Predicting Readmissions: Cohort Selection

Ayzhamal Zhamangaraeva
Dr. Ioannis A. Kakadiaris, University of Houston

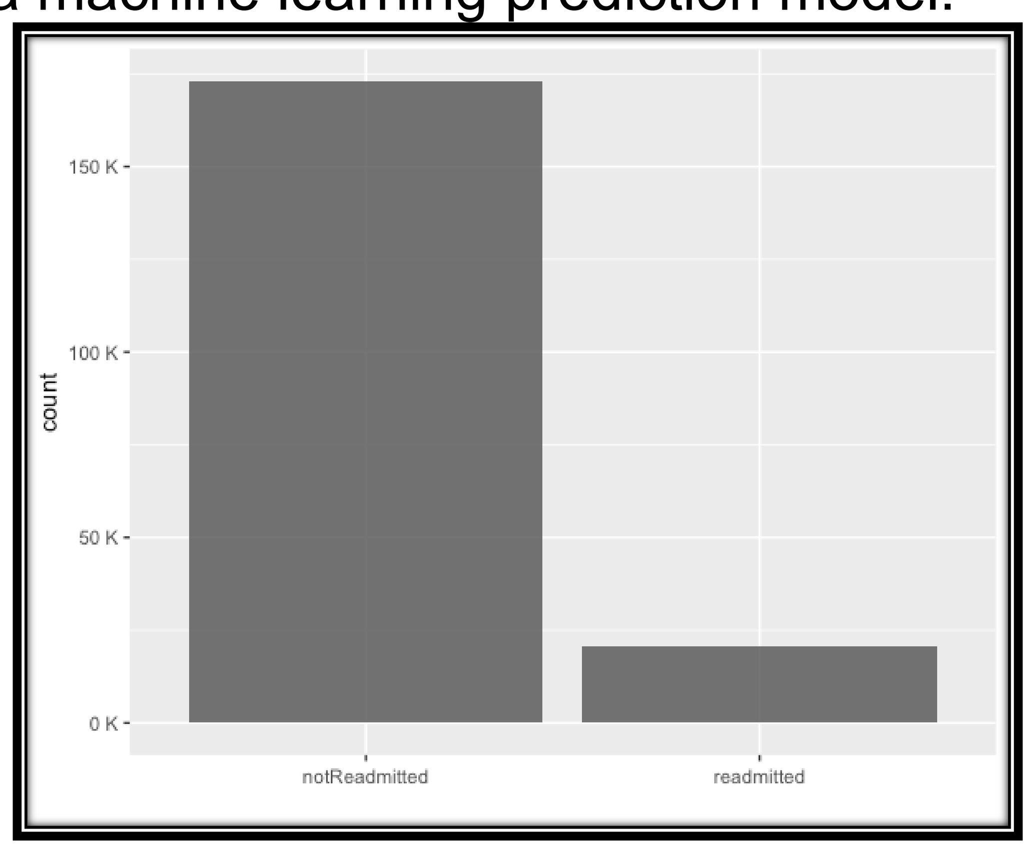
Motivation

- "Hospital Readmission Reduction Program" established by Medicaid since October 1, 2012.
- About half of US registered hospitals receive reduced payments for **excess** readmission rates.
- Historically, **nearly 20%** of Medicare discharges have a readmission **within 30 days**.
- **Heart Failure is the most expensive diagnosis** for hospitalizations in the US accounting for 70% of total costs for heart failure. **CHF is the most frequent diagnosis for 30-day readmissions**.

Accurate prediction model can help to readjust health care, reduce readmission rate and health care costs.

