Democratizing & Accelerating AI Through Automated Machine Learning

global.ainights.com

AI Platform Team
#AIGlobalNight

Agenda

- Welcome Video
- Why Automated Machine Learning
- Automated ML Capabilities
- How to Get Started

Michał Górnik

- Team Leader Data Scientist at Objectivity, previously working at Tooploox, KRUK, Eurobank
- Lecturer – conducting classes on statistics, data analysis, machine learning, predictive analytics at Wrocław University of Economics
  - Post-graduate Data Science studies
  - Data Science Summer School
  - Extramural studies

Welcome Video

Machine Learning on Azure

Domain Specific Pretrained Models
To reduce time to market

- Vision
- Speech
- Language
- Search

Familiar Data Science Tools
To simplify model development

- PyCharm
- Jupyter
- Visual Studio Code
- Command line

Popular Frameworks
To build machine learning and deep learning solutions

- PyTorch
- TensorFlow
- Scikit-Learn
- ONNX

Productive Services
To empower data science and development teams

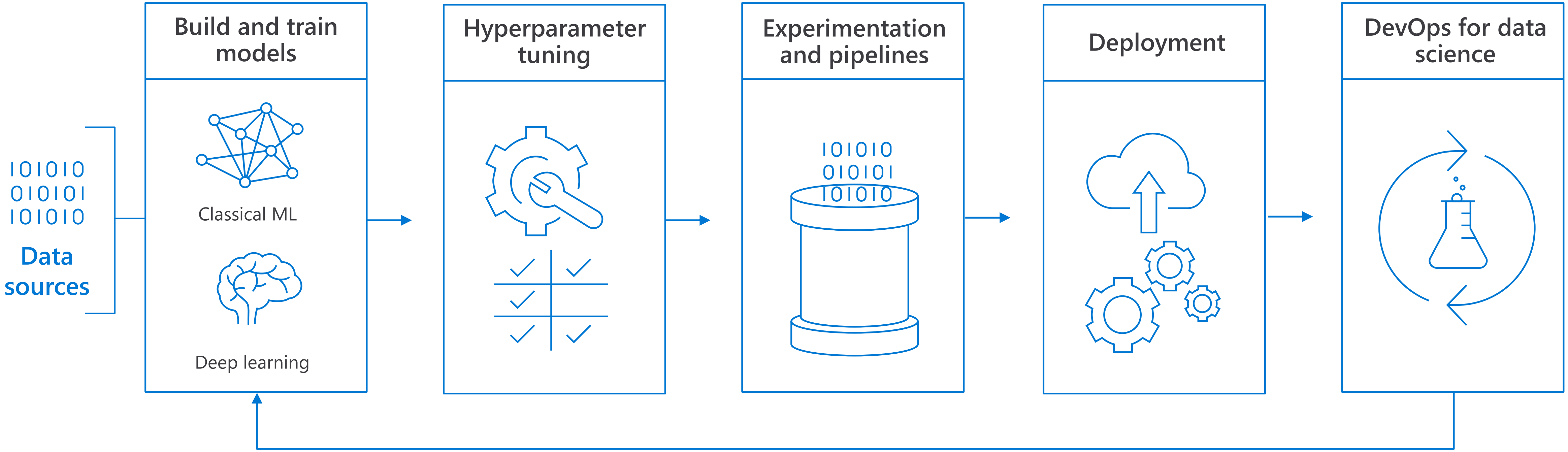
- Azure Databricks
- Azure Machine Learning
- Machine Learning VMs

Powerful Hardware
To accelerate deep learning

- CPU
- GPU
- FPGA

From the Intelligent Cloud to the Intelligent Edge

Building blocks for a Data Science Project

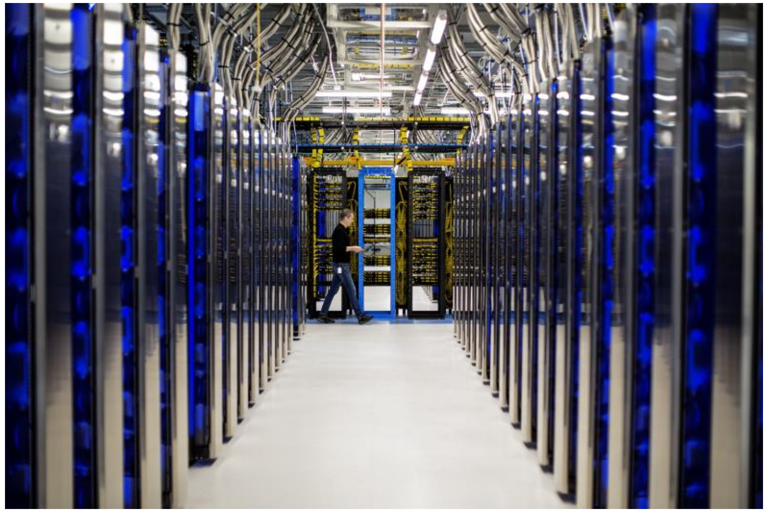


What is automated machine learning?

Automated machine learning (automated ML) automates feature engineering, algorithm and hyperparameter selection to find the best model for your data.

Automated ML Mission

Enable automated building of machine learning with the goal of accelerating, democratizing and scaling AI



Democratize AI

Enable Domain Experts & Developers to get rapidly build AI solutions

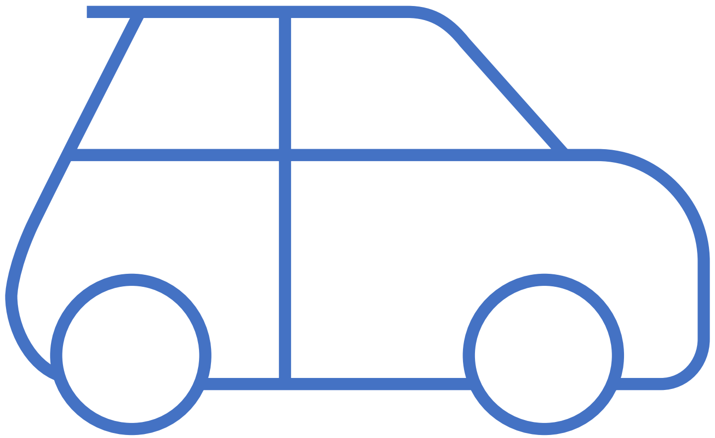
Accelerate AI

Improve Productivity for Data Scientists, Citizen Data Scientists, App Developers & Analysts

Scale AI

Build AI solutions at scale in an automated fashion

Machine Learning Problem Example



How much is this car worth?

