# Course 3: Data Science Methodology

# Week 1:

## Lesson 1: Welcome

Course Introduction

- Methodology: System of methods used in a particular study or activity
- *10 Questions to be Answered:*
    - **Problem to Approach**
        - What is the problem you are trying to solve?
        - How can you use data to answer the question?
    - **Working With Data**
        - What data do you need to answer the question?
        - Where is the data coming from (identify all sources) and how will you get it?
        - Is the data you collected representative of the problem to be solved?
        - What additional work is required to manipulate and work with the data?
    - **Deriving the Answer**
        - In what way can the data be visualized to get to the answer that is required?
        - Does the model used really answer the initial question or does it need to be adjusted?
        - Can you put the model into practice?
        - Can you get constructive feedback into answering the question?

Introduction to CRISP-DM

- Process aimed to increase the use of data mining over a wide variety of business applications and industries
- Six steps with an entity that has to implement
    - **Business Understanding**
        - Intention of project is outlined
        - Communication and clarity
        - Clear, concise, complete perspective necessary of what project goals are
    - **Data Understanding**
        - Data is collected
        - Wants and needs of business determine data collected, sources, and by what methods
    - **Data Preparation**
        - Data is transformed into a useable subset unless it is determined more data is needed

        - Once a dataset is chosen, must be checked for questionable, ambiguous, or missing cases
    - **Modeling**
        - After data preparation, data is expressed through appropriate models
        - Find meaningful insights or new knowledge
        - Models reveal patterns and structures within the data
    - **Evaluation**
        - Model is tested
        - Pre-selected test used to run the trained model
        - Shows effectiveness of model on a set it sees as new
        - Foreshadows its role in the final stage
    - **Deployment**
        - Model is used on new data outside the scope of the dataset and by new stakeholders
        - New interactions may reveal new variables and needs for the dataset and model
        - Could initiate revision for business needs and actions, model and data, or both
- Flexible and cyclical model
- Communication very necessary

## Lesson 2: From Problem to Approach

Business Understanding

- Seek clarification
- Clearly defined question is vital in order to approach it analytically
- What's the goal?
- Support the goal
- Stakeholder "buy-in" and support
- Apply the concepts
    - Best way to distribute healthcare budget
    - Define goals
        - Provide quality care without increasing costs
    - Define objectives
        - Review process to identify inefficiencies

Analytic Approach

- How to use data to answer the question?
- Pick approach based on type of question
    - Descriptive
        - Current status
        - Show relationships
    - Diagnostic (Statistical Analysis)
        - What happened?

        - Why is this happening?
        - Yes/No answer
    - Predictive (Forecasting)
        - What if these trends continue?
        - What will happen next?
        - Determine probability of an action
    - Prescriptive
        - How do we solve it?
- Machine Learning
    - Learning without explicit programming
    - Identifies relationships and trends in data that might otherwise not be accessible or identifiable
    - Uses clustering association approaches
- Case Study
    - Predict an Outcome
        - Decision tree classification
        - Explicit decision path showing conditions leading to high risk
        - Likelihood of classified outcome
        - Easy to understand and apply

Lab:

- Business Understanding
    - Clarify the goal of the entity in question
- Analytic Approach
    - Identify the type of patterns which will be used to most appropriately address a question
- Decision Trees
    - Pros
        - Easy to interpret
        - Can handle numeric or categorical features
        - Can handle missing data
        - Uses only the most important features
        - Can be used on very large or small data
    - Cons
        - Easy to overfit or underfit the model
        - Cannot model interactions between features
        - Large trees can be difficult to interpret

Lesson 2 Summary

- The need to understand and prioritize the business goal.
- The way stakeholder support influences a project.
- The importance of selecting the right model.
- When to use a predictive, descriptive, or classification model.

# Lesson 3: From Requirements to Collection

Data Requirements

- Cooking with data
- Each step is important to cooking the meal
- Identify ingredients, how to work with them
- Identify conditions which are essential, and which could skew results unfairly (if a patient readmits to the hospital for a different heart related reason than the one being studied, it will skew results, so those particular patients will be excluded from the study)

Data Collection

- After ingredient collection, ingredients are revised and analyzed to advise quality and useability of ingredients
- Know the source or where to find the data elements
- Deferring data is okay if it is unacquireable at the time and to acquire it at a later stage

From Requirements to Collection (notebook)

- Data Requirements stage
    - Identify necessary data content, formats, and sources for initial data collection

Lesson 3 Summary

- The significance of defining the data requirements for your model.
- Why the content, format, and representation of your data matter.
- The importance of identifying the correct sources of data for your project.
- How to handle unavailable and redundant data.
- To anticipate the needs of future stages in the process.

