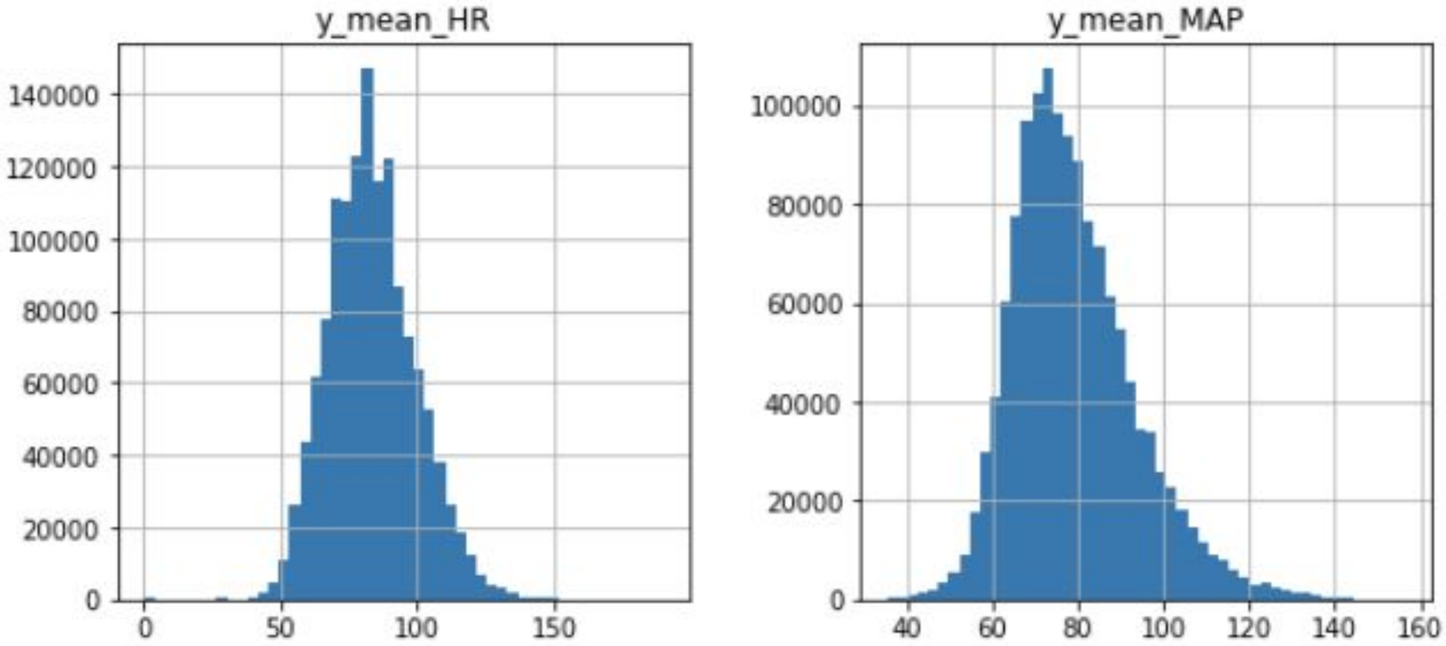
Kaggle Competition
Advanced Machine Learning