Tasks

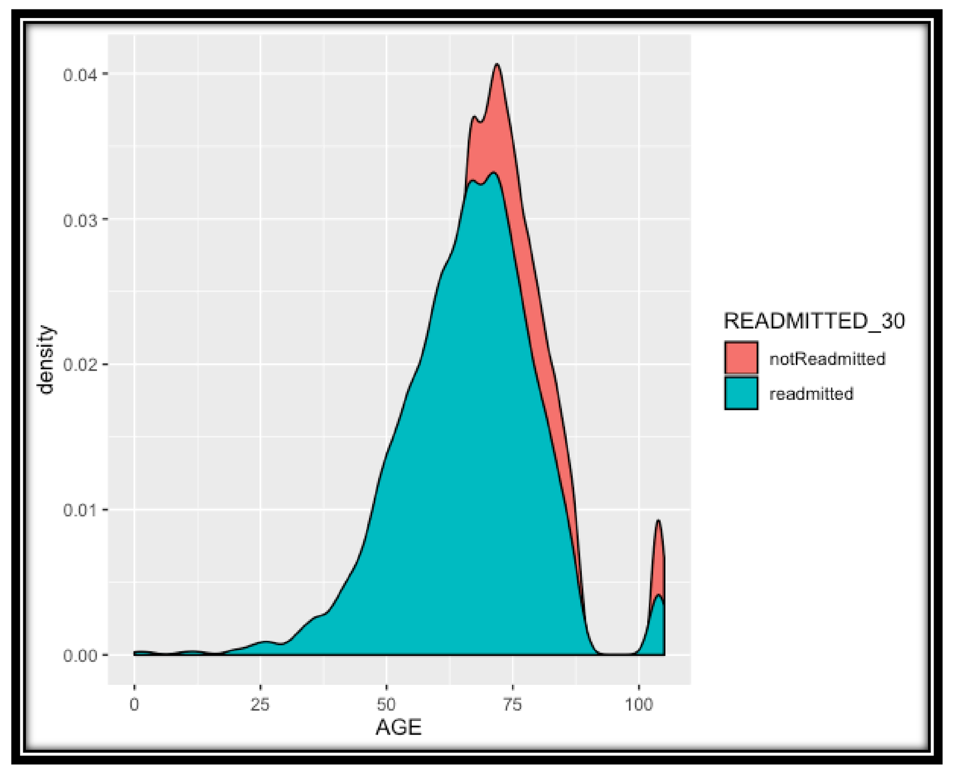
1. Define the cohort of CHF patients. Criteria: length of stay>0, non-deceased patients with discharge status code "to home or self-care", ICD-9 codes, and CHF diagnosis.
2. Data pre-preprocessing: compute new features. Original Humana dataset: medical, lab, patient, and pharmacy records of 12.9 million patients for 2013-2015. Total 20,415 (10.55%) positive readmissions out of 193,419 CHF hospital admissions.
3. Partition to training, testing, and validation datasets.
4. Build a machine learning prediction model.

