**Capstone Project**

**Project proposal**

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

**1.1.** Group name

|  |
| --- |
| ***Readmit None*** |

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

|  |
| --- |
| A: Shalabh Agarwala: background in chemistry, worked as an analytical chemist for 3 years. Healthcare or tech as target industry.  B: Matthew Harris: education in economics/business, former health insurance agent; healthcare or finance/private equity as target industry |

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

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | |  |  |  | | --- | --- | --- | | **Student** | **Data science** | **Project team** | | A | - data engineering  - Exploratory data analysis  - data visualization  -machine learning | - project proposal submission  - presentation  - team leader | | B | - data engineering  - Exploratory data analysis  - data visualization  -machine learning | - time management  - presentation | | |

**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?

|  |
| --- |
| **Context**: The hospital in which you work has a goal to reduce hospital readmission rates for diabetic patients. Both the hospital finance and clinical care teams are interested in how the data science team may help these departments reach this goal.  **Objective**:  - Create a model which predicts whether a patient will be readmitted in <30 days  - Model evaluated on AUC (Area Under Curve), False positive rates, and true positive rates  - Display how rates will be affected, payment, and the bottom line  - Model will be designed to direct the hospital’s intervention team  **Measure of success**:  - Was model able to predict if patient would be readmitted vs hold out data?  - Business Outcomes: Using hold out data as a test, what cost savings would the hospital have expected during the period by adopting our model to lower readmission penalties? |

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

|  |
| --- |
| **Population**: clinical care data between the years 1999 to 2008. Data is based off “diabetic” encounter, which includes one during which any kind of diabetes was entered to the system as a diagnosis. The length of stay was at minimum 1 day and at most 14 days  **Timeframe**: Dataset includes 10 years of data from 1999 to 2008. The length of stay per patient varies between 1 day at minimum to 14 days at most.  **Target variable**:  Readmission(<30 days) |

**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 1: diabetes data**  The dataset represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. It includes over 50 features representing patient and hospital outcomes. Information was extracted from the database for encounters that satisfied the following criteria. (1) It is an inpatient encounter (a hospital admission). (2) It is a diabetic encounter, that is, one during which any kind of diabetes was entered to the system as a diagnosis. (3) The length of stay was at least 1 day and at most 14 days. (4) Laboratory tests were performed during the encounter. (5) Medications were administered during the encounter. The data contains such attributes as patient number, race, gender, age, admission type, time in hospital, medical specialty of admitting physician, number of lab test performed, HbA1c test result, diagnosis, number of medication, diabetic medications, number of outpatient, inpatient, and emergency visits in the year before the hospitalization, etc.  [Source: https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008](https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008) |

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

|  |
| --- |
| **Data preparation**  - explore and clean data  - compute missingness, determine which entries are relevant to model.    **Target variable**  - hospital readmission of patient (<30 days)  **Tools**  - Data preparation in python  - Predictive model development  - Feature engineering  **Analysis**  - Exploratory data analysis: insights to see which features will determine best fit, calculate pass rate, visualization, statistical test for significance  - Prediction model: XG Boost , Logistic Regression, Logistic Regression Reduced, Random Forrest |

**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**: schedule a 30 minute meeting before project declaration in order to approve the project proposal.

- **Milestone 1**: schedule 30 minutes with your stakeholders to present initial results and insights and to validate any assumptions or definitions needed to move forward.

- **Milestone 2**: schedule 30 minutes to go over the final results and proposed presentation before the final presentation in front of the whole team.

If you think additional discussions will be required, feel free to add secondary milestone(s) in your project plan.

**Kick off**

Project proposal and timeline

**Milestone 1**

Check on progress

**Milestone 2**

Final results Preparation for presentation

**Delivery**

Final presentation

(to be filled in as information becomes available) >>

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **December** | | | | | | | | | |
|  | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | | 18 | 19 | 20 |
| **Kick off** |  |  |  |  |  |  |  |  |  |  | |  |  |  |
| Project declaration |  |  |  |  |  |  |  |  |  |  | |  |  |  |
| Review data & Proposal |  |  |  |  |  |  |  |  |  |  | |  |  |  |
| Begin EDA |  |  |  |  |  |  |  |  |  |  | |  |  |  |
| Complete EDA |  |  |  |  |  |  |  |  |  |  | |  |  |  |
| Review EDA |  |  |  |  |  |  |  |  |  |  | |  |  |  |
| **Milestone 1** |  |  |  |  |  |  |  |  |  |  | |  |  |  |
| Begin Model Development |  |  |  |  |  |  |  |  |  |  | |  |  |  |
| Update Model |  |  |  |  |  |  |  |  |  |  | |  |  |  |
| Complete Model |  |  |  |  |  |  |  |  |  |  | |  |  |  |
| **Milestone 2** |  |  |  |  |  |  |  |  |  |  | |  |  |  |
| Create Presentation |  |  |  |  |  |  |  |  |  |  | |  |  |  |
| **Delivery** |  |  |  |  |  |  |  |  |  |  | |  |  |  |