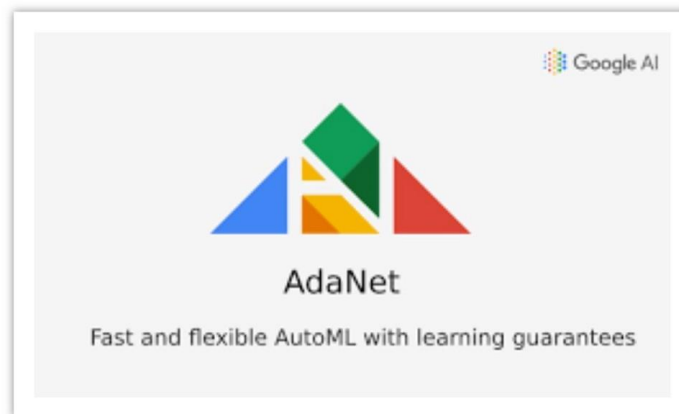


AutoML

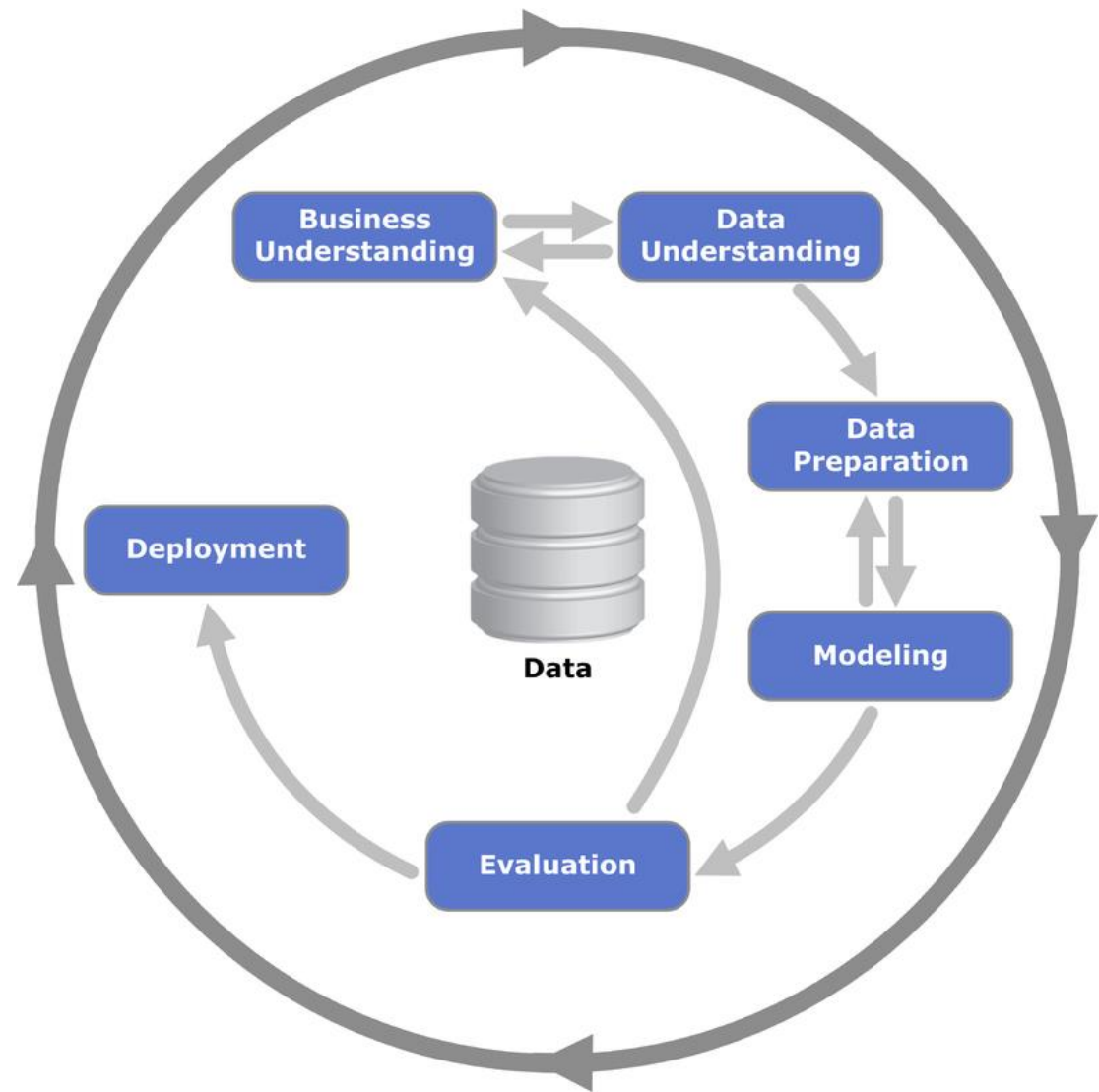
**Auto-
Sklearn**



TransmogrifAI



AutoML — это метод, который автоматизирует процесс применения методов машинного обучения к данным. Как правило, специалист по обработке данных тратит большую часть своего времени на предварительную обработку, инженерию признаков, выбор и настройку моделей, а затем оценку результатов. AutoML может автоматизировать эти задачи, предоставляя базовый результат, а также может обеспечить высокое качество моделей при решении определенных задач прогнозирования (как правило это регрессия, бинарная и мультиклассовая классификация) и дать понимание того, с какими моделями можно продолжить исследование.



CRISP-DM (*Cross-Industry Standard Process for Data Mining*) — наиболее распространённая методология по [исследованию данных](#).

Легко создавайте высокоточные модели машинного обучения

Предоставьте профессионалам по обработке и анализу данных (Data Scientist) и специалистам из других областей возможность быстро создавать модели машинного обучения. Автоматизируйте времязатратные и повторяющиеся задачи разработки моделей с помощью результатов революционного исследования и сократите время выхода на рынок.

Feature Selection

| 1 | 1 | 1 | 1 |
|----|----|----|----|
| 2 | 2 | 2 | 2 |
| 3 | 3 | 3 | 3 |
| 4 | 4 | 4 | 4 |
| 5 | 5 | 5 | 5 |
| 6 | 6 | 6 | 6 |
| 7 | 7 | 7 | 7 |
| 8 | 8 | 8 | 8 |
| 9 | 9 | 9 | 9 |
| 10 | 10 | 10 | 10 |

Основные цели и задачи AUTOML

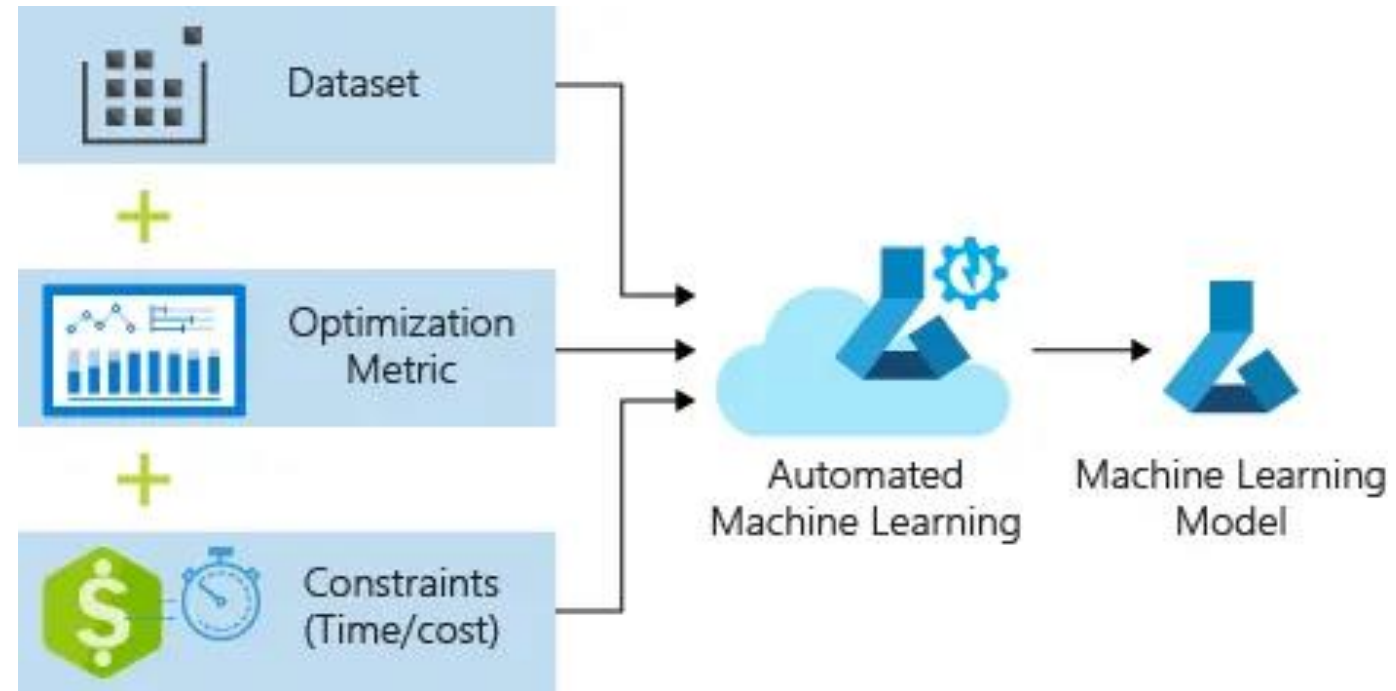
1. Реализуйте автоматическое создание и развертывание моделей прогнозирования
2. Оперативно создавайте точные модели, настроенные в соответствии с данными и уточненные с помощью широкого набора алгоритмов и гиперпараметров
3. Повысьте производительность благодаря удобному способу анализа данных и интеллектуальному конструированию признаков
4. Создавайте решения ML для ответственного использования с интерпретируемостью моделей и настраивайте модели для повышения точности
5. Точно прогнозируйте будущие бизнес-результаты с использованием популярных моделей временных рядов и глубокого обучения

