|  |  |
| --- | --- |
| Image result for carlson school of management university of minnesota | **PREDICTIVE ANALYTICS**  **MSBA 6420** |

|  |  |
| --- | --- |
| **Deliverable #2 - Project Proposal** |  |

Hospital-Acquired Infection (HAI) Patient Classification and Risk Prediction

Team:

Xiurong Lin

Ananya Mishra

Lavanya Basava Raju

Ryan Borowicz

**Project Proposal**

**Business Problem**

In the United States, the Centers for Disease Control and Prevention estimated roughly 1.7 million hospital-associated infections (HAI’s), from all types of microorganisms, including bacteria and fungi combined, cause or contribute to 99,000 deaths each year. [1] HAI’s are infections contracted from the environment or staff of a healthcare facility.

In addition to the serious patient health consequences, these infections cost an estimated ~$10 billion a year to treat, which is ~$30,000 per patient on average. As part of recent healthcare regulations, hospitals will no longer be reimbursed by Medicare for errors such as HAI’s. Because of this, hospitals are making further efforts to prevent these infections from occurring by identifying high-risk patients and pre-emptively admitting them to the hospital for antibiotics prior to their surgery.

**Data Details**

We will utilize the four datasets below provided by the Informs Data Mining Contest website. [2] A fairly substantial amount of data preparation and cleansing will need to be performed prior to modeling, but no assistance should be necessary for our team.

1. Patient conditions (213,018 records – 4 variables): Provides a listing of all conditions experienced by any of the patients at the hospital during the three-year period.
2. Patient demographics (68,619 records – 18 variables): Includes a variety of demographic information including sex, race, marriage status, education year, income, poverty (primarily nominal categorical)
3. Hospital information (38,956 records – 57 variables): Includes details related to the patient’s stay at the hospital and their diagnosis codes, as well as the associated cost and payment information.
4. Prescription information (629,571 records – 74 variables): Provides details of each patients’ prescription information, linking back to the hospital and patient datasets.

**Modeling Ideas**

**Part 1 – Data Cleaning & Preparation**

The four datasets – Patient demographics, Patient conditions, Hospital and Prescription information will be merged. The data will be cleaned to remove redundancy, missing values and outliers.

**Part 2 – Classification**

For an unseen patient dataset, probability of each patient to be diagnosed with an HAI will be determined.

We shall use multiple algorithms such as Decision Trees, Support Vector Machines and Propensity Score Matching to score each patient with a probability of that patient to be diagnosed with an HAI. We will compare the results from these models and pick the model that gives the best accuracy.

**Part 3 – Prediction**

A predictive model will be developed to maximize the total cost savings of the proposed strategy.

**Predict cost**

Run prediction algorithms to predict the cost of treatment for patients who are most probable to be affected by the HAI.

**Score each patient**

Run multiple algorithms - Support vector machines, Propensity score matching, etc.

Evaluate results

Select best prediction. model

**4 datasets -**

1. Patient Demographics

2. Patient Condition

3. Hospital

4. Prescription

**References**

[1] <https://en.wikipedia.org/wiki/Hospital-acquired_infection>

[2] <https://sites.google.com/site/informsdataminingcontest/home> (Data Source)

[3] <http://www.cdc.gov/mrsa/tracking/>