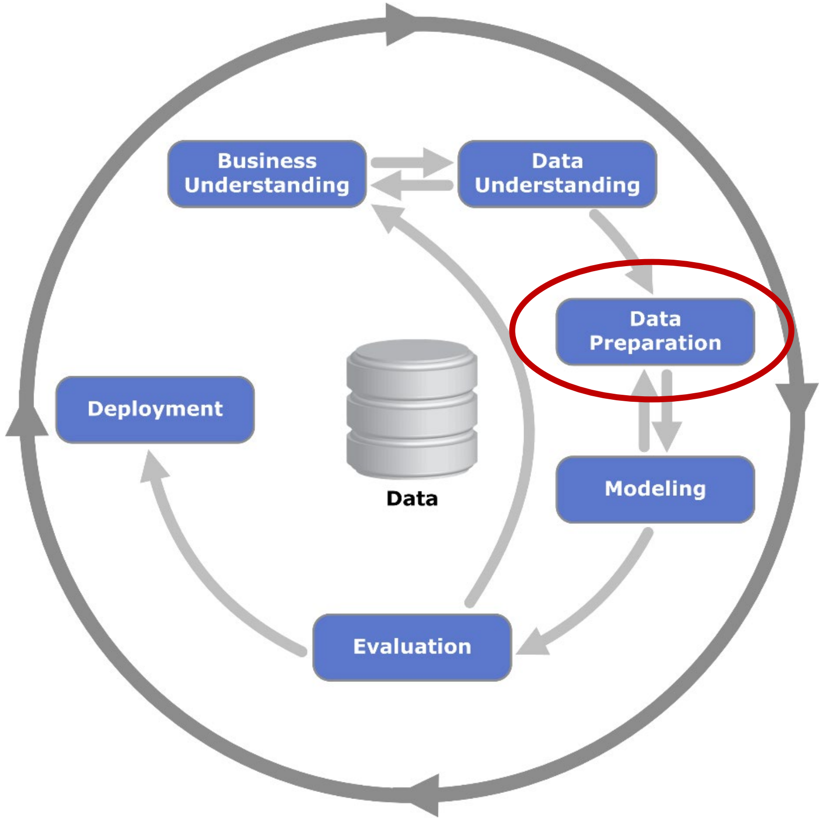
- To get an idea about what pre-processing is needed before we can start training a churn prediction model, we will first do Exploratory Data Analysis.
- Go do Task 5 and 6 in the notebook “Churn case.ipynb”
- **The target variable**
  - Unbalanced classes: More than twice as many non-churners than churners
  - There is a class “Joined” we might not want (check out “Tenure in Months” for those)
- **The feature variables**
  - We look to check, among other things:
    - Is there too many categorical values?
    - Are the variable very unbalanced, or contain only limited relevant information?
    - Is there need for data cleaning?
    - Are the missing values?
    - Are there outliers?