Основные цели и возможности

1. Удобное создание моделей

Ускорьте создание моделей с помощью AUTOML

Быстро настраивайте модели и применяйте параметры управления к итерациям, порогам, проверкам, заблокированным алгоритмам и другим критериям экспериментов. Для обработки больших наборов данных и получения более точных оценок моделей используйте встроенные возможности для распространенных задач машинного обучения, таких как классификация, регрессия и прогнозирование временных рядов.

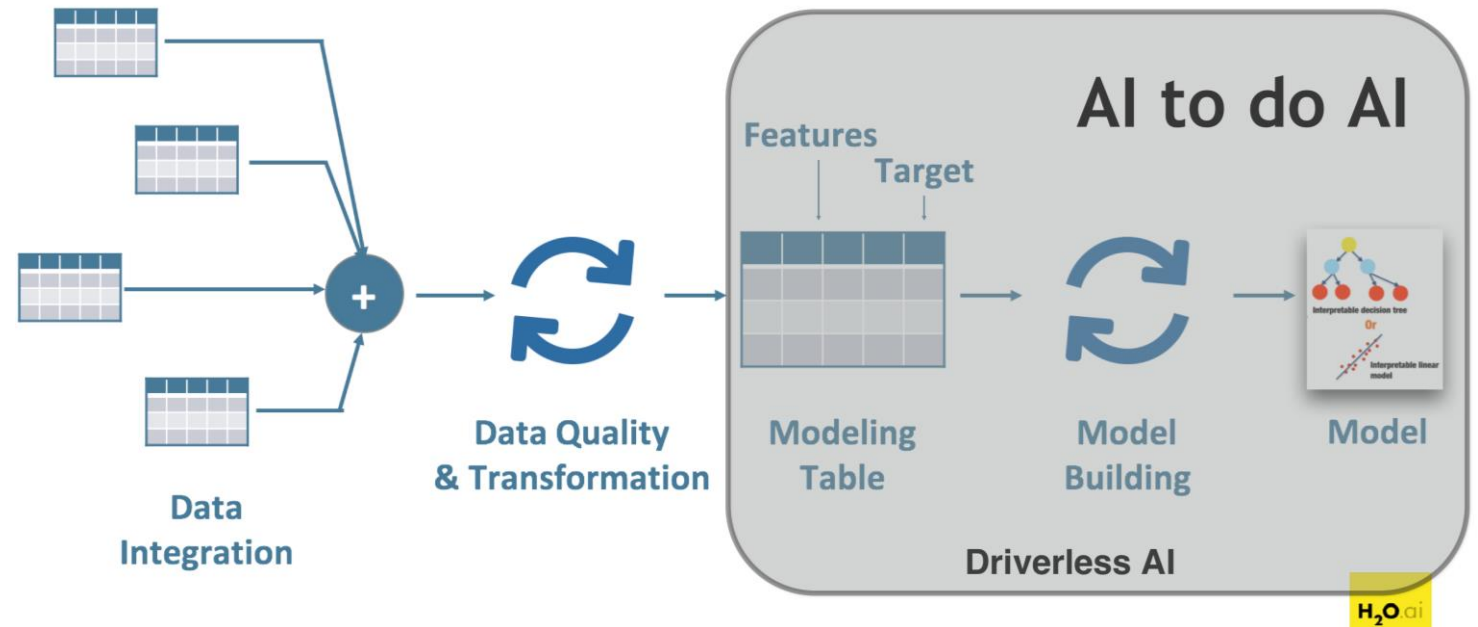


Основные цели и возможности

2. Управление созданием моделей

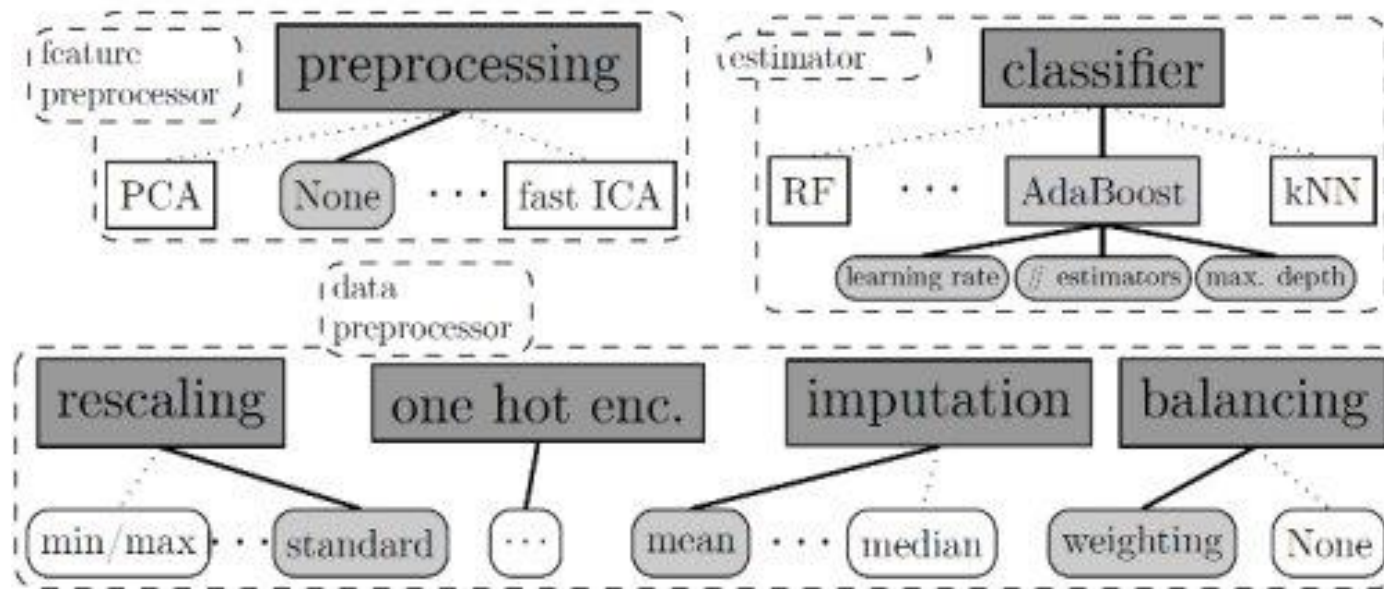
В процессе автоматизированного машинного обучения выполняется интеллектуальный выбор из широкого набора алгоритмов и гиперпараметров, что помогает создавать высокоточные модели.

Обнаруживайте распространенные ошибки и несоответствия в данных с помощью ограничений, а также получайте больше информации о рекомендуемых действиях и применяйте их автоматически. Используйте интеллектуальную остановку, чтобы сократить время на вычисление и выявление основной метрики для ускоренного получения результатов.



