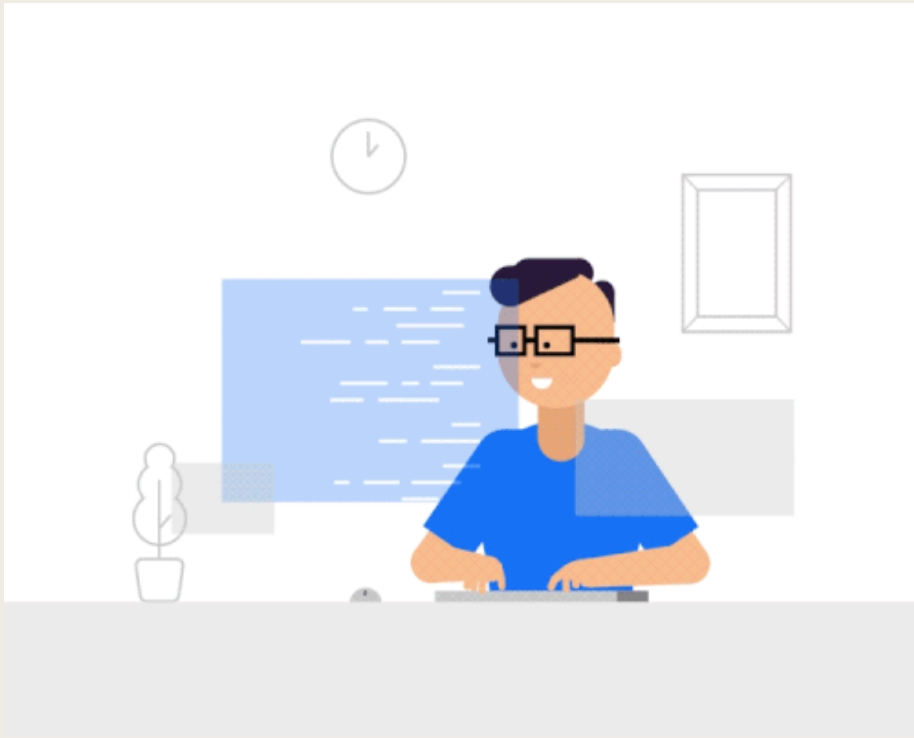
SYNTHETIC HEALTH DATA HACKATHON 2020

David Tandio
Marius Dioli

The Diabetes datasets

- Synthetic data is skewed relative to the general population
- Aggregating drugs into drug classes
- Binary classification for drugs (yes, no)
- Removed glucose clinical parameters



- Variables omitted
  - *Only have 1 value: Citoglipton, Examide*
  - *Don't aid prediction: diag_1, diag_2, diag_3*
- Citoglipton – an upcoming blockbuster drug?

Our goals and models used

- Can we discriminate between real and synthetic data?
- To what extent can we predict hospital readmission rate?
- How well can we predict variables such as number of lab procedures or time in hospital given limited demographic data?

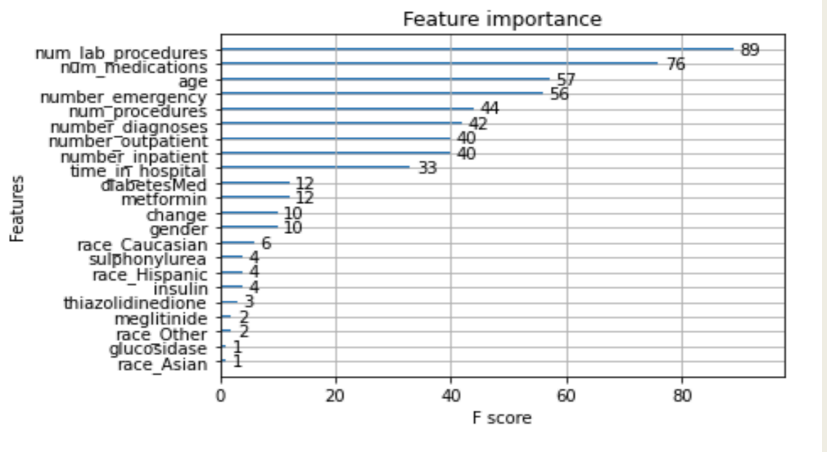
- We used logistic regression and boosted decision trees (XGBoost) on classification tasks.
- On regression tasks we used multiple linear models (standard linear regression, ridge, lasso, elasticnet), random forest, and boosted decision trees (XGBoost).