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

Pre-processing techniques

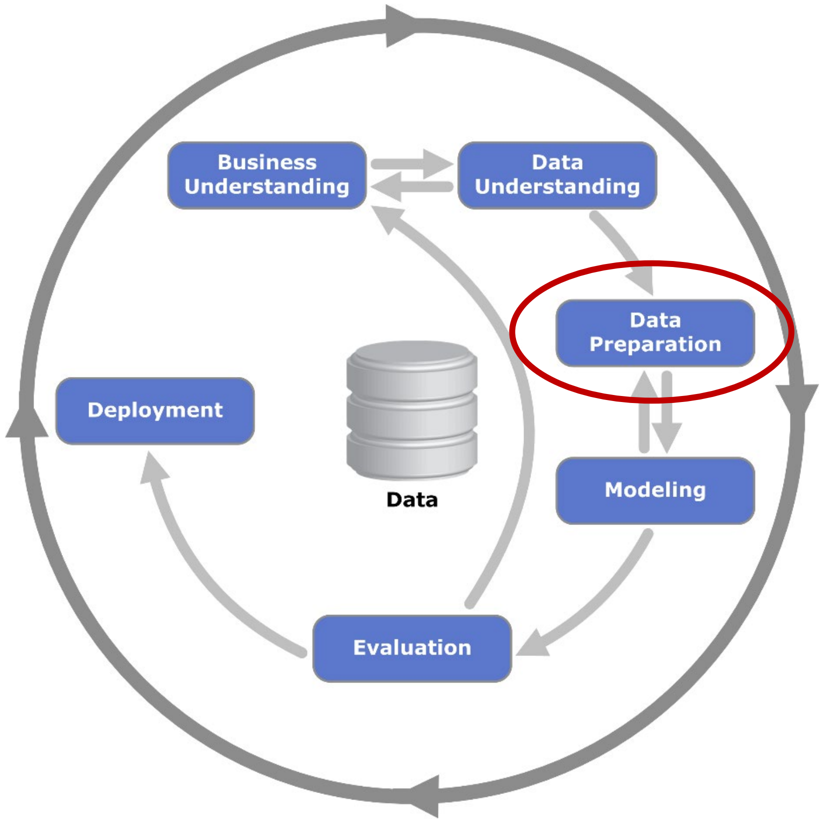
- There seems to be no need for changing data types…
- However, we need to deal with missing values
- We need to do the transformation of the target variable as previously discussed.
- We need to turn categorical variable into dummy variables
- Go do Task 7, 8, 9, and 10 in the notebook “Churn case.ipynb”
- **Missing values**
  - In this case, we might suffice by just replacing missing values with “No”.
- **Data transformation of the target variable**
  - We should probably remove all data point labeled “Joined”, which corresponds to all costumers that have three or less months of tenure.
  - Note, how this also change the business objective!
    - We will predict only predict churn of costumers that have been with the company for more than 3 months



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

Pre-processing techniques

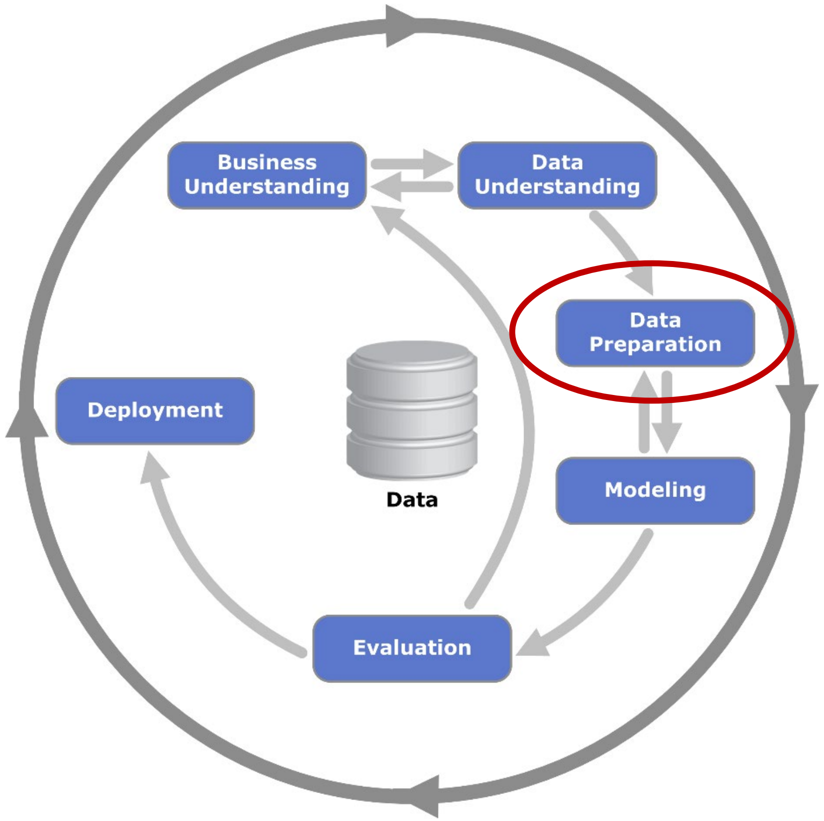
- **Dealing with missing values**
  - Drop rows or columns
  - Replace by constant values (0, “No”, … etc.)
  - Replace by simple calculated values (mean, median, mode, … etc.)
  - More advanced imputation methods
    - K-Nearest Neighbor imputation
    - Predicting missing values using other supervised learning models
  - See for instance, https://scikit-learn.org/stable/modules/impute.html