Основные цели и возможности

3. Повышение производительности благодаря автоматическому конструированию признаков

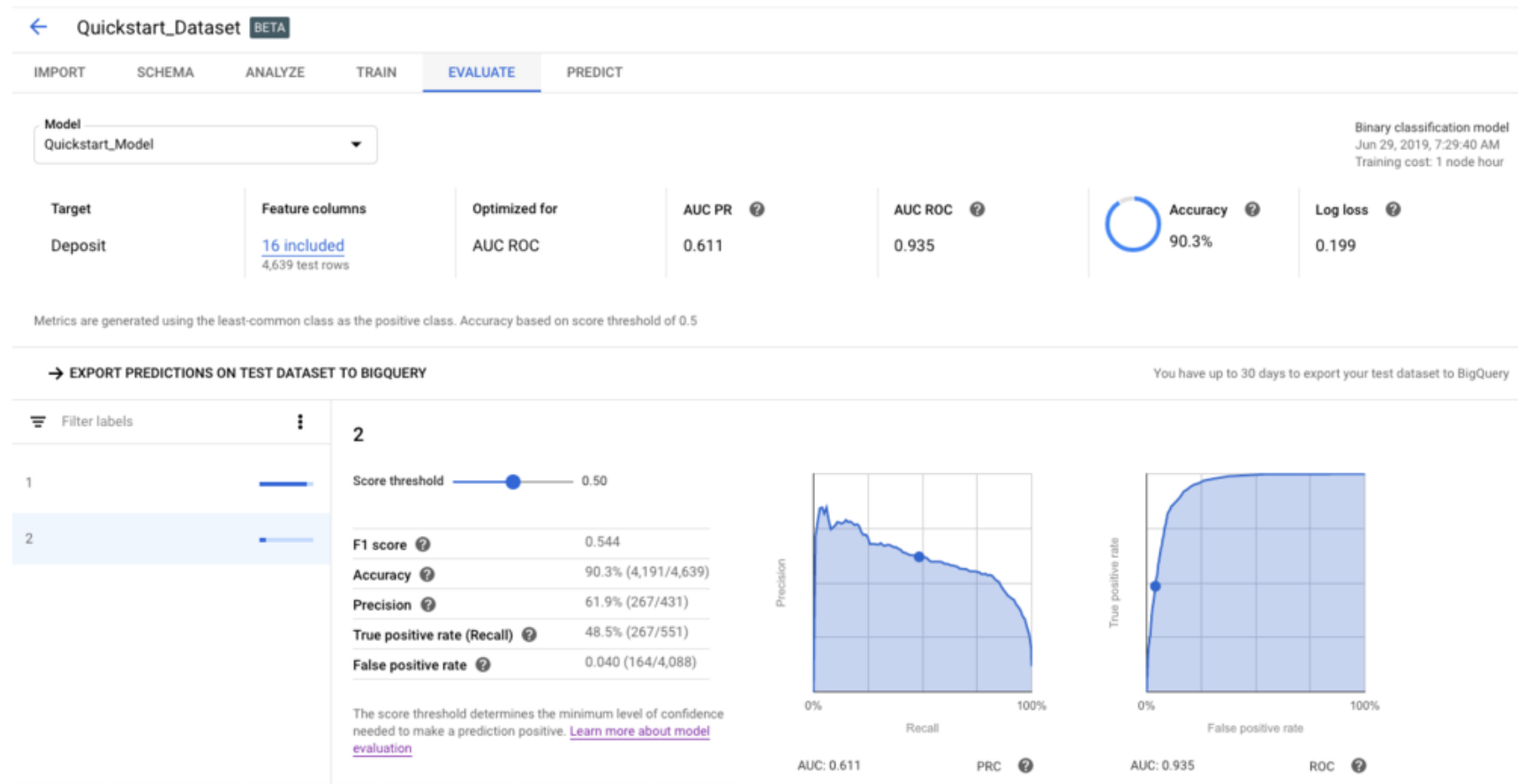


Для обработки больших наборов данных и получения более точных оценок моделей используйте встроенные возможности для распространенных задач машинного обучения, таких как классификация, регрессия и прогнозирование временных рядов, включая поддержку глубокой нейронной сети. Используйте автоматический выбор признаков и новые возможности их создания, чтобы экономить время и создавать высокоточные модели. Автоматизированное машинное обучение теперь включает архитектуру глубокого обучения BERT для конструирования признаков текстовых данных на 100 языках, которое доступно через пользовательский интерфейс, а также в записных книжках.

Основные цели и возможности

4. Более полное представление о моделях

Благодаря встроенной поддержке сводок запуска экспериментов и подробных визуализаций метрик вы можете больше узнать о моделях и сравнить их производительность. Интерпретируемость модели помогает оценивать модель в отношении необработанных и сконструированных признаков, а также позволяет получить ценные сведения о важности признаков.



Выявляйте закономерности, анализируйте возможные варианты и получайте более глубокое представление о моделях для поддержания прозрачности и доверия в бизнесе.

AutoGluon: AutoML для текста, изображений и табличных данных

Табличное предсказание:

прогнозирование значений в столбце
таблицы данных на основе
значений других столбцов



| Region | Country | Population | GDP | Unemployment | Life Expectancy | Healthcare | Education | Environment |
|---------------|-----------|---------------|------------------|--------------|-----------------|------------|-----------|-------------|
| North America | USA | 328,217,000 | \$21,450,000,000 | 4.7% | 78.4 | 1000 | 15.1 | 70.0 |
| Europe | Germany | 82,270,000 | \$3,850,000,000 | 3.5% | 81.2 | 1000 | 13.5 | 75.0 |
| Asia | China | 1,395,280,000 | \$14,580,000,000 | 5.3% | 77.3 | 1000 | 12.0 | 65.0 |
| South America | Brazil | 212,500,000 | \$1,740,000,000 | 11.5% | 75.3 | 1000 | 11.0 | 60.0 |
| Africa | Nigeria | 198,960,000 | \$51,000,000 | 23.5% | 53.5 | 1000 | 5.0 | 30.0 |
| Oceania | Australia | 23,262,000 | \$1,350,000,000 | 5.2% | 84.7 | 1000 | 14.0 | 70.0 |

Dog



Прогнозирование изображения:

распознать главный объект
на изображении

Dog



Dog



Cat



Обнаружение объектов:

обнаружение нескольких объектов
с их ограничивающими рамками
на изображении

Прогнозирование текста:

делайте прогнозы на
основе текстового содержимого



Подход AutoGluon-Tabular к AutoML

В то время как приложения машинного обучения в изображениях и видео привлекают все внимание, люди десятилетиями применяли статистические методы к табличным данным (например, строки и столбцы в электронной таблице или базе данных) либо для построения прогнозных моделей, либо для сбора сводной статистики. В эту категорию попадает большое количество задач науки о данных - например, прогнозирование продаж на основе данных о запасах и спросе, обнаружение мошенничества на основе данных транзакций и создание рекомендаций по продуктам на основе предпочтений пользователей.

AutoGluon-Tabular предоставляет вам доступ ко всем передовым методам, используемым экспертами в области данных, через удобный API и был разработан с учетом следующих ключевых принципов:

- 1.Простота: пользователи должны иметь возможность обучать модели классификации и регрессии и развертывать их с помощью нескольких строк кода.
- 2.Надежность: пользователи должны иметь возможность предоставлять необработанные данные без какой-либо разработки функций или манипулирования данными.
- 3.Predictable-time: пользователи должны иметь возможность указать временной бюджет и получить лучшую модель при этом временном ограничении.
- 4.Отказоустойчивость: пользователи должны иметь возможность возобновлять обучение в случае прерывания и иметь возможность проверять все промежуточные этапы.

Подход AutoGluon-Tabular к AutoML

Если вы уже являетесь опытным специалистом в области науки о данных и хотите знать, полезен ли для вас AutoGluon-Tabular, ответ - да. Даже для эксперта AutoGluon-Tabular может сэкономить время за счет автоматизации трудоемких ручных шагов - обработки пропущенных данных, ручного преобразования функций, разделения данных, выбора модели, выбора алгоритма, выбора и настройки гиперпараметров, объединения нескольких моделей и повторения процесса, когда произошли изменения в данных.

AutoGluon-Tabular также включает в себя новые методы объединения многослойных стеков, которые значительно повышают точность модели. Поскольку AutoGluon является полностью открытым исходным кодом, прозрачным и расширяемым, у вас есть полная видимость того, что он делает на каждом этапе процесса, и вы даже можете вносить свои собственные алгоритмы и использовать их с AutoGluon.

