Introduction to Data Science
Semester-Long Project

Overview

This term you will undertake a group research project (**2+ people**). For the team's grade, each person will receive the same grade if each person has contributed equally. At the end of the semester, we will ask each student on the team to assess how much of the final project each team member was responsible for. **Lack of participation will result in a lower grade.** Great teams have great contributors, each contributing equally. Within the team, you must negotiate on how much and what each person will contribute. Think carefully about your team members: Where do people live? What courses are they taking? What hours do they work? Where will you meet? What skills do the different individuals bring to the group (computing, programming, design, evaluation, statistics, etc.)? We strongly encourage you to form a heterogeneous team full of individuals with varying types of skills.

Team GitHub Repositories & Reports

Each team should create a project GitHub repository which includes:

1. Project title & one-paragraph abstract/description of the project on the main README page.
2. Team name, members, and member roles
3. Reports and files for project deliverables 1-4
    a. Reports should be professionally prepared, expressive, grammatically sound, illustrative of your efforts and process, and easy to understand. **A good design effort can easily be hampered by a poor communication of what was done**.
    b. Reports should be formatted as a Research-style paper, with the project **layers** as an appendix in the SAME document, until the final paper is completed.

All required files will be posted to the GitHub Repository.

Part 0 – Life Cycle of a Typical Data Science Project Explained
(adapted from https://www.analyticsvidhya.com/blog/2021/05/introduction-to-data-science-project-lifecycle/)

1. **Understanding the Problem**
   In order to build a successful model, it's very important to first understand the problem that is being addressed, whether in a business or research setting. Suppose your client wants to predict the customer churn rate in the retail store. You may first want to understand the business, its requirements and what is to be achieved from creating a model and making a prediction. In such cases, it is important to consult with domain experts and fully understand the underlying problems that are present in the system. A Business Analyst is generally responsible for gathering the required details from the client and forwarding the data to the data scientist team for further speculation. Even a minute error in defining the problem and understanding the requirement may be very crucial for the project hence it is to be done with maximum precision.

   After asking the required questions to the company stakeholders or clients, we move to the next process which is known as data collection.

2. **Data Collection**
   After gaining clarity on the problem statement, relevant data needs to be collected to break the problem into smaller components.

   The data science project starts with the identification of various data sources, which may include web server logs, social media posts, data from digital libraries such as the US Census datasets, data accessed through sources on the internet via APIs, web scraping, or information that is already present in an excel spreadsheet. Data collection entails obtaining information from both known internal and external sources that can assist in addressing the business issue.

   Normally, the data analyst team is responsible for gathering the data. They need to figure out proper ways to source and collect data and collect to achieve the desired results.

   There are two common ways to source the data:
   1. Through web scraping with Python
   2. Extracting Data with the use of third-party APIs

3. **Data Preparation**
   After gathering the data from relevant sources, now the data must be prepared for use in analysis and modelling. This stage helps us gain a better understanding of the data and prepares it for further evaluation.

Additionally, this stage is referred to as Data Cleaning or Data Wrangling. It entails steps such as selecting relevant data, combining it by mixing data sets, cleaning it, dealing with missing values by either removing them or imputing them with relevant data, dealing with incorrect data by removing it, and also checking for and dealing with outliers. By using feature engineering, you can create new data and extract new features from existing ones. It is imperative to format the data according to the desired structure and delete any unnecessary columns or functions. Data preparation is the most time-consuming process, accounting for up to 90% of the total project duration, and this is the most crucial step throughout the entire life cycle.

Exploratory Data Analysis (EDA) is critical at this point because summarizing clean data enables the identification of the data's structure, outliers, anomalies, and trends. These insights can aid in identifying the optimal set of features, an algorithm to use for model creation, and ultimately model construction.

4. **Data Modeling**

   Throughout most cases of data analysis, data modeling is regarded as the core process. In this process of data modeling, the prepared data is used as the input to generate the desired output.