Model Creation Is Typically Time-Consuming

Which features?

- **Mileage**
- Condition
- **Car brand**
- **Year of make**
- Regulations
- …

Which algorithm?

- **Gradient Boosted**
- Nearest Neighbors
- SVM
- Bayesian Regression
- LGBM
- …

Which parameters?

- Criterion / Parameter 1
- Loss / Parameter 2
- Min Samples Split / Parameter 3
- Min Samples Leaf / Parameter 4
- Others / …

30%

Model

Model Creation Is Typically Time-Consuming

Which features?

- **Mileage**
- **Condition**
- **Car brand**
- **Year of make**
- Regulations
- ...

Which algorithm?

- **Gradient Boosted**
- **Nearest Neighbors**
- SVM
- Bayesian Regression
- LGBM
- ...

Which parameters?

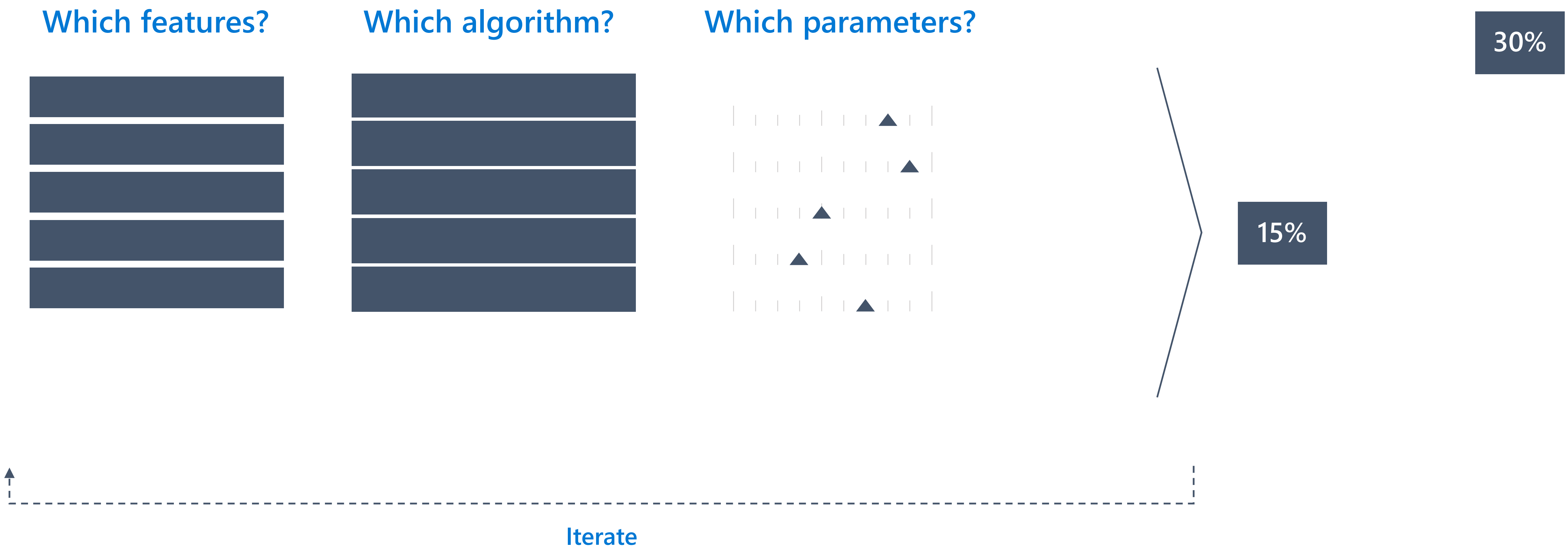
- N Neighbors
- Weights
- Min Samples Split
- Min Samples Leaf
- Others

