Data Scientist **Hackathon**

Data Science Venn Diagram v2.0
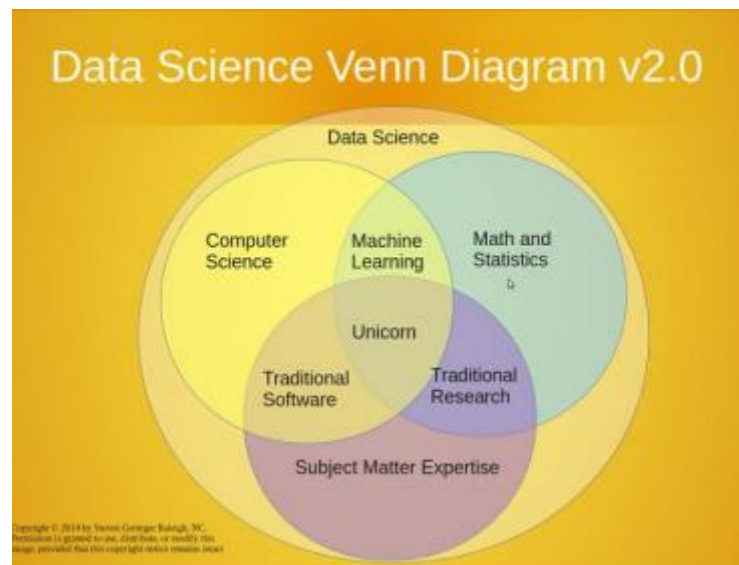
1. **Hard Programming Skills (computer science)**
    a. Able to collect data.
    b. Able to process and format data.
    c. Able to study data structure, analyze and re-format according to the objective.
    d. Code clean Objects/Functions with efficient algorithms to produce output information according to the objective.

2. **Math and Statistics**
    a. Core mathematics
    b. Understanding of statistical and probability concepts and theory.
    c. Analytical Skills (Subject Matter Expertise)
    d. Inference capability given information and general guidelines of propose or objective.
    e. Move from outputs to valuable and actionable insights.

3. **Soft Skills**
    a. Proactive

     b. Resilient

     c. Detail Oriented

     d. Able to manage client accounts

     e. Able to communicate insights from constructed models

1. CASE:

Security is a security service company that offers a variety of products to its clients. In specific, their star product, Safety", has been in the market since 2012 and has been a total success in the European Market. For this product, a client must sign a contract, pay the installation costs, and a monthly fee for the service. However, in recent years, Security has seen that a lot of its clients have canceled Safety" before two years. The biggest problem with this phenomenon is that Safety" has a significant fixed cost. 80% of this cost is assumed by Security in order to be able to match the prices of its competition. Therefore, if a client leaves before two years, the company is not able to recover their investment.

Thus, Security has contacted you to help them develop a classification model that allows them to predict if a client will stay less than two years with the product. With this model, they plan to construct commercial campaigns that benefit those clients that will stay more than two years and penalize the ones who will not.

Security has recently bought information from a credit bureau because they think that this information will be valuable to build the classification model. Specifically, they asked for each client's credit score 6 months before and 6 months after he/she purchased " Safety". Additionally, Security asked the client for all the transactions in their bank account. For this project, we will assume that each client has only one bank account. Also, most of the clients that apply for " Safety" usually have another product from Security.

1. **Objective:**

     a. Develop a two-year churn predictive model.

2. **Desired Population:**

     a. Contracts from 2015 onwards.

     b. Operations in Italy were closed this year (2019).

     c. Take out clients with more than 75% of their info missing.

     d. Each client should only have one contract in the database, no more.

     e. Remember that for a client to be eligible, it must have at least two years of information within the company (even if it canceled the product).

3. **Databases:**

Delivered on 30/11/2019 (day/month/year).

a. Clients:
   i. Historical database of clients that have purchased the product " Safety" from 2012 to the delivery date.

b. Products per client:
   i. Historical of the products the client had with the Security at the moment he/she applied to the new product.

c. Transactions in the client's bank account:
   i. Assume that the client opened the account with 0 balance and that all have only 5 transactions.
   ii. All transactions are pre-application moment.

d. Score from credit bureau:
   i. We asked the bureau for the score of our clients 6 months before and 6 months after the client applied to the new product.

2. INSTRUCTIONS AND DELIVERABLES:

In order to assess your work, Security is asking for you to construct a presentation of at most 7 slides with the following information:

- Clean Client Database: Apply filters to find the final desired population (look at section "Desired Population" above):
  o Include the number of records in the "Client" database after each filter.
  o State clearly the number of records in the "Client" database after applying all filters.
- Obtain relevant variables: use the different data sets to find the following variables:
  o Products data set: Number of products per client at the moment of application.
  o Transactions data set: Balance in the bank account at the moment of application.
  o Credit bureau Scores data set: Score of the credit bureau at the moment of application. Consider it is a monthly score.
  o Clients data set: Age of the client at the moment of application.
  o Append these new variables to the Clients data set.

- o Find descriptive Statistics of these new variables: Mean, standard deviation, maximum, and minimum.
- Build a predictive statistical Model: use all the available variables to build a statistical model that predicts the probability of a client dropping/canceling the product before 2 years:
  - o Explanation of why you choose that model.
  - o What insights did you find about the model?
  - o Accuracy of the model (compared to a baseline).