AutoGluon API

Пользователи AutoGluon-Tabular нужно только знать , как использовать три функции Python:

Dataset(), **fit()** и **predict()**

Магия функции `fit()`

Когда вы передаете свой набор данных в функцию `task.fit()`, она выполняет две вещи: предварительную обработку данных и обучение модели. Теперь давайте узнаем, что происходит за кадром.

Предварительная обработка данных

AutoGluon-Tabular сначала проверяет столбец меток и определяет, есть ли у вас проблема классификации (прогнозирование категорий) или проблема регрессии (прогнозирование непрерывных значений). Затем он инициализирует шаги предварительной обработки данных, которые преобразуют данные в форму, которая будет использоваться многими различными алгоритмами машинного обучения на `fit()` этапе.

Магия функции `fit ()`

Предварительная обработка данных

Текстовые столбцы преобразуются в числовые векторы характеристик n-грамм (непрерывная последовательность из n элементов или слов); дата и время преобразуются в подходящие числовые значения. Чтобы справиться с отсутствующими дискретными переменными, AutoGluon-Tabular создает дополнительную категорию Неизвестно, а не заменяет их (заменяя его константой, например, средним). В реальных наборах данных значения могут отсутствовать по разным причинам - например, из-за повреждения данных, сбоев датчиков и человеческих ошибок - и это не означает, что там не было ничего интересного. Отнесение его к категории "неизвестно" позволяет AutoGluon-Tabular обрабатывать ранее невидимые категории при создании прогнозов с новыми данными. На этапе подбора модели AutoGluon-Tabular также выполняет дополнительные шаги предварительной обработки данных, которые зависят от модели.

Магия функции `fit()`

Обучение модели

Когда вы вызываете функцию `fit()`, AutoGluon-Tabular обучит серию моделей машинного обучения на предварительно обработанных данных. Затем он объединяет несколько моделей, используя ансамблирование и стэкинг.

AutoGluon-Tabular обучает отдельные модели в специально выбранной последовательности. Сначала он обучает надежно работающие модели, такие как Random Forest, а затем постепенно обучает более затратные в вычислительном отношении, но менее надежные модели, такие как k-ближайшие соседи. Преимущество этого подхода заключается в том, что вы можете определить ограничение по времени для функции `fit()`, и она вернет лучшие модели, которые она может обучить с учетом ограничений по времени. AutoGluon-Tabular дает вам гибкость, чтобы решить, нужна ли вам максимальная точность без ограничений или лучшая точность при определенных затратах или временном бюджете.

Магия функции `fit ()`

Обучение модели

AutoGluon-Tabular в настоящее время поддерживает следующие алгоритмы и обучает их все, если не наложено ограничение по времени:

1. Random Forests
2. Extremely Randomized trees
3. k-nearest neighbors
4. LightGBM boosted trees
5. CatBoost boosted trees
6. AutoGluon-Tabular deep neural networks

Магия функции `fit ()`

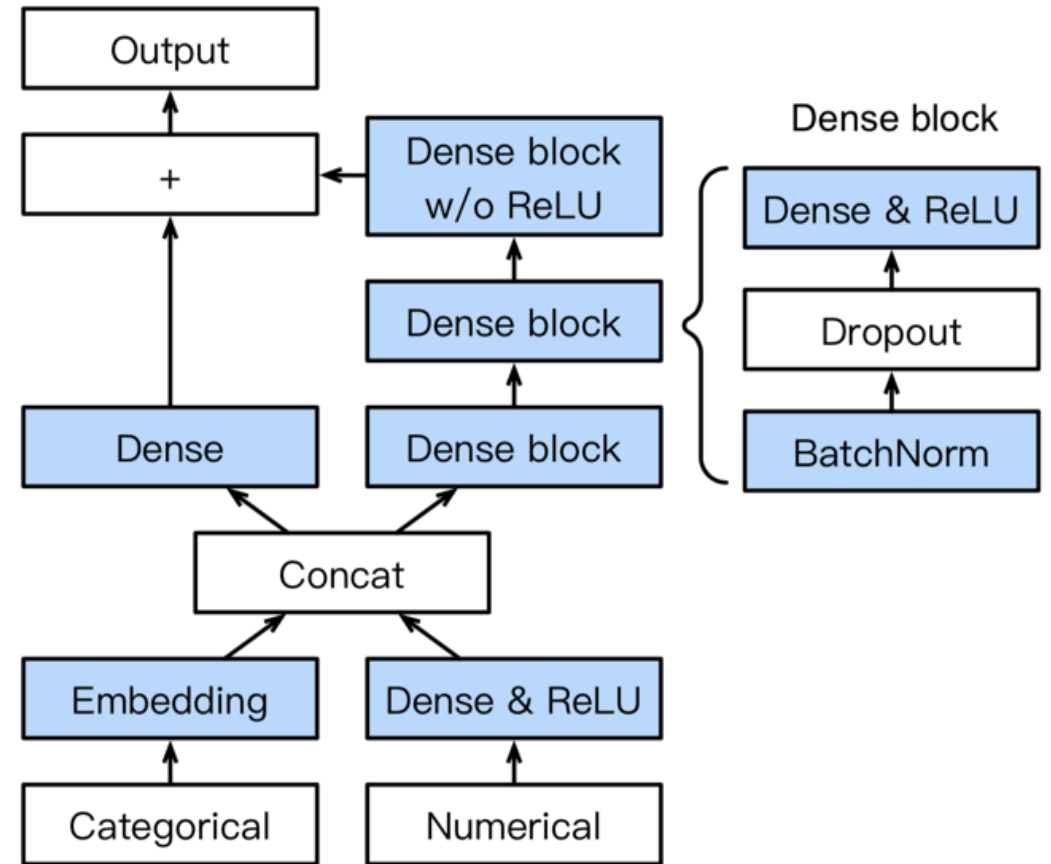
Новинка в архитектуре глубокой нейронной сети **AutoGluon-Tabular**

В сообществе специалистов по науке о данных распространено заблуждение, что подходы к глубокому обучению плохо работают с табличными данными. Для такого мышления есть причина: свертки были введены в нейронные сети из-за их свойств инвариантности перевода через распределенные весовые коэффициенты (все участки изображения будут обрабатываться одинаково). И это отлично работает для наборов данных, которые представляют собой одномерные сигналы, двухмерные или трехмерные изображения или видео, где каждая выборка сигнала или значение пикселя сами по себе имеют низкую предсказательную силу. Во многих приложениях с табличными наборами данных каждая функция уникальна и имеет более высокую предсказательную силу, чем отдельные пиксели изображения. В этих ситуациях архитектуры нейронных сетей с прямой связью или сверточные нейронные сети, как правило, работают хуже, чем модели, основанные на деревьях решений.

Магия функции fit ()

Новинка в архитектуре глубокой нейронной сети AutoGluon-Tabular

Для решения этих проблем AutoGluon-Tabular использует новую архитектуру нейронной сети, показанную на рисунке. Эмпирические исследования показывают, что тщательно спроектированные нейронные сети могут обеспечить значительное повышение точности, особенно при создании ансамбля с другими типами моделей.



В отличие от обычно используемых сетевых архитектур с прямой связью, AutoGluon-Tabular вводит слой embedding (внедрения) для каждой категориальной функции, где размер embedding выбирается пропорционально количеству уникальных категорий в функции. Преимущество уровня embedding состоит в том, что он вводит обучаемый компонент для каждой категориальной функции до того, как он будет использован последующими уровнями прямой связи. Вложения категориальных признаков затем объединяются с числовыми признаками в большой вектор, который одновременно подается в трехуровневую сеть с прямой связью, а также напрямую подключается к выходным прогнозам через линейное пропускное соединение, как остаточное семейство сетей.