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

Pre-processing techniques

- **Common pre-processing techniques**
  - Changing data format
  - Dealing with missing values
  - Dealing with outliers
  - Creating dummy variables
  - Filtering the data
  - Scaling of numeric variables
  - Binning, … etc., for categorical variables
  - Dimensionality reduction (-like PCA)
  - Feature selection
  - Feature engineering



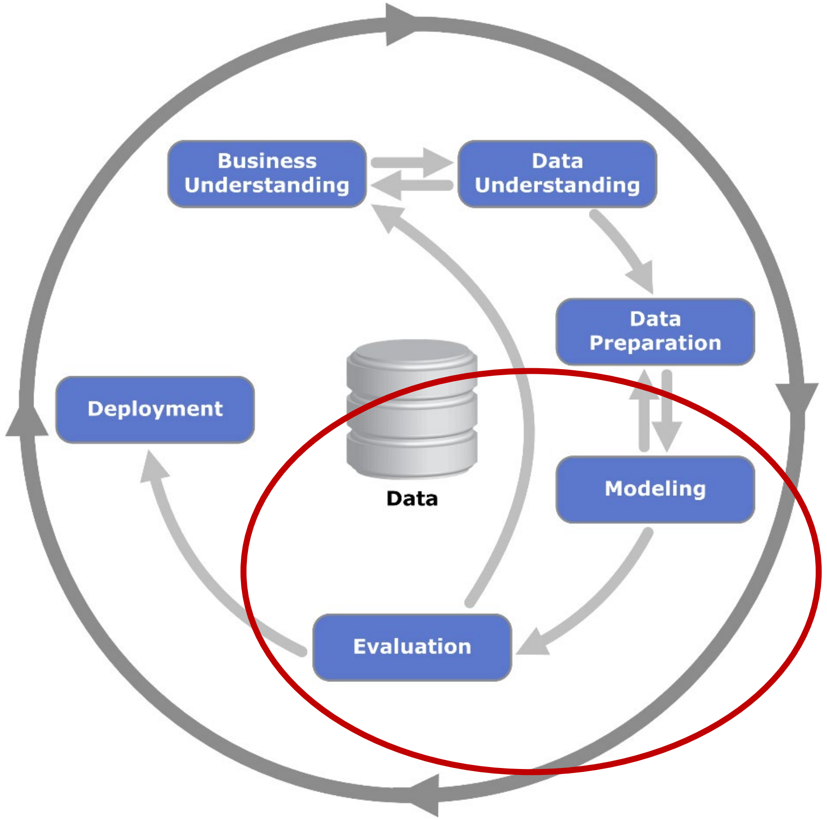
https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

Outline of this lecture

- Hand-in for the exam
- The case of today
- The data science process revisited
- The business problem and the data
- Pre-processing techniques
- The model training process
- Model evaluation
- Hyper-parameter tuning

