

Utilizing automation for mass model creation to improve the economics of machine learning projects

Alex Tsui

Alex.Tsui@SAS.com



Contents



Economics of
automaton



Healthcare
use case



Modeling at
scale

Caveats

- The predictive models generated by this project have not been used in real-world settings.
- We addressed a specific use case in the lab test process of antibiotic drug resistance in this project. More valuable predictive models can potentially be produced if we had additional patient data. If your organization has relevant data and is interested in collaboration, please reach out to me.
- This presentation assumes the audience are aware of the common machine learning algorithms and understand general programming concepts. Knowledge of Python and SAS Viya platform is not required.



As data scientists, we are always the students of history. So let's start by watching a piece of history.

[Play video clip](#)



**making more of them in a
short amount of time**

“His whole business model is to sell a cheap product. And the way you create cheap products is by making more of them in a short amount of time.”

Antimicrobial Resistance (AR) is a major issue in the world

COVID-19

World-wide death count

830,205

As of 8/27/2020

700,000

estimated deaths/year globally

10 million

death per year by 2050
in worst case projection by WHO

2.8 million

infections per year in US

Source: [World Health Organization's report to United Nations 2019](#) ; [ANTIBIOTIC RESISTANCE THREATS IN THE UNITED STATES](#);

John Hopkins COVID-19 dashboard

Copyright © SAS Institute Inc. All rights reserved.

The use case



Day1

Empirical treatment

Doctors prescribe antibiotics based his/her empirical knowledge.



