Accelerating the Delivery of ML Based Products

•••

Twin Tech Labs

85% of all data science projects fail

(hint: this is not a technology problem)

Why?

Process, Organizational, Security Impediments

Disconnect between reality and desired product capabilities

Data Availability/Access

Data Security (PII)

Organizational Alignment and Priorities

Proficiency with Data Science and Machine Learning

SDLC Does not Apply to Data Science

Key Takeaways

Path to Machine Learning Success

Work iteratively and follow agile practices

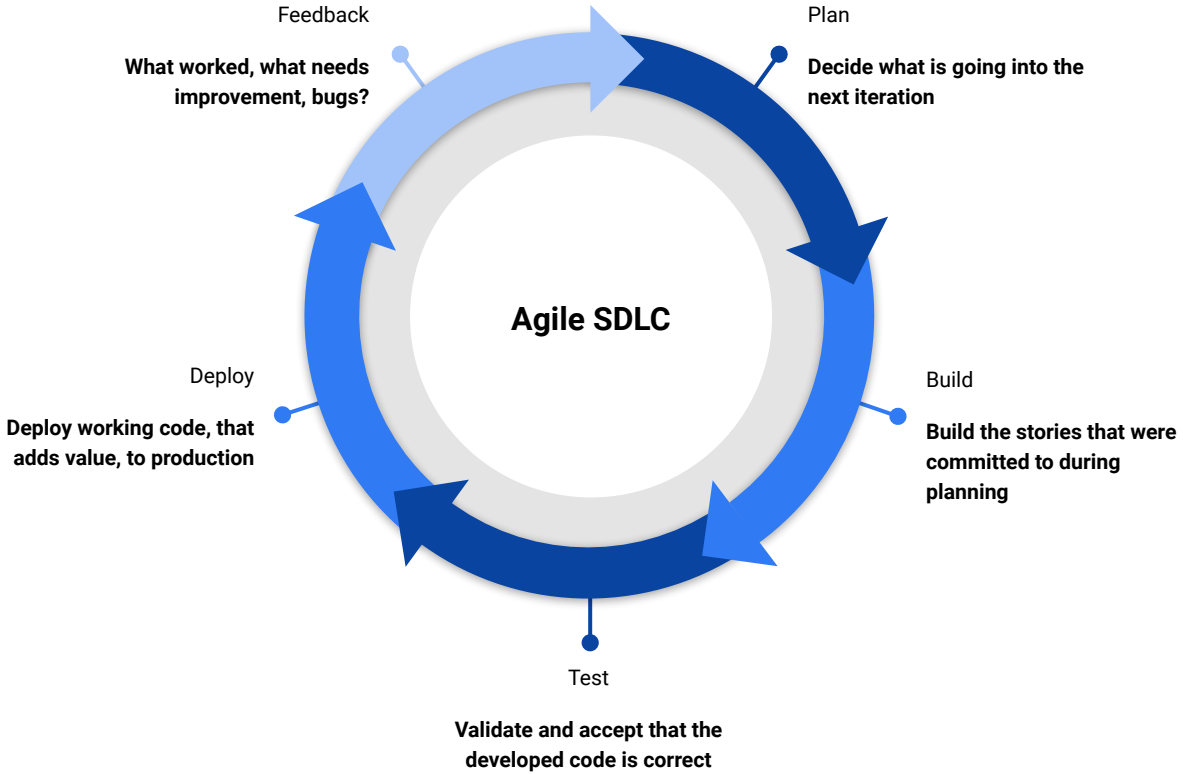
Fail fast

Do not let perfection become the enemy of good

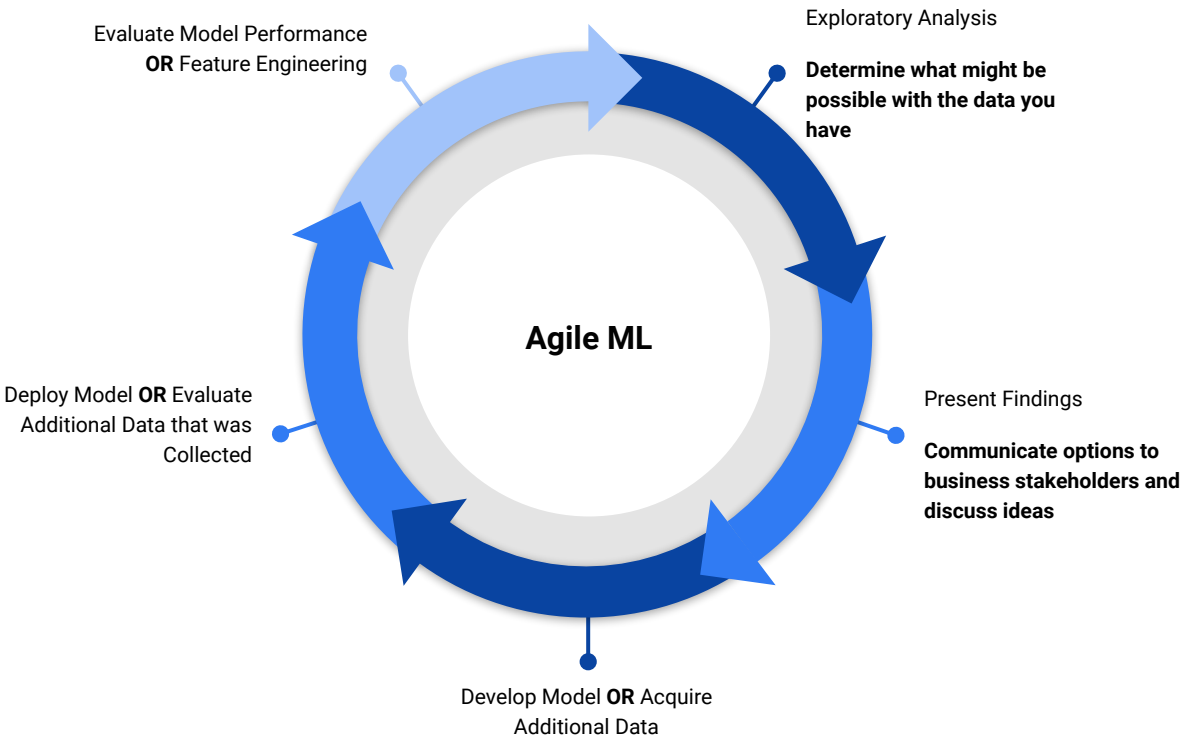
Use all the tools in your toolbelt to bridge gaps (SMOTE, Cluster Analysis, data engineering/management tools, etc.)