The model training process

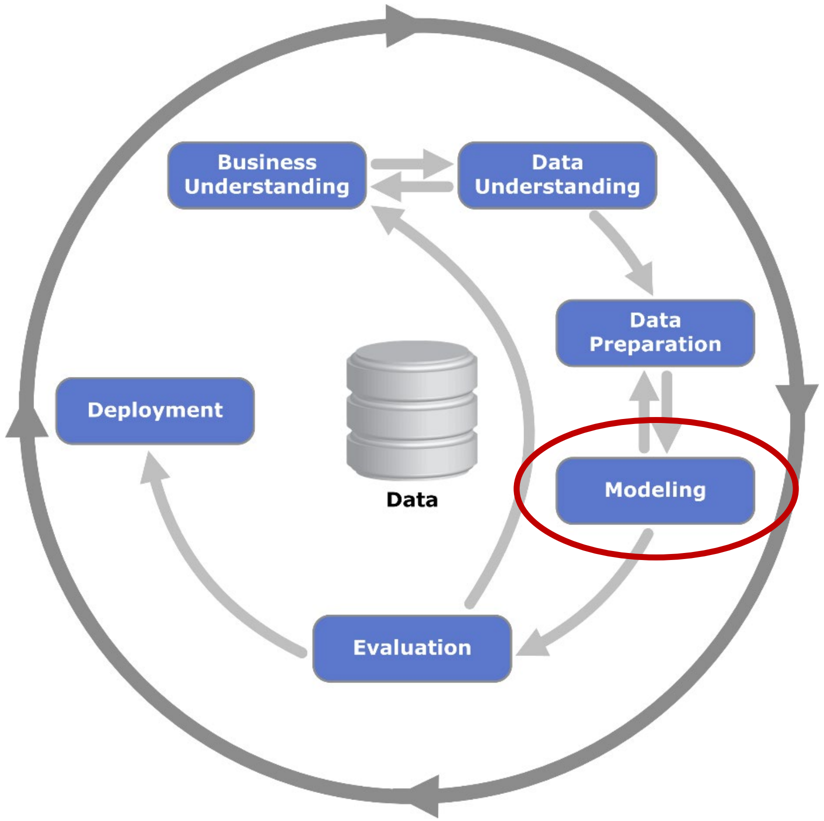
- Before we start training machine learning models, we need to decide how we will evaluate the models
- Go do Task 11 in the notebook “Churn case.ipynb”
- **Statistical performance metrics**
  - ***Regression***: Root Mean Squared Error (RMSE), Mean absolute Error (MAE), Mean Squared Error (MSE), R-squared
  - ***Classification***: Accuracy, Precision, Recall, F1-score, …. ROC, AUC, …. Etc.
  - ***Unsupervised learning***: …? (- for Clustering, Silhouette score, … etc.)
- **Business performance metrics**
  - Particular business applications might suggest particular statistical performance metrics
  - However, particular business applications often come with their own additional measures of success



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

The model training process

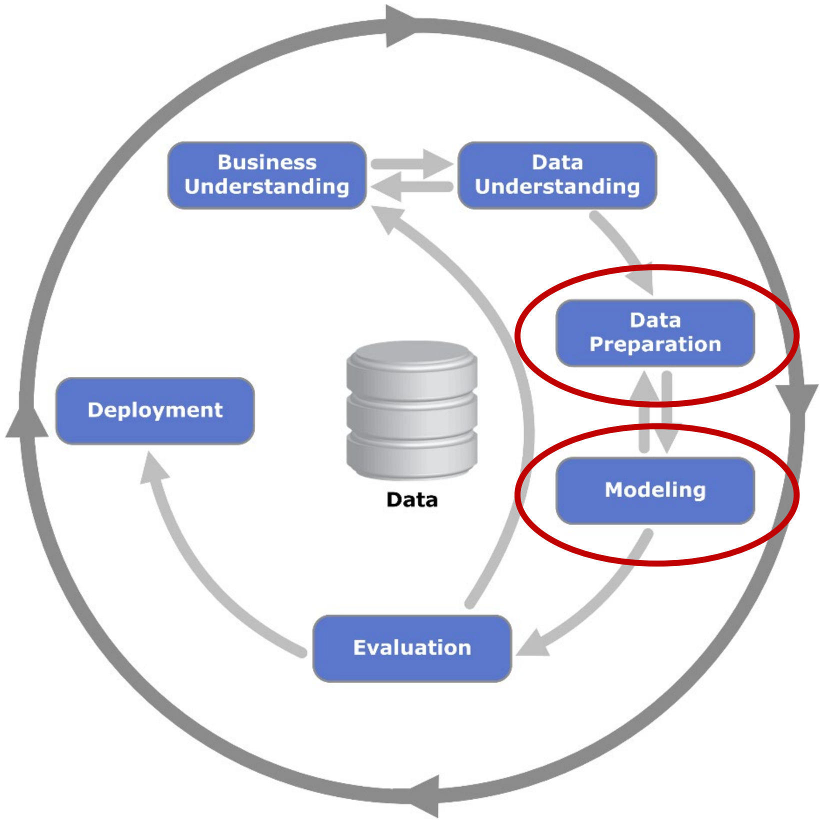
- We always start by training the most simple baseline model we can think of
  - This will allow us to compare any future more complex model we train – if it cannot beat the baseline model, there is no need for it
  - More, any model should be better than obvious "cheating" models such as
    - Randomly assigning a class (for binary classification it will have an accuracy of 50%)
    - Constantly assign the most prominent class (for unbalanced dataset this will often have an accuracy much higher than 50%)
  - A baseline model will also give Proof of concept model, we can use to test further downstream applications that will use the final model
- Go do Task 12 in the notebook "Churn case.ipynb"