Day 1 -3

Lab test

Continue empirical treatments



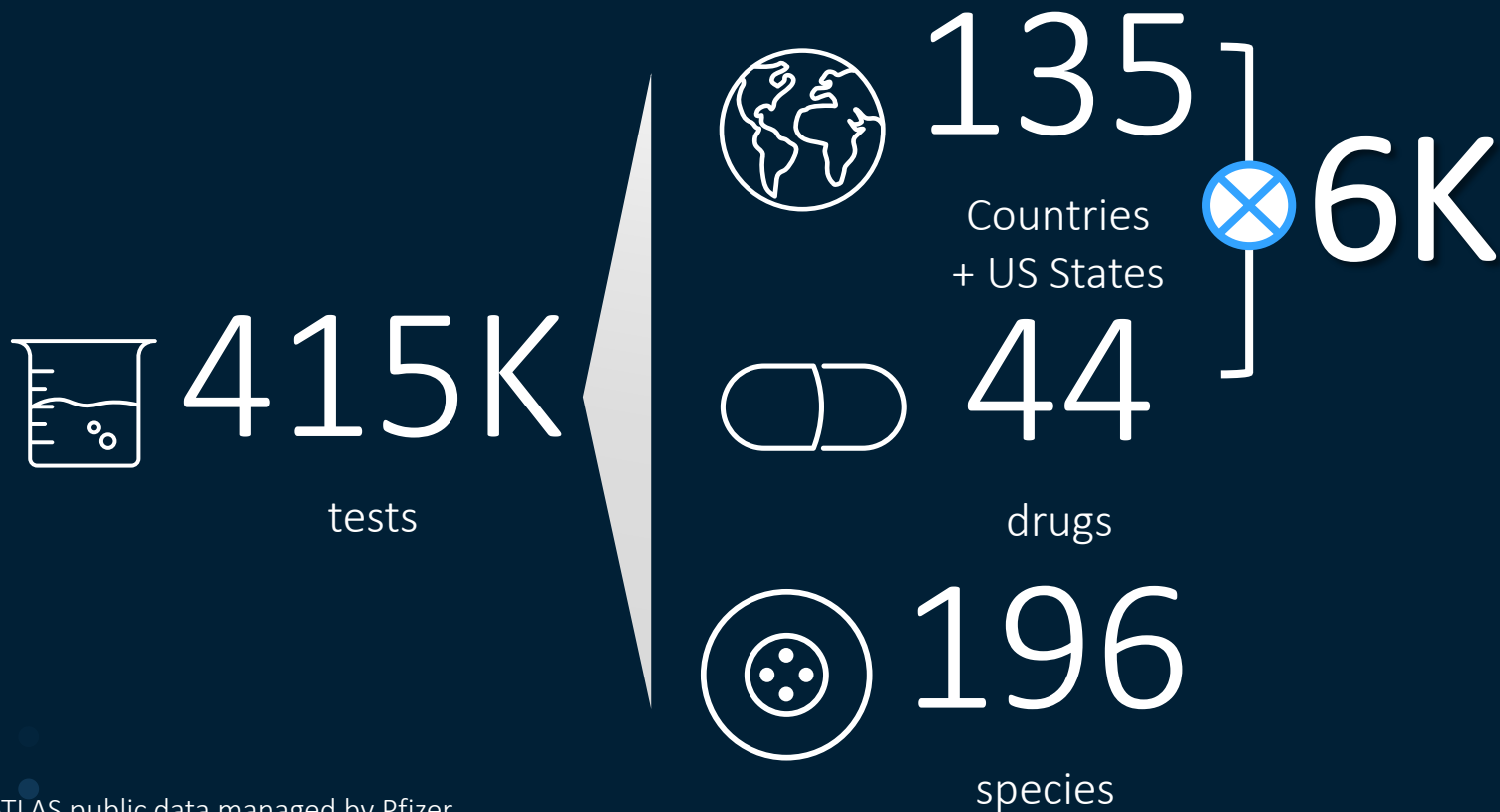
Day 3-4

Direct treatment

Predict drug resistance for partially successful lab tests

Use case

Analysis dataset



Source: ATLAS public data managed by Pfizer

Data fields in the dataset

Patient and test information

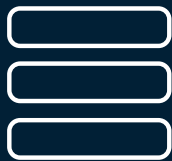
Isolete_id	Age_Group
Bacteria species	Speciality
Country	Source
State	InOut_Patient
Gender	Phenotype

Drug responses

Drug 1
⋮
Drug N

■ Data fields excluded from modeling

Modeling process for this project



Data acquisition



Data exploration
w/ SAS Visual
Analytics



ETL
(Python on
SAS platform)



Base table
for ML



Loop to generate a model for each drug in every country



Evaluate autoML results.
Iterate. Potentially
manual modeling for
some cases



Score test data and
generate model
performance
metrics



Generate multiple
models with autotuning
using training data and
select best model via
cross-validation