30%

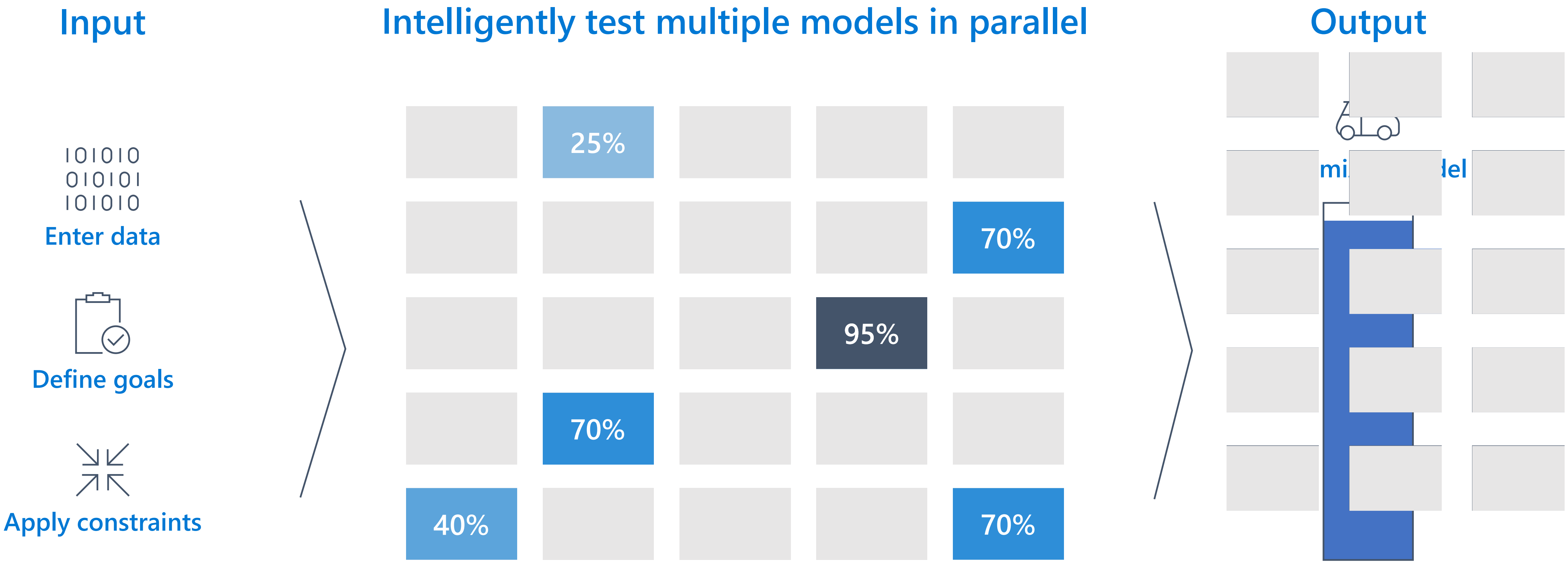
Model

Iterate

Model Creation Is Typically Time-Consuming

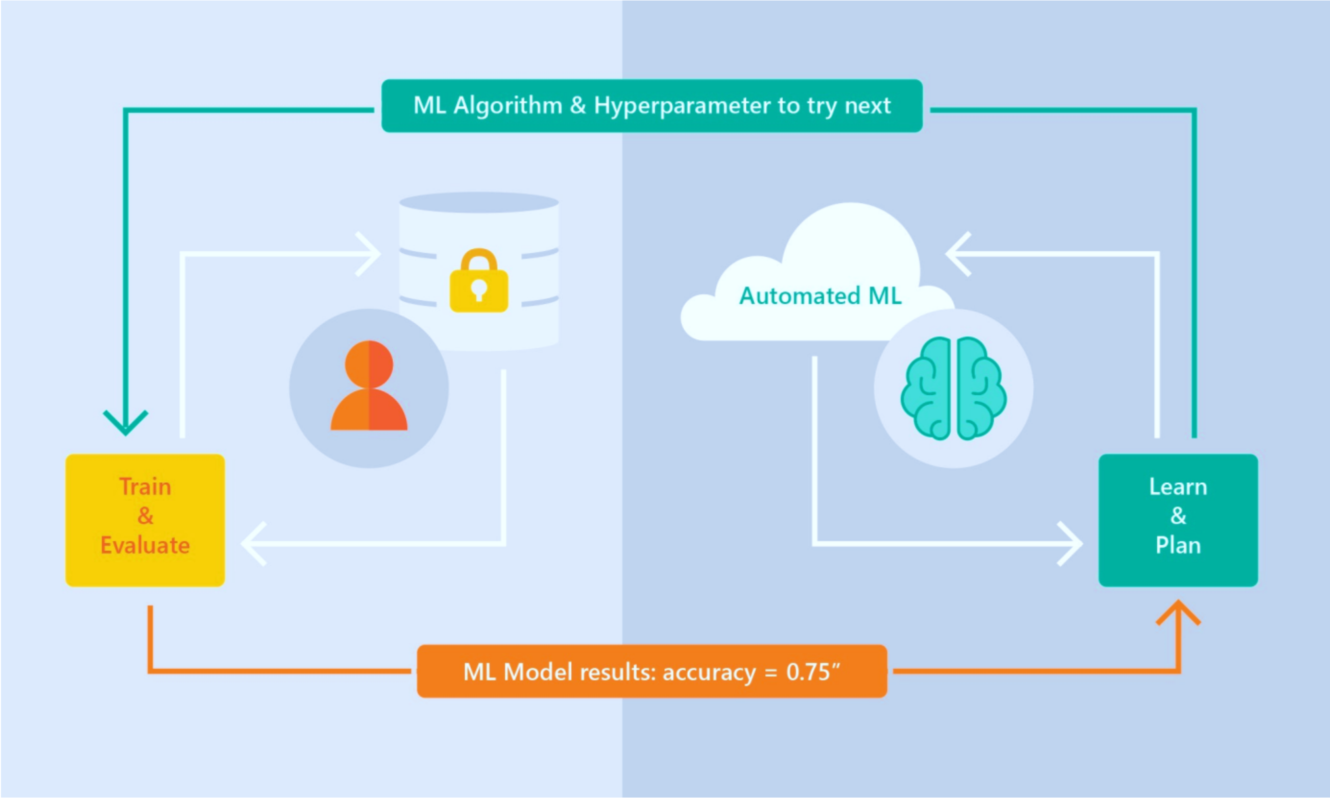


Automated ML Accelerates Model Development



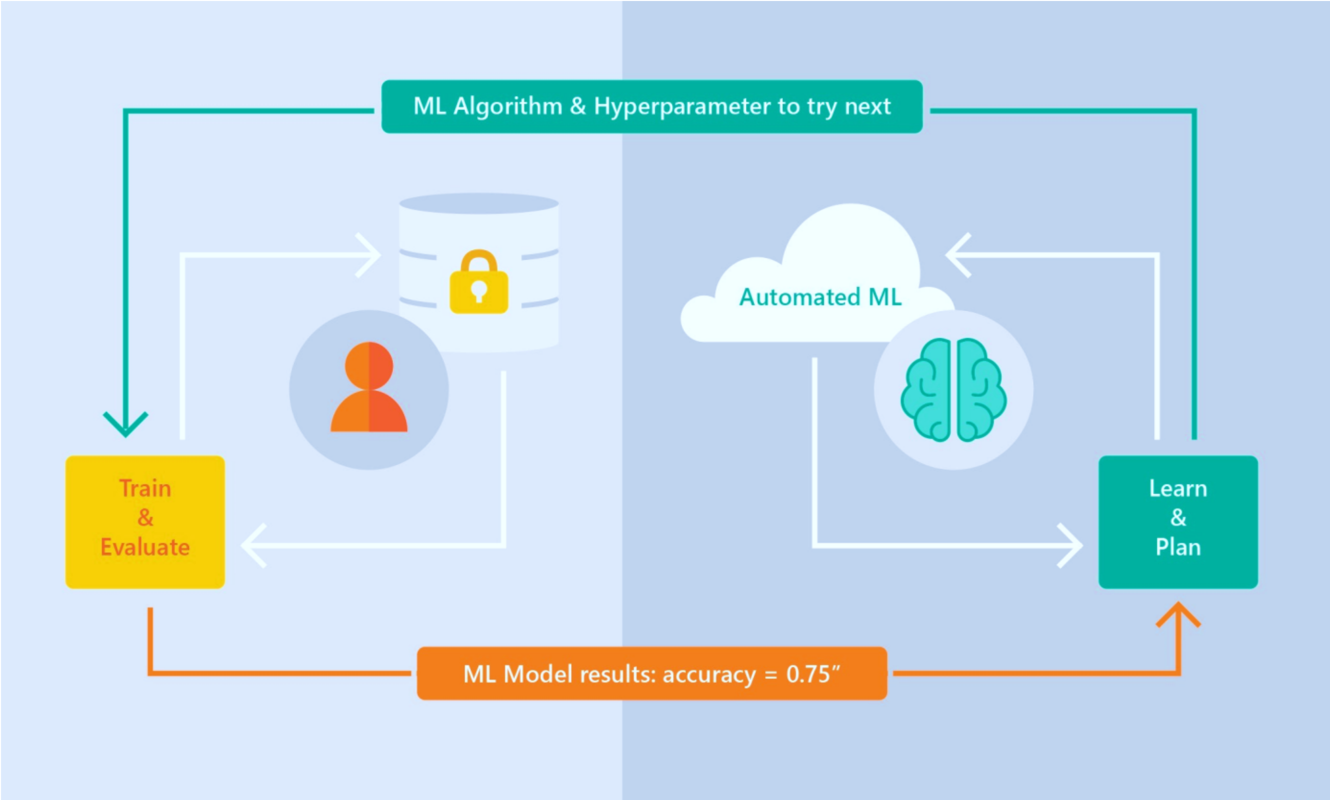
Automated ML Capabilities

- Based on Microsoft Research
- Brain trained with several million experiments
- Collaborative filtering and Bayesian optimization
- Privacy preserving: No need to "see" the data



Automated ML Capabilities

- ML Scenarios: Classification & Regression, Forecasting
- Languages: Python SDK for deployment and hosting for inference – Jupyter notebooks
- Training Compute: Local Machine, AML Compute, Data Science Virtual Machine (DSVM), Azure Databricks*
- Transparency: View run history, model metrics, explainability*
- Scale: Faster model training using multiple cores and parallel experiments



* In Preview

Automated ML

1. **Data Preprocessing**

Automated ML currently supports automated data cleaning

2. **Feature Engineering**

Most time-consuming part when done manually can now be done within minutes.

3. **Algorithm Selection**

Testing many different algorithms at once.

4. **Hyper-parameter Tuning**

Hyperparameter tuning what to include what to leave out

5. **Model Recommendation**

Having an overview of the best performing models based on accuracy & speed.

6. **Interpretability & Explaining**

Being able to explain what created an outcome and what features had the most significant impact

Guardrails

- Class imbalance
- Train-Test split, CV, rolling CV
- Missing value imputation
- Detect high cardinality features
- Detect leaky features
- Detect overfitting
- Model Interpretability / Feature Importance

Customer feedback

With one line of code, it runs through different algorithms within the prediction family and the different parameter (or variable) combos that previously were manually tested by the scientists. The power of the cloud comes in here. The results are comparable to what the data scientists produced.

Manish Naik, BP, Digital Innovation

Auto ML's execution of different models was an impressive that enabled data scientists to work iteratively on machine learning experiments to increase auction sales by 10% and optimize the time auction cars are kept in the showroom to less than 30 days.

Farika Maharani, PT. Serasi Autoraya, Data Platform Supervisor

The CBRE AI and Data Engineering Team have successfully deployed a complete Azure Machine Learning model to their new API gateway leveraging the Azure AutoML solution in Azure Databricks. The API Gateway plus the model deployment goes into production this March.

