Data Science Project Workflow

### Available Resources

1. This document
2. Online presentation

### Business Problem

1. What is the business problem the project is trying to solve?
2. Where will it appear in the product? How will it affect our users?
3. GET: 5+ examples with associated labels for each of the classes

### Engineering Constraints and Interfaces

1. Hold a design meeting with an Engineering architect
2. What is the API?
   1. Input
   2. Output
3. Is the model generic for all users? Per company? Per user?
4. List cases that would make the problem much harder. These are assumptions you're making regarding the API (for example, only a single class for each observation)
5. When will engineering get to it? Who will be responsible in engineering?
6. What are the engineering constraints?
   1. Running speed
   2. Memory
   3. Size of files

### Project Evaluation

1. Dataset selection:
   1. Is there an existing company dataset you can use for training/testing? If so, [describe it](#_enul3v7d2p9d)
   2. Is there an external dataset that can be used for training/testing? If so, [describe it](#_enul3v7d2p9d)
   3. Otherwise, define the requirements for a new dataset
2. Define the objective function. Consider error cost for different classes
3. What is the expected performance? Get Product to provide a number that they think will make them happy
4. REVIEW: Get a research lead to review and approve the project evaluation

### Initial Research - Getting to know the problem & literature

1. List resources (articles, blog posts etc) that give proper introduction to the field
2. List resources that describe a similar research. This is also where you list any other useful resource that helped along the way

### Exploratory Data Analysis

1. Create a Jupyter Notebook with basic statistics and visualizations of the data
2. The notebook should also show how to preprocess and clean the dataset, if needed, by importing code from .py files
3. Check for outliers
4. Tidy the data
5. Check the units of all data points to make sure they are in the right range
6. Initial Exploratory Data Analysis (EDA) targeted to understanding the underlying dataset

### Basic Model

1. Is the model general / per company / per user etc.?
2. Define the simplest rule-based model
3. Define the simplest ML-based model
4. Define and quickly develop the simplest solution to the problem
5. COMMIT: I can't think of a simpler solution to implement that will do
6. ANSWER: What do you expect the result to be?
7. LIST: Potential confounders
8. Describe a set of product requirements that are similar but will be much simpler to answer from a Research perspective
9. LIST: Similar projects done in the company (for reusing code/ideas)
10. Create a basic\_model.ipynb notebook which shows results of a basic model
11. Report model performance for train & test sets
12. Perform error analysis of the basic model
13. Send an email to the Research group, with a link to the basic\_model.ipynb notebook

### Basic Model Notebook

1. Read data
2. Run basic model
3. Show train+test performance (confusion matrix, classification report)
4. Show important features
5. Show learning curve
6. Show ROC curve
7. Show precision/recall curve
8. Plot Show performance vs. confidence
9. Pre-mortem - discuss when this model can fail miserably/embarrassingly

### Model Error Analysis

1. What characterizes the errors made by the model?
2. What are the main sources of errors made by the model?
3. In what cases does it perform better/worse than the simple baselines?

### Model Performance Analysis

1. Describe the engineering requirements of the model, in terms of:
   1. Running speed
   2. Memory
   3. Size of files

### In production

1. Take into account considerations of product
2. Describe the data you will collect on an going basis in production to improve the model performance, and the ways the data will be saved
3. Describe how often the model needs to be retrained. If retraining is needed, link to a document that describes how retraining should be done
4. Describe the steps that should be taken when the model is applied to a new company/user
5. Describe how you will monitor the performance of the model in production, to ensure that performance doesn't deteriorate with time

### Project Documentation

1. Fill the following information:
   1. Project title
   2. Programming Language
   3. Folder in the code
   4. Jira ticket
   5. Research branch
   6. Docker name

### Tests

1. Develop the Data Sanity check
2. Develop performance tests
3. Develop end2end test for python (assuming you get the input in the format defined, testing that you create output in the format defined)
4. Clearly state all of the assumptions/hypothesis and document whether they have been verified or not and how they can be verified
5. If engineering imposed constraints on memory, run time etc., assert that they are met

### Wrap Up

1. Does it make more sense to write the code in the language Engineering use?

### 

### Checklist

1. I moved code from notebooks to py files. My notebooks only include code relevant for visualization, and import all other code
2. The following parts of the research were done by hand and can't be recreated automatically: (PROVIDE LIST)
3. I wrote a meaningful README file, that is sufficient to introduce a new person to the project
4. I checked for Leakage in my code
5. I am not using data from my test set in my code for preprocessing and feature engineering. I have set my test set to random values and got the exact same performance on the train set.
6. All my figures in my notebooks have labels for their axes
7. All my figures in my notebooks have legends where relevant
8. All my figures in my notebooks have proper titles
9. My notebooks are divided into small blocks with explanatory texts
10. Relevant functions are moved to gong\_utils
11. REVIEW: I have reviewed the notebooks, and am able to understand them and follow up on them

### Dataset Description

1. # observations
2. diversity of data (# companies, time period etc.)
3. was it cleaned from missing values, outliers etc?
4. who annotated it?
5. was it reviewed for labeling accuracy?
6. what's the planned split for train-test?
7. Do you plan to reserve part of the dataset for a dev set?
8. How long does it take to create the dataset?
9. Write guidelines for the labeler
10. Schedule a meeting with the labeler after up to 100 labels or 4 hours of work, to review labels
11. pre-mortem - why would this dataset fail in providing good results?

### Useful resources:

1. <https://datasciencevademecum.wordpress.com/2015/11/12/the-complete-18-steps-to-start-a-new-agile-data-science-project/>
2. <https://blog.k2datascience.com/essential-checklist-for-any-data-analysis-or-science-project-7c4fa924e563>