Магия функции `fit ()`

Ансамбли и `multi-layer stacking`

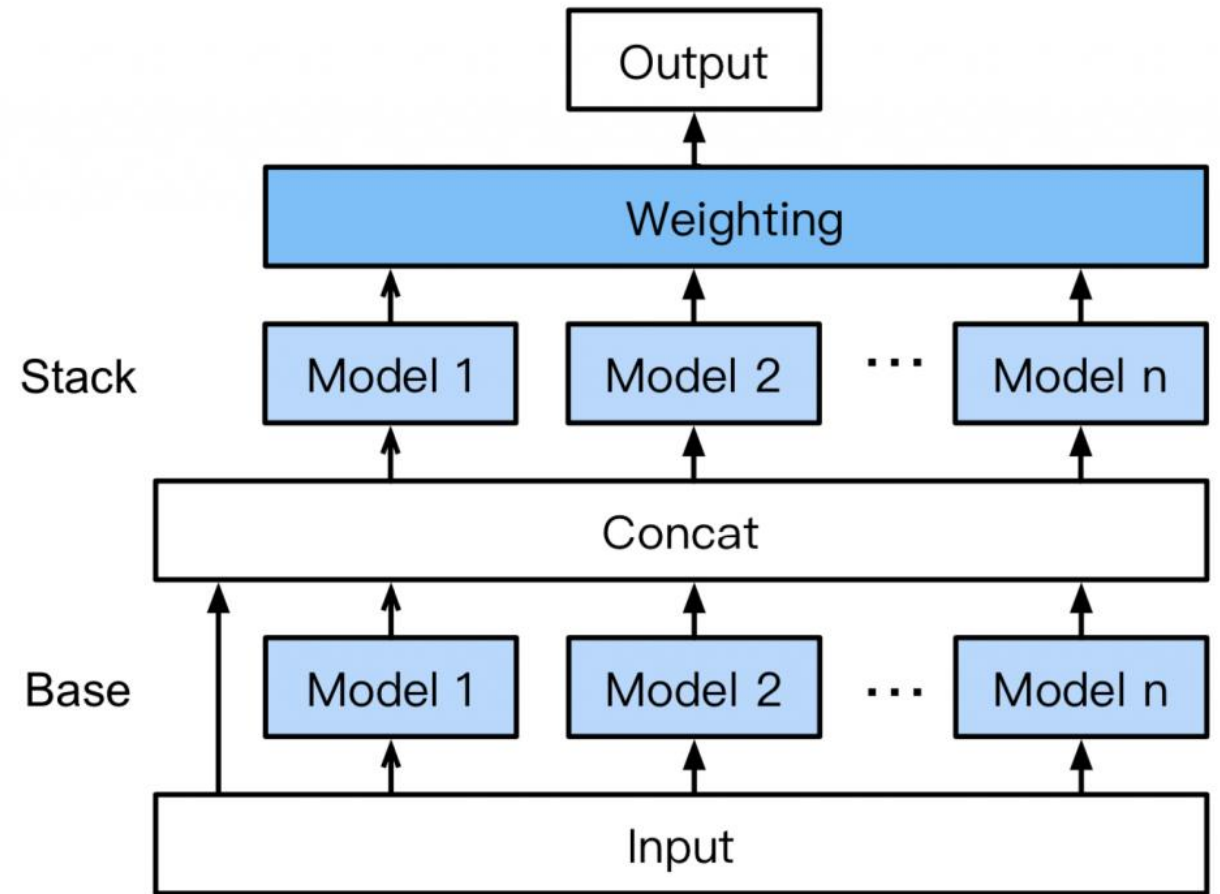
Идея объединения нескольких моделей для создания «ансамбля», который имеет более высокую точность прогнозов, чем каждый из его участников, не нова. Самые ранние реализации ансамблевых методов относятся к началу 1990-х годов, когда были изобретены подходы повышения (и алгоритма AdaBoost) и бэггинга (агрегация начальной загрузки). Эти методы создали ансамбли деревьев решений, которые являются слабыми учениками (не намного лучше, чем случайность) и нестабильными (чувствительными к изменениям в наборе данных).

Но когда объединяются многие деревья решений, они создают модели с высокой предсказательной способностью, устойчивые к чрезмерной подгонке. Эти ранние работы лежат в основе популярных пакетов машинного обучения, таких как LightGBM, CatBoost и RandomForest от scikit-learn, которые используются AutoGluon.

Магия функции fit ()

Ансамбли и multi-layer stacking

Если вам интересно, можете ли вы комбинировать выходные данные RandomForest, CatBoost, k-ближайших соседей и других для дальнейшего повышения точности модели, ответ - да, вы можете. Опытные практики машинного обучения занимаются этим уже много лет и умеют изобретать умные способы комбинировать несколько моделей. Проверьте соревнование на Kaggle с датасетом Otto Group (классификация продукта). Решение, занявшее первое место, включало 33 модели, результаты которых затем используются для обучения еще трех моделей (суммирование), за которыми следует средневзвешенное значение.



Магия функции `fit()`

Ансамбли и `multi-layer stacking`

С AutoGluon-Tabular вам не нужно иметь навыки `stacking and ensembling`. AutoGluon-Tabular автоматически сделает это за вас. AutoGluon-Tabular представляет новую форму `multi-layer stack ensemble`. Вот как это работает:

- Базовый уровень: индивидуальное обучение нескольких базовых моделей.
- Слой `Concat`: выходные данные первого слоя объединяются вместе с входными объектами.
- `Stacker layer`: несколько моделей укладчика обучаются на выходе слоя `concat`. Новинка, представленная AutoGluon-Tabular, заключается в том, что слой `Stacker` повторно использует те же самые модели в базовом слое, включая их гиперпараметры. Поскольку входные объекты объединяются с выходными данными базового слоя, `stacker models` также получают возможность просматривать входной набор данных.
- Слой взвешивания: реализуется подход к выбору ансамбля, в котором модели укладчика вводятся в новый ансамбль, так что точность проверки максимальна.

Чтобы каждая модель “видел” весь набор данных, AutoGluon-Tabular выполняет `k`-кратную перекрестную проверку. Для дальнейшего повышения точности прогнозов и уменьшения переобучения AutoGluon-Tabular будет повторять `k`-кратную перекрестную проверку `n` раз на `n` различных случайных разделах входных данных. Число `n` выбирается путем оценки того, сколько раундов может быть выполнено за указанные временные ограничения при вызове функции `fit()`.

AutoGluon: AutoML для текста, изображений и табличных данных

AutoGluon обеспечивает простой в использовании и легко расширяемый AutoML с акцентом на автоматическое стекинг моделей, глубокое обучение и реальные приложения, охватывающие текст, изображения и табличные данные. AutoGluon, предназначенный как для начинающих, так и для экспертов в области машинного обучения, позволяет:

- Быстро создавать прототипы решений глубокого обучения и классического машинного обучения для необработанных данных с помощью нескольких строк кода.
- Автоматически использовать самые современные методы без специальных знаний.
- Использовать автоматическую настройку гиперпараметров, выбор/объединение моделей, поиск архитектуры и обработку данных.
- Легко улучшать/настраивать свои индивидуальные модели и конвейеры данных или настраивайте AutoGluon для вашего случая использования

AutoGluon: AutoML для текста, изображений и табличных данных

Пример использования AutoGluon для обучения и развертывания высокопроизводительной модели в табличном наборе данных:

```
>>> from autogluon.tabular import TabularPredictor
```

```
>>> predictor =  
TabularPredictor(label=COLUMN_NAME).fit(train_data=TRAIN_DATA.csv)
```

```
>>> predictions = predictor.predict(TEST_DATA.csv)
```

AutoGluon так же легко может применяться для задач прогнозирования с изображениями или текстовыми данными

AutoGluon: AutoML для текста, изображений и табличных данных

Выберите свои предпочтения ниже и выполните соответствующие команды установки:

| | | |
|----------|---|---------------------------------------|
| OS: | <input checked="" type="button" value="LINUX"/> | <input type="button" value="MAC"/> |
| VERSION: | <input checked="" type="button" value="PIP"/> | <input type="button" value="SOURCE"/> |
| BACKEND: | <input checked="" type="button" value="CPU"/> | <input type="button" value="GPU"/> |

```
python3 -m pip install -U pip
```