Francis Dogbey, Microsoft CSA

In evaluating Azure Automated ML we discovered real potential in shortening the time to market for producing predictive models. The availability of the Automated ML UI also holds the promise of opening the ML space to non data science trained resources which in turn allows the democratizing of the predictive work without the pain of hiring expensive/ hard to retain staff.

Bogdan Rosca, Senior Director, Principal Information Architect

We see advantages moving over to Azure AutoML because we think we will be able to increase our speed to create models significantly and do more with less in terms of labor hours.

Dan Metzendorf, Data Science Manager, The Sherwin Williams Company

AutoML resulted in a significant improvement in model performance (1) Consistently produced better models than other automated ml libraries (TPOT) (2) Also outperformed hand-tuned models. AutoML explored a solution space larger than what was plausible to do manually

David Robinson, Devon Energy Data Scientist

The reason we see the sharp uplift in sales is the customers are getting content that really connects with them, and they're getting offers for things that are truly relevant and relevant at that moment in time.... Microsoft—they are really wanting to be our partners and were really going to help us on this journey, which was very differentiating

Daniel Humble, Chief Data and Analytics Officer, Walgreens Boots Alliance

If I have 200 models to train—I can just do this all at once. It can be farmed out to a huge computer cluster, and it can be done in minutes so I'm not waiting for days or setting experiments to run over the weekend anymore.

Dean Riddlesden, Senior Data Scientist, Walgreens Boots Alliance

With automated machine learning in Azure Machine Learning service, we can focus our testing on the most accurate models and avoid testing a large range of less valuable models, because it retains only the ones we want. That saves months of time for us.

Matthieu Boujonnier: Analytics Application Architect and Data Scientist, Schneider Electric

Models evolve over time. And we use automated machine learning to speed that process, from four months for our first-generation models to a day for our newest models.

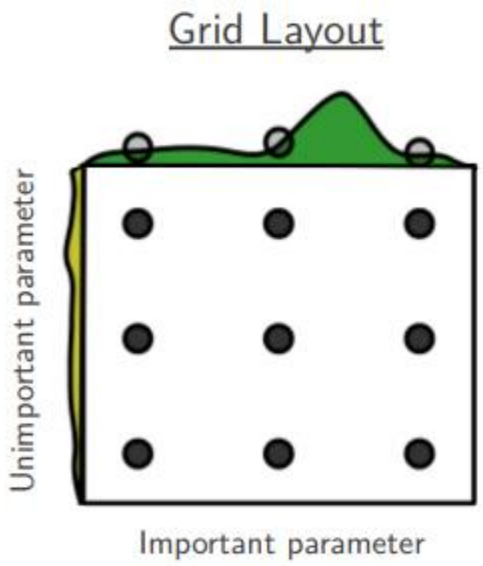
Loryne Bissuel-Beauvais: Data Scientist, Schneider Electric

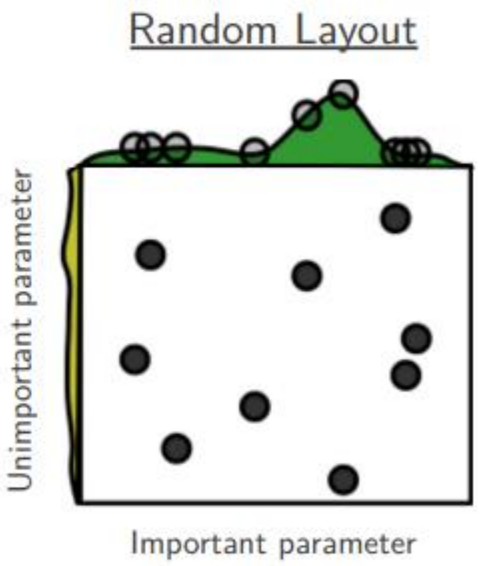
We tried AutoML for aspect ratio model and pleased to see AutoML produced much better model than our baseline. We need to build almost 50 models and are looking forward to the productivity boost we will get by not hand tuning each one of them!

