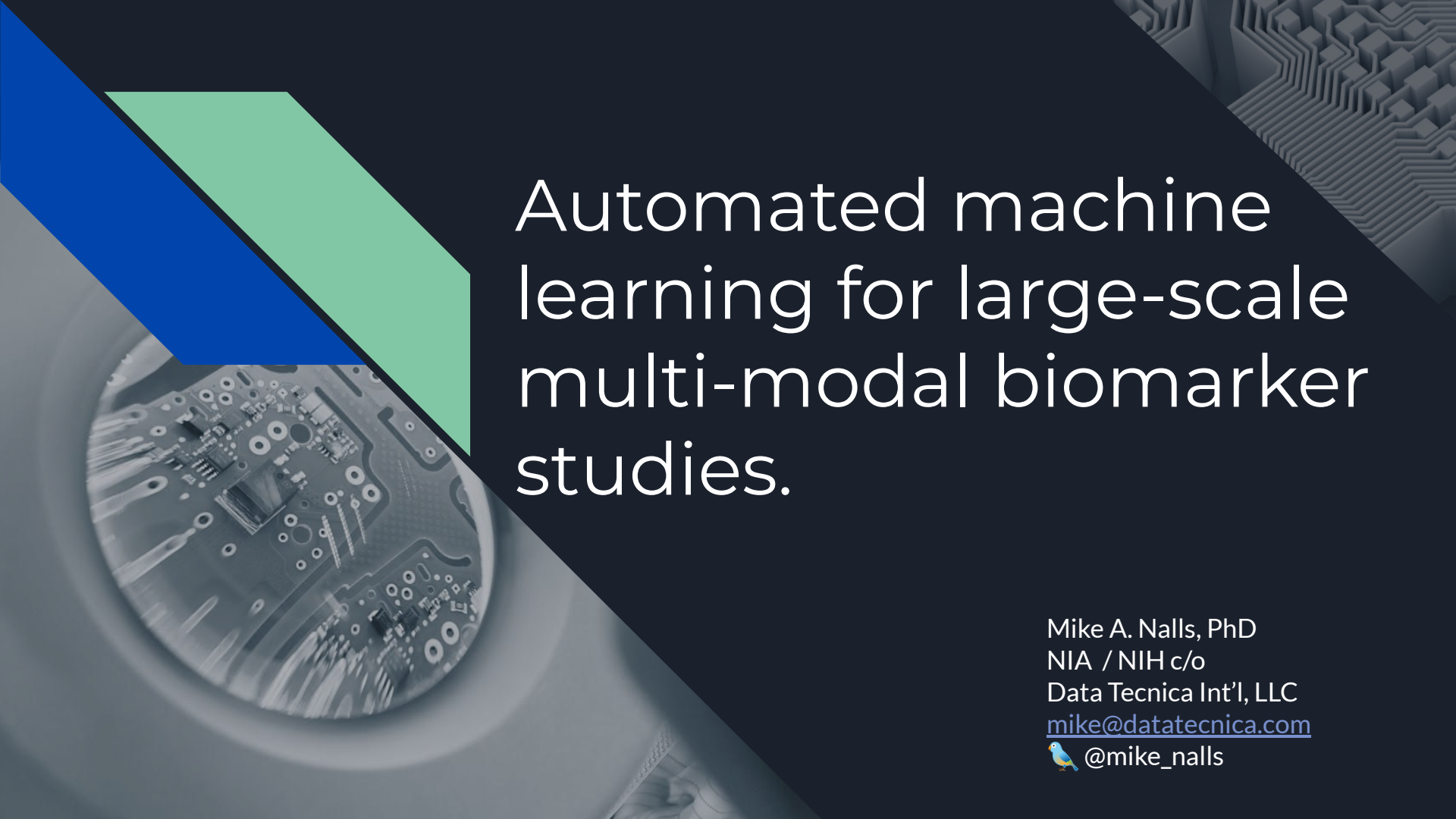


# Automated machine learning for large-scale multi-modal biomarker studies.

Mike A. Nalls, PhD
NIA / NIH c/o
Data Tecnica Int'l, LLC
mike@datatecnica.com
@mike_nalls

# Quick intro to key concepts in machine learning (ML).

Images from this great review I had no part of.