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

The model training process

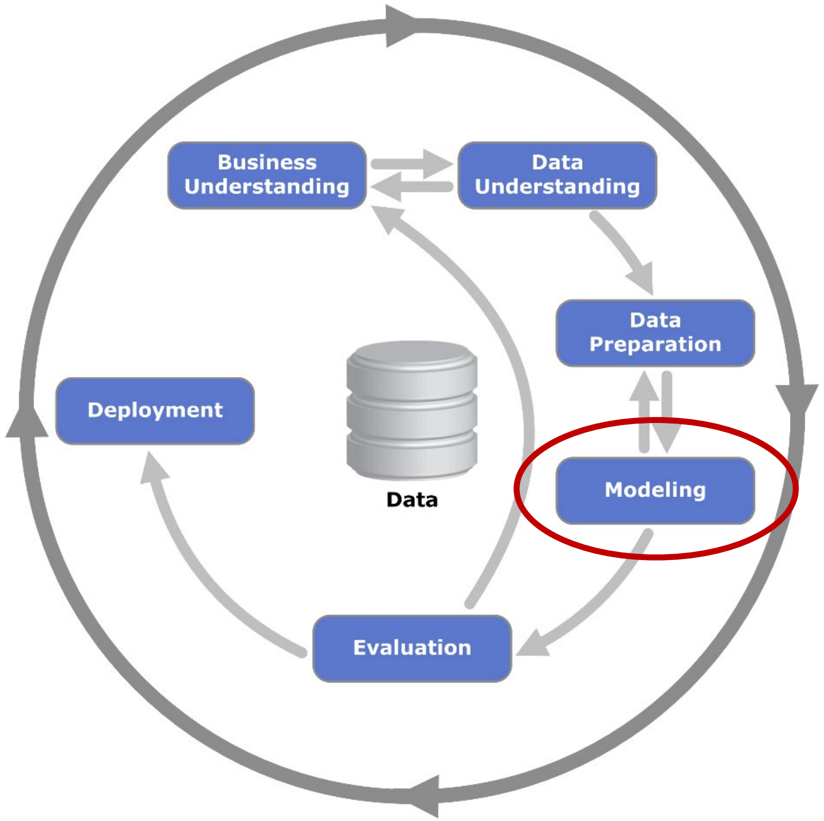
- The baseline model, confirm our observation that the dataset is unbalanced and that it likely have a negative effect on the ability to train a good model – thus, we need to do something about the unbalanced dataset…
- We go back to data preparation…
- **Dealing with imbalanced classes**
  - Under-sampling of the majority class
  - Over-sampling of the minority class
  - Combination of the two
  - More advanced technique like SMOKE
    - Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357. https://www.jair.org/index.php/jair/article/view/10302
- Go do Task 13 in the notebook "Churn case.ipynb"