# Week 2:

## Lesson 1: From Understanding to Preparation

Data Understanding

- Descriptive statistics
- Pairwise correlations
- Histogram
    - Good way to understand how values or a variable are distributed
- Data quality
    - Missing values
    - Invalid or misleading values
- Iterative data collection and understanding
    - Refined definition of CHF(congestive heart failure) admission

Data Preparation

- Washing freshly picked vegetables
- Get rid of imperfections
- Most time consuming phase (70-90% of project time)
    - Can be reduced to 50% with certain techniques
- Transforming data: get data in a state where it will be easier to work with
- Address missing or invalid values
- Feature engineering
    - Using domain knowledge of data to create features that make ML algorithms work
- Case Study
    - Define CHF
    - Define readmission
    - Aggregating records
        - Inpatient and outpatient records
        - All put together for 1 record per patient
    - Literature review addresses if more or less data is needed and to loop back
    - Creating new variables based on existing data

Lesson 1 Summary

- The importance of descriptive statistics.
- How to manage missing, invalid, or misleading data.
- The need to clean data and sometimes transform it.
- The consequences of bad data for the model.
- Data understanding is iterative; you learn more about your data the more you study it.

# Lesson 2: From Modeling to Evaluation

Modeling - Concepts

- Sampling the food (data)
- What is the purpose?
- Characteristics?
- Descriptive or predictive models
    - If person likes this, they might like that
    - Try to see what will happen in the future
- Using training / test sets
- Understand the question
- Select an analytic approach or method to solve the problem
- Obtain, understand, prepare, and model the data
- Make sure the question is answered
- Training sets are used to build predictive models

Case Study

- Parameter tuning
- Initial decision tree classification model
- Low accuracy on “yes” outcome
- Weight for yes and no can be different

Evaluation

- Done iteratively
- Performed during model development
- Does the model used actually answer the question? Or does it need to be adjusted?
- Diagnostic measures
    - Predictive model
    - Descriptive model
- Statistical significance

Lesson 2 Summary

- The difference between descriptive and predictive models.
- The role of training sets and test sets.
- The importance of asking if the question has been answered.
- Why diagnostic measures tools are needed.
- The purpose of statistical significance tests.
- That modeling and evaluation are iterative processes.

# Week 3:

## Lesson 1: From Deployment to Feedback

Deployment

- Are stakeholders familiar with the new tool?
- Can be rolled out to limited users
- Case Study
    - Assimilate knowledge for business
        - Practical understanding of meaning of model results
        - Implications of model results for designing intervention actions
    - Automated near real time risk assessments of CHF inpatients
    - Easy to use
    - Automated data preparation and scoring
    - Up-to-date risk assessment for clinicians
    - Training for clinical staff
    - Tracking / monitoring processes

Feedback

- Problem solved?
- Question answered?
- If not, back to modeling stage
- Once model is evaluated and data scientist is confident it will work, it is deployed and put to the ultimate tes
    - Real time use in the field
- Define review process
    - Measure results of applying the risk model to CHF patients
    - Track patients (readmission outcomes)
    - Measure effectiveness of interventions (compare readmission rates before and after model implementation
- Refinement
    - Initial review after first year of implementation
    - Based on feedback data and knowledge gained
    - Participation in intervention program
    - Possibly incorporate detailed pharmaceutical data (originally deferred)
    - Other possible refinements
- Redeployment
    - Continue modeling, deployment, feedback, and improve iteratively

Course Summary

- Thinking like a DS
    - Forming a concrete business / research problem
    - Collecting and analyzing data

        - Building a model
        - Understanding feedback after model deployment
- Importance
    - Understanding the question
    - Picking the most effective analytic approach
- Working with data
    - Determine data requirements
    - Collect data
    - Understand data
    - Prepare data for modeling
- Modeling data
    - Evaluate and deploy the model
    - Get feedback on it
    - Use feedback to improve the model
- Stages of the methodology are iterative!

Lesson 1 Summary

- The importance of stakeholder input.
- To consider the scale of deployment.
- The importance of incorporating feedback to refine the model.
- The refined model must be redeployed.
- This process should be repeated as often as necessary.