CloudFrame Analytics

Introduction

This document is meant to be a guide for data science teams' initial interactions with customers. The questions suggested here can be used to define and refine project objectives. In our experience following this line of questioning has resulted in product definitions that are useful and worthwhile. It is a general guide, and should be customized with knowledge of the prospective customer and domain.

Definitions

Consumers: people or organizations who will use the recommendations to inform decisions.
Customers: people or organizations that are experiencing the problem with a vested interest in improving the outcome.
Data: information emanating from the problem as it manifests itself within its context.
Domain: the field, business, or subject matter in which the problem exists.
Model: a statistically or computationally learned (i.e. resulting from an algorithm) object that accepts an observation as an input and produces an inference or classification.
Observation: data for individual manifestations within the phenomenon.
Outcome: a state of an observation resulting from a phenomenon.
Phenomenon: a fact or situation that is observed to exist or happen, especially one whose cause or explanation is in question.
Predictors: data (in the form of fields or columns) that can be observed at or prior to the time of inference.
Problem: a challenge or opportunity to improve an outcome that may be addressable with data science methods.
Product: A quantitative result that is defensible, repeatable, and useful in addressing the problem. Could be a predictive model or a data summary.
Project: a data science effort to address the problem.
Recommendation: a finding that provides the organization with defensible information with which to inform decisions.
Response: data (one or more fields or columns) that records some aspect of an outcome that is to be inferred or predicted.
Signal: differences in data that distinguish a particular class of the response variable from others, or observable changes in a regressed response variable that are indicated by predictors.

The CloudFrame Method

I Define the problem

II Connect the problem to data

III Assess the available data

IV Frame the potential solutions

CloudFrame Analytics

Start with a problem in mind

Questions that aim to define the customer(s), problem, and outcomes. Most often, and especially in the commercial sector, you should come to this interaction with an idea of what problems exist. It is NOT recommended to start cold with no knowledge of the domain.

Sample Questions
• What are the outcomes that your organization is trying to improve or affect?
• Are you currently measuring those results, and if so, how?
• Who will be positively impacted by the improvement? (could be individuals, communities, business units, or the entire organization)
BONUS: What have you tried in the past to improve these results?
Answer for yourself: Are there any stakeholders who are not incentivized to see the results improve?

Now you should be able to write a concise problem statement. For example:

<customer> is experiencing <outcome> results, and the organization desires to improve <outcome> by the following measurements...

CloudFrame Analytics

Connect the problem with data… maybe

Questions that aim to determine whether or not the problem is addressable by data science methods.

Sample Questions
• What events or pieces of information do you believe influence the problem?
• Do you currently collect that information, and if so, how is it stored?
• Is your organization amenable to collecting new data for the purpose of this project?
Answer for yourself: If the data is collected, does it actually relate to the problem? If the data is not collected, is it collectable?

You should now be able to list available, or potentially available, data sources that may affect the problem. If this list is insufficient, the recommendation from this process should be a report to that effect. That is:

> <problem> is not addressable by a data science approach at this time due to a <reasons>. We recommend…

Further, a possible result is a data collection plan. If so, you should proceed with defining the product with an explicitly stated assumption that data collection will be successful.

CloudFrame Analytics

Assess the data

Questions aimed at assessing the state of the data; including location, availability, veracity, and accessibility.

Sample Questions
• How well does the data you collect represent reality? Can you estimate the percentage of data that is somehow faulty?
• How far back does the data go and how often is it updated?
• Is it archived or deleted?
• What format are the various data stored in? (Relational DB, text blobs, et cetera)
• Is the information you collect stored locally or virtually (e.g. cloud storage)? How much of each type of information?
• What approvals are needed to share the information?
• Is any of the data considered sensitive? Why? (e.g. Personally Identifiable Information or privacy or business sensitive)
• Can you release a comprehensive data sample?

You should now be able to produce a table of available data sources with assessments of its characteristics. We recommend a determination of how accessible, how complete, how realistic, how frequent, how big, and how painful it is. How painful is an individual assessment of how much effort will be required to utilize the data to address the problem; that is, how difficult will it be to munge each data source.

CloudFrame Analytics

Identify the desired characteristics of the product

Questions aimed at determining the correct data science approach to addressing the problem. The product should exist at the confluence of the data and the problem.

Sample Questions
• Transparency: How important is it to you to understand the reasons why specific recommendations or inferences are made?
• Control: How much interpretation of the data should be left to the consumers?
• Decision authority: How involved should consumers be in the final decision? Consider the analogy of driving a car. A speedometer is a data science solution that is transparent and gives the user full control over the recommendation (how fast should I be going?) and the decision (how fast will I go?). A lane departure warning is a data science solution that is transparent and a controlled decision, but not a controlled recommendation. A self-driving car is not transparent, not a controlled recommendation, and not a controlled decision. Answering this question should help scope the technical effort.
• Does your organization have an infrastructure for acting on the results of this project? If not, is the organization amenable to the development of said infrastructure (e.g. cloud platforms)? Infrastructure could be a human organization, a computational architecture, or some combination thereof.
• Is the information you collect stored locally or virtually (e.g. cloud storage)? How much of each type of information?
• What approvals are needed to share the information?
• Is any of the data considered sensitive? Why? (e.g. Personally Identifiable Information or privacy or business sensitive)
• Can you release a comprehensive data sample?
Answer for yourself: Are the consumers of the results capable of ingesting or interpreting the results?

After this step you should be able to define the product definition. For instance:

> The organization requires <data science method> to address <problem>. It will be measured by sustained changes in <outcome>, and will be deemed successful if it achieves <percentage> improvement.

Conclusion

In closing, arriving at a product definition is more complicated than merely defining what is mathematically or computationally feasible. A deep learning model may perform the best, but may not provide the appropriate transparency. Likewise, a pivot table may accurately summarize the data, but may not provide a result that allows for useful inference.