Automate everything (or use tools that automate everything for you)

The dialog between product and your data science team needs to be a two-way discussion


Feedback
What worked, what needs improvement, bugs?
Plan
Decide what is going into the next iteration
Agile SDLC
Deploy
Deploy working code, that adds value, to production
Build
Build the stories that were committed to during planning
Test
Validate and accept that the developed code is correct

The traditional agile SDLC flow can not work for machine learning projects (at least at conception)


Agile ML
Exploratory Analysis
Determine what might be possible with the data you have
Present Findings
Communicate options to business stakeholders and discuss ideas
Develop Model OR Acquire Additional Data
Deploy Model OR Evaluate Additional Data that was Collected
Evaluate Model Performance OR Feature Engineering

You've got to work with what you have

(don't let your data science initiative become a multi-million dollar, multi-year failed science experiment)

(remember, you have an 85% chance of this happening to you)

Enabling Practices of Agile ML

What to do or what to do more of to drive success

Data science/statistics techniques that can accelerate research

Organizational streamlining to enable the cross functional work necessary for data science

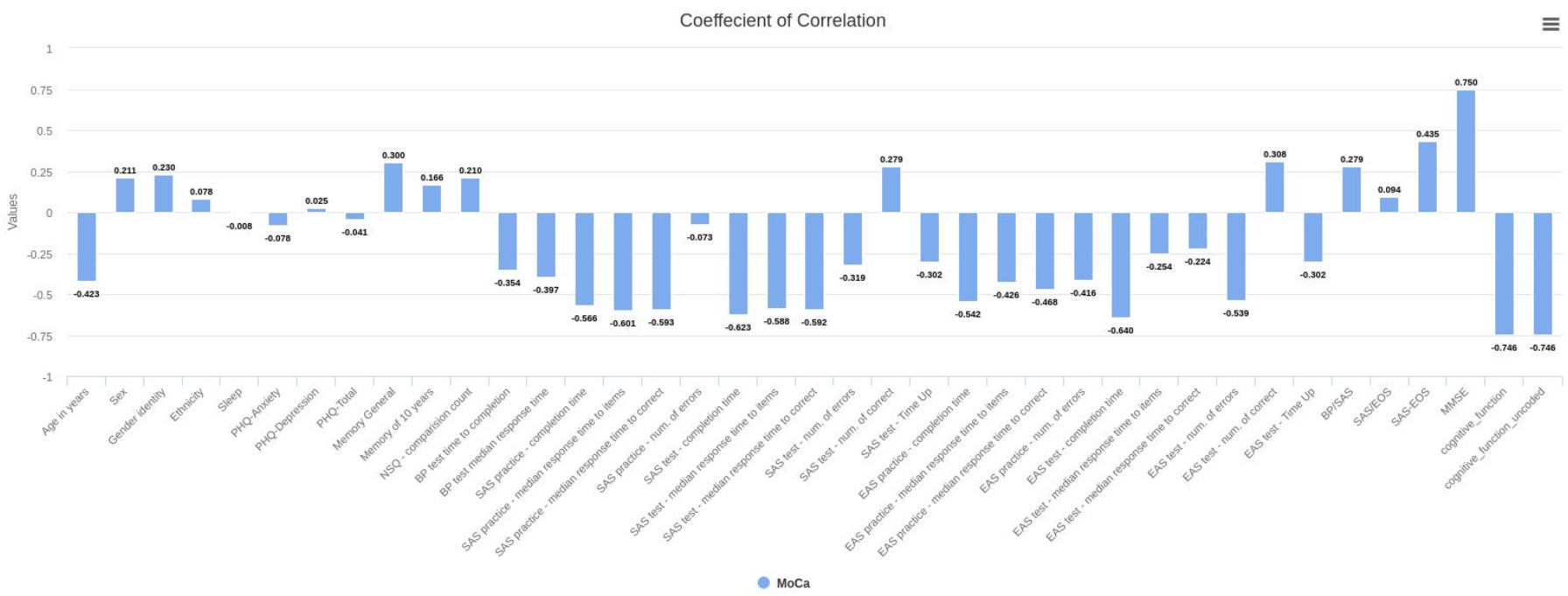
Follow a disciplined, agile model development life cycle

Small (right) data

Accelerating the pace of research and development

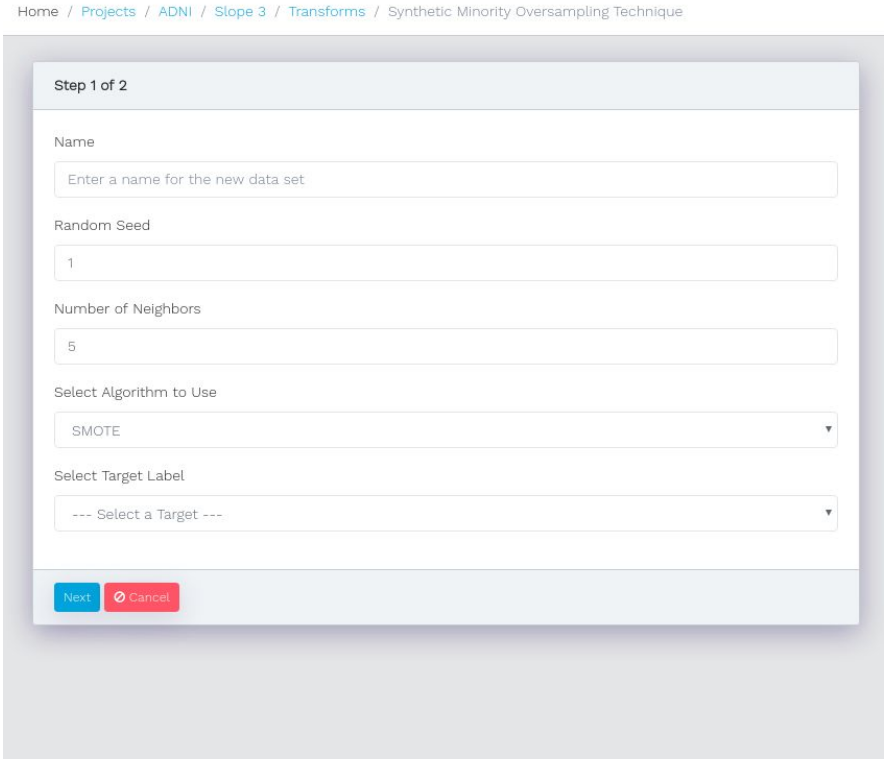
(use all the shortcuts at your disposal)