Saurabh Naik, Sr. Software Engineer

Teams uses machine learning to analyze, gain insights and improve the quality of calls. We use AutoML to significantly scale up the application of ML solutions by semi-automating model train tasks

How does it work?






```
params_grid = {
    'n_estimators': [10, 20, 30, 50],
    'max_features': [1,2,3,4],
    'max_depth': range(1,7,1)
}
estimator = RandomForestClassifier()
gs = modsel.GridSearchCV(estimator, params_grid)
gs.fit(X_train, y_train)
```

```
params_grid = {
    'n_estimators': stats.randint(low=5, high=400),
    'max_features': stats.randint(low=1, high=40),
    'max_depth': range(1,7,1),
    'criterion': ['gini', 'entropy']
}
estimator = RandomForestClassifier()
gs = modsel.RandomizedSearchCV(estimator, params_grid)
gs.fit(X_train, y_train)
```

http://jmlr.csail.mit.edu/papers/volume13/bergstra12a/bergstra12a.pdf

What is a Machine Learnin pipeline?

$$\underset{\mathcal{M}, \mathcal{P}, \theta_m, \theta_p}{\arg\min} \; \mathscr{L}(\mathcal{M}(\mathcal{P}(\mathbf{x}; \boldsymbol{\theta}_p); \boldsymbol{\theta}_m), \mathbf{y}),$$

- ML pipeline is an unique combination of:
  - Preprocessors - MinMaxScaler(), OneHotEncoder()…
  - Models - ElasticNet, RandomForest, SVM …
  - Hyperparametrs
    - For KNN – k, weights, metric
    - For RandomForest – n_estimators, max_depth, criterion

https://arxiv.org/pdf/1705.05355.pdf

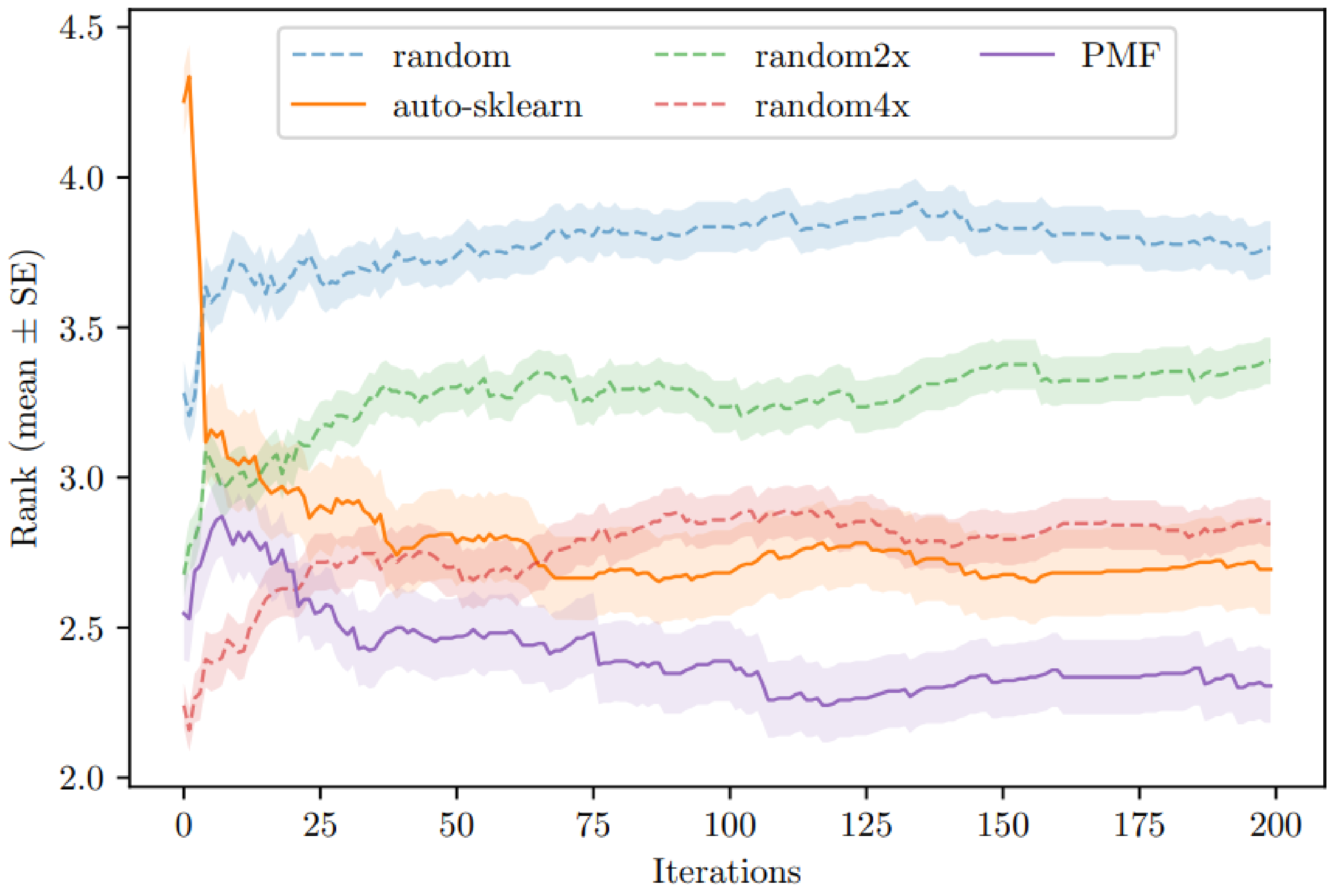
How does it work?

- Selecting models is performed in an intelligent way
- Under the hood the Bayesian Optimization with the recommender system is used
- For trainign auto-ml MS Researchers used:
    - 500 datasets
    - 42000 different ML pipelines

https://www.youtube.com/watch?v=jtRPxRnOXnk&t=298s

https://arxiv.org/pdf/1705.05355.pdf

Performance



What’s new?

