## Scenario



<https://www.pexels.com/photo/ambulance-architecture-building-business-263402/>

You are part of data science & analytics team at Great River Medical Center. Your team has been asked to create a readmissions model for diabetic patients. When patients are admitted, the hospital collects information related to their ailment along with medicine intake, & physician information. First, hospital administrators want to know if there are interesting cohorts within patient groups that may increase the probability of readmission. Further, you will need to create a machine learning model that identifies the patients most likely to return to the hospital after treatment. Administrators need to limit readmissions because it negatively impacts patient quality of life, causes undue strain on the medical staff and insurance and government supplementals may deny payments if a patient is readmitted within 30 days of discharge.

The cost of readmissions is [large](https://www.hcup-us.ahrq.gov/reports/statbriefs/sb278-Conditions-Frequent-Readmissions-By-Payer-2018.jsp) for the patient and hospital, ranging from $10,900 to $15,200 and diabetes is cited as the third leading cause of hospital readmissions. In fact, “The top four principal diagnoses at index admission—septicemia, heart failure, diabetes, and COPD—combined accounted for one in five readmissions.” Taken holistically, hospital readmission costs the US [$52.4B per year](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9113654/#:~:text=In%20addition%2C%20over%20%2452.4%20billion,for%20a%20previously%20treated%20condition.). All information pre-dates the impact of Covid hospitalization.

After you have built model(s) and evaluated their outputs, you will need to identify the top 100 patients among the test set that are expected to be return. Hospital administrators will examine the results, and work with staff to improve treatment outcomes for these pilot 100 patients.

Once your team has modeled and identified the customers, you must present your findings to the hospitals chief analytics officer *and* chief patient advocate. Once they are comfortable with your insights, model and evaluation, the pilot program can commence.

**You are asked to examine the historical data from 7500 diabetic patients. Using this historical data, and any supplemental data, perform EDA, create a propensity model, evaluate it and identify by unique identifier the top 100 patients from the test set of 2500 patients to be included in the pilot for improved treatment plans. Additionally, hospital executives are eager to learn more about the diabetic patient profile for historical records and the identified most likely to readmit. As a result, variable importance and sound EDA will aid the presentation. You will need to turn in code and PowerPoint slides.**

## Data

Source: <https://www.kaggle.com/kondla/carinsurance>

There are 6 files included in the case.

* **diabetesHospitalInfoTrain.csv** – 7500 patients’ data to be used for EDA and model training. This data has 16 variables include a unique identifier and consists of information gathered at the hospital such as the number of lab procedures and diagnosis descriptions.
* **diabetesMedsTrain.csv** – 7500 patients medicine information to be used for EDA and model training. This data has 23 variables such as whether or not the patient is taking a specific drug like Metformin.
* **diabetesPatientTrain.csv** – 7500 patients’ demographic and insurance information to be used for EDA and model training. **\*\*\*This data set contains the depdentant variable `readmitted\_y` which is a Boolean indicator of whether the patient was readmitted within 30 days of hospital discharge. \*\*\*** Other variables include race, gender, age, weight and how they patient paid for services.

## **diabetesHospitalInfoTest.csv, diabetesMedsTest.csv, diabetesPatientTest.csv** – These files are like the above except represent 2500 test patients. Once a model is built using the previous 3 files, the patients in the test set must be evaluated for hospital readmissions and the top 100 most likely need to be identified.

## Example *Abridged* Data

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| tmpID | race | gender | age | wgt | … | readmitted\_y |
| 2162 | Caucasian | Male | 47 | 272 | … | TRUE |
| 6271 | AfricanAmerican | Female | 46 | 183 | … | FALSE |
| 3495 | AfricanAmerican | Male | 48 | 176 | … | TRUE |
| 5835 | Caucasian | Female | 66 | 179 | … | TRUE |
| 1228 | AfricanAmerican | Female | 31 | 190 | … | TRUE |
| 3019 | Caucasian | Female | 69 | 175 | … | FALSE |

## The Submission

* The submission will include business analyst slides covering the problem, data, methods, model explanation and any insights. Without a presentation, the “organization” section of the rubric will be 0. Exceptional submissions are well ordered and provide a coherent narrative covering data, methods, modeling and any insights that may be of interest to the audience.
* The submission will include a supplemental identifying the top 100 patients by ID and corresponding probabilities. This can be in a CSV or similar file format. In addition, any insights identified in the presentation will be included in a written supplemental. The insights written portion can be 3-5 sentences for each insight in a bulleted list format. Exceptional submissions include statistics from external credible sources that support the identified personas or insights. For example, “…according to [some medical journal] patients on more than 2 diabetes drugs are less likely to have severe complications…”. Without a the top 100 patients file AND a written supplemental that coincides with the narration and supported by code the “written supplemental” section will be 0.
* The submission will include either a recorded screen narration of the presentation, a text file with a URL to a recording (like youtube video) or audio that is embedded into the slide deck. Tone, volume, cadence, use of filler words and pronunciation will be accounted for in this section. No points will be deducted based on English proficiency (ie ESL) but technical descriptions that are incorrect will be detrimental such as “Logistic Regression is used for predicting continuous outcomes”. Failure to submit a narration, the “delivery” section of the rubric will be 0.
* An R script covering all data munging, modeling, evaluation and visualization construction used to create the presentation artifacts (you do not need to use R to construct the slides but it is possible) and come to the case outcomes. Your code must use the following R functions at least once throughout your code, group\_by, aggregate & subset in addition to modeling code. Make sure to that your code contains ample comments. Failure to turn in an R script will result in a “Documentation” score of 0.

## Criteria for Success

The presentation will be evaluated on an equal weighted scale with the following criteria. For example 20 points per each category [depends on the individual course weighting found in Canvas]

* **Organization** – Was the presentation well organized?
* **Delivery** – Was the content delivered clearly and persuasively with the audience in mind?
* **Code Documentation** – Was the data mined to support the conclusion?
* **Written Supplemental** – Are the bullets clear and supported in narration and code? Were the top 100 households identified?
* **Data Mining & Modeling Proces**s – Overall, as a complete portfolio of work, is the topic interesting, organized, researched, supported and delivered effectively? Was CRISP-DM, SEMMA or similar followed to organize the work?