Coefficient of Correlation



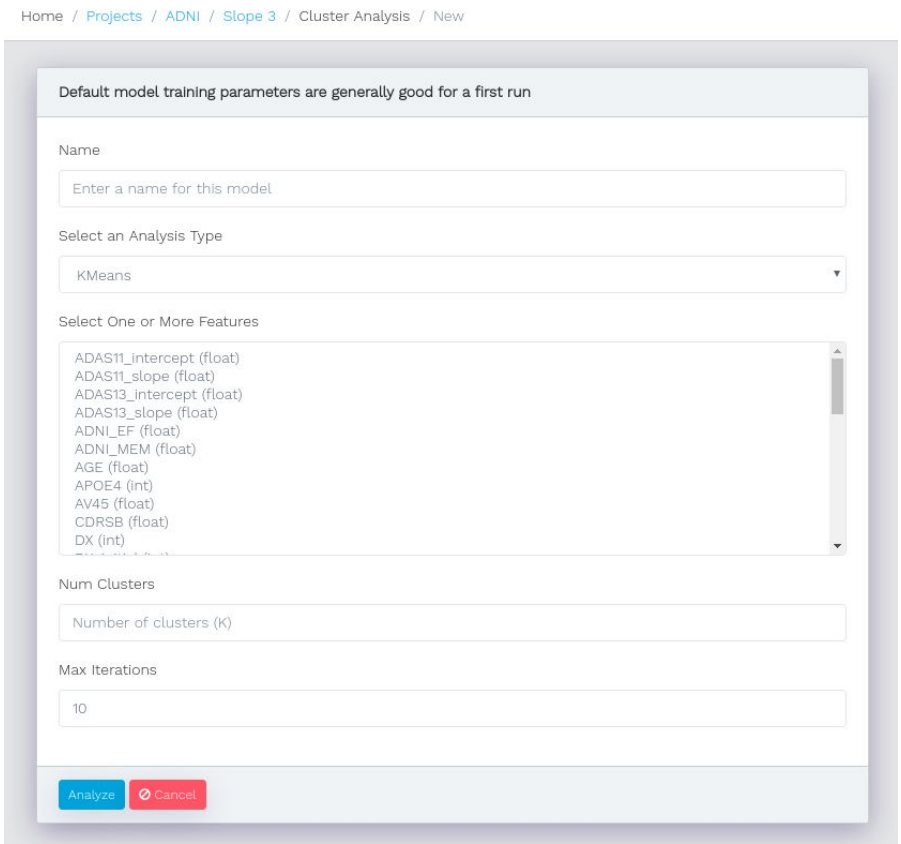
Statistical Minority Oversampling Technique

Use this when the data is largely representative, but you'd like more examples of a case where you only have a few



Cluster Analysis

Generate labels that are mathematically sound when your data is not already classified



Organization Streamlining
(how to become an R&D powerhouse)

Most Common Organizational Pitfalls

And how you can avoid them

Data access (aka, data silos)

Network connectivity (we can't get there, from here)

PII/Regulation (data obfuscation and encryption)

Big data (aka use small data)

Hire some data engineers (and listen to them)

Use AN Agile Process

(you want your work to be repeatable and predictable)

Agile ML

Key differences from agile development

You don’t have the requirements up front

If you do have the requirements, the likelihood that your requirements line up with reality is extremely low

Models require constant evaluation to stay relevant - data drift is real and will invalidate the best models in time

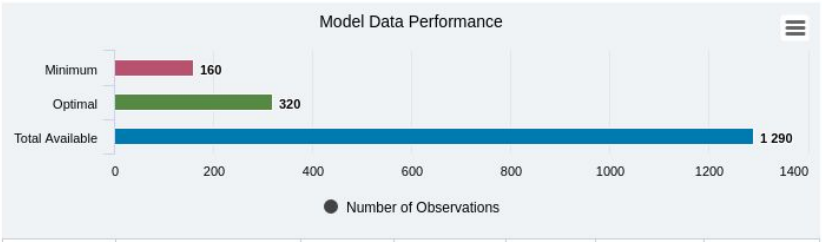
Forget about big data

(use only what you need)

You don’t/rarely need BIG DATA

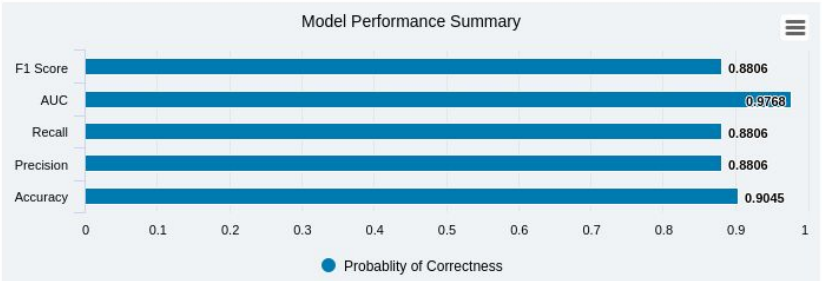
There are some exceptions...