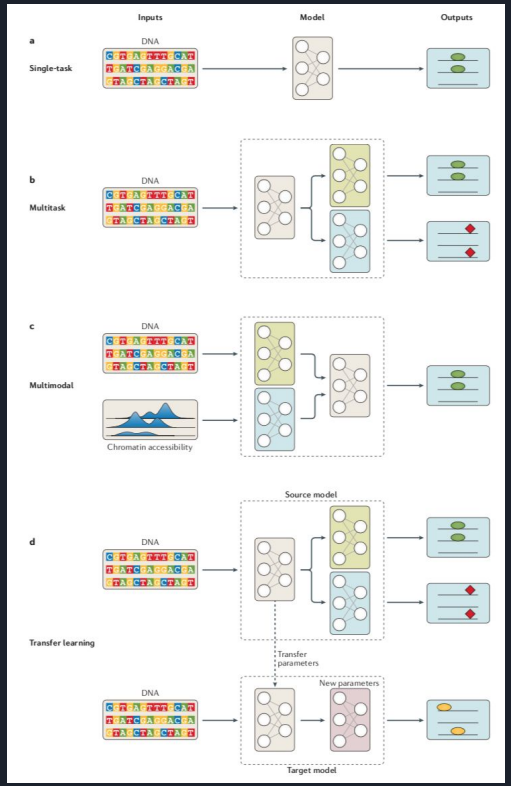
REVIEWS

## Deep learning: new computational modelling techniques for genomics

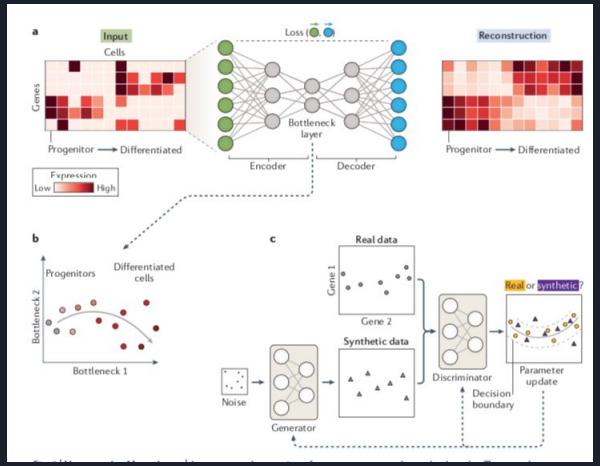
*Gökcen Eraslan[1,2,5], Žiga Avsec[3,5], Julien Gagneur[3]* and Fabian J. Theis[1,2,4]**

Abstract | As a data-driven science, genomics largely utilizes machine learning to capture dependencies in data and derive novel biological hypotheses. However, the ability to extract new insights from the exponentially increasing volume of genomics data requires more expressive machine learning models. By effectively leveraging large data sets, deep learning has transformed fields such as computer vision and natural language processing. Now, it is becoming the method of choice for many genomics modelling tasks, including predicting the impact of genetic variation on gene regulatory mechanisms such as DNA accessibility and splicing.

