Analytics Lifecycle

Jonathan Grundy

Project Description:

Mental health issues are a problem for any city, and New York City is no exception. Overall trends indicate that 1 in 5 people will experience acute anxiety or depression at some point in their lives, which in New York means almost 1.8 million New Yorkers are likely to experience these issues. Furthermore, this covers just depression and anxiety, and doesn't account for other mental illnesses or substance abuse problems.

Combating these problems requires a significant investment of time and resources on the part of the mental health industry. Mental health professionals are needed to treat those experiencing symptoms, and greater education is needed to better inform the public about the dangers of mental health issues, recognition, and where and how to seek treatment. The city of New York is well situated to deliver these services and, through ThriveNYC, has already promised to do so. However, for this to be effective, it's important to determine how best to allocate those resources.

This project will seek to best determine where to allocate mental health professionals where they will have the greatest impact, as well build a demographic model of the areas with greatest mental health issues for purposes of education and prediction of future mental health hotspots.

Phase 1: Discovery

- Problem Framing
  - The client, in this case the city of New York, requires information that will allow them to better allocate their resources to combat mental health problems in the city. Key stakeholders in this case would be New York City (particularly the Department of Health and Mental Hygiene), mental health professionals, hospitals, schools, mental health facilities, those suffering from mental illnesses, and community members. The current situation is fair, in that services are already provided, but a major pain point is that DOHMH is planning to introduce several thousand more trained professionals and they need proper allocation. Additionally, it is not clearly understood where educational services should be offered.
- Domain
  - Solving this problem requires deep knowledge of the mental health system, as well as existing literature and research about mental health. Knowledge of city structure and how resources are allocated in the city can help determine the most effective ways of using resources, or methods that decidedly would not work.
  - The objectives are threefold:

        - Determine the most efficient allocation for mental health professionals in New York City
        - Determine the areas of the city most in need of mental health educational services
        - Building a predictive model of the kinds of neighborhoods that suffer most from mental health problems
    - Success conditions will be the satisfaction of the first two, with the third as a stretch goal. Failure would be failing to accomplish the first two goals.
- Resources
    - The main data source is semistructured mental health survey data. Most responses are ratings on a 1-5 scale, though there are some text-based responses as well. Other data sources include Census demographic data, and locations of hospitals and schools. CDC mental health survey data for New York state may be added if time and structure permits.
    - Available technology includes programming languages, GIS products, and team members adequate the fill the necessary roles for the project, drawn from city workers and analytics professionals.
    - The scope in timeframe will be 4 months, so as to provide time to determine strategy for the distribution of professionals at the end of calendar year 2016. Approximately 100 person hours are expected to finished this project
    - Currently, resource requirements for the project appear to be met.
- Hypotheses
    - IH0: Areas with low median household incomes or high crime rates are places that have higher rates of mental illness incidence.
    - IH1: Areas with higher income will have better access to health care services and insurance.
- Roles
    - Business User/Project sponsor/Project Manager: Director from Department of Health and Mental Hygiene
    - Business Intelligence Analyst: Lower-level analyst from DOHMH
    - Data Engineer/Database Administrator: IT support person from DOITT
    - Data Scientist: Analyst from MODA

Phase 2: Data Preparation

- Phase Objectives: Prepare the data for analysis. The survey data is the main information source, but it needs to be joined to the demographic data in order to evaluate neighborhood parameters. Shapefiles for the city will likely need to be brought in to provide geospatial context for the information. Locations of facilities can be joined here as well.
- Roles

    - Business User/Project sponsor/Project Manager: Provide context for the information and help identify further data sets. Will also advise about how to incorporate CDC level data with the more local data.
    - Business Intelligence Analyst: Assist the project manager in determining if all the information available is appropriate, finding new data sources, helping with queries, and preparing descriptive statistics.
    - Data Engineer/Database Administrator: Set up the analytic workspace, and do the ETLT of the data to make sure that it's analytics ready, fine-tune SQL queries.
    - Data Scientist: Examine data to see evaluate how well it can be joined, whether it needs normalization/regularization. Begin visualization of the data in Tableau, or Python (with matplotlib) if necessary.
- Tools used: Tableau, Python (potentially), SQL, GIS

Phase 3: Model planning

- Phase Objectives: Explore the data further to determine which methods would be best for this particular data structure and volume.
- Roles
    - Business User/Project sponsor/Project Manager: Advise about variable selection, work with data scientist to determine which methods most appropriate for testing each hypothesis.
    - Business Intelligence Analyst: Work with data scientist to determine best methods while keeping an eye to how the methods are meeting overall project goals. Do exploratory linear modeling.
    - Data Engineer/Database Administrator: Prepare the data to be processed at scale, especially given use of ML techniques. Advise data scientist about proper analytic workflow to ensure efficient operation of analytic methods.
    - Data Scientist: Determine appropriate testing methodology and start planning scripts for the analytics phase to ensure proper workflow. Do relevant GIS work.
- Relevant evaluation methods: Clustering (for neighborhood attributes), classification (to create prediction model for future neighborhood problems), NLP (for parsing survey text responses)
- Tools for this phase: GIS, Python, SQL, Multiple regression

Phase 4: Model building

- Phase Objectives: Build relevant models based on methods identified in Phase 3, focusing on reproducible code. Develop training and test data sets.
- Roles
    - Business User/Project sponsor/Project Manager: Make sure project is on track and meeting overall project objectives.

    - Business Intelligence Analyst: Assist data scientist in evaluating validity of results from model, continue to offer domain support. Offer suggestions for data visualization. Work on production of training and test sets.
    - Data Engineer/Database Administrator: Continue to make sure that the data is being processed smoothly and efficiently. Make suggestions about algorithmic efficiency. Work on training and test sets.
    - Data Scientist: Produce (or oversee), production of the model code and evaluate results of the training data sets, making appropriate adjustments to the models based on the outcomes from the test data. Produce visualizations.
- Relevant tools: Python, GIS, SQL, H2O

Phase 5: Communicate results

- Phase Objectives: Determine if the model results fit with the objectives of the project and present the results to key stakeholders (in this case the Mayor and relevant city agencies). Identify key findings.
- Roles
    - Business User/Project sponsor/Project Manager: Evaluate if the project goals were met, present to key stakeholders, summarize findings and quantify project value.
    - Business Intelligence Analyst: Assist in the identification of key findings. Help with result interpretation. Assist in determining if the hypotheses were verified or not (possibly via statistical testing).
    - Data Engineer/Database Administrator: Prepare the systems to be operationalized, if necessary.
    - Data Scientist: Interpret the results of the model, prepare presentation visualizations.
- Relevant tools: Python (matplotlib), PowerPoint

Phase 6: Operationalize

- Phase Objectives: Run a pilot (or wider implementation) of the program
- Roles
    - Business User/Project sponsor/Project Manager: Identify key implementation issues, oversee delivery of final deliverables.
    - Business Intelligence Analyst: Assess the benefits of the trial/pilot, report back on relationship to original goals and KPIs.
    - Data Engineer/Database Administrator: Prepare the analytic environment for greater scale, communicate technical methodology to those for responsible for wider implementation.
    - Data Scientist: Oversee model execution on production data, retrain models as necessary. Produce friendly code for others to follow if necessary.