Training Data Summmary



Total Features	Num of Training Features	Total Rows	Training Rows	Test Rows	Training Loss	Optimal Rows
45	16	1726	1290	335	5.8517%	160 - 320

Model Performance



F1 Score	AUC	Recall	Precision	Accuracy

Case Study

(how we built a model to forecast the onset of dementia that has practical, clinical application and can improve health outcomes)

10% of the population will develop dementia

And the number of people who are aging into the risk bracket is peaking with baby boomers

Dementia is like diabetes - catch it early enough, and you can slow or halt the progression of the disease

Modifiable risk factors are key

Most existing (and currently under investigation) research is focused on biomarkers and the search for a pill

What we know about the disease

Significant research exists on both biomarkers and (to a lesser extent) cognitive function

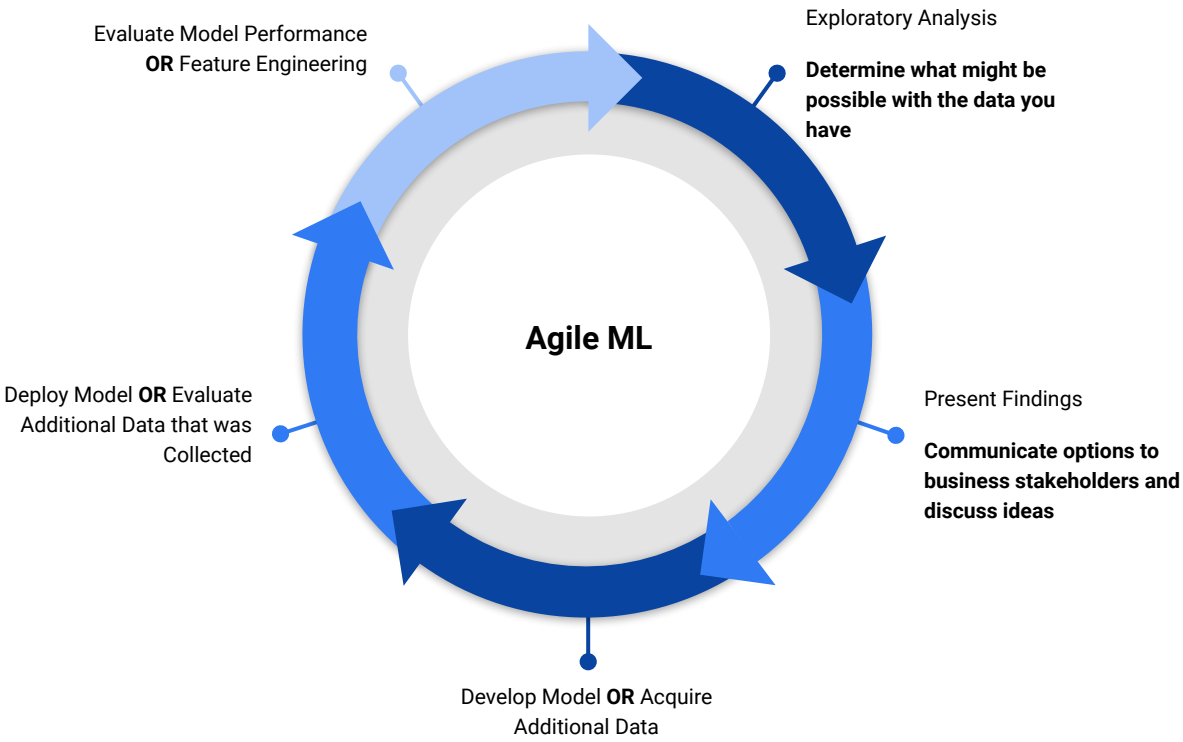
The search for a magic pill is not going well, but industry continues pouring money into it

The slope of cognitive decline for individuals developing dementia is significantly different than the slope of decline associated with healthy aging

We know that modifiable risk factors, improved before the onset of symptoms, can delay or halt progression

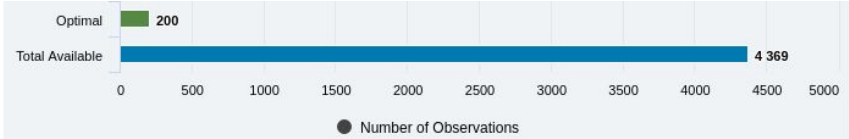
Target state:

An instrument that can be easily leveraged in a clinical setting to collect longitudinal data and improve patient outcomes


Agile ML
Exploratory Analysis
Determine what might be possible with the data you have
Present Findings
Communicate options to business stakeholders and discuss ideas
Develop Model OR Acquire Additional Data
Deploy Model OR Evaluate Additional Data that was Collected
Evaluate Model Performance OR Feature Engineering

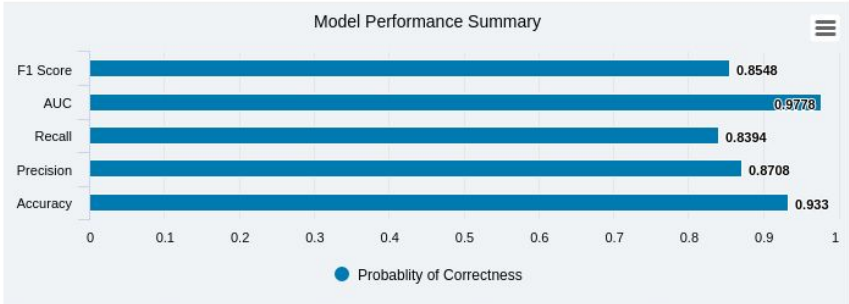
Can we replicate an existing study?

