Meeting_1: Meeting minute [Nov 25,2019] – With Brian

Team Attendee: Paul, Kevin, Xing, Emanuel, Stephen

- Project Roles:
  - Kevin will do financial projections
    - ○ Savings to insurance company
    - ○ Reduction of financial penalty to hospital as a result of re-admission
  - Emanuel will also work on the financial projections
  - Stephen will take care of the data visualization and EDA
  - Paul will take care of the data visualization and presentation preparation
  - Xing will take care of model compiling
- Target Variables:
  - 0 is >30 days and No admission
  - 1 is readmission < 30 days
- Possibly explore how different models affect the lift, and ultimately the finances. Whether a simple model is sufficient or whether using a more complex model is justified based on the lift/economics
- Purpose of the project:
  - To get a model in CSV file for diabetes prediction:
    - Taking 10 readmission patients for each model and average it to represent for a group:
    - **<u>Population: Readmission- 2.5* sample group</u>**
    - Using 2.5 as a lift factor to calculate to a population of the large group.
    - Normally, Lift model Avg. =1 for a sample group.
  - Model:
    - Use "AUC" or "C-stat" and sensitivity scores to evaluate the accuracy of the model.
    - Brian expects to have 0.6 with target 0.68-0.69 for a group readmission rate.
- Building a model from the diabetes readmission:
  - The purpose is to help the hospitals to not be penalized.
- Brian will provide a lift chart, going from AUC to lift
- Financial Model [Correlated to reducing readmission]:
  - Differences of AUC score will affect the financial part.
  - Brian helps us to link the readmission rate to finance in order to show how much impact the model will affect to improve a business.
- Reason to use the old published dataset:
  - Availability & privacy of the patients
- 10% hold out for the test data, 90% for the train [Xing]
- Q&A about the dataset.
  - Refer to Meeting_1 Q&A Below

Questions about dataset:

1. 'num_lab_procedures' : how does number of lab procedures affect readmission? Link to error?

Ans: The numbers of the lab procedure refer to a complexity of doctor diagnosis. The column will link directly to the prediction result.

2. 'num_procedures' : What's 'num_procedures'? Patient readmission?

Ans: The numbers of procedure will add up the complexity and will affect to the prediction model

3. 'number_outpatient' : What's outpatient in the healthcare terminology?

Outpatient (OPD) = Clinic / Don't have to admit in the hospital

4. 'number_inpatient' : What's in patient in the healthcare mean?

Inpatient= patients that require to stay in the hospital after the treatment

5. 'medical_specialty' : How to interpret **Surgery-Cardiovascular/Thoracic ?**

Ans: Each type of medical specialty will count as one category for each type.

Meeting_2: Meeting minute [Nov 26,2019]

Team Attendee: Paul, Kevin, Xing, Emanuel, Stephen

- Everyone will stick with the same plan:
    - Data cleaning
    - EDA
    - Modeling
- Focus on features which affect the outcome of readmissions.
- Waiting for Brian to get information about finances.

Meeting_3: Meeting minute [Dec 2,2019]

Team Attendee: Paul, Kevin, Xing, Emanuel, Stephen

- Summarize about the update on the progress.
    - Everyone has 5 cleaning datasets.
        - Each presents and debate about their data cleaning work & modeling.
    - Be ready to work on modeling
- Paul will make a summary note for statistical explanation.
- The team practiced on Github and uploaded their files on the Github.
- Restructure the work schedule:
    - EDA : 12/2-12/4
        - Assigned: Kevin, Stepehn
    - Modeling : 12/4 – 12/5
        - Assigned: Paul , Emanuel
    - Finance : 12/4 – 12/6
        - Assigned: Xing
    - PPT : 12/6/12/8
        - Assigned: All
    - Brian will provide financial part in two days.
        - Brian doesn't want us to spend more time on the financial part.
        - May try to use average cost and turn it in to savings.
    - Suggesting playing with rebalance and technique. AUC/ Patients
        - SMOTE
        - XGBoost
    - Model needs to be able to handle correlated data.

    - Try NLP or Machine learning.
- Model concern: Should we work on more complex model and increase more accuracy on data prediction?
    - Who will be the consumer of the model - > best performing models.
    - [increase or decrease to affect final outcomes.] – Impact in savings.
    - How to interpret the data? – Create a profile □ high risk/low risk.
    - Diabetes code compare
    - LIME Model – to interpret ML model.: How to add interpretability to higher complex model.
    - https://towardsdatascience.com/understanding-model-predictions-with-lime-a582fdff3a3b
    - Finance team interested in savings.
- AUC – We will see some trade off that the higher model we are not gonna get anything.
- AUC is a quick way to identify a large number of models. Easily to implement at the end of packages. Top ten percent individuals -classify as high risk.
- AUC correlates with true positive rates.
- We have to perform the EDA and make one dataset that we all agree to proceed with
- Explain calculattion of “Lift”

Meeting_4: Meeting minute [Dec 5,2019] - Brian

Team Attendee: Paul, Xing, Emanuel, Stephen

- Financial Part:
    - Xing found 2 dataset that we can use for the financial dataset.
    - Hard to find the soft numbers correlation with the hospitalized cost.
- Progress:
    - Paul/Xing / Emanuel are working on modeling
    - In the past two days, we worked on data cleaning.
        - Stuck on some column
            - New features
            - Suggested to throw typical numerical columns
            - Suggested to drop 'diag_2' and 'diag_3'
            - On the repeated encounters – Use the "last" encounter.
            - If we drop the use the same test, but on train can manipulate.
    - SVM – not working- no predict proba
    - Try more higher model.
- Presentation on Wed.
- Check up call at 6 pm. [Wait for a confirmation]
    - Agenda – Presentation/ model

Meeting_5: Meeting minute [Dec 6,2019] - Brian

Team Attendee: Paul, Emanuel, Xing, Kevin, Stephen

- Update on the progress:
  - Choose the dataset. [just made another final decision]
  - Model:
    - Logistic Regression
    - Tree Model
    - Gradient Boosting
    - Etc.
  - Class balance
    - Oversampling (our team goes with this)
    - under sampling
  - We have to present on Thursday at the school.
    - Brian will confirm if we have to present in front of the class
- Presentation:
  - If we have questions on the lift chart,
  - High level: Some data in background
    - Background for business case
    - Methodology
      - Iteration
      - How to sample
      - Numbers of observation
      - Targeting
      - Lift table to see how the model performing
    - Share the word doc to draft our plan
    - Meet in person on Tuesday at 6 pm.