Machine Learning project

Vivian Chu, Samarth Inani

Agenda

- Features created & Feature engineer techniques
- Model selections
- Experiment result
- Lessons and learned

Dataset Description

- Times series data
- 5 numerical columns
  - xx1, xx2, xx3, xx4, xx5
- 8 categorical columns
  - gender, age, x1, x2, x3, x4, x5, x6

Feature Engineering & Feature Created

- Apply Aggregation function on continuous variables
  - Max & Min--capture outliers
  - Last-latest data -- can help with the prediction (time series)
  - Median--avoid skewness on the distribution of continuous variables, capture non-linearity
- Select all the 30 records of each key and transform them into features.

Model Selections & Experiments

Model	Validation Test Score		Public R^2	Private R^2
	y_mean_MAP	y_mean_HR		
Stacked Model (186)	0.892	0.9543	0.91895	0.92844
Deep Learning (88)	0.8898	0.9523	0.91843	0.92787
XGBoost Regressor (88)	0.893	0.9562	0.91898	0.92785

Stacked Model ---Why?

Selected Model			
Input	**Models - Stack**	**Weights**	**Prediction**
X	Linear Regression	0.2	0.2 * (lr.predict) +0.2 * (rf.predict) +0.6 * (xgb.predict)
	Random Forest	0.2	
	XGBoost	0.6	

Lessons & Learned

- More keys for a few patients, could have sampled better
- Time series data , the latest data is more predictive
- More features in XGB might lead to overfitting