Most research was being done on biomarkers linked to dementia

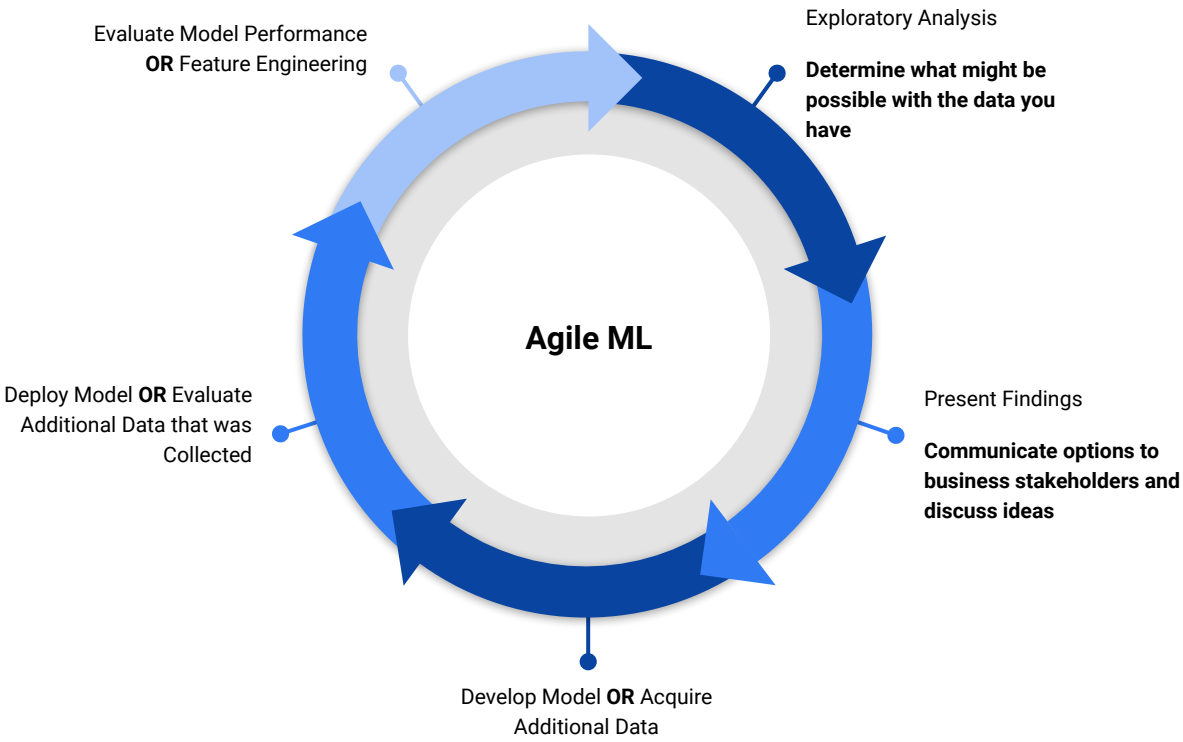


Total Features	Num of Training Features	Total Rows	Training Rows	Test Rows	Training Loss	Optimal Rows
36	10	12660	4369	1059	57.1248%	100 - 200

Model Performance

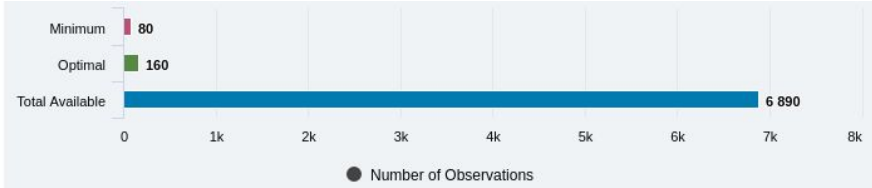


F1 Score	AUC	Recall	Precision	Accuracy
0.8548	0.9778	0.8394	0.8708	0.933


Agile ML
Exploratory Analysis
Determine what might be possible with the data you have
Present Findings
Communicate options to business stakeholders and discuss ideas
Develop Model OR Acquire Additional Data
Deploy Model OR Evaluate Additional Data that was Collected
Evaluate Model Performance OR Feature Engineering

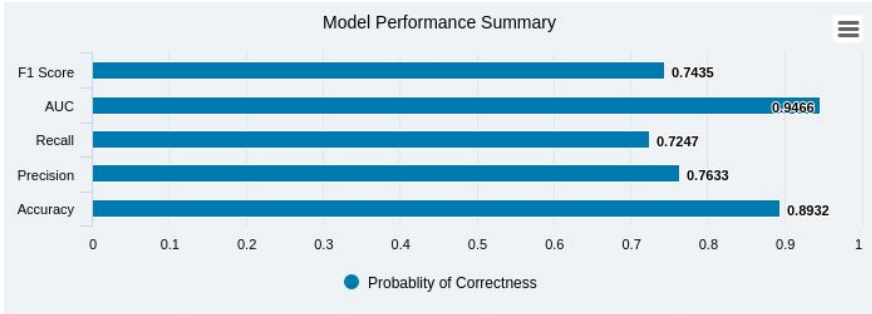
Can we do it with only cognitive test scores?

The answer should be yes, but we should test it

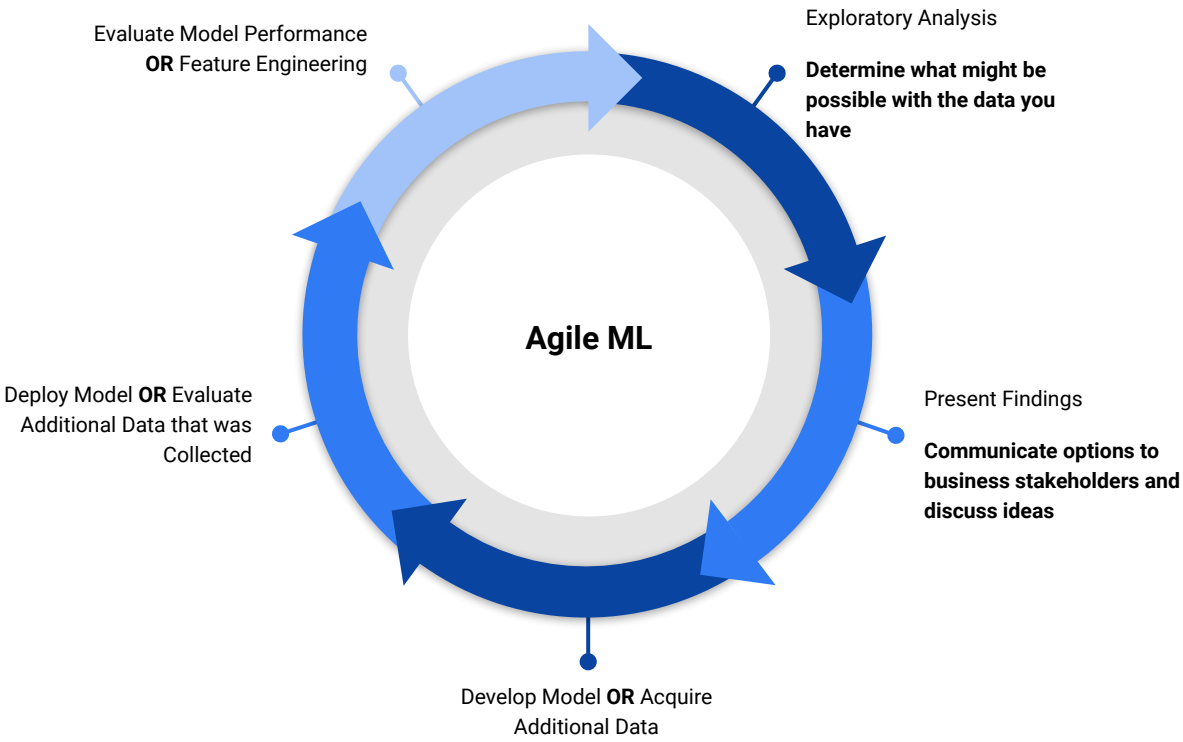


Total Features	Num of Training Features	Total Rows	Training Rows	Test Rows	Training Loss	Optimal Rows
36	8	12660	6890	1667	32.4092%	80 - 160