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

The model training process

- Now we can train other more advanced machine learning models on the dataset, such as K-Nearest Neighbor, Decision Trees, Random Forest, AdaBoost, and XGBoost, …
- Go do Task 14 in the notebook “Churn case.ipynb”



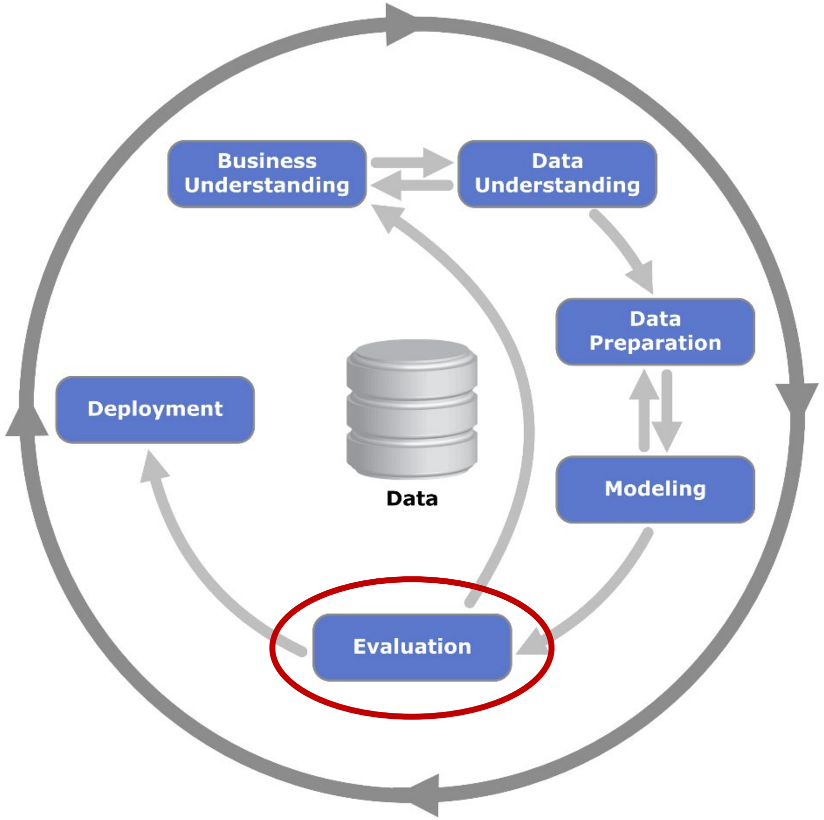
https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

Outline of this lecture

- Hand-in for the exam
- The case of today
- The data science process revisited
- The business problem and the data
- Pre-processing techniques
- The model training process
- Model evaluation
- Hyper-parameter tuning

Model evaluation

- Recall the different ways of evaluating a model, previous discussed (Task 11)
  - **Statistical performance metrics**
    - ***Classification***: Accuracy, Precision, Recall, F1-score, …. ROC, AUC, …. Etc.
  - **Business performance metrics**
    - Particular business applications might suggest particular statistical performance metrics
    - However, particular business applications often come with their own additional measures of success
    - Example: Customer value lost as: FN*CLV
- Go do Task 15 in the notebook “Churn case.ipynb”



