Capstone Project Proposal - Group 1: Wendy Xu, Qinyun Cai, Jiarui Chen, Ze Yu

Introduction that lays out the background, the research questions, and a general explanation of the data

- SpotDX is a healthcare technology company that focuses on developing and deploying rapid diagnostic tests for infectious diseases, including COVID-19. With a focus on remote care, SpotDX can personalize products based on user's needs and provide individualized customer experience. The process of remote care is simpler compared to traditional healthcare: customers will receive their test kit, collect the sample, ship the sample, and get results via mobile devices. Currently, SpotDX is figuring out ways to improve the remote care process to reduce the cost of operations. Specifically, SpotDX wants to know "How can we increase the success rate of sample testing?" Based on this question, our team will develop statistical models that predict test failure/rejections. We will also explore what factors/variables could potentially damage the sample and lead to test failure/rejections.
- SpotDX provides the dataset. It contains information about customers, demographics, test kits, customer behaviors, and test results. For more details regarding the dataset, please refer to Appendix 1 as attached.

A methodology section on the initial choice of analytical methods you will try to use for the first research question. You are free to change the course on the methodology as you work on the project.

- Data preprocessing
    - Remove N/A values
    - Convert characters into factor / categorical variables
    - Use regular expression to standardize date and time
    - Create additional behavioral variables
- Visualization
    - Bar chart (categorical variables)
    - Histogram (age)
    - Map (zip code)
    - Timeseries (time related variables)
- Classification Models
    - Logistic regression
    - Decision tree
    - Neural Network
    - TBD

Ideas for the second research question (creative value-adding advice)- just a list of brainstormed ideas is fine for now, and you can add new approaches as you work on the project.

- Identify the reasons for rejection, then back forward to identify the factors which are significant to be improved.
- Whether the time gap between delivery time and register time is matter to the rejection?
- Whether the gap between the sent-in and the time of lab receiving the sample matter? Do we need to consider that the sample may have an expiration date?
- Should we suggest the customer complete the test in a certain period?
- Should we set up a suggested period for samples to be sent back (after the test)? For example, the sample should be sent back by a quicker carrier compared to winter. Or perhaps the lab can collaborate with some reliable carrier.

A week-by-week timeline of how your group will progress through the project. Indicate important milestones and plans for regular out-of-class meetings if any.

- Week 10
    - Background research - Wendy
    - Data preprocessing - Claire / Jason
    - Data visualization -
        - Bar chart (categorical variables) - Ze
        - Histogram (age, time gaps) - Claire
        - Map (zip code) - Wendy
        - Timeseries (time related variables) - Jason
- Week 11
    - Build models
    - Verify the accuracy and error rate
- Week 12
    - Model selection
    - Start presentation slides
- Week 13
    - Finalize the presentation deck
    - Start writing report
- Week 14
    - Finalize the report
    - Mock presentation

Appendix 1 Data dictionary

Demographics

• Patient ID (hash of patient MRN)
  - Based on common sense, we believe that patient ID won't have any significant influence on the test result. Can be deleted, with no meaning.
  - The unique identifier of individual patient
• Patient ZIP - breakdown in 9 categories https://en.wikipedia.org/wiki/ZIP_Code
  - The geological location of the patient
• Patient Age
  - The age of the patient
• Patient Sex
  - dummy variable (0 or 1)
  - The biological sex of the patient

Kit Information

• Type of Kit (hash of kit type field)
  - Categorical variable
  - 48 levels
• Specimen Type - (Saliva, Blood Spot, Buccal Swab, Fecal Swab, Urine, etc.)
  - Categorical variable
  - 7 levels
  - The substance being sampled or tested
• Client ID (hash of the Spot client field) - distributor (Costco, Walmart etc…)
  - Categorical variable
  - 18 levels
  - The unique identifier indicates the distributor of the kit.

Behavioral Information

• Datetime Kit Delivered
  - The date and time when the kit is delivered to the patient
• Datetime Kit Registered (when they used the kit)
  - When customers register their accounts online
• Datetime Sample Sent In (they are supposed to return right away)
  - The date and time when the patient sent the sample to the lab
• Datetime Sample Received
  - The date and time when the lab received the patient's sample
• Datetime of End Status
  - The date and time when the lab produces the end status
• End Status
  - Rejected or not
  - Independent variables (results)
• Reject reason
  - The reason that the sample was rejected from the end status.