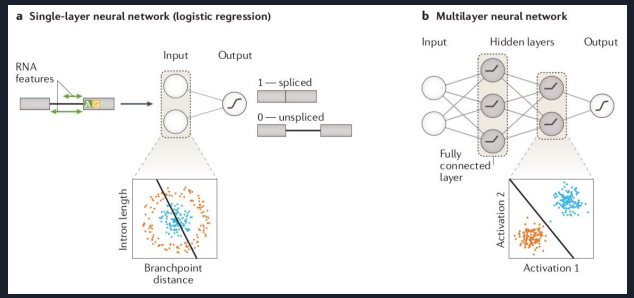
## Supervised learning



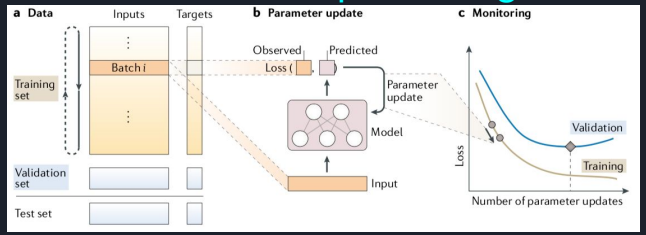

## Unsupervised learning




## Linear models (regression) are ML.

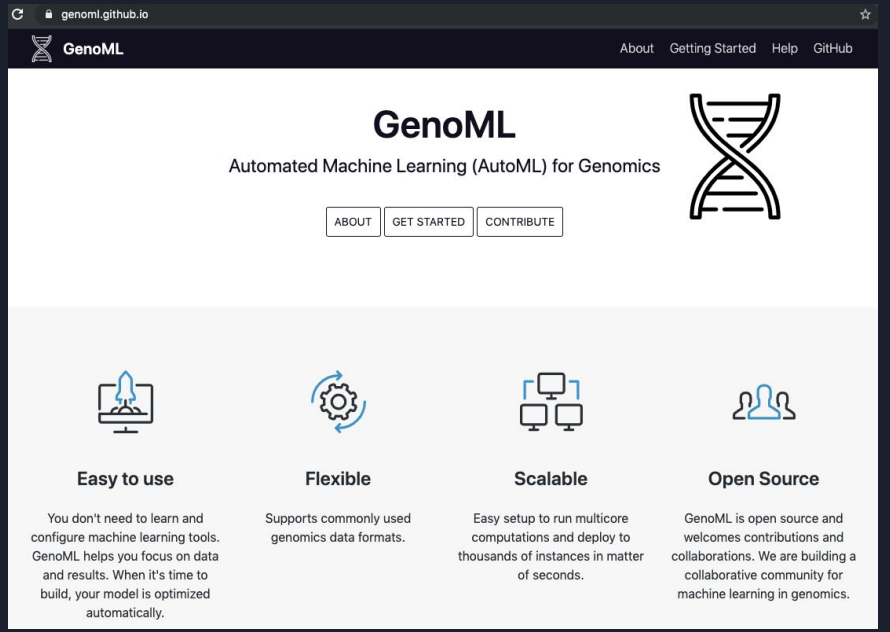



## Cross-validation plus tuning



# Democratizing and automating ML
# GenoML - genoml.github.io.

- Older beta version available (R)
- Current version (python)in October ‘19
- Automated ML tuned for clinical and genomics applications
- Feature selection
- Compete dozens of most popular algorithms for classification, regression and clustering
- Longitudinal and meta/federated learning methods under development
- Model zoo
- Terra ready!




Current python version is quite fast and memory efficient from preprocessing to training and tuning... competing 12 models on 1,000 samples @ 100,000 features takes ~1 hour using 6 cores and 64GB RAM.

# Why use ML as an analytic framework for biomarker studies?

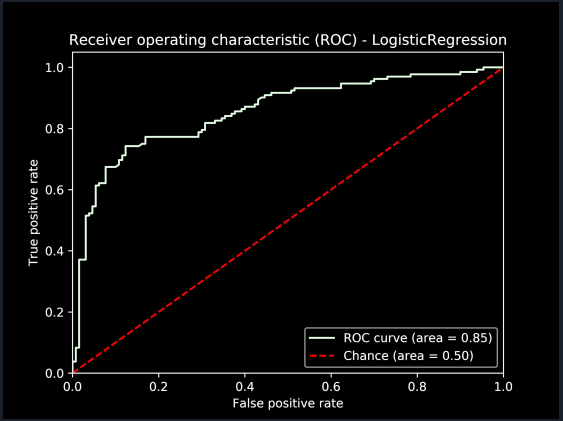
- PRS are great but you need a huge GWAS, got one applicable to your trait or even your population of interest?
- Maybe problems you are dealing with can't be solved by a line?
- Hey, are you trying to include things that aren't SNPs?
- Need to run production scale analytics?
- Did you know business is really far ahead of science in terms of analytics?
- Maybe just looking at one predictor at a time isn't a great idea?