   First, an appropriate type of model needs to be selected that would generate the desired output. For example, is it a regression problem, classification, or a clustering-based problem? Depending on the type of data received will help to identify a choice for an appropriate machine learning algorithm that is best suited for the model. Once this is done, there will likely be iterations in reviewing the model and its parameters to develop the most robust prediction model given the inputs, knowledge of the system, and the desired output.

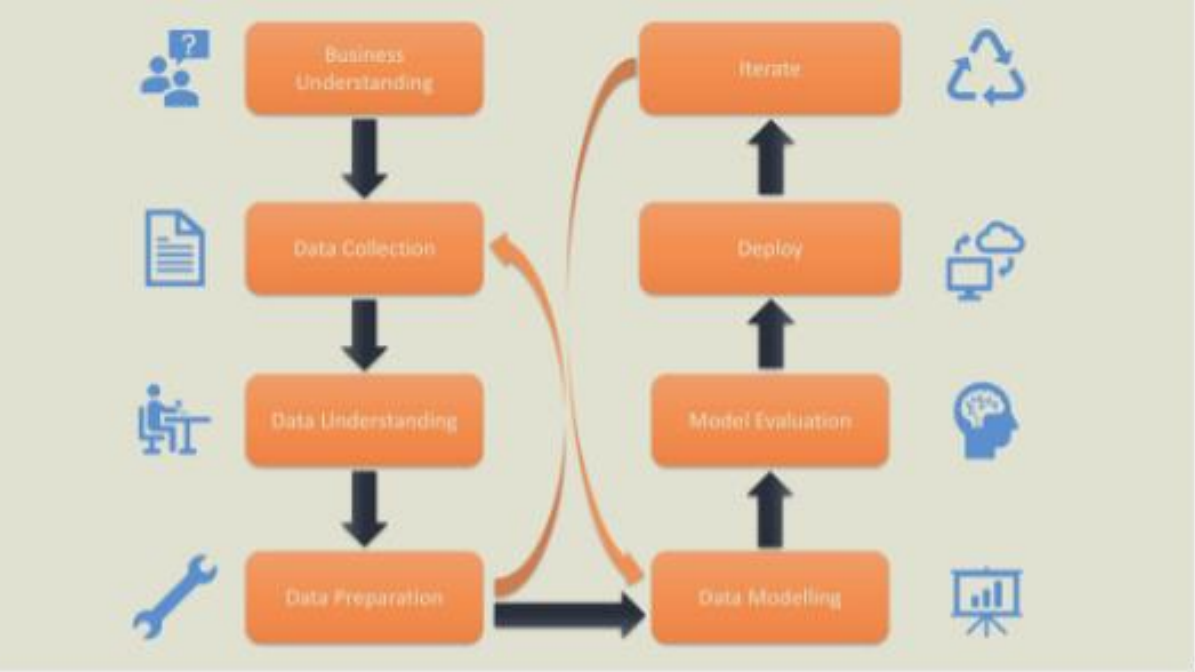
   Finally, the model should be evaluated on its accuracy and relevance. In addition to this project, it is important to make sure there is a good balance between specificity and generalizability (i.e., creating an unbiased model).

5. **Model Deployment**

   Before the model is finalized and deployed, it is important to conduct a rigorous evaluation to ensure the right modelling choice has been made. Once evaluation is complete, it is then deployed in the desired channel and format for a client to use. This is naturally the last step of the life cycle of data science project. Extra caution is always warranted before deploying a prediction model to avoid unwanted errors. For example, if the wrong machine learning algorithm is chosen for modelling, then it is not likely the desired accuracy will be achieved and will be difficult to get final approval for the project from the client (or other important stakeholders). Additionally, if the data have not been cleaned properly, there will be a need to handle missing values or noise present in the dataset later on. Hence, in order to make sure that the model is deployed

properly and accepted in the real world as an optimal use case, rigorous testing should be done at every step of the process.