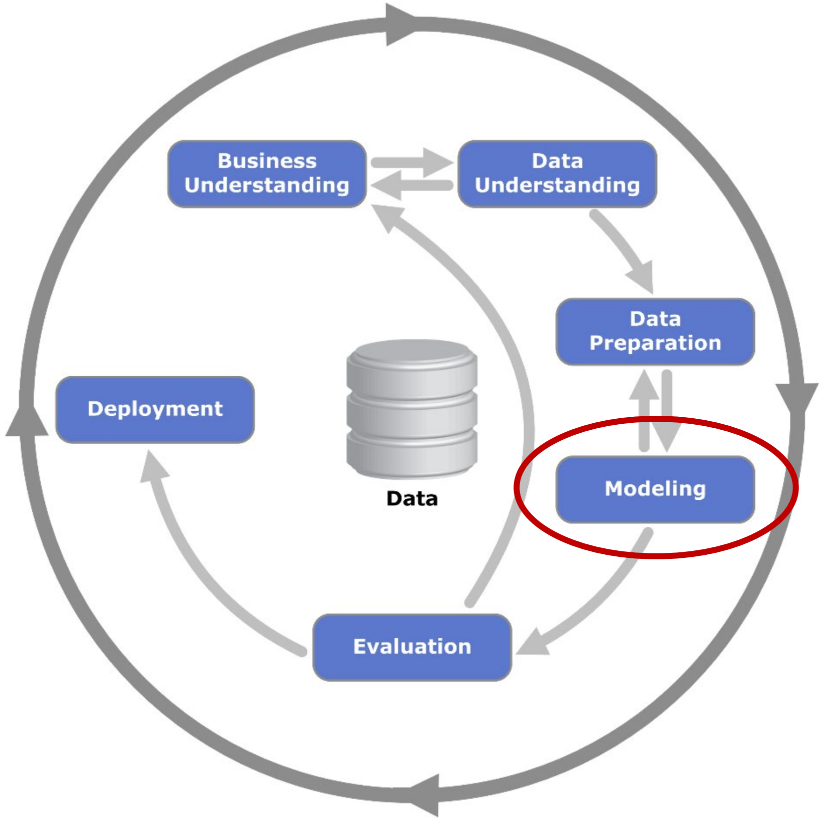
https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

Outline of this lecture

- Hand-in for the exam
- The case of today
- The data science process revisited
- The business problem and the data
- Pre-processing techniques
- The model training process
- Model evaluation
- Hyper-parameter tuning

Hyper-parameter tuning

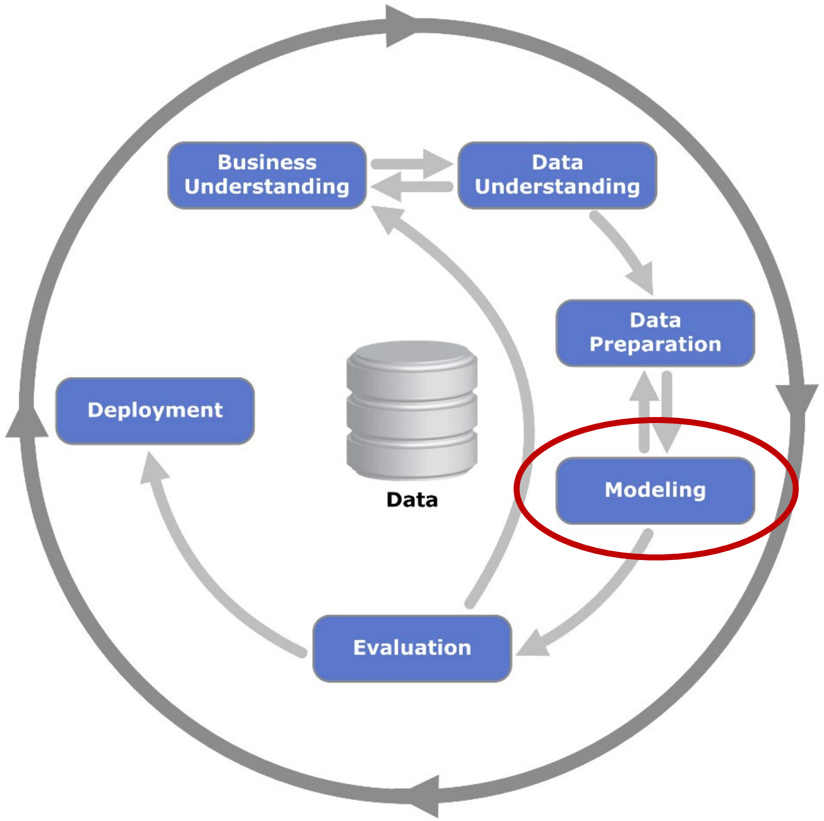
- For the K-Nearest Neighbor model that trains fast and only have one hyper-parameter (K), we can easily search a range of potential values
- If we had two hyper-parameters, the natural extension would be to do “grid search”, that is searching the grid consisting of the product of the two ranges
  - Hyper-parameter hp_1, might take values in [1, 2, 3, 4] – 4 values
  - Hyper-parameter hp_2, might take values in [0.001, 0.01, 0.1, 1.0, 10.0] – 5 values
  - Searching the grid of these hyper-parameters mean searching 20 values