# Multimodal ML in AMP-PD: pilot work.

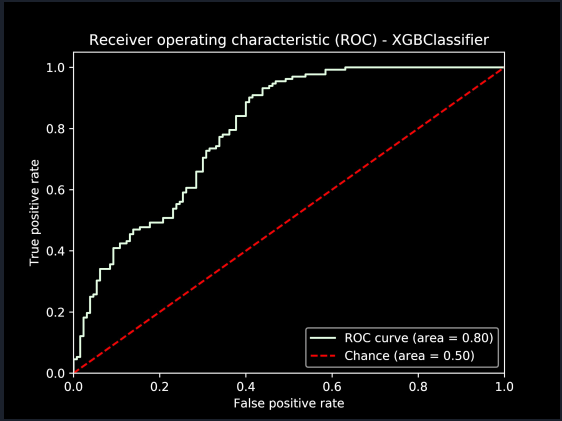
- 872 samples from PPMI passing QC
  - “Idiopathic PD” with no other NDD
  - Controls
- Genome sequencing data
  - Reduced to ~50K variants
    - GWAS hits (90 SNPs) - Nalls et al 2019
    - LD pruned tag SNPs at MAF > 5%
- Clinical candidate predictors
  - UPSIT
  - Family history
  - Age
  - Genetic sex
- Transcriptomic data
  - ~50k normalized transcripts
  - Pre-release
  - Jensen and Craig labs
  - Thanks David and Kendall!

# clinical VS genotype VS omic Performance in 30% test samples

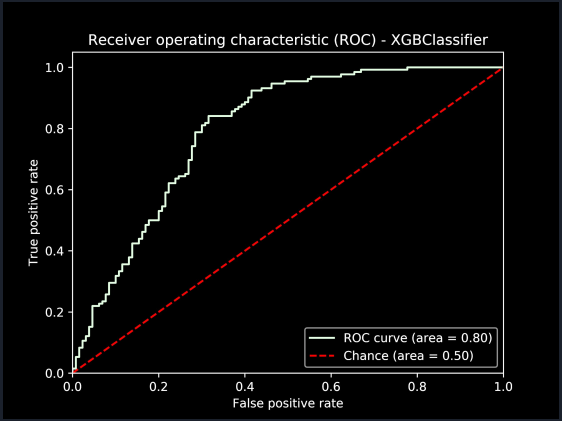
| Dataset | Algorithm | AUC_Percent | Accuracy_Percent | Balanced_Accuracy_Percent | Log_Loss | Sensitivity | Specificity | PPV | NPV | Runtime_Seconds |
|---|---|---|---|---|---|---|---|---|---|---|
| *Omic* | XGBClassifier | 79.639 | 75.191 | 75.152 | 0.540 | 0.803 | 0.700 | 0.731 | 0.778 | 248.344 |
| *Genotype* | XGBClassifier | 79.557 | 70.611 | 70.594 | 0.522 | 0.727 | 0.685 | 0.701 | 0.712 | 83.112 |
| *Clinical* | LogisticRegression | 85.487 | 79.389 | 79.452 | 0.477 | 0.712 | 0.877 | 0.855 | 0.750 | 0.012 |




