Data Science Project Scoping Worksheet[1]

1. **Project Name:** Developing predictive risk models to support the decisions of the child maltreatment hotline in the County of Allegheny (Pennsylvania)[2]

2. **Organization Name:** Department of Human Services (DHS) of Allegheny County

1. Problem Description:

1.1 What is the problem you are facing?

Allegheny County has 2 emergency hotlines where child maltreatment or abuse can be reported: the county's emergency hotline and a specialized emergency hotline, "Childline". Complaints can be classified as CPS (Child Protection Services) and GPS (General Protection Services). When a complaint is classified as CPS, a thorough investigation is immediately generated to see whether the child and their environment must be intervened by the county, and that household is considered as already part of the system. When a complaint is classified as GPS, the county's emergency hotline personnel must decide on one of the following actions: not to investigate or intervene, to investigate the complaint in a general manner or to investigate the complaint in depth. A household is said to enter the system at the time that a thorough investigation is mandated.

---

[1] Created by the Center for Data Science and Public Policy at the University of Chicago and GobLab at Universidad Adolfo Ibáñez. This project description was created by GobLab at Universidad Adolfo Ibáñez as part of a curriculum that is available here. You may use it quoting its creators and adapt it following the Creative Commons Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0) License. The terms and conditions are available here.

[2] This worksheet was created by GobLab UAI for teaching purposes. It is a retrospective exercise conducted using public information that doesn't necessarily represent the project's initial formulation.

From the total number of complaints classified as GPS, 48% are entered into the system (April 2010-May 2016).

The problem is that from the complaints that did not end in an intervention, for 53% of the cases there was a new complaint within 2 years. Therefore, the current decision system leaves out children who need support from the county. Given the high call rate and the fact that family data must be accessed manually from other sources, the operators of the emergency hotlines must make decisions in a short time and they are not always able to make an adequate risk profile for the child.

1.2 **Who/what is affected by this problem?** (Certain type of people, organizations, neighborhoods, environment)

Children who did not receive an adequate intervention from the authority. Families who suffered a negative stigma because they were intervened when it wasn't really necessary or who did not receive an adequate intervention. Employees of the county's child protection system who had to investigate low-risk cases, casting aside other riskier cases. Neighbors of Allegheny, whose taxes are not used efficiently.

1.3 **How many people/organizations/places/etc. are affected and how much?** (e.g. mean wait time for surgery, number of students dropping out of school, cost due to tax fraud, etc.)

The Census currently estimates that Allegheny County has 1,218,452 inhabitants, from which 18.7% are under the age of 18 (227,850) and 5.2% are under the age of 5 (63,359)[3].

15,678 calls were received in 2017, from which 11,751 were classified as GPS.

1.4 Why is solving this problem a priority for your organization?

The Department of Human Services (DHS) of Allegheny County has the mission of creating an accessible, culturally competent and comprehensive human services

---

[3] U.S Census Bureau. Estimations as of July 1st, 2018

system that ensures services for the population of Allegheny County, especially those who are most vulnerable.

A process was initiated in 2014 that seeks to improve the ability to make decisions based on data efficiently and consistently, ensuring that public resources are allocated equitably according to vulnerability, and to improve general aspects of the health and wellbeing of Allegheny County residents.

2. Goals (in order of priority)

- The technical solution (e.g. predictive model) is not the goal.
- The goal must be measurable.
- Achieving the goal should help solve the problem.
- Typical goals include improving/maximizing/increasing or decreasing/mitigating/reducing some outcome or metric.
- Typical constraints include budget, lack of human capital, legal restrictions, political will and social license.
- Consider tradeoffs between conflicting goals.

	Goal	Constraints
1	To reduce the number of cases of children who are not visited after a complaint and who are separated from their family after a judicial decision in the following 2 years	Only households that have been reported are visited
2	To increase the proportion of cases visited after a complaint that do require an intervention from the county	Not all cases reported to the county can be visited; approximately 50% is visited
3	To decrease the proportion of cases visited after a complaint that do not require an intervention from the county	

3. Actions

- Actions are what institutions can do to address a problem by allocating resources, for instance, inspect facilities, provide preventive services, outreach, etc.
- Ideal actions should help you achieve the goal defined above.

	Action 1	Action 2	Action 3
Action: *e.g. inform the owner of a vehicle on how to renew their permit*	To decide whether a complaint must be investigated	To enter a child into the child protection system	
Who is executing the action? *e.g. IT department (sends email)*	Allegheny Child Services (call center)	Allegheny Child Services	
Who/what is the action being taken on? e.g. vehicle owner	Children in households from which a hotline complaint is received and their family environment	Children in households from which a hotline complaint is received and their family environment	
How often is the decision to take this action made? e.g. annually	After each telephone call	After the hotline complaint (CPS)/after the investigation (GPS)	

What channels are/can be used to take this action? e.g. email	Face-to-face meetings of operators and their supervisors	On-site visits to the reported household	
Other useful information about the action	Other information sources that may be useful for decision-making are researched at the face-to-face meetings	There is no standard protocol for decision-making	

4. Data

- The data has to connect to the actions it informs so the organization can achieve its goal.
- Typical data science projects use administrative data as the primary data source and enhance it with publicly available data sources (Census, other open data). Partnering with the private sector or non-profits could be a way to obtain data you might be missing internally.