Latest announcements ([Blog post with all the announcements])

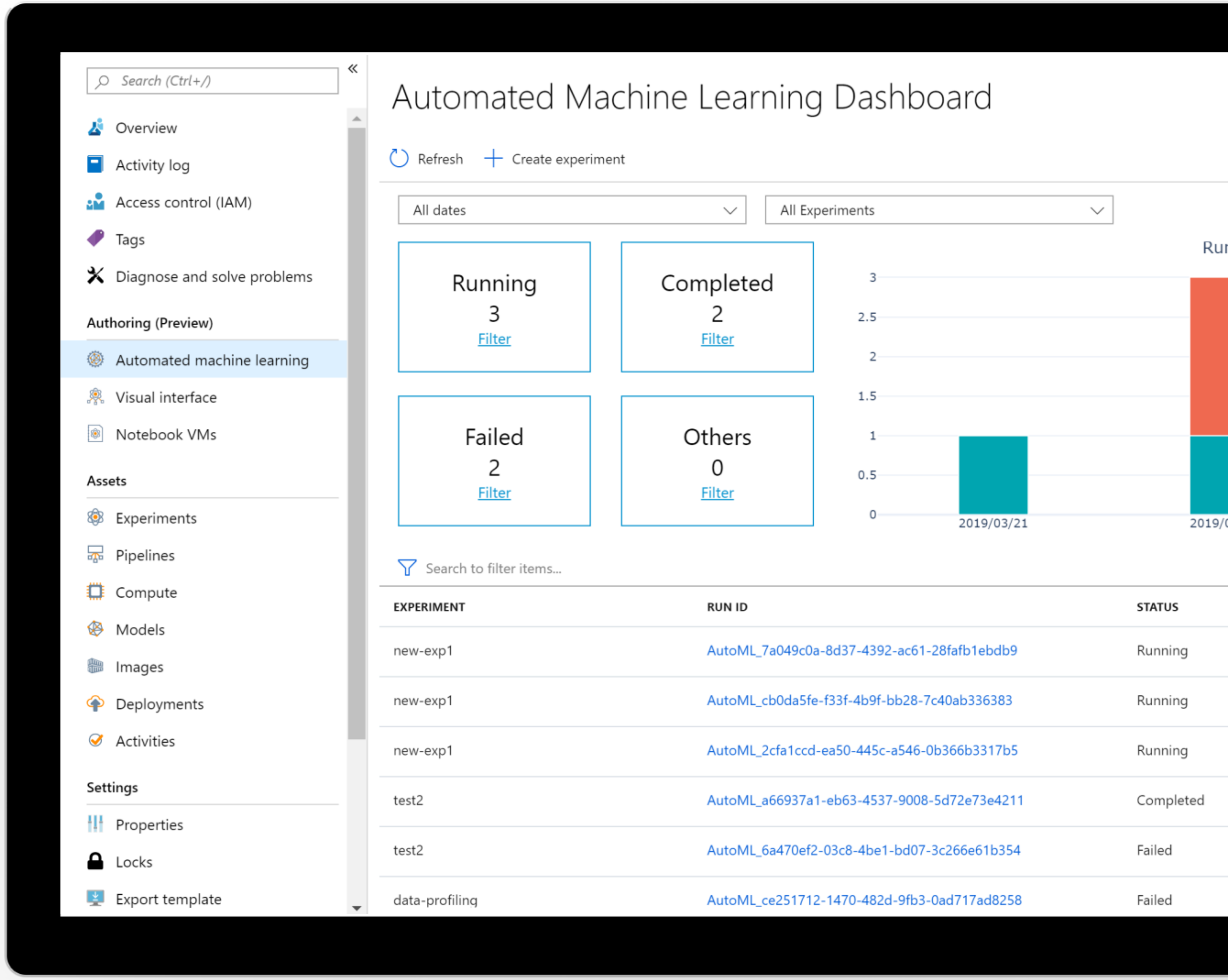
Automated ML UI in Azure portal (Preview)

- End-to-end no-code experience for non-data scientists to train ML models
- Classification, Regression, Forecasting
- Deploy models easily and quickly
- Advanced settings for power users to tune the training job

  [Blog post]

Coming up next

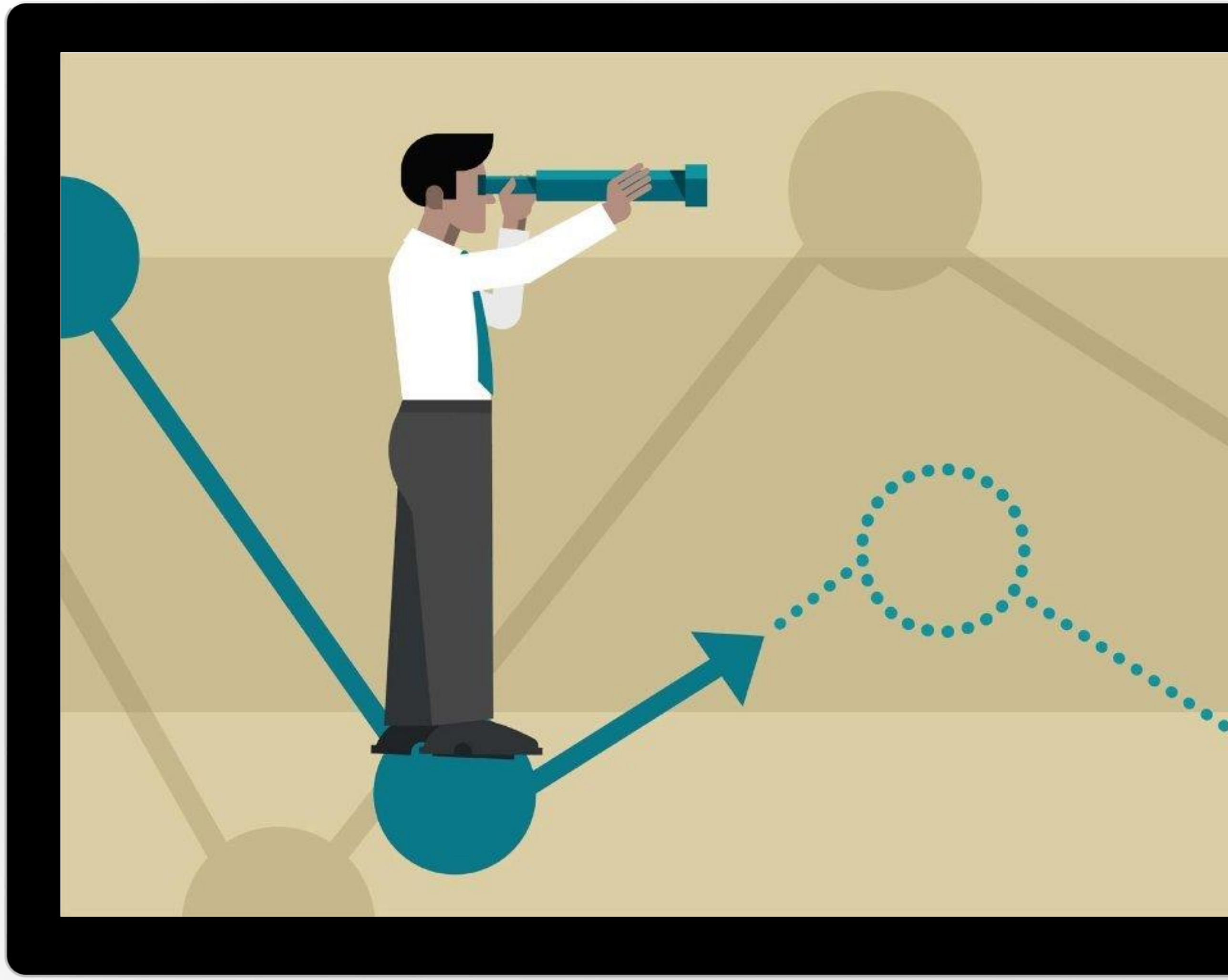
- Model explainability
- Additional data sources (with Datasets)
- Re-run experiments