clinical




genotype

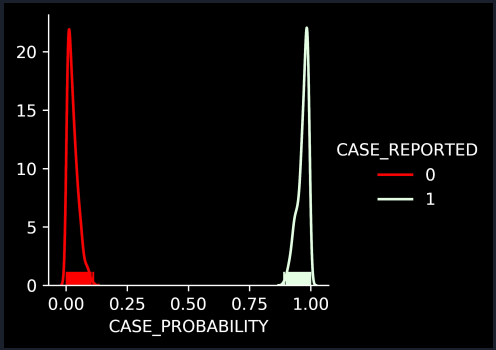


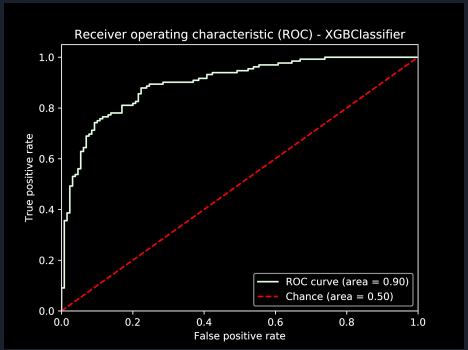

omic

# Multiple modalities in the same model perform best.

Performance in 30% test samples

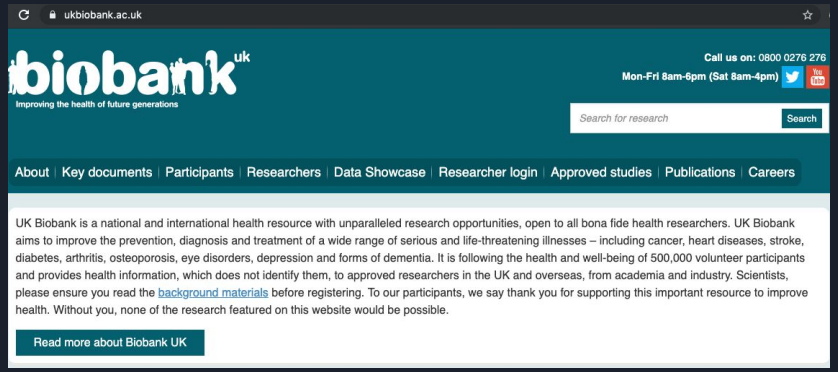
| Algorithm | AUC_Percent | Accuracy_Percent | Balanced_Accuracy_Percent | Log_Loss | Sensitivity | Specificity | PPV | NPV | Runtime_Seconds |
|---|---|---|---|---|---|---|---|---|---|
| *XGBClassifier* | 89.8776 | 80.5344 | 80.5536 | 0.4078 | 0.7803 | 0.8308 | 0.8240 | 0.7883 | 308.8823 |
| *GradientBoostingClassifier* | 89.7786 | 82.8244 | 82.8613 | 0.4103 | 0.7803 | 0.8769 | 0.8655 | 0.7972 | 482.1020 |
| *BaggingClassifier* | 84.3590 | 75.1908 | 75.2331 | 0.6044 | 0.6970 | 0.8077 | 0.7863 | 0.7241 | 158.2600 |
| *AdaBoostClassifier* | 83.8054 | 79.7710 | 79.7844 | 0.6553 | 0.7803 | 0.8154 | 0.8110 | 0.7852 | 223.0328 |
| *LogisticRegression* | 81.2238 | 73.6641 | 73.6364 | 0.7095 | 0.7727 | 0.7000 | 0.7234 | 0.7521 | 19.0327 |
| *LinearDiscriminantAnalysis* | 78.3625 | 72.1374 | 72.1212 | 0.5645 | 0.7424 | 0.7000 | 0.7153 | 0.7280 | 36.2314 |
| *MLPClassifier* | 77.9720 | 72.1374 | 72.1620 | 0.8206 | 0.6894 | 0.7538 | 0.7398 | 0.7050 | 229.5861 |
| *SVC* | 74.4814 | 69.4656 | 69.5280 | 0.5967 | 0.6136 | 0.7769 | 0.7364 | 0.6645 | 360.2292 |
| *SGDClassifier* | 72.8147 | 72.9008 | 72.8147 | 9.3597 | 0.8409 | 0.6154 | 0.6894 | 0.7921 | 2.9697 |
| *RandomForestClassifier* | 71.9930 | 60.3053 | 60.4138 | 0.6056 | 0.4621 | 0.7462 | 0.6489 | 0.5774 | 1.5003 |
| *KNeighborsClassifier* | 63.6946 | 59.9237 | 60.0233 | 1.5562 | 0.4697 | 0.7308 | 0.6392 | 0.5758 | 146.4244 |
| *QuadraticDiscriminantAnalysis* | 44.6911 | 44.6565 | 44.6911 | 19.1150 | 0.4015 | 0.4923 | 0.4454 | 0.4476 | 26.8305 |