All the steps mentioned above are equally applicable for beginners as well as seasoned data science practitioners. The key is to learn the process first, practice and deploy smaller projects and build to more complex, real-world problems. Finally, the key to success in any data science project (or really any project at all) is that it is not a clear shot from beginning to end but is iterative and often cycles between different aspects of a process to develop the most robust solution to the problem that is faced. Below is a graphic representation of this iterative structure within the context of a data science project.



https://editor.analyticsvidhya.com/uploads/22454new%201.png

Part 1 - Proposal + Design Outline

There are five draft projects to choose from that can be slightly added to form your semester-long project. Once you choose your topic, you'll need to work as a group to craft a formal proposal that includes one, and absolutely no more than two, research questions that will be answered as a result of your project. Additionally, as part of the proposal you'll need to document an initial outline of the work you'll do to accomplish the proposal goals.

Advice on your proposal:

1. **Think Small:** Most projects build off of a body of code developed by a company -- they're not starting from scratch. If you are, then you need to design a very small project. Take into account your teammates' availability, and plan accordingly.

   ML Dataset Repository: https://archive.ics.uci.edu/ml/index.php

2. **Do One Thing Well:** To do well in this project, you need to do one thing well. In this case, you need to answer ONE **research question**. Communicating the answer to a single, focused research question is a preferable outcome than partial answers to several such questions.
3. **Understand the Affordances of Your Chosen Tools And Methods:** The tools that are available for data science and analytics have different strengths and weaknesses. Use your chosen tool for what it's best for. Choose tools that integrate with the data types we already have.
4. **Plan in Layers**: You can't accurately anticipate how long each step in your project is going to take. Consequently, you need to make a detailed development schedule that is organized in functional deliverable layers. Structure your development so that you complete each layer before going on to the next. Plan exactly what is entailed in each layer, and which team member is going to do each component. There will be times where your work straddles between layers and iteration occurs; this is common and expected.

   Layer 1: Understanding the Problem: Given a short prospectus on a project, further develop your understanding through researching and asking questions to clarify what the desired outcomes are to properly plan to find data that will adequately describe the problem to fulfill the project outcomes. Make preliminary choice for what type of type of modelling will be done.

   Layer 2: Data Collection, Understanding, and Preparation: Collected needed data and more to further understand the problem. Prepare datasets for use in modelling through needed data wrangling. Confirm choice of initial modelling effort based on refined understanding of the problem and the data.

Layer 3: Modelling and Evaluation: Use collected data to create an initial model and evaluate it based on separate test data. Refinement of data may be needed during modelling and evaluation progress.

Layer 4: Finalize Model and Deploy: Once model produces sufficient predictive power on test data, refinement of model through tuning parameters and developing how the model will be deployed by the end user.

Part 1 Deliverables

For this first part there are two deliverables that will need to be turned in and posted to your GitHub project repository. The deliverables and brief descriptions should appear as their own folder within the project repository. Completing the deliverables will fulfill layer 1 of the planning outlined above.

1. Proposal (50 points)
   a. Project Description (10 points)
      Describe the project in detail (approximately 1 – 2 pages of text with any needed accompanying images) that addresses Layer 1: Understanding the Problem, outlined above.
   b. Assessment (20 points)
      - What is the research question?
      - What is your hypothesis and why is it important?
      - How will you be able to tell/measure the success of the research question and hypothesis?
   c. Schedule (20 points)
      Use the described layers to plan and outline a work schedule to complete all of the elements of this project, paying particular attention to the future project deliverables and their relationship to the various layers. Be sure to include which group member is responsible for each element of the schedule. The person who is responsible for a particular element is leading the work on that bit of effort and all members of the group are expected to contribute in some way to each aspect of the project, at least in some small way (e.g., reviewing draft work).
2. Presentation (50 points)
   Create a presentation using slides (e.g., Powerpoint, Google Slides) that documents your research proposal. Maximum presentation time is eight (8) minutes, with an additional two (2) minutes for questions. Post a PDF of your slides to your GitHub project repository for Part 1 deliverables. The following must be included in some fashion within the presentation:
   a. Overview of Project (10 points)
   b. Describe research question and hypothesis in detail (20 points)
   c. Plan for data and methods used to answer your research question (10 points)
   d. Schedule (10 points)

Part 2 – Data Collection, Cleaning, and Exploratory Data Analysis

In this part of the project, you'll retrieve the data you intend to use in your project, work on cleaning the data (if needed), and conducting an in-depth Exploratory Analysis of your dataset. You should demonstrate the progress made on any data collection, your detailed and deep data analysis, and make some initial choices about the predictive modelling you plan to do to fulfill your project question(s).