Team Toxic

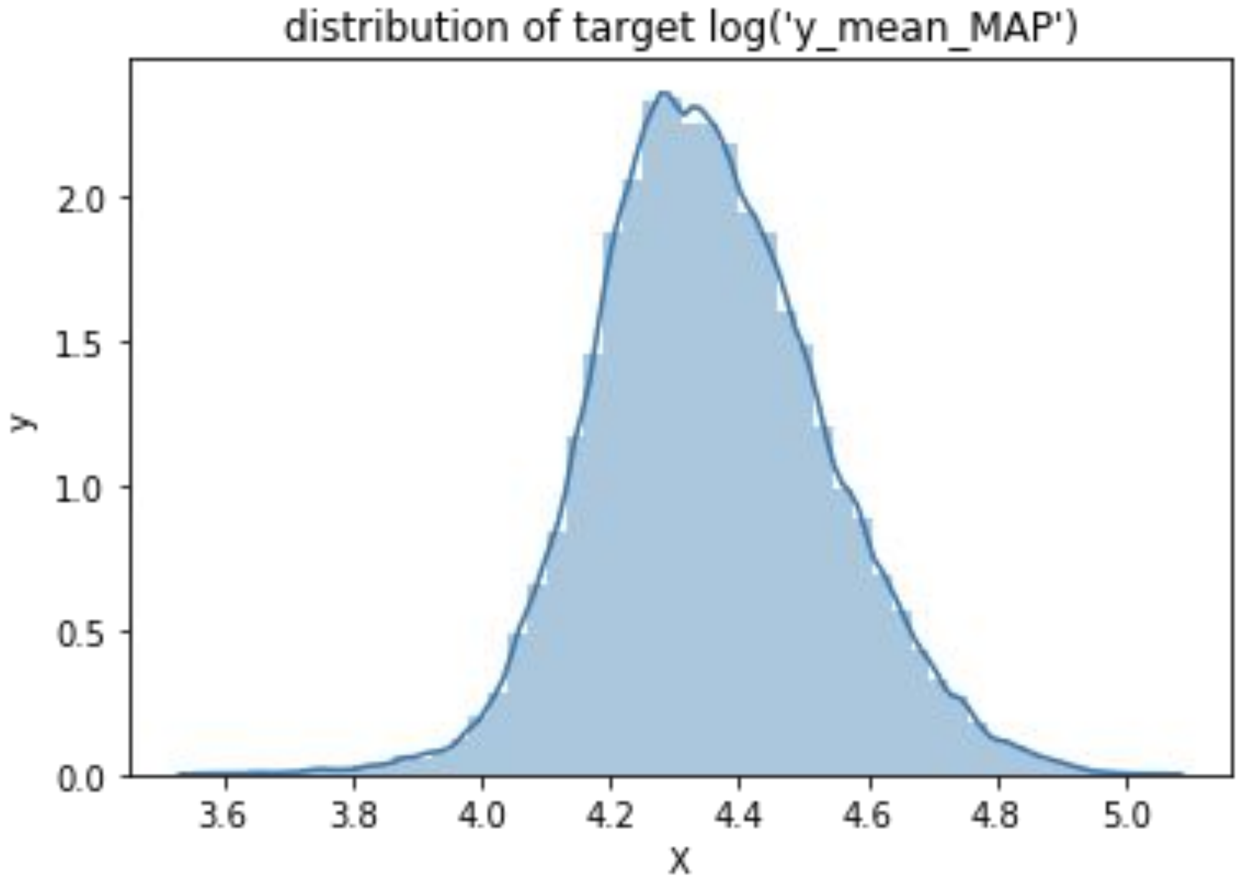
Todd Zhang & Lexie Sun

• Data & EDA

- Time series data
- Categorical features: patient_id, key, gender, x1, x2, x3, x4, x5, x6
- Target variable distribution:



log(y_mean_MAP)



• Feature Engineering

- Log transformed y_mean_MAP
- Created new features: take the ‘xx1, xx2, xx3, xx4, xx5’ of the last 5 records of each ‘key’, add them to every records of the ‘key’
- Randomly sampled the data on the ‘key’ level, keep at most 4 keys per patient
- 80/20 Train-test split with shuffling

• Modeling: y_mean_hr

Model	Parameters	R^2
Linear Regression	normalize=True	0.9398

Random Forest	n_estimators=200,
	min_sample_leaf=2500,
	max_depth=20,
	bootstrap=True,
	criterion=’mse’
		0.9410

- For linear regression model, the most important feature is ‘xx5’.

- **Modeling: y_mean_MAP**

Model	Parameters	R^2

Random Forest	n_estimators=200,
	min_samples_leaf=2500,
	max_depth=25,
	criterion='mse',
	bootstrap=True
		0.9069

Neural Network	hidden_layer_sizes=(6, 5, 1),
	activation=’tanh’,
	solver=’adam’,
	max_iter=30,
	alpha=1e-05,
	warm_start=True
		0.8317

XGBoost	eta=0.05, max_depth=25	0.8989

- **Takeaway**

- Taking a log transformation is helpful when dealing with skewed distributions
- Creating new columns for existing columns that contain important information is helpful
- Using more complex machine learning models does not always result in higher accuracy
- Aggregating over-represented data is helpful in improving generalizability of the model

Thank You!