**Data Science Workflow Template**

1. Identify project roles:
   1. Sponsor: business interest championing the project (get a name and contact info)
   2. Clients: all end users at various levels (both end users and managers of end users)
   3. Data scientists: ideally at least 2 data scientists should pair up at the beginning of every project.
   4. Data architect: individual(s) who manage the data to be accessed for analysis and development. These can be Bis, Application developers/managers, etc...
   5. Operations engineer: individual(s) in charge of deployment and model’s API maintenance within the “live” system.
2. Data Science Project Lifecycle: this is a cycle which means you may go back or forth at any point as the project is refined.
   1. Define the goal
   2. Collect data
   3. Exploratory Data Analysis/Data Management
   4. Build Model and/or create recommendations
   5. Assess whether model/recommendations solve the problem and meet original goal
   6. Present your findings
   7. Deploy model
   8. Live model continuous evaluation (ideally an operations engineers will do this).
3. Defining the goal: it must be measurable, quantifiable, and finite. Some questions to answer during this phase are:
   1. Why do the sponsors want this project? Do they lack something? What do they need to accomplish?
   2. What is it being done to solve the problem now? Is there a current system/application addressing this need? Why is this not good enough?
   3. What resources are available: internal data tables, external data, domain experts, servers, etc.…?
   4. How will the project be deployed: batch mode or on-demand (“live”)?
   5. Identify project roles (step 1) during this phase.
   6. Define metrics identifying when the goal is met. This is a stopping condition or acceptance criteria. For example, “we want to increase acceptance of Spot Shipper offers to 25%”, or “we want to have 20% acceptance of bid counter offers at first counter”, “we want to have a constant rate of 93% of on-time-and-in-full for Walmart deliveries month after month”.
4. Collect data: you must identify what data are needed and what data are available.
5. EDA and Data Management: ideally create a list of questions and use an R notebook or Jupiter notebook to conduct this step, so that you can share it with your team.
   1. Explore the data
   2. Determine the meaning of categorical and nominal values
   3. Determine the meaning of continuous values
   4. Have a general sense how the data are collected into your source table
   5. Is the data quality good enough?
   6. Do you have too many missing values?
   7. Are there errors in the data?
      1. Can they be corrected?
      2. Do you need to impute values? Can you do this without introducing much bias?
   8. What is your unit of analysis:
      1. Orders
      2. Bids
      3. Drivers
      4. Other
6. Build model and/or Create recommendations:
   1. If data were not good enough, make recommendations as what and how to collect going forward.
   2. If a simple solution was found during your EDA, present this to manager and project sponsors. You can work on a more robust solution long term instead.
   3. Determine what can of problem you have:
      1. Classification
      2. Scoring
      3. Ranking
      4. Clustering
      5. Finding relations: correlations and/or potential causes of effects.
      6. Characterization: general visualizations and summarization of the current system.
   4. Determine if you already know an algorithm that may provide a solution or whether you must conduct further research of newer algorithms.
   5. Implement your modeling pipeline:
      1. Sampling
      2. Data transformation/Feature engineering
      3. Training
      4. Evaluation
      5. Cross-validation