A. What data sources do you have internally?

(add columns for more sources if applicable)

	Data of calls entered into the system	**Data Source 2**	**Data Source 3**
What does it contain? e.g. hospital admission and discharge records at a national level	Historical information of complaints and household demographics from April 2010 to April 2014		
What level of granularity? e.g. transaction, person, organization, location	At a family level, members of the family		

How frequently is it collected/updated once it is captured? e.g. in real time, daily, weekly, monthly, yearly, one-off	In real time		
Does it have reliable and unique identifiers that can be linked to other data sources? e.g. national identifier	Yes, identifier of the household members at the time of the complaint		
Who is the internal owner of the data? e.g. hospitals	Allegheny Child Services		
How is it stored? e.g. in a database, PDF, Excel	Database (Allegheny Datawarehouse)		
Additional comments			

B. What data can you get from external, private or public sources?

	County Prison	**Juvenile Probation**	**Public Welfare Service**
What does it contain?	Information of past interactions with the county jail	Historical information of interactions with the Allegheny juvenile jail	Information of people who have been part of a welfare program (temporary assistance, general assistance, income support, food subsidy, etc.)
What level of granularity?	At an individual level	At an individual level	At an individual level
How frequently is it collected/updated once it is captured?	Monthly	Monthly	Monthly
Does it have reliable and unique identifiers that can be linked to other data sources?	Yes, personal identifiers	Yes, personal identifiers	Yes, personal identifiers

Who is the internal owner of the data?	Allegheny prison	Allegheny County Office of Juvenile Probation	Allegheny County
How is it stored?	Excel/SQL	Excel/SQL	Excel/SQL
Additional comments			

	Behavioral Health Programs	Census Poverty Estimation
What does it contain?	Historical information of patients who have received assistance from the office or have been diagnosed	Poverty level in each zip code (a little more than a block)
What level of granularity?	At an individual level	At the level of a little more than a block, less than a neighborhood
How frequently is it collected/updated once it is captured?	Monthly	Every 10 years
Does it have reliable and unique identifiers that can be linked to other data sources?	Yes, personal identifier	Yes, zip code

Who is the internal owner of the data?	Allegheny County Office of Behavioral Health	Central Government
How is it stored?	Excel	Databases
Additional comments	This database includes alcohol and drug services, services related to the judicial power, mental health services, among others	

C. In an ideal world, is there additional data you would want to obtain/gather that would be relevant to this problem? (Surveys, CCTV, phone records, DNA, different frequency or granularity for currently available data, etc.)

5. Analysis
- Typical data science projects include a combination of analyses.
- The analysis is not the goal of the project.
- Choose the right analysis for the right problem.
- You must validate the analysis, and the validation process must match your goal.

	Analysis 1:	Analysis 2:	Analysis 3:
Type of analysis (e.g. description, prediction, detection, behavior change)	Prediction		
Purpose of the analysis (e.g. understand historical behavior of individuals, estimate a patient's risk of disease)	To predict a child's risk of maltreatment upon receiving the call in order to reinforce or dismiss the decision of a household intervention		
Which action will this analysis inform?	Action 1: the decision to investigate (enter them into the system)		
How will you validate this analysis using existing data? (e.g. using historical data, running an RCT)	Training of a predictive model based on historical data		

1. Ethical Considerations

Privacy Are you working with personal and/or sensitive data that is individually identifiable?	Yes, there is personal and sensitive data, and we must use innominate indicators.
Transparency Which stakeholders should know about which parts of the project? (Stakeholders typically include policymakers, frontline workers, people who will be affected by the actions, etc.)	Regional government of Allegheny, users of the system
Discrimination/Equity Are there any specific groups for whom you want to ensure equity of outcomes?	There is a racial bias present in the historical data. Afro-American children are almost 3 times more likely to interact with a child protection service than white children because the population tends to report Afro-American families more. They are also more likely to be entered into the system once the call is made. We will contemplate reviewing the effect that the model has on the system's disparities.
Social License If the entire population of the country finds out about your project, will they be ok with it?	Careful outreach work must be conducted. It is always complex to talk about algorithms in interventions with people.

Accountability Who are the people responsible for all the things above?	Director of the Allegheny Department of Human Services
Other considerations such as consent, legal, etc.	When working with sensitive data, it is important to use innominate identifiers to safeguard people's privacy.

6. What field trial or randomized controlled trial can you design to validate the project in the field? The outcomes you will measure should match your goals.

To cross-reference the data of calls with the information of the Pittsburgh Hospital and measure whether children with a high estimated risk have had to go to the emergency system. This way, we can empirically measure whether a child with a high estimated risk is also at a real risk.

7. Who are the external organizations and internal departments that will need to be involved?
(Typically, data science projects need involvement from data owners, IT infrastructure owners, the problem owner, analytics people)

Organization/Department	Description of desired involvement	Name/role of counterpart
Allegheny Office of Human Services	Coordinate internal work	Marc Cherna/Director

IT	Real-time implementation of the algorithm in the computers	
Statistics Department	Coding of variables for the analysis; permanent updating	
Child Hotline	Training of telephone operators in the new tool	
Communities of interest (e.g. representatives of the Afro-American community, academics, etc.)	Provide their opinions for consideration in the design and implementation of the system	