**Capstone Project**

**Project proposal**

**1.** **Group description**

**1.1.** Group name

|  |
| --- |
| YAM |

**1.2.** Students names, background and target industry if any

|  |
| --- |
| Team members:   * Yi Cao: background in policy analysis; open to various industries, leaning toward public/nonprofit/mission-driven companies * Matt Hope: background in biology; interested in applied data science in genetics * Ava Park: background in education tech/curriculum designing; interested in online education and the healthcare industry |

**1.3.** Group structure: roles and responsibilities

|  |
| --- |
| Roles/responsibilities   * Proposal submission – Yi * GitHub repo owner – Matt * EDA – all members * Model building – all members * Presentation – all members |

**2. Why** do we want to develop a data science project?

**2.1 Objective**: what problem do you want to solve? What questions are you trying to answer? How will you **measure the success** of your analysis from a business/user perspective?

|  |
| --- |
| * Problem: Potential hospitalization inefficiencies manifested in the rate of inpatient readmissions * Goal: Building a classification model to predict readmission within 30 days * Questions:   + Which features are correlated with early readmission? Knowing this, which features can be adjusted to lower early readmission?   + Which subgroups of patients are more likely to be readmitted? * Based on our project findings, hospitals may:   + Make recommendations on inpatient care/procedures during their hospital stay that can lead to a lower chance of them being readmitted within 30 days   + Advise preventative health measures to patients who are more at risk of early readmission * Measure of success: the recommended improvement informed by our prediction leads to a reduction in early patient readmissions and in hospital costs |

**2.2.** **Scope** of application: what population and timeframe will your analysis/model be applied to or used for?

|  |
| --- |
| * Population: diabetic patients from 130 hospitals across the U.S. from 1999 to 2008 * Potential application to: future diabetic inpatients |

**3. How** do you translate the objective and scope in terms of data?

**3.1.** What **dataset**(s) do you plan to use? Initial description: source, granularity, number of observations, variables list…

|  |
| --- |
| * Dataset: Diabetes 130-US hospitals for years 1999-2008 Data Set * Source: the data are submitted on behalf of the Center for Clinical and Translational Research, Virginia Commonwealth University, a recipient of NIH CTSA grant UL1 TR00058 and a recipient of the CERNER data * Granularity: encounter-level * Number of observations: 101766 * Number of variables: 50   + Features: 'encounter\_id', 'patient\_nbr', 'race', 'gender', 'age', 'weight', 'admission\_type\_id', 'discharge\_disposition\_id', 'admission\_source\_id', 'time\_in\_hospital', 'payer\_code', 'medical\_specialty', 'num\_lab\_procedures', 'num\_procedures', 'num\_medications', 'number\_outpatient', 'number\_emergency', 'number\_inpatient', top 3 diagnoses, 'number\_diagnoses', 'max\_glu\_serum', 'A1Cresult', medications that the patient is on, 'change', 'diabetesMed'   + Target: 'readmitted' |

**3.2.** What **data treatment and analysis** do you plan? Data Aggregation, target variable definition, tools, analysis/machine learning, ...

|  |
| --- |
| Data treatment/analysis   * Target variable: readmission * Data cleaning/preprocessing * EDA   + Data visualization   + Statistical testing * Feature engineering: simplifying categorical variables with high levels, for example * ML building:   + Supervised learning: classification methods including logistic regression and decision trees, and if time permits, other types of classification models   + Potential unsupervised learning (if time permits): clustering/PCA * Train-test split: randomized split on final dataset * Model evaluation (confusion matrix)   + Coarse-screened metric: AUC-ROC   + Fine-grained metric: different types of error metrics such precision and recall |

**4. Project plan**

Please submit a project plan proposal broken down by a few significant steps. Plan at least three meetings with your stakeholders.

* **Kick-off meeting:** met with Brian on 12/7
* **Initial findings review**: met with Brian on 12/9
* **Model** **review**: meeting with Brain on 12/14 and ? (TBD)
* **Project submission and presentation:** 12/18 – 12/21

**Project plan and schedule**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dec 7th | Dec 8th | Dec 9th | Dec 10th | Dec 11th | Dec 12th | Dec 13th |
| Meeting with Brian | Internal  meeting | Meeting with Brian; finalizing dataset | Working on building initial model | Initial baseline model discussions  Friday EoD: internal meeting  Sunday afternoon (~2pm): internal meeting | | |
| Dec 14th | Dec 15th | Dec 16th | Dec 17th | Dec 18th | Dec 19th | Dec 20th |
| Meeting with Brian |  |  |  | Submitting project |  |  |
| Dec 21st | Dec 22nd |  |  |  |  |  |
| Presenting project |  |  |  |  |  |  |