Latest announcements (Blog post with all the announcements)

Time Series Forecasting Generally Available

- Rolling cross validation splits for time series data
- Configurable lags
- Window aggregation
- Holiday featurizer



Latest announcements (Blog post with all the announcements)

Feature engineering updates

- Additional data guardrails and synthetic features
- Added XGBoost algorithm
- Improved transparency retrieving the engineered features

Coming up next

- Improved feature sweeping, text featurization
- Transparency: Get auto-featurized data

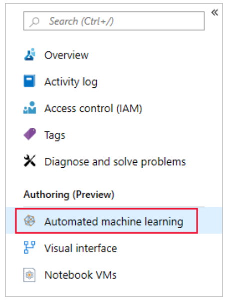
Preprocessing steps	Description
Drop high cardinality or no variance features	Drop these from training and validation sets, including features with all values missing, same value across all rows or with extremely high cardinality (for example, hashes, IDs, or GUIDs).
Impute missing values	For numerical features, impute with average of values in the column. For categorical features, impute with most frequent value.
Generate additional features	For DateTime features: Year, Month, Day, Day of week, Day of year, Quarter, Week of the year, Hour, Minute, Second. For Text features: Term frequency based on unigrams, bi-grams, and tri-character-grams.
Transform and encode	Numeric features with few unique values are transformed into categorical features. One-hot encoding is performed for low cardinality categorical; for high cardinality, one-hot-hash encoding.
Word embeddings	Text featurizer that converts vectors of text tokens into sentence vectors using a pre-trained model. Each word's embedding vector in a document is aggregated together to produce a document feature vector.
Target encodings	For categorical features, maps each category with averaged target value for regression problems, and to the class probability for each class for classification problems. Frequency-based weighting and k-fold cross validation is applied to reduce over fitting of the mapping and noise caused by sparse data categories.
Text target encoding	For text input, a stacked linear model with bag-of-words is used to generate the probability of each class.
Weight of Evidence (WoE)	Calculates WoE as a measure of correlation of categorical columns to the target column. It is calculated as the log of the ratio of in-class vs out-of-class probabilities. This step outputs one numerical feature column per class and removes the need to explicitly impute missing values and outlier treatment.
Cluster Distance	Trains a k-means clustering model on all numerical columns. Outputs k new features, one new numerical feature per cluster, containing the distance of each sample to the centroid of each cluster.

Workshop

Prerequisites
1. Azure Subscription
2. Azure ML workspace

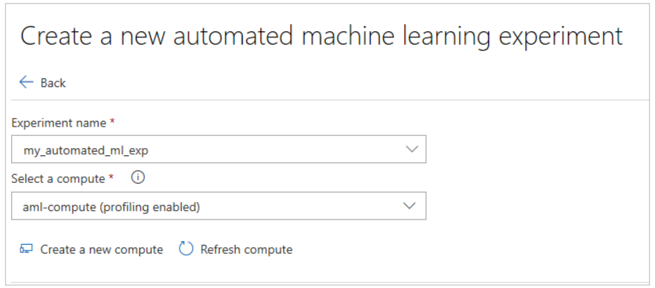
Automated ML using Azure Portal UI: Energy demand Forecasting

Follow instructions: https://docs.microsoft.com/en-us/azure/machine-learning/service/how-to-create-portal-experiments

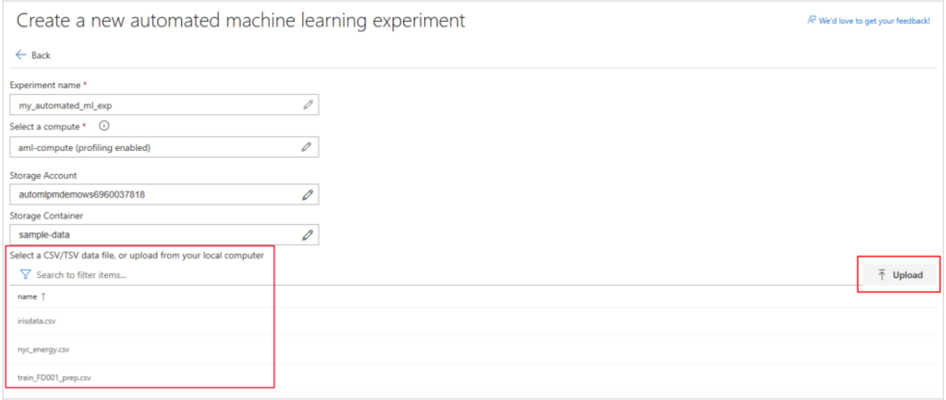


Navigate to the left pane of your workspace. Select Automated Machine Learning under the Authoring (Preview) section.