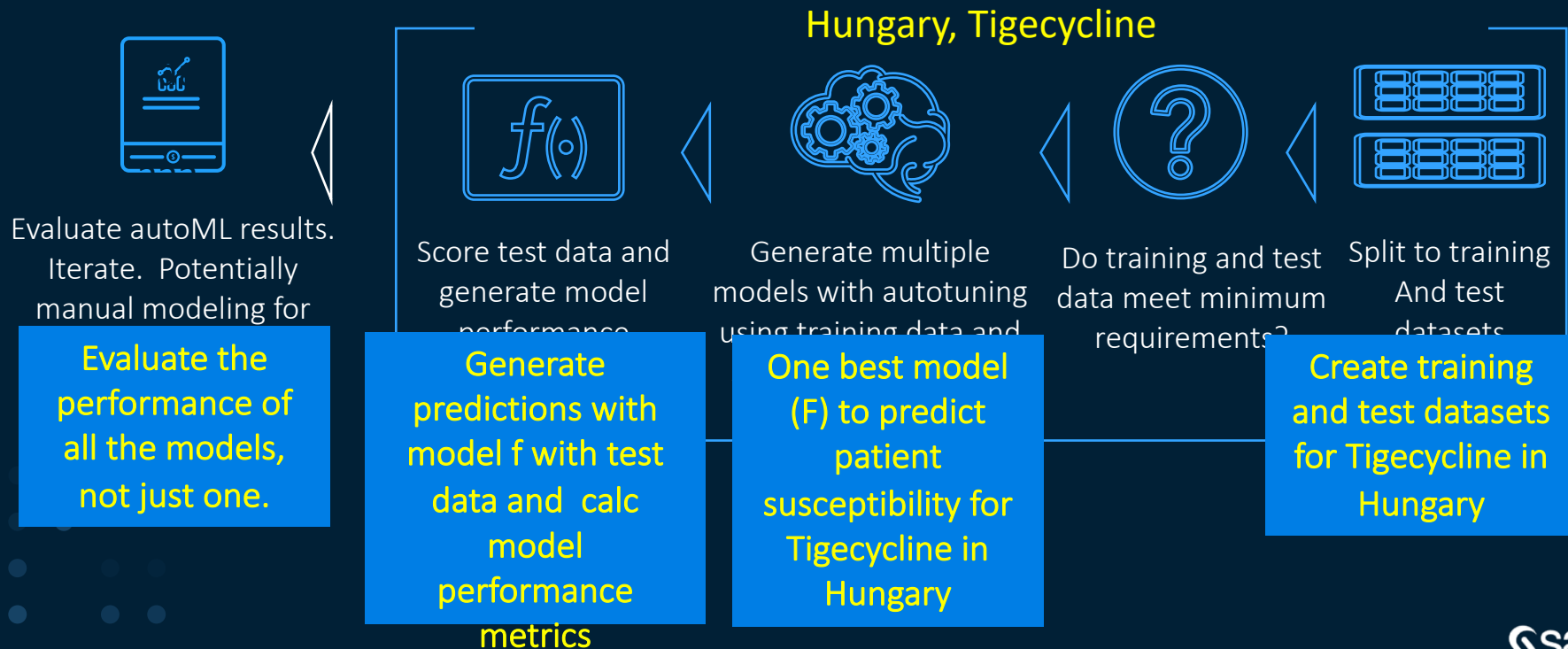


Do training and test
data meet minimum
requirements?