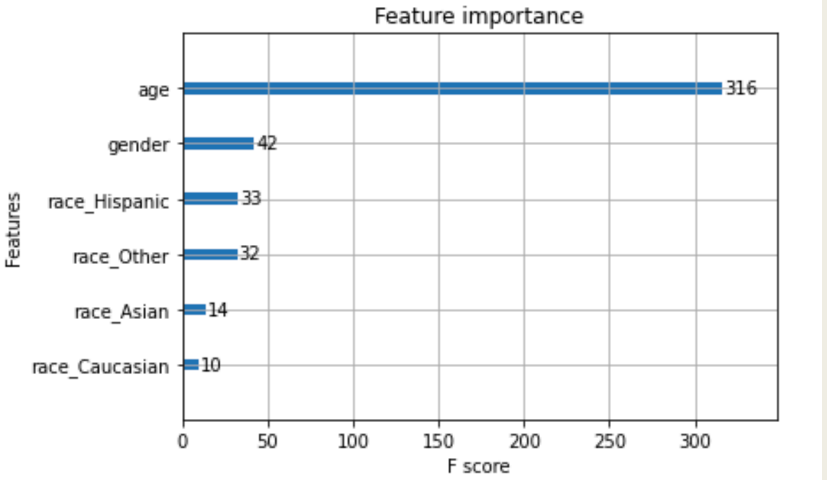
Key findings from our best models

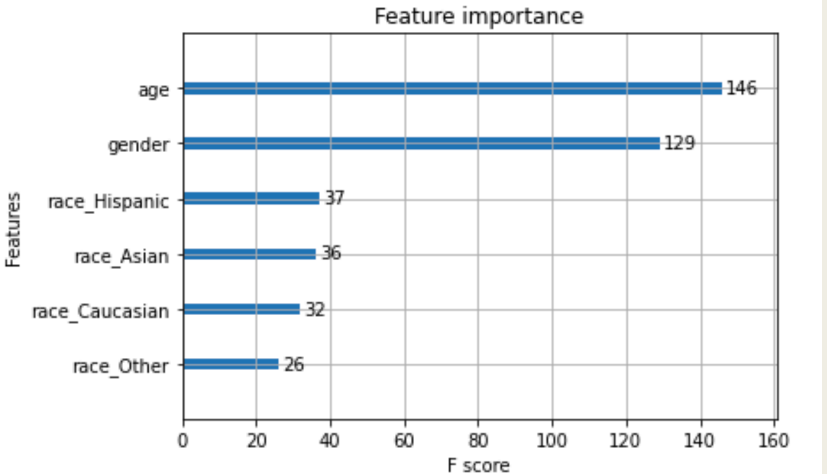
- We can differentiate between synthetic and real data with around 60% accuracy
- Predicting binary hospital readmission rate was more accurate than multi-class (73% vs 66%)
- Big difference in hospital readmission accuracy depending on whether we use the real data (~63%) or synthetic data (~71%)
- Assuming a patient arrives at the hospital and you only have demographic data, can you accurately predict continuous variables?
  - *Our model achieved high accuracy on all regression tasks (low MSE)*
  - *Generally weaker performance when using synthetic data*

Interesting findings from our decision trees

- Number of drugs more important than which drugs
- Race and gender relatively unimportant for readmission
- Highly skewed feature importance depending on whether the dataset is synthetic or real.







THANK YOU FOR YOUR TIME :D