You'll also want to comment on what has proved harder than expected to this point, what revisions you've made to your expected modelling efforts, and what you have learned about your data and the topic in general. You may be starting to explore different modelling techniques at this point, but it should remain preliminary and primarily focused on potential predictor variables and necessary data reduction.

In addition, you'll update and expand the details of the schedule of the work for the remaining layers of the project. Once your project question(s) are well defined, you'll need to work to find adequate data to further understand, refine, and ultimately answer the question(s). Your outline should be sufficiently detailed to allow for useful feedback about the proposed use of data, analysis, and modelling.

Part 2 Deliverables

For this part of the project there are two deliverables that will need to be turned in and posted to your project repository in a separate directory for Part 2 and include brief descriptions along with the files. Completing the deliverables will substantially complete layer 2 of the planning outlined above. In addition to including the report in your GitHub repository, document your exploratory data analysis with figures and short descriptions directly within the part 2 deliverables folder.

1. Report (50 points)
   The report should document the data collection and exploratory data analysis you have conducted to this point, document the demonstrated progress on understating your collected data, the choices you are making in terms of potential predictor variables, and any data reduction methods used. In addition, you can use the original schedule from Part 1 and update to include more details on remaining project elements and indicate what is complete and still remains to be done in layers 2 – 4.
2. Short Presentation (25 points)
   Prepare a short update that would be given in approximately four minutes covering the highlights of your exploratory data analysis that is covered in your report.

Part 3 – Draft of Model and Research Paper

In this part of the project you'll be working with your collected data in light of your exploratory data analysis to develop a predictive model based on your research question. Through this process you'll like create new features derived from your collected data, test and choose the final modelling framework that best answers your research question.

Be sure to separate your data into training and evaluation sets, so that the evaluation of your model performance is done on independent samples and will more accurately measure the predictive power of your modelling choices.

Part 3 Deliverables

There are two deliverables for this portion of the project, a draft of your research paper (detailed below) and a demonstration of how your model works including preliminary results on its predictive power.

Paper Draft (50 points): The draft will need to be turned in and posted to your project repository in a separate directory for Part 4 and include brief descriptions along with the files. The draft of the final report should primarily focus on the introduction, data and methods, and results sections. If you are able to include drafts of the discussion and conclusion sections, that is an added bonus, but not required for your Part 3 submission.

Demonstration (25 points): You'll perform a brief demonstration of your draft modelling efforts for the class, where you should highlight the important challenges and trade-offs you have faced to this point in your effort to address your project question(s).

Requirements for a Research-style paper:

1. **Introduction**:
    a. Describe the context needed to understand the project questions.
    b. Why is this important to ask and develop a prediction for your project question(s)?
    c. Describe at least three papers of related work that support some of the following:
        i. This project/objective is important
        ii. Similar projects or products have not addressed this need (directly)
        iii. Methods such as those you are using are typical for answering the client's questions or evaluating the impact of the results.

    iv. Methods for measuring your project's impact are typical or valuable
d. State your project question(s) (e.g., A project that teaches middle-school aged children about nutrition while they perform exercises that simulate activities like jumping rope will increase scores on a survey about ways to be healthy and perceptions that exercise is fun)