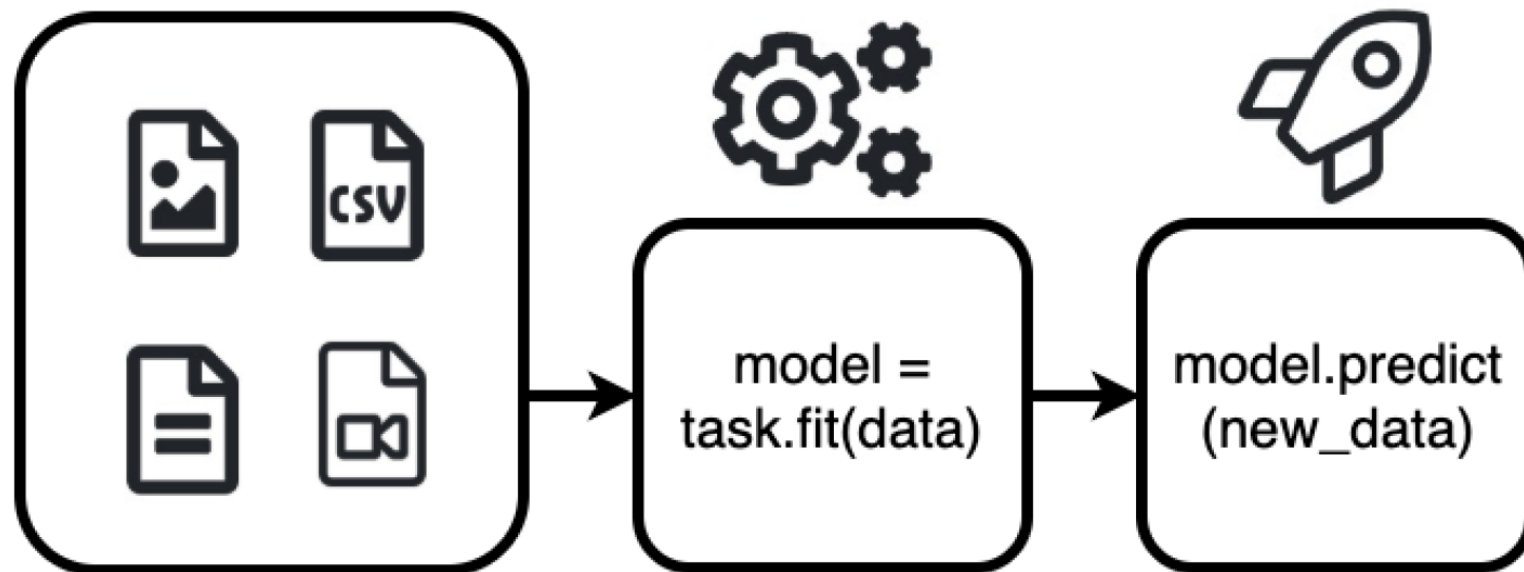
```
python3 -m pip install -U setuptools wheel
```

```
python3 -m pip install -U "mxnet<2.0.0"
```

```
python3 -m pip install autogluon
```

AutoGluon: AutoML для текста, изображений и табличных данных

AutoGluon состоит из модулей в [sub-modules](#) специализированных для данных в табличной форме, текста или изображения. Можно уменьшить количество требуемых зависимостей, установив только определенный подмодуль с помощью: `python3 -m pip install <submodule>`, где `<submodule>` может быть одним из следующих вариантов:



- **autogluon.tabular** - функциональность только для табличных данных (TabularPredictor)
 - доступные дополнительные зависимости: **lightgbm, catboost, xgboost, fastai**.
 - экспериментальная необязательная зависимость: **skech**. Это ускорит обучение и вывод на CPU моделей KNN в 25 раз.

AutoGluon: AutoML для текста, изображений и табличных данных

- `autogluon.vision` - только функционал для компьютерного зрения (ImagePredictor, ObjectDetector)
- `autogluon.text` - только функциональность для обработки естественного языка (TextPredictor)
- `autogluon.core` - только базовая функциональность (поисковик/планировщик), полезная для настройки гиперпараметров произвольного кода/моделей.
- `autogluon.features` - только функциональность для конвейеров генерации признаков/предварительной обработки признаков (в первую очередь связанных с табличными данными).
- `autogluon.extra` - различные дополнительные функции, такие как эффективный поиск нейронной архитектуры.
- `autogluon.mxnet` - прочие дополнительные функции для MXNet.

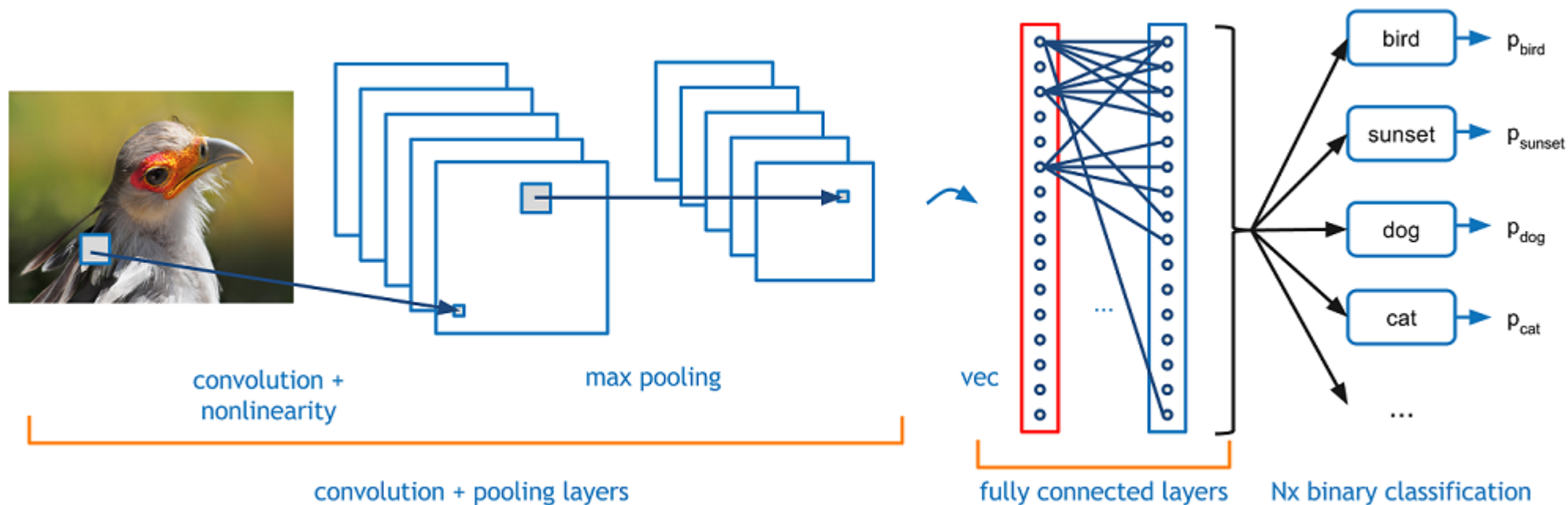
Табличное предсказание

Для стандартных наборов данных, представленных в виде таблиц (сохраненных в виде файла CSV, паркета из базы данных и т.д.), AutoGluon может создавать модели для прогнозирования значений в одном столбце на основе значений в других столбцах. С помощью всего лишь одного вызова **fit ()** вы можете достичь высокой точности в стандартных контролируемых задачах обучения (как классификация, так и регрессия), не решая таких громоздких проблем, как очистка данных, разработка функций, оптимизация гиперпараметров, выбор модели и другие.

Таблицы мультимодальных данных: сочетание BERT / трансформаторов и классических табличных моделей:

- AutoGluon Tabular для работы с табличными данными, которые содержат текстовые, числовые и категориальные столбцы. В AutoGluon необработанные текстовые данные рассматриваются как важный признак. AutoGluon Tabular может обучить и объединить разнообразный набор моделей, включая классические табличные модели, такие как LightGBM/RF/CatBoost, а также нашу предварительно обученную мультимодальную сеть на основе модели NLP