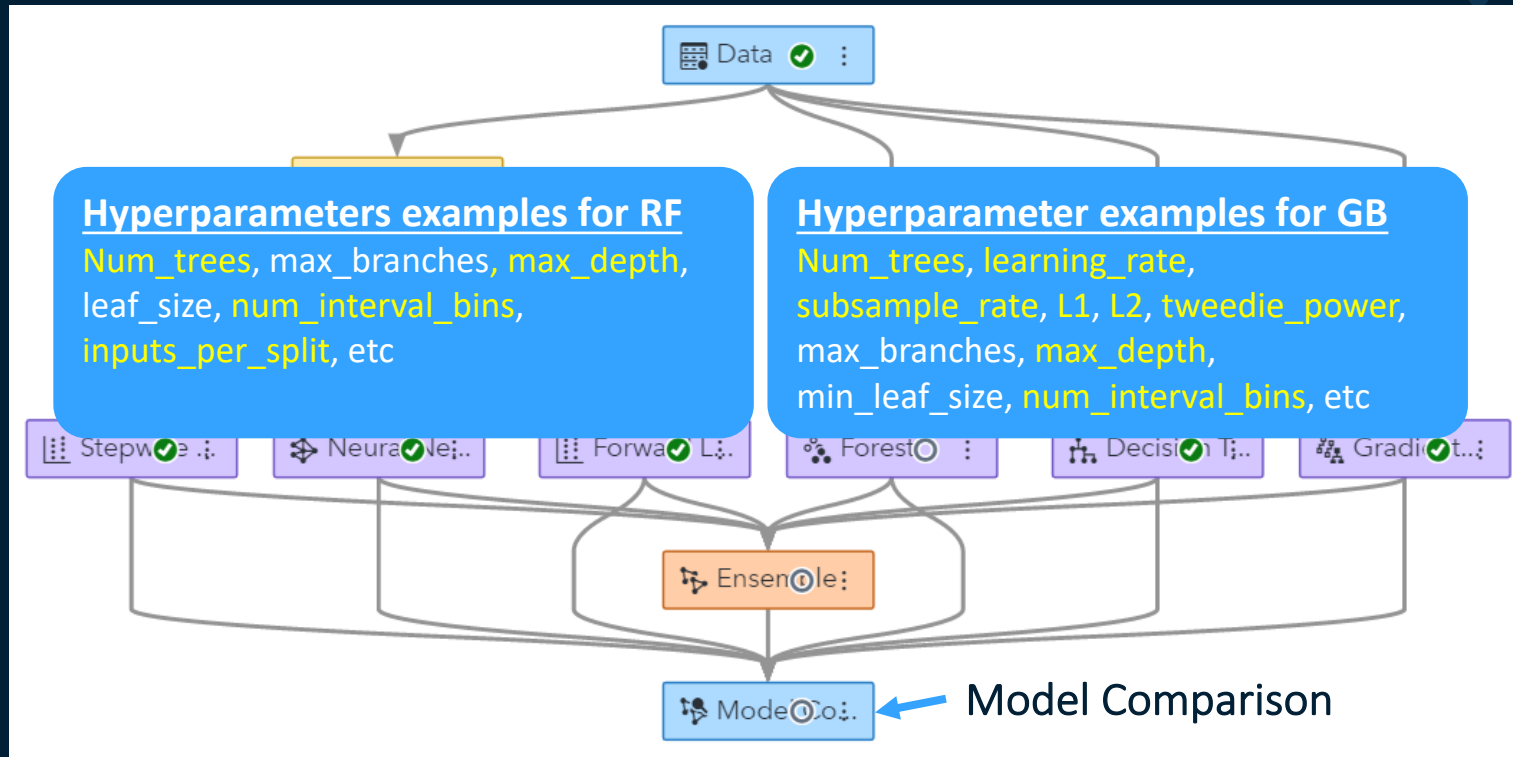


Split to training
And test
datasets

Modeling process for this project



Machine learning algorithms and autotuning



For each drug/country, only one function is needed to generate all candidate models and select the campaign one



illustration

```
result = cas.autotune.modelComposer (  
  table=training_table,  
  target=target_variable,  
  inputs=[input_variables],  
  tunerOptions={"seed":1, "objective":"F05", 'nCrossValFolds':5, ...},  
  modelTypes=[  
    {'modelType':'FOREST', 'tuningOptions': {...}},  
    {'modelType':'GRADBOOST', 'tuningOptions': {...}},  
  ],  
  nRounds=5, nEvalsPerRound=16)
```

Why F05?