2. **Data and Methods**:
    a. Describe your data and the technology used to analyze your project question(s)
    b. Describe your techniques and/or experimental design to determine whether your project was successful at answering your research question
    c. Describe the measures of success (e.g., the pre- and post-survey questions, heart rate, correctness of in-project actions)
3. **Results**:
    a. Provide appropriate table/chart of the results of your project
    b. Provide graphs and plots that highlight the things you've learned through your exploratory data analysis and modelling efforts.
    c. If performing machine learning, provide a confusion matrix, or other quantification of success/failure with your algorithms.
4. **Discussion**
    a. Describe the data in a way that helps the reader interpret the results.
    b. Tell the reader why the results are or are not as expected. Let us know if there were any issues that came up in testing that may have impacted the results.
    c. What did you learn from your project? Report at LEAST ONE CHANGE you made to your project as a result of the testing.
5. **Conclusion**
    a. Summarize what was done.
    b. State the overall results succinctly.
    c. State the implications of this result.
    d. Reflect on the ethical implications related to studying and modelling the project questions with the data sources you used. What potential biases exist within your prediction framework?
6. **References**
    a. Include appropriate references to support your project methods/ideas. Each team member should be responsible for contributing at least 1 reference.

Part 4 – Final Model and Research Paper

Using the feedback from part 3, and additional work completed in the meantime, you will complete your **research paper**. Your final paper should reside in the top directory of your GitHub repository, along with your final presentation slides.

Final Demo/Presentation Instructions:

Presentations:
Each team is required to give a 12 min presentation with 3 minutes for Q&A during which the next team will get ready to start their presentation. YOU WILL BE TIMED AND MAY NOT RUN OVER.

What to say:
The first half of your presentation should showcase the purpose of your project, the design, and your achievements. The second half should focus on the results. It can be safely assumed that you could use more time to improve your project, so you can skip mentioning that as part of this presentation.

Demos:
Each team will have an additional five minutes to demonstrate and run their project code to illustrate how it works for the class. Make sure your project is in a runnable/testable state so the rest of the class is able to participate in or view a working demonstration.

Additional REQUIRED separate items:

1. Final Presentation file with your final presentation slides
2. Layers containing your work plan document, listing what was completed/not completed and by whom.
3. Final Project Report including:
    a. Executive Summary Page: A home-page level executive summary of your project, process and final results that appears on the main GitHub homepage of the project repository.
    b. Background Page: A page with links and discussion of any background needed to understand your project's technical/science aspects
    c. Methods Page: A page which details the final methods you used to reach any conclusions.
    d. Results Page: A page detailing any conclusions and results you found from your analysis. This SHOULD include at least some charts, tables, or graphs. You might want to imbed interactive binder link here with use of static graphics!
    e. Full Final Report Document: should be in the top-level repository folder.
4. Code:

a. Project runnable files (Jupyter Notebook or Python Script File)
b. readme (.txt or .pdf) file consisting of all the steps required for a new user to test/reproduce the project, and ALL necessary requirements to run your project code.

Overall your project GitHub should have the following folders:

- reports
  - Include folders for each part of the project (1-5).
  - Each part folder should contain a brief summary and description of the content of that folder and the accompanying files
- data
  - Include any data file(s) that are small enough to be stored on GitHub, otherwise within the README file be sure to include the necessary information about obtaining the data (e.g., links to external datasets)
- code
  - Include the final production code that could be used by someone outside of the project to complete your projects objectives.
  - Old code may be included in a separate sub-folder, labeled appropriately.
- Readme file (txt or pdf) to provide any other information like online link, repository, etc.

Timeline:

Part 1: Due 23 September 2021

- All documents uploaded to correct places within the GitHub repository
- Presentations conducted during lab

Part 2: Due 21 October 2021

- All documents uploaded to correct places within the GitHub repository
- Presentations conducted during lab

Part 3: Due 16 November 2021

- All documents uploaded to correct places within the GitHub repository
- Modelling demonstration conducted during lab
- Peer review of draft paper conducted during lab on 18 November 2021

Part 4: Due 16 December 2021

- All documents uploaded to correct places within the GitHub repository
- Presentations conducted during the final exam time for the course