Model Performance



F1 Score	AUC	Recall	Precision	Accuracy
0.7435	0.9466	0.7247	0.7633	0.8932

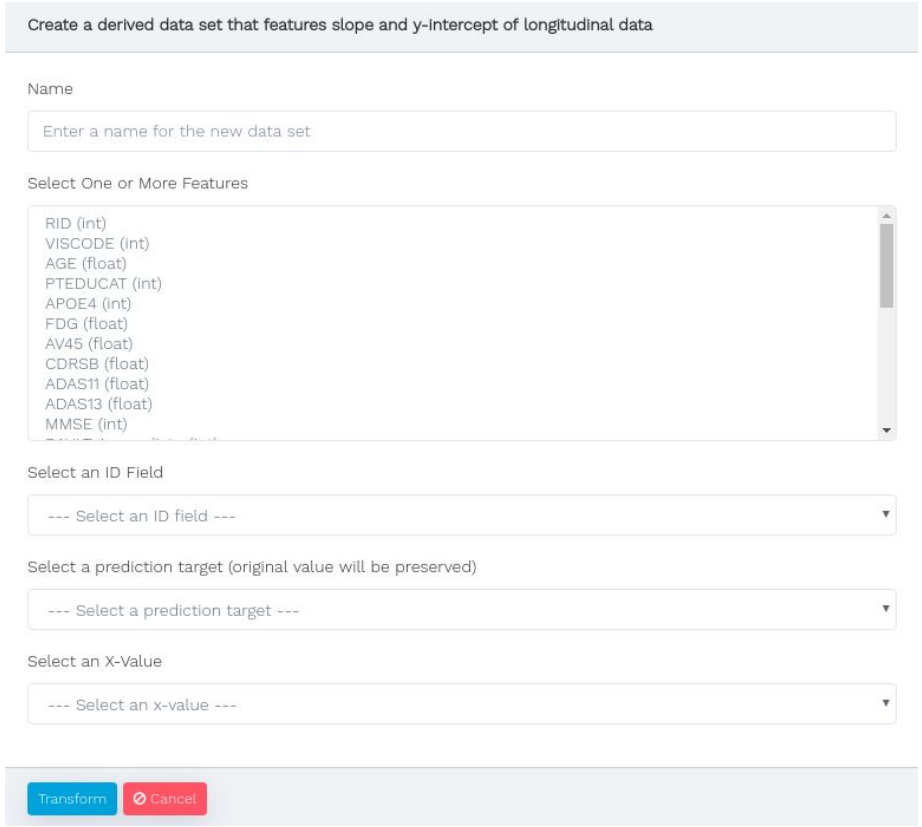

Agile ML
Exploratory Analysis
Determine what might be possible with the data you have
Present Findings
Communicate options to business stakeholders and discuss ideas
Develop Model OR Acquire Additional Data
Deploy Model OR Evaluate Additional Data that was Collected
Evaluate Model Performance OR Feature Engineering

What about the steeper slope of cognitive decline for individuals developing dementia?

The ADNI data set has longitudinal data...

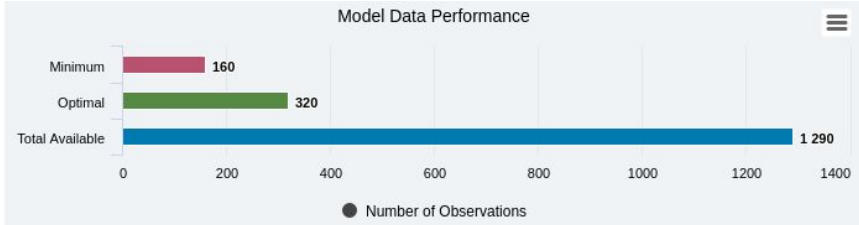
First, let's figure out the slope for each score

This will give us slope scores for both normal and AD patients to train a model



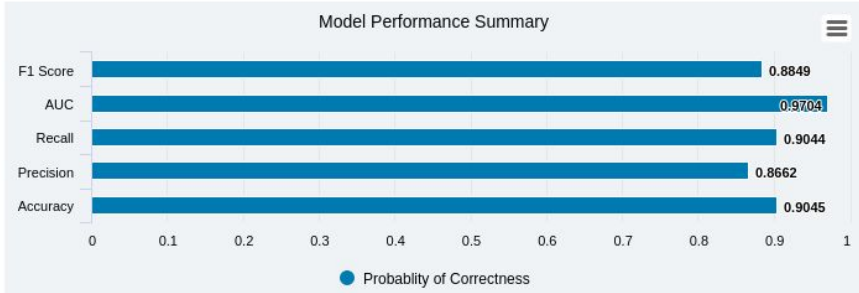
Now we have a new, derived data set

Can we train a model that is capable for predicting dementia based only on rate of decline for cognitive function?