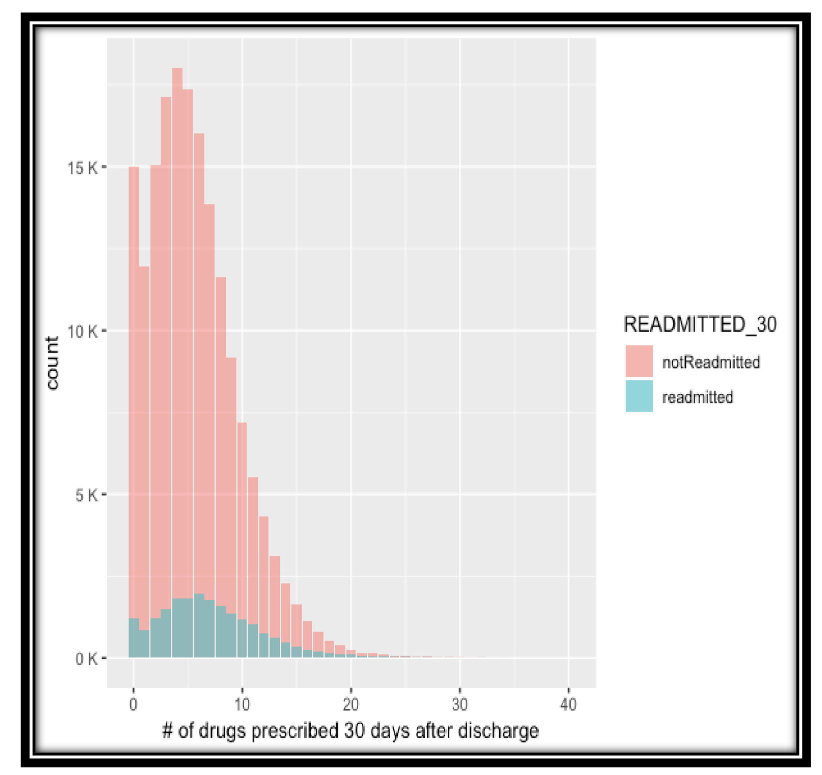


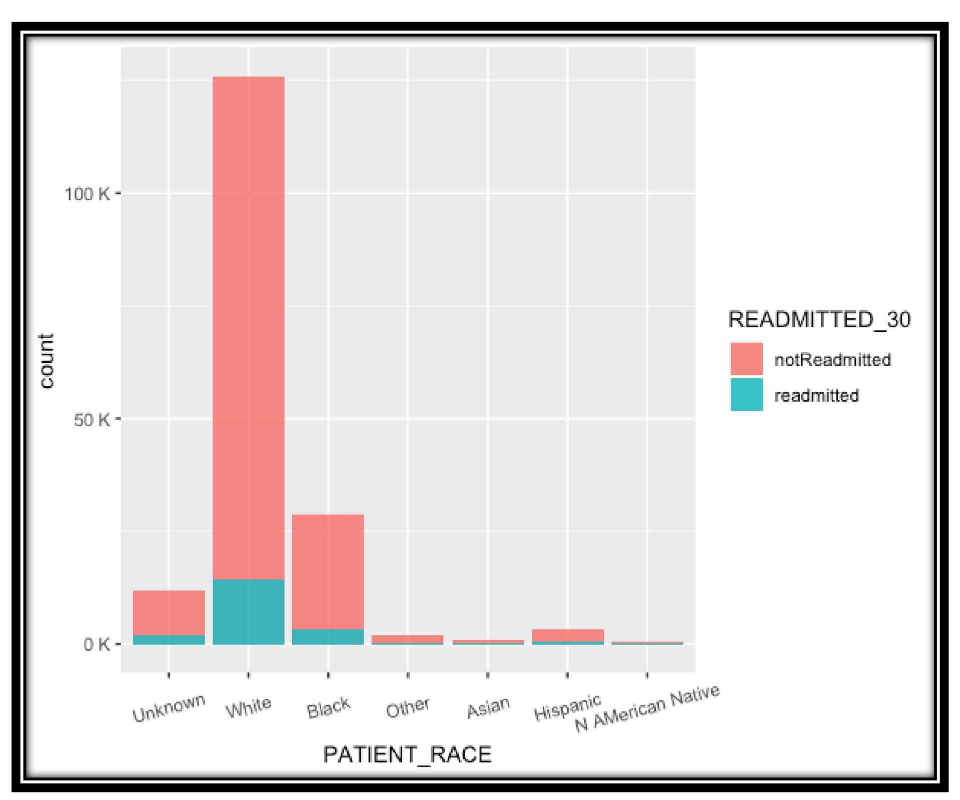
Data Pre-Processing

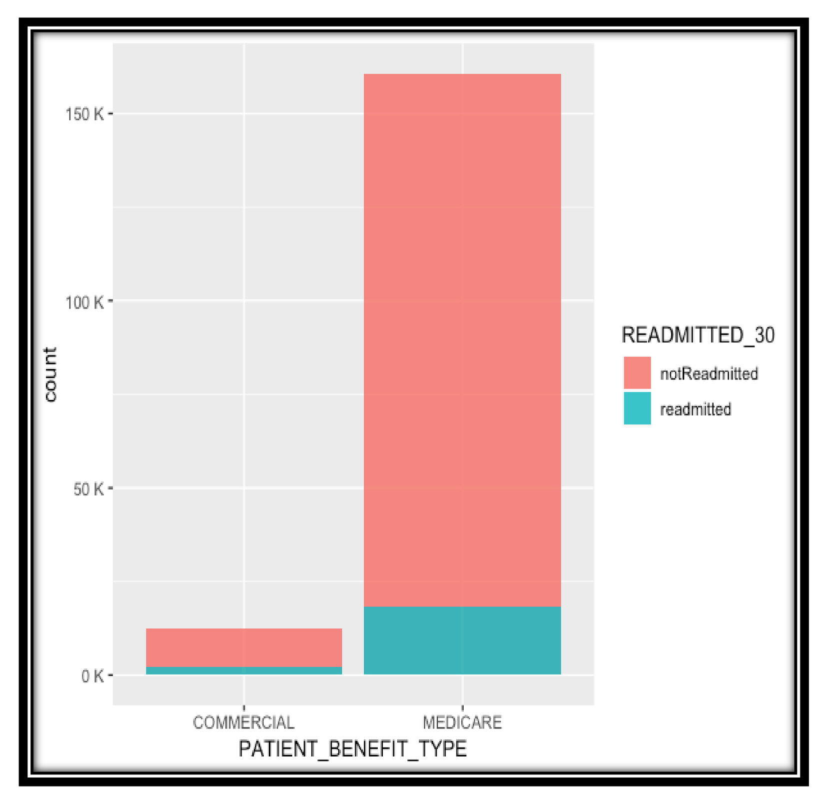
1. Compute readmission interval
2. Compute a target column READMITTED_30
3. Compute the number of unique diagnosis, and the number of chronic conditions for the period 2013-2015. Mean(chronic)= 41.45.
4. Compute the number of unique generic drug names prescribed 30 days before admission, during hospital stay, and 30 days after discharge. Mean(rxAfter)=5.8, Max(rxAfter)=40.
5. Compute the number of unique lab orders 30 days before admission, during hospital stay, and 30 days after discharge. Mean(labDuring)=6.7, Max(labDuring)=240

Visualization









Challenges

- Presence of both ICD-9 and ICD-10 coding classification systems. Overlap during the period 2014-2015. Necessary to build an equivalence table to avoid counting diagnoses twice.
- Negative readmission intervals caused by multiple records related to the same admission event. Need to clarify with experts and, probably, concatenate them into one.
- Lab file is limited (records of 4M patients instead of 12.9M), has parsing problems. Eighty-five percent of CHF admissions have zero lab orders during hospital stay. Probably, the data is incomplete.

Discussion

- Needs further clarifications from the creator of dataset.
- Would benefit from incorporating physician's intuition on the differences of positive and negative readmissions.
- Upon completion of data pre-processing, Random Forest and Support Vector Machines can be applied to the whole dataset first, then on subsets based on gender, different race, and age groups.

Acknowledgements

This research was made possible by a NSF grant to the University of Houston Computer Science Department (NSF CNS-1551221). We thank Peggy Lindner (Hewlett Packard Enterprise Data Science Institute), and Dr. Dan Price (UH Honors College) for providing the data and all their support.