# Production scale analyses in the UK Biobank.

- > 400K participants
  - Sampling to reduce resource use
  - 1:1 case to control
  - 5K samples per trait maximum
- > 7M common imputed variants
  - LD pruned to ~31K variants
- Demographics
  - Age
  - Genetic sex
  - Townsend score
  - All analyses stratified by ancestry
  - Sex stratification as well when needed
- ICD10 codes
  - Granular and parent level analyses
  - Only codes with > 1K cases



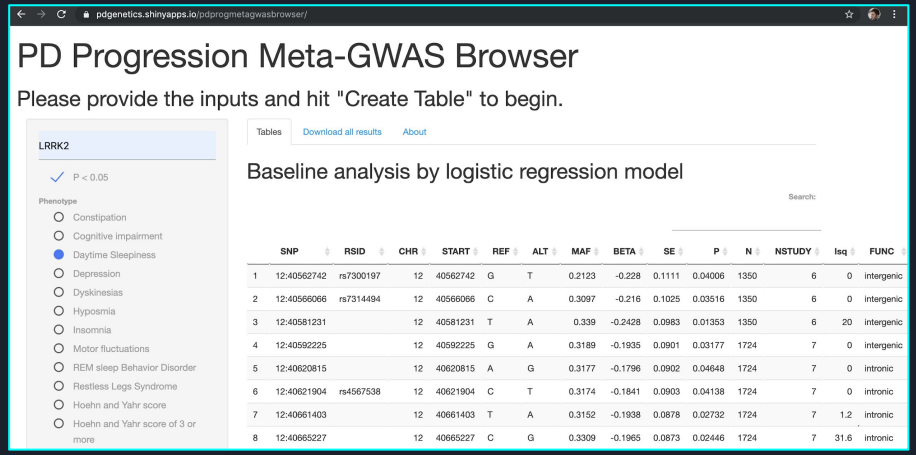
# UKB results browser and model zoo.