$$Precision = \frac{TP}{TP + FP} = \frac{TS}{TS + FS} \leftarrow \text{MIN}$$

$$Sensitivity = \frac{TP}{TP + FN} = \frac{TS}{TS + FR}$$

$$Accuracy = \frac{TP + TN}{Total\ cases} = \frac{TS + TR}{Total\ cases}$$

T: true; F:false; P:positive, N:negative;
R:resistant; S:Susceptible

F-Score

$$F_1 = \frac{2 * Precision * Sensitivity}{Precision + Sensitivity}$$

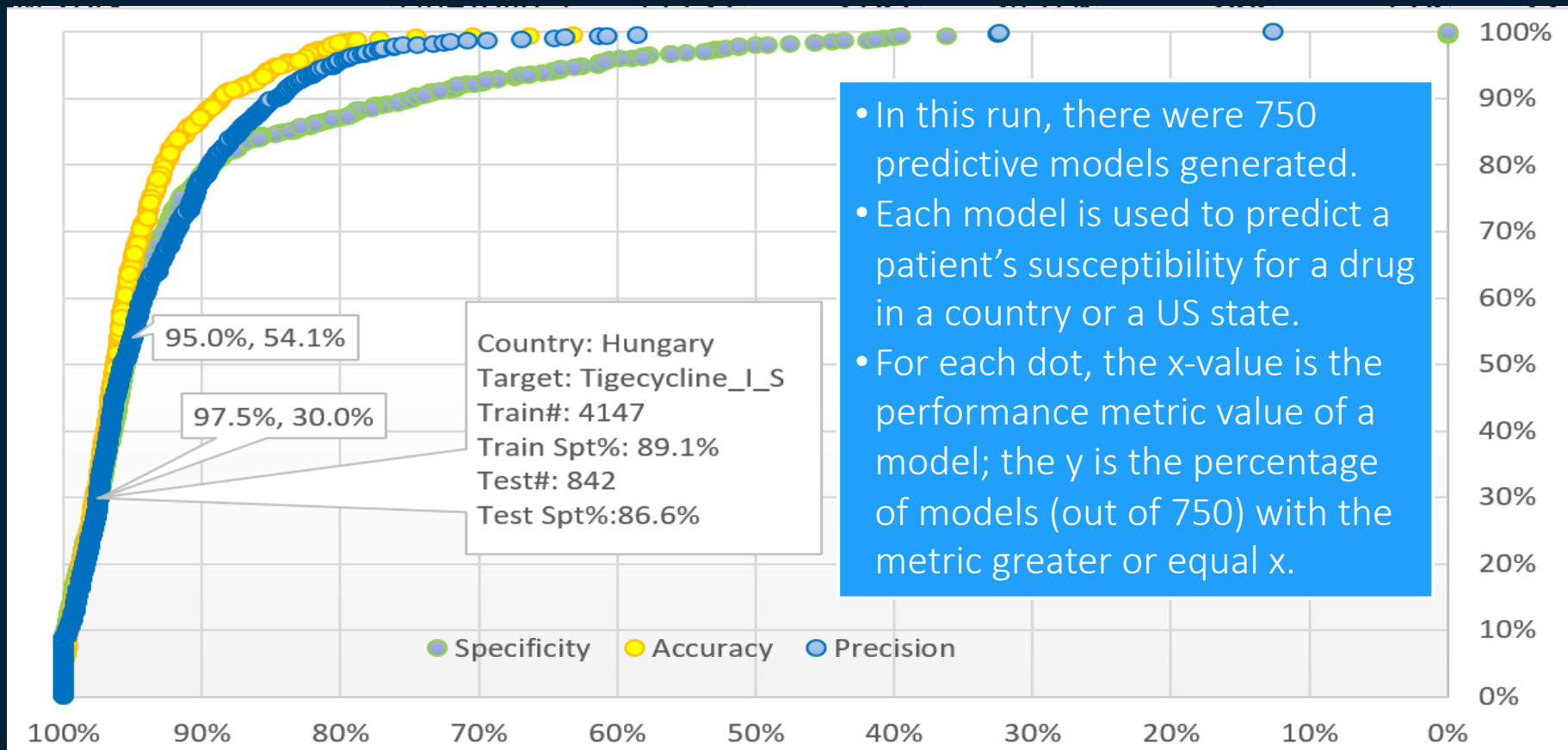
$$F_{0.5} = \frac{1.25 * Precision * Sensitivity}{0.25 * Precision + Sensitivity}$$

Scoring the test dataset is also just one function call



```
forestScore(  
    assess=True,  
    modelTable={'name': bestModelOutputTable},  
    table=testDataTable,  
    target=tgtVar,  
    copyVars=tgtVar,  
    casOut={"name": "ATLAS_MODEL_SCORE", "replace": True},  
    ...  
)
```

This chart shows the performance metrics of 750 models



Summary

- Large economic value can be gained with the creation of batch programs to automatically generate machine learning models. However, it's important to note that not every model is suitable for such batch processes due to the characteristics of data. So there can be models generated automatically and model created manually in a given project.
- At this point, programming is often needed to build such batch processes. Although Python is used here, one can accomplish the same with other programming languages, SAS or R, for example.
- I hope this presentation provided useful information for your next machine learning project and raise awareness of Antimicrobial Resistance issues.