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

Hyper-parameter tuning

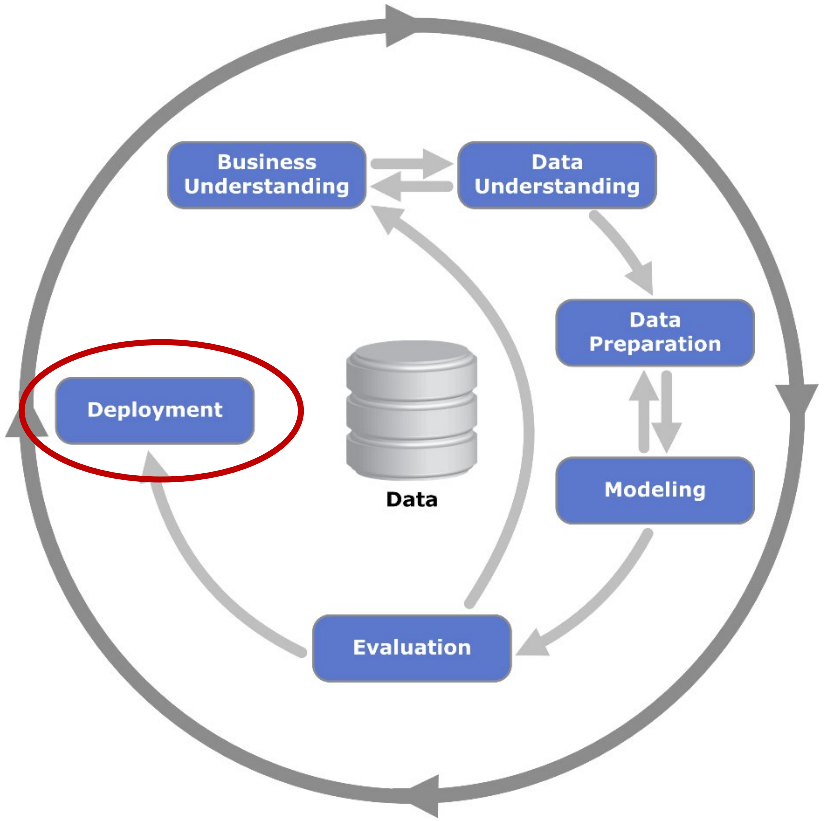
- Grid search should be combined with cross-validation
  - In Scikit-learn, grid search can be used together with cross-validation in the class GridSearchCV:
    - https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
  - Grid search quickly becomes to computational demanding due to its multiplicative increase in the grid size for new hyper-parameter values included
- Randomly searching the grid
  - An alternative, that might be good enough and computational much faster
  - In Scikit-learn, random search can be combined with cross-validation in the class RandomizedSearchCV:
    - https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html
- Other more advanced ways of searching the parameter space exists as well, of course…



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

Deployment…

- …much more about this next, when Shabab talks about Machine Learning Operations (MLOps)!



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