| Algorithm | AUC_Percent | Accuracy_Percent | Balanced_Accuracy_Percent | Log_Loss | Sensitivity | Specificity | PPV | NPV | Runtime_Seconds |
|---|---|---|---|---|---|---|---|---|---|
| LogisticRegression | 60.15840336469242 | 64.22222222222223 | 55.58821744729337 | 1.260644814993547 | 0.2891566265060241 | 0.8226077224398433 | 0.4543889845094665 | 0.6937234544596508 | 44.19177460670471 |
| RandomForestClassifier | 54.77621939147954 | 64.29629629629629 | 51.38541651982096 | 0.8484929278328721 | 0.1150054764512595 8 | 0.9127028539451596 | 0.4022988505747127 | 0.6687166871668717 | 5.575837373733520 5 |
| AdaBoostClassifier | 82.79248141776037 | 74.18518518518519 | 72.59656727331567 | 0.6582476586558149 | 0.6768893756845564 | 0.7750419697817571 | 0.6058823529411764 | 0.8244047619047619 | 162.17812204360962 |
| GradientBoostingClassifier | 84.22849458576025 | 76.74074074074075 | 77.527058940345 | 0.41648210408422703 | 0.7995618838992333 | 0.7509792949076665 | 0.6212765957446809 | 0.88 | 879.8812704086304 |
| SGDClassifier | 51.43202917995429 | 65.74074074074075 | 51.43202917995429 | 11.83272895010829 | 0.07228915662650602 | 0.9563514269725796 | 0.4583333333333333 | 0.6686228482003 13 | 8.285813570022581 |
| SVC | 60.32508729530729 | 66.18518518518518 | 50 | 0.6259577616963824 | 0 | 1 | 0.6618518518518518 | 13463.533816814424 |
| MLPClassifier | 59.20561730056003 | 65.07407407407408 | 55.48178367435249 | 1.1383166462206389 | 0.2584884994523549 | 0.8511471740346949 | 0.4701195219123506 | 0.6919927206551411 | 588.8015115261078 |
| KNeighborsClassifier | 51.660342341028155 | 60.25925925925926 | 50.50516968418008 | 2.142168131767036 | 0.20372398685651694 | 0.8063794068270845 | 0.3496240601503759 5 | 0.6646678966789668 | 5293.71445941925 |
| LinearDiscriminantAnalysis | 55.3258871575226 | 65.03703703703704 | 51.40932044809446 | 0.6614189365910975 | 0.0930996714129244 4 | 0.9350867375489648 | 0.4228855721393035 | 0.6686674669867947 | 558.4927585124971 |
| QuadraticDiscriminantAnalysis | 50.45438915962982 | 49.555555555555564 | 50.45438915962982 | 17.422893870321612 | 0.5323110624315444 | 0.4767767207610 52 | 0.3420126671358198 5 | 0.6661454261141517 | 416.3209097385407 |
| BaggingClassifier | 82.47958512587256 | 74.22222222222223 | 71.71386875272367 | 0.7197249156180204 | 0.6396495071193866 | 0.7946278679350868 | 0.6140904311251314 | 0.8118925100057176 | 301.46415305137 64 |
| XGBClassifier | 84.09285511583904 | 76.33333333333333 | 77.43355780552132 | 0.4165772632122712 | 0.8083242059145673 | 0.7403469501958589 | 0.6139767054908486 | 0.8831775700934 58 | 473.8226733207703 |

Should be up in a couple weeks for multiple ancestry groups and > 2K codes.
Trained and tuned model files will be available for download and application to your data

# Application of UKB results to PD studies.

- Progression cohorts are generally small and underpowered
- Predict comorbidities and ancillary outcomes from UKB to PD cohorts
- Predict aspects of progression markers that overlap with ICD10 codes
  - For example … constipation, depression etc from UPDRS
- Endpoint prediction in trial design




https://pdgenetics.shinyapps.io/pdprogmetagwasbrowser/

- >4k samples
- 25k obs @ ~4 years
- Progression model predictions will be included
  - In the works ;-)
  - Basis for federated learning work

# Next steps...

- UKB
  - Models across ancestry groups
  - Fit relevant models to PD data to predict comorbidities
  - Browser and model zoo go live
- PD progression
  - Predict PD specific outcomes
- PD-AMP
  - Integrate more molecular datatypes
  - Include more cohorts
- GenoML
  - Python release
  - Currently optimizing
  - Expand ecosystem
    - LSTM
    - Federated learning
    - Outlier detection
    - Unsupervised learning

# Thanks!

If you have any questions, please email mike@datatecnica.com.

Shout outs to:

Andy Singleton - Good idea guy

**Hirotaka Iwaka - Lead, clinical data and progression studies**

**Jeff Kim - Lead, open data and browser development**

**Mary Makarious - Lead, notebooks for multimodal analysis in AMP-PD**

**Hampton Leonard - Lead, production scale analyses in UKB (and everywhere else)**

Cornelis Blauwendraat - Data management for literally everything

Faraz Faghri - Co-lead, GenoML ecosystem development

Biowulf and Terra contributors - Resources to get all these analyses done!