Total Features	Num of Training Features	Total Rows	Training Rows	Test Rows	Training Loss	Optimal Rows
45	16	1726	1290	335	5.8517%	160 - 320

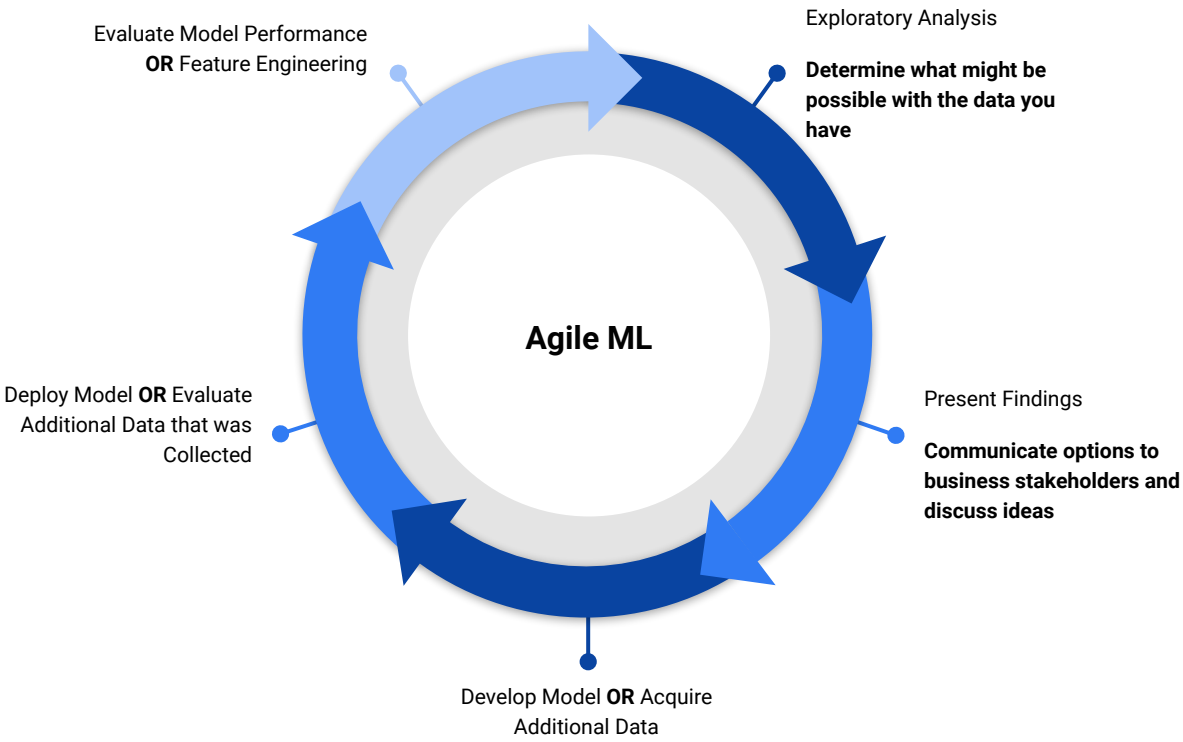
Model Performance



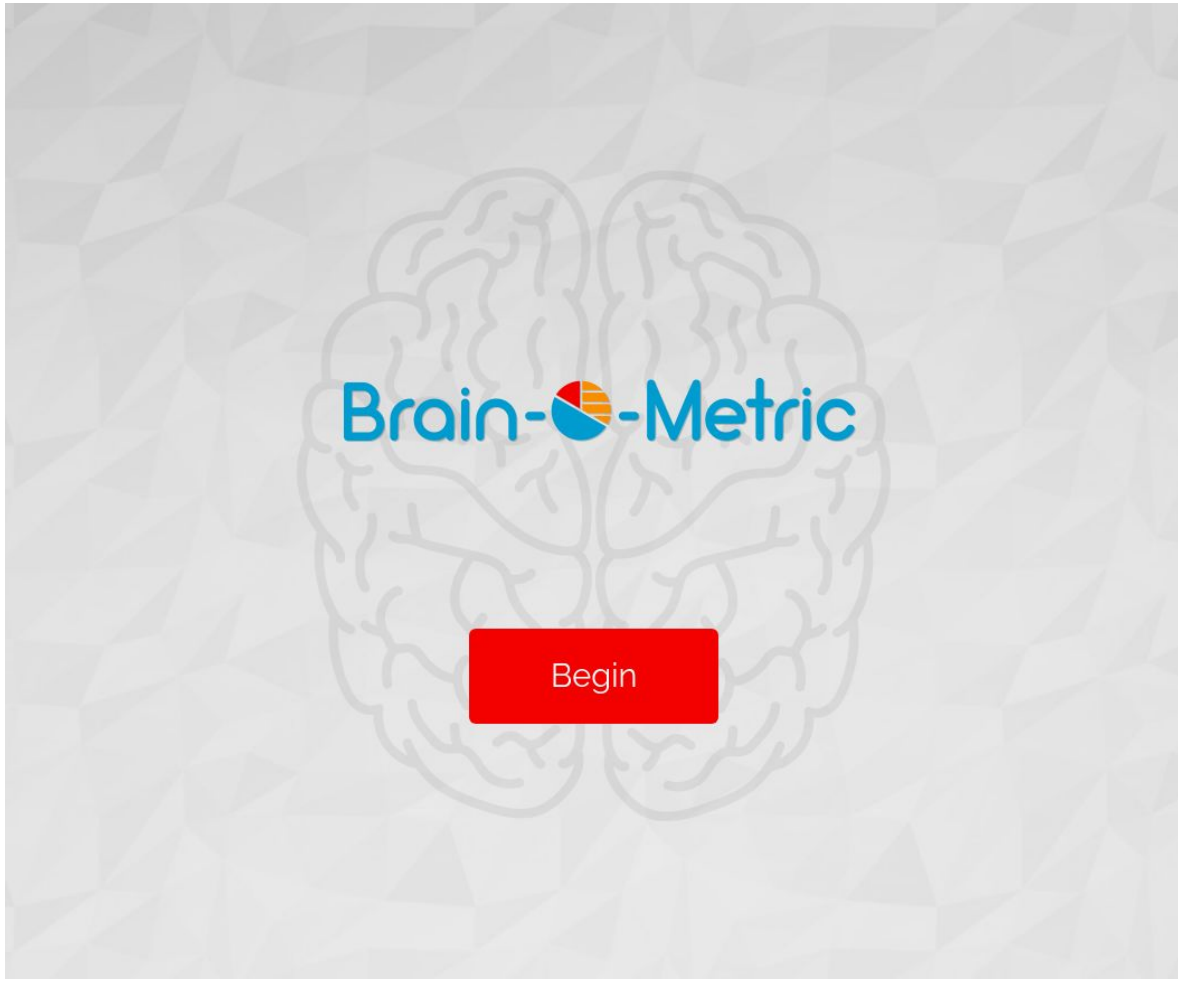
F1 Score	AUC	Recall	Precision	Accuracy
0.8849	0.9704	0.9044	0.8662	0.9045

Using ioModel, we created a new derived data set that incrementally measured the slope of decline between office visits