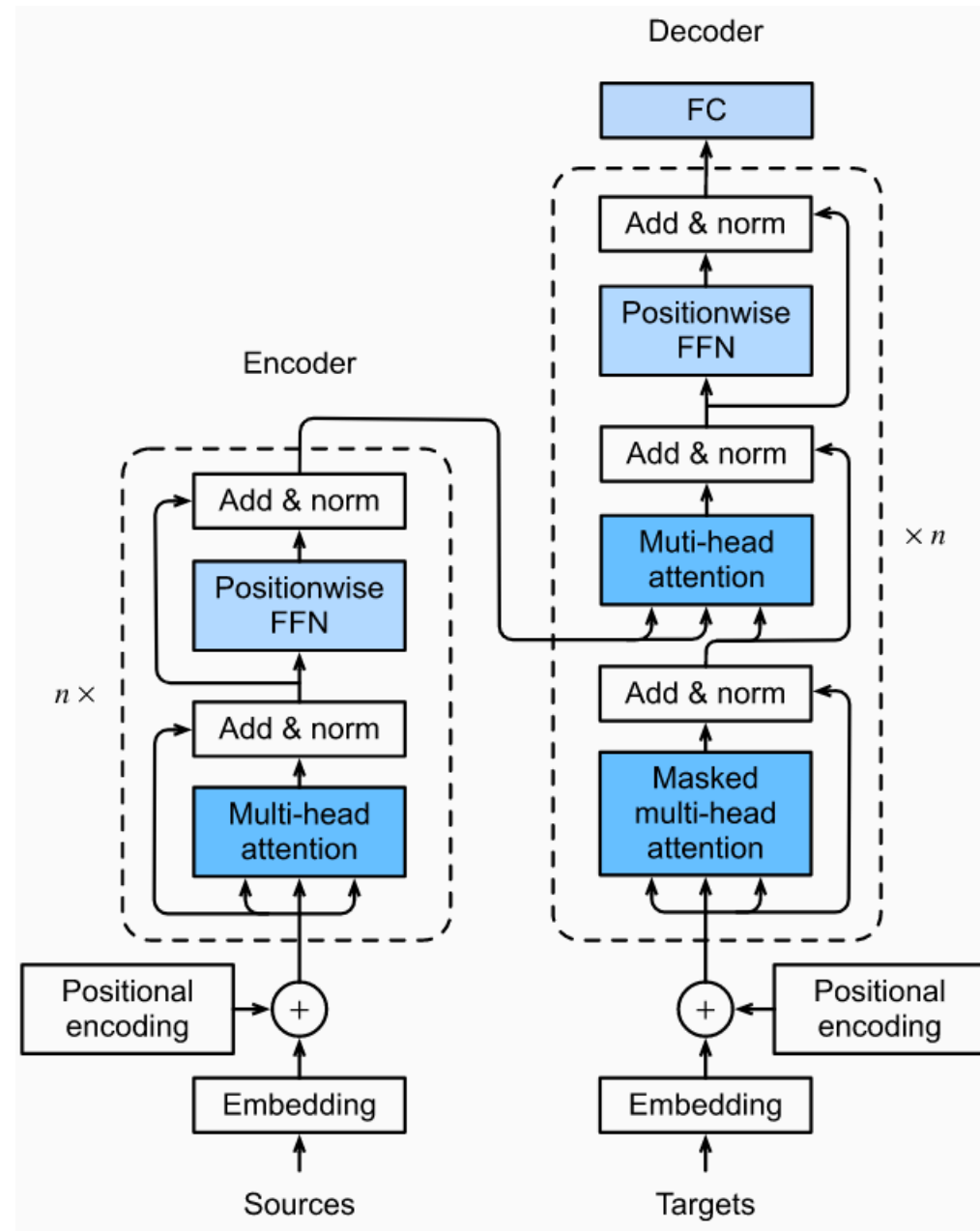
Классификация изображений, Обнаружение объекта



Для классификации изображений на основе их содержимого AutoGluon предоставляет простую функцию **fit ()**, которая автоматически создает высококачественные модели классификации изображений. Один вызов **fit ()** обучит высокоточные нейронные сети на предоставленном вами наборе данных изображений, автоматически используя методы повышения точности, такие как трансферное обучение и оптимизация гиперпараметров.

Текстовое предсказание

Для контролируемого обучения с текстовыми данными AutoGluon предоставляет простую функцию `fit()`, которая автоматически создает высококачественные модели предсказания текста (нейронные сети Transformer). Каждый обучающий пример может быть предложением, коротким абзацем, состоящим из нескольких текстовых полей (например, предсказывающих, насколько похожи два предложения), или может даже содержать дополнительные числовые / категориальные особенности помимо текста. Целевые значения (метки) для прогнозирования могут быть непрерывными значениями (регрессия) или дискретными категориями (классификация). Один вызов `predictor.fit()` обучит высокоточные нейронные сети на предоставленном вами наборе текстовых данных, автоматически используя методы повышения точности, такие как точная настройка предварительно обученной модели НЛП (трансферное обучение) и оптимизация гиперпараметров.



архитектура трансформатора

H2O AutoML

Масштабируемый AutoML в H2O-3 с открытым исходным кодом

H2O AutoML с открытым исходным кодом

- Обучите лучшую модель за минимальное время, чтобы сэкономить время
- Уменьшите потребность в экспертных знаниях в области машинного обучения, сократив время написания кода вручную
- Повысьте производительность (точность и скорость обучения) моделей машинного обучения
- Повысьте воспроизводимость и установите основу для научных исследований или приложений
- Масштабирует набор обучающих данных до кластеров (**Hadoop**, **Spark**, **Kubernetes**)



H2O AutoML

Масштабируемый AutoML в H2O-3 с открытым исходным кодом

Аспекты AutoML

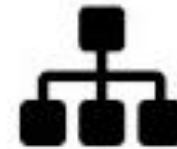
- Imputation, one-hot encoding, standardization
- Feature selection and/or feature extraction (e.g. PCA)
- Count/Label/Target encoding of categorical features
- Поиск по декартовой сетке или поиск по случайной сетке
- Байесовская оптимизация гиперпараметров
- Индивидуальные модели можно настроить с помощью validation set (набора для проверки)
- Ансамбли часто превосходят отдельные модели
- Stacking / Суперобучение (Вольперт, Брейман)
- Выбор ансамбля (Каруана)



Data
Preprocessing



Model
Generation



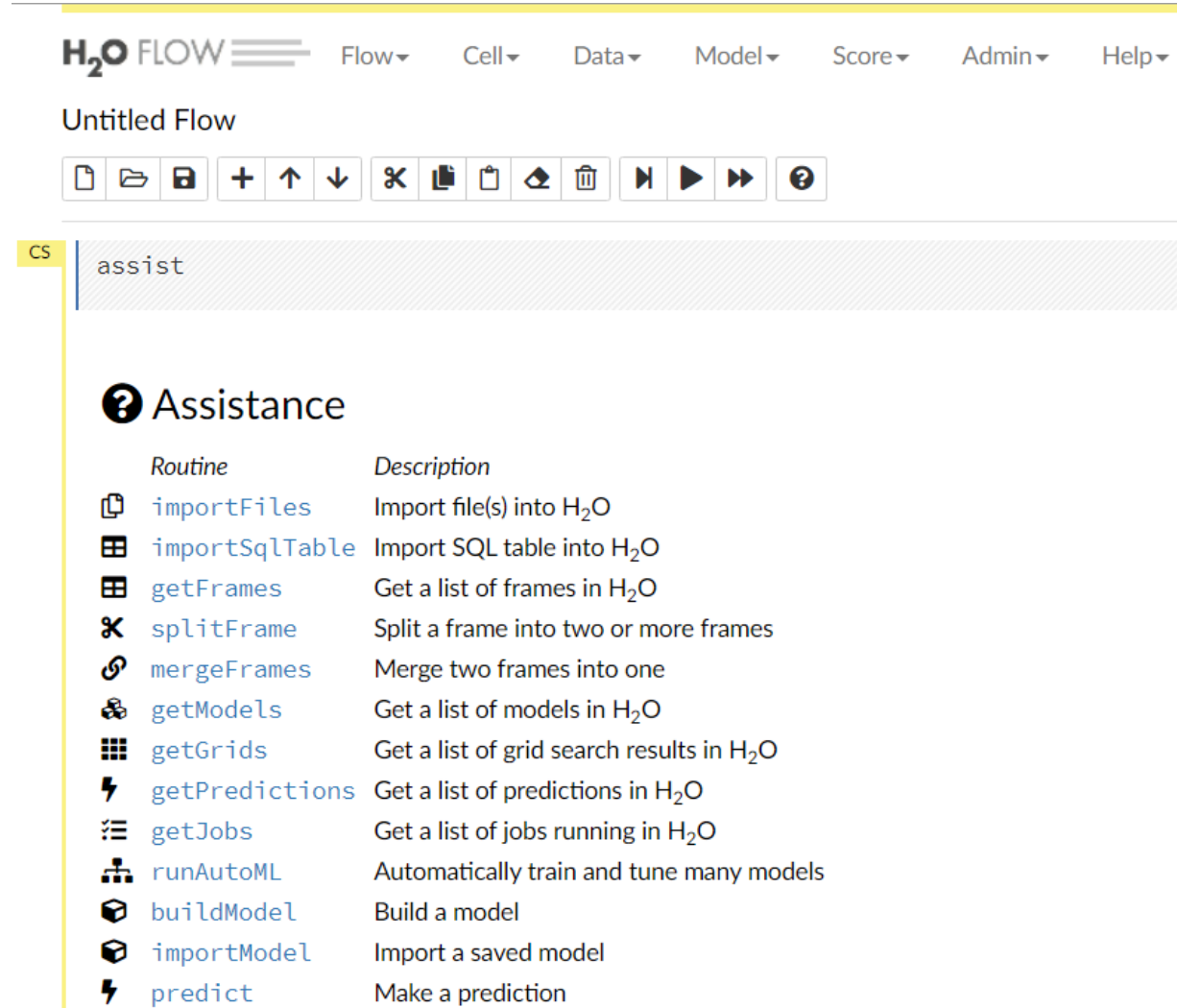
Ensembles

H2O AutoML














Масштабируемый AutoML в H2O-3 с открытым исходным кодом

H2O AutoML использует веб-интерфейс

H2O Flow - это пользовательский интерфейс с открытым исходным кодом для H2O. Это интерактивная веб-среда, которая позволяет объединить выполнение кода, текст, математические данные, графики и мультимедийные данные в одном документе. С H2O Flow вы можете захватывать, повторно запускать, комментировать, представлять и делиться своим рабочим процессом. H2O Flow позволяет вам использовать H2O в интерактивном режиме для импорта файлов, построения моделей и итеративного их улучшения. На основе ваших моделей вы можете делать прогнозы и добавлять форматированный текст для создания виньеток своей работы - и все это в среде на основе браузера Flow.