Enter your experiment name, then select a compute from the list of your existing computes or create a new compute



Select a data file from your storage container, or upload a file from your local computer to the container



Preview data and keep all columns selected for training

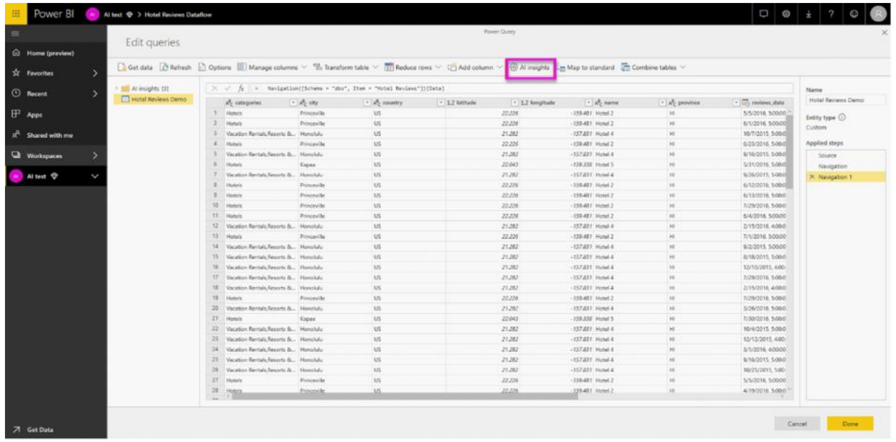
Select the training job type: **forecasting**
Select target column: **demand**
Select time column: **timeStamp**
Select forecast horizon: 168 (forecast a week ahead on an hourly basis)

Open "Advanced settings", set **training job time** to **10** minutes (for the sake of the workshop)

Hit Start

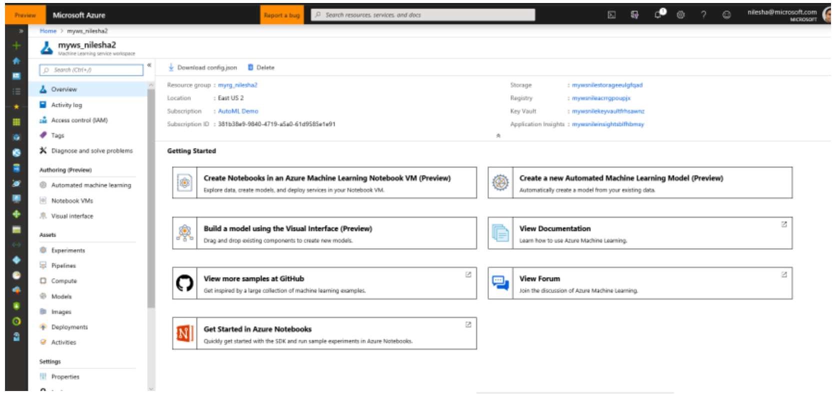
Once the run is completed, click **deploy the best model,** then **register model.**
Follow the instruction to deploy, but use the provided **scoring script** and **conda file** to enable consumption from Power BI

Once deployed, follow instructions to **consume from Power BI**

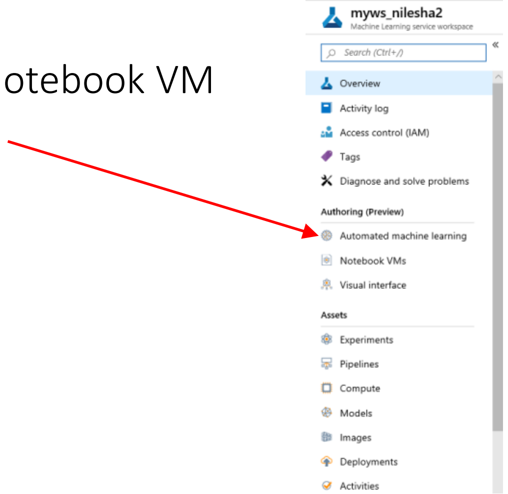


Automated ML using Notebook VM : Energy Demand

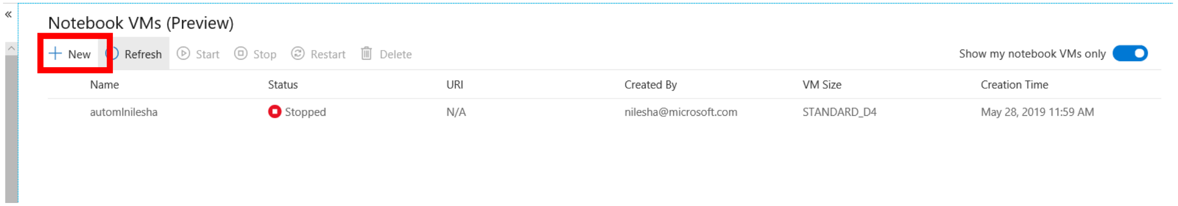
1. Login into Azure Portal : https://ms.portal.azure.com/#home
2. Open your Machine Learning Service Workspace.



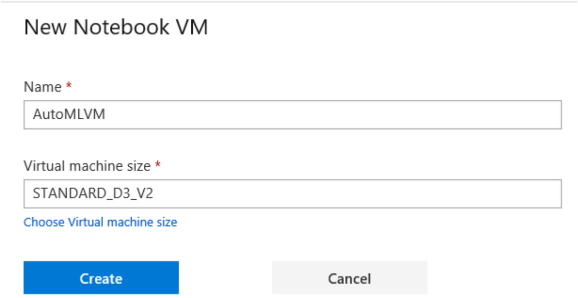
3. Click on Notebook VM



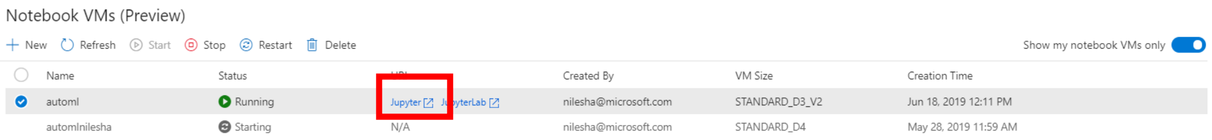
4. Click “New”. If you have an existing VM go to step



5. Name your VM and select machine size and click on create



6. After VM has started (~10mins) click on Jupyter link



7. Click on root folder > Samples-x.x.xx > how-to-use-azureml > automated-machine-learning > forecasting-energy-demand > /auto-ml-forecasting-energy-demand.ipynb

8. Follow instructions in notebook executing each cell.

Resources

http://aka.ms/amlfree

Learn more : https://aka.ms/automatedmldocs
Notebook Samples : https://aka.ms/automatedmlsamples
Blog Post : https://aka.ms/AutomatedML
Product Feedback : AskAutomatedML@microsoft.com

Microsoft

Thank You!

AI Platform Team

Microsoft