Our new model can forecast the onset of dementia between 1 and 10 years before a doctor will diagnose it based on symptoms and cutoff scores

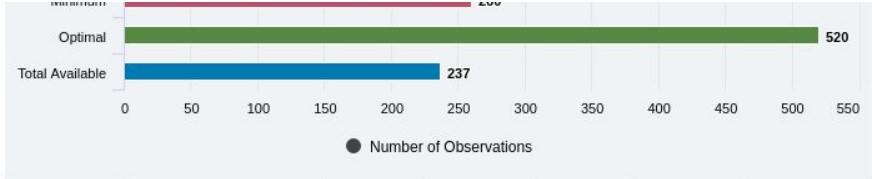

Agile ML
Exploratory Analysis
Determine what might be possible with the data you have
Present Findings
Communicate options to business stakeholders and discuss ideas
Develop Model OR Acquire Additional Data
Deploy Model OR Evaluate Additional Data that was Collected
Evaluate Model Performance OR Feature Engineering

What if we had a computerized test that accurately measured cognitive function, can be applied in a clinical setting and done at each office visit?


Brain-e-Metric
Begin

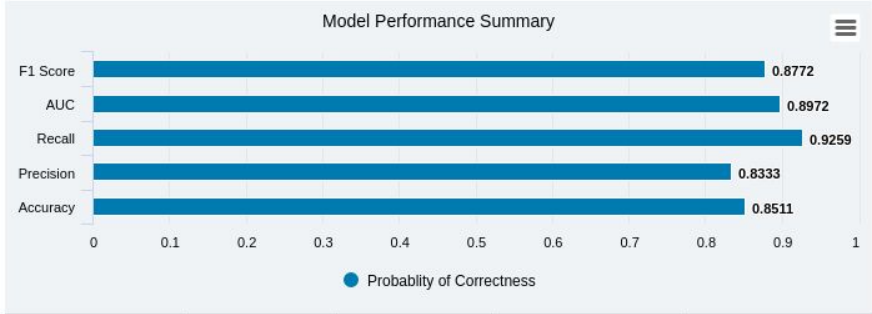
We need to validate the correctness of Brain-o-metric

We can verify that we can predict MoCA scores from Brain-o-metric's cognitive scores

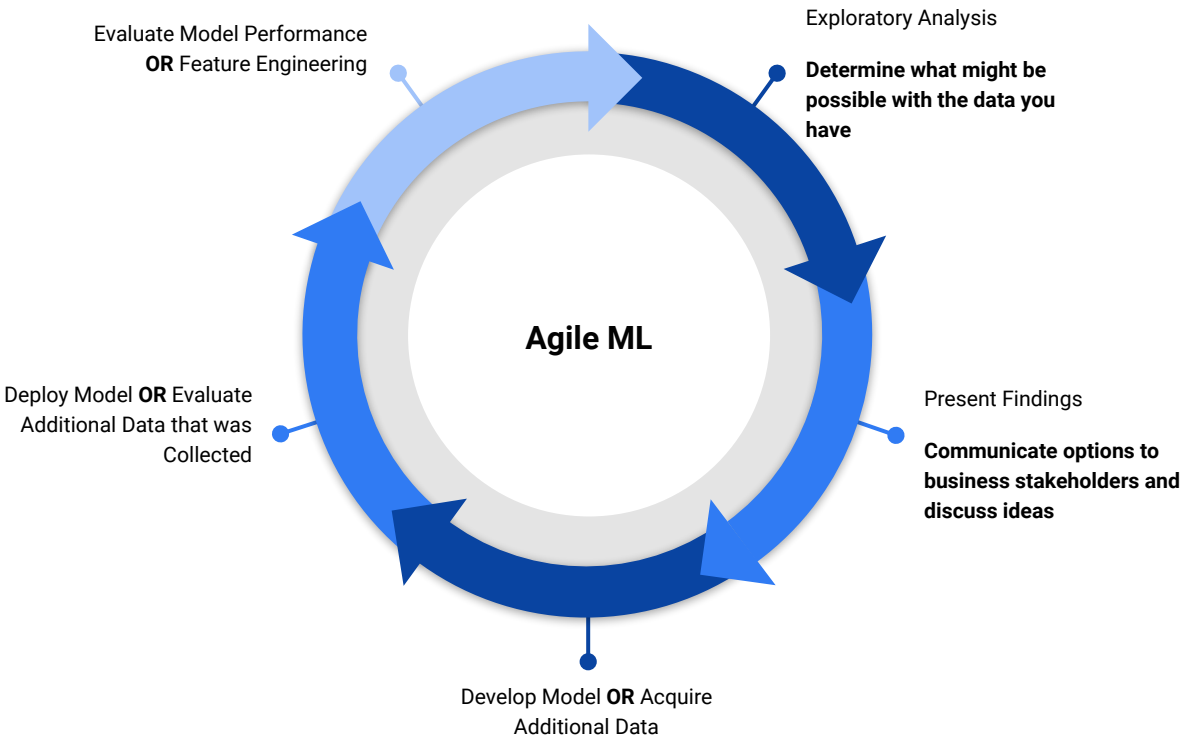


Total Features	Num of Training Features	Total Rows	Training Rows	Test Rows	Training Loss	Optimal Rows
28	26	284	237	47	0.0%	260 - 520

Model Performance



F1 Score	AUC	Recall	Precision	Accuracy
0.8772	0.8972	0.9259	0.8333	0.8511


Agile ML
Exploratory Analysis
Determine what might be possible with the data you have
Present Findings
Communicate options to business stakeholders and discuss ideas
Develop Model OR Acquire Additional Data
Deploy Model OR Evaluate Additional Data that was Collected
Evaluate Model Performance OR Feature Engineering

We are currently gather longitudinal data with Brain-o-metric. Once we have enough data, we'll train a new model that can forecast the onset of the disease using a simple computerized test.

Key Takeaways

Path to Machine Learning Success

Work iteratively and follow agile practices

Fail fast

Do not let perfection become the enemy of good

Use all the tools in your toolbelt to bridge gaps (SMOTE, Cluster Analysis, data engineering/management tools, etc.)

Automate everything (or use tools that automate everything for you)

Thank you for coming

matt@twintechlabs.io

Slides are here:

http://github.com/twintechlabs/talks