The screenshot displays the H2O Flow web interface. At the top, there is a navigation bar with the H2O FLOW logo and several dropdown menus: Flow, Cell, Data, Model, Score, Admin, and Help. Below the navigation bar, the title 'Untitled Flow' is visible. A toolbar contains various icons for file operations, navigation, and execution. The main content area is divided into two sections. The top section, labeled 'CS', contains a text input field with the word 'assist'. The bottom section, titled '? Assistance', lists various routines with their descriptions.

| Routine | Description |
|--|---|
|  importFiles | Import file(s) into H ₂ O |
|  importSqlTable | Import SQL table into H ₂ O |
|  getFrames | Get a list of frames in H ₂ O |
|  splitFrame | Split a frame into two or more frames |
|  mergeFrames | Merge two frames into one |
|  getModels | Get a list of models in H ₂ O |
|  getGrids | Get a list of grid search results in H ₂ O |
|  getPredictions | Get a list of predictions in H ₂ O |
|  getJobs | Get a list of jobs running in H ₂ O |
|  runAutoML | Automatically train and tune many models |
|  buildModel | Build a model |
|  importModel | Import a saved model |
|  predict | Make a prediction |

Model▼

Score▼

Admin▼

Run AutoML...

Aggregator...

Cox Proportional Hazards...

Deep Learning...

Distributed Random Forest...

Gradient Boosting Machine...

Generalized Linear Modeling...

Generalized Low Rank Modeling...

Isolation Forest...

K-means...

Naive Bayes...

Principal Components Analysis...

RuleFit...

Stacked Ensemble...

TargetEncoder...

Word2Vec...

Import MOJO Model

List All Models

List Grid Search Results

Import Model...

Export Model...

CS

runAutoML

Run AutoML

PARAMETERS

training_frame*

train_2_hex

response_column*

class

validation_frame

(Choose...)

blending_frame

(Choose...)

exclude_algos

Search...

☐ GLM

☐ DRF

☐ GBM

☐ DeepLearning

☐ StackedEnsemble

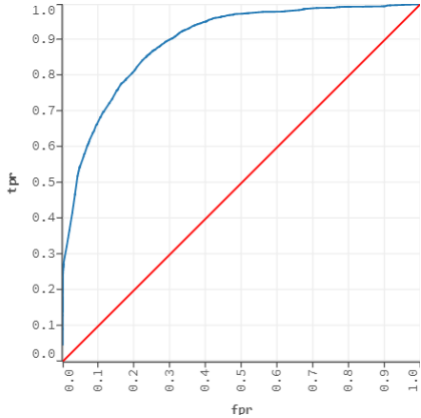
☐ XGBoost

☒ All

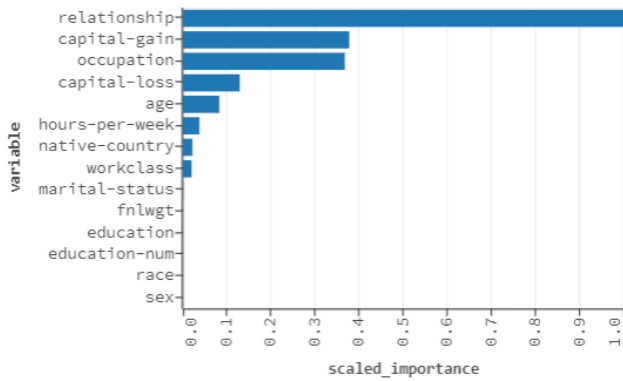
☐ None

Threshold: Choose... Criterion: Choose...

ROC CURVE - CROSS VALIDATION METRICS - AUC =



VARIABLE IMPORTANCES



CROSS VALIDATION METRICS - CONFUSION MATRIX

| | <=50K | >50K | Error | Rate | Precision |
|--------|-------|------|--------|------------------|-----------|
| <=50K | 26357 | 3347 | 0.1127 | 3B 347 / 29B 704 | 0.90 |
| >50K | 2840 | 6529 | 0.3031 | 2B 840 / 9B 369 | 0.66 |
| Total | 29197 | 9876 | 0.1583 | 6B 187 / 39B 073 | |
| Recall | 0.89 | 0.70 | | | |

Leaderboard

Monitor Live

MODELS

models sorted in order of auc, best first

| model_id | auc | logloss |
|---|--------------------|---------------------|
| 0 GBM_3_AutoML_20210508_192210 | 0.8956911002806587 | 0.5081623548133803 |
| 1 GBM_2_AutoML_20210508_192210 | 0.8954934946856876 | 0.508095736447805 |
| 2 GBM_4_AutoML_20210508_192210 | 0.8953389654072025 | 0.5070591018162857 |
| 3 StackedEnsemble_BestOfFamily_AutoML_20210508_192210 | 0.8932309262540649 | 0.35754414263342527 |
| 4 GBM_1_AutoML_20210508_192210 | 0.89260016975547 | 0.5092247840563243 |
| 5 GLM_1_AutoML_20210508_192210 | 0.8452700346769378 | 0.5462063927209736 |
| 6 DRF_1_AutoML_20210508_192210 | 0.8281883617652832 | 2.9489488969833735 |

H2O AutoML

Масштабируемый AutoML в H2O-3 с открытым исходным кодом

Следующие элементы являются частью конвейера **H2O AutoML**:

- Базовая предварительная обработка данных (как и во всех алгоритмах H2O).
- Обучает случайную сетку из GBM, DNN, GLM и др., используя тщательно подобранное пространство гиперпараметров. Команда H2O.ai потратила много времени на размышления об используемых алгоритмах, а также о том, сколько времени и параметров следует использовать вместе с ними. Это своего рода - если можно - «умная грубая сила», позволяющая избежать типичных ошибок.
- Индивидуальные модели настраиваются с использованием перекрестной проверки, чтобы избежать переобучения.
- Обучаются два составленных ансамбля:
 - ❑ - «Все модели»: обычно наиболее эффективный ансамбль - содержит ансамбль всех обученных моделей.
 - ❑ - «Лучший из семейства»: лучший из каждой группы (например, лучший GBM, лучший XGBoost, лучший RF и др.), обычно более легкий, чем подход, основанный на всех моделях, рассмотрите случай, когда у вас может быть 1000 моделей: вы хотели бы лучше экспорт для увеличения производства.
- Все модели легко экспортируются в продакшн формате Java.

H2O AI Hybrid Cloud (коммерческое)

Prepare

H2O ускоряет процесс подключения, предварительной обработки, очистки и преобразования данных, чтобы помочь специалистам по данным и инженерам создавать высококачественные наборы данных и функции для машинного обучения.

Data Connectors

H2O AI Hybrid Cloud предоставляет более 200 коннекторов данных, чтобы упростить прием данных из популярных хранилищ данных, включая Hadoop HDFS, сервисы хранения объектов и базы данных.



Automatic Data Preprocessing and Cleaning

H2O AI Hybrid Cloud автоматически обрабатывает и очищает как табличные, так и текстовые данные. Он обнаруживает пропущенные значения и помечает их, обрабатывает несбалансированные данные и обнаруживает повторяющиеся записи. Для текстовых данных H2O автоматически удаляет стоп-слова, обеспечивает выделение корней и лемматизацию, а также автокоррекцию опечаток. Платформа также предоставляет более 40 уникальных скриптов Python, называемых рецептами, которые разработали инженеры и специалисты по обработке данных, чтобы пользователи могли быстро получить доступ к классическим функциям подготовки данных, таким как связывание, разбиение и вычисления.

H2O AI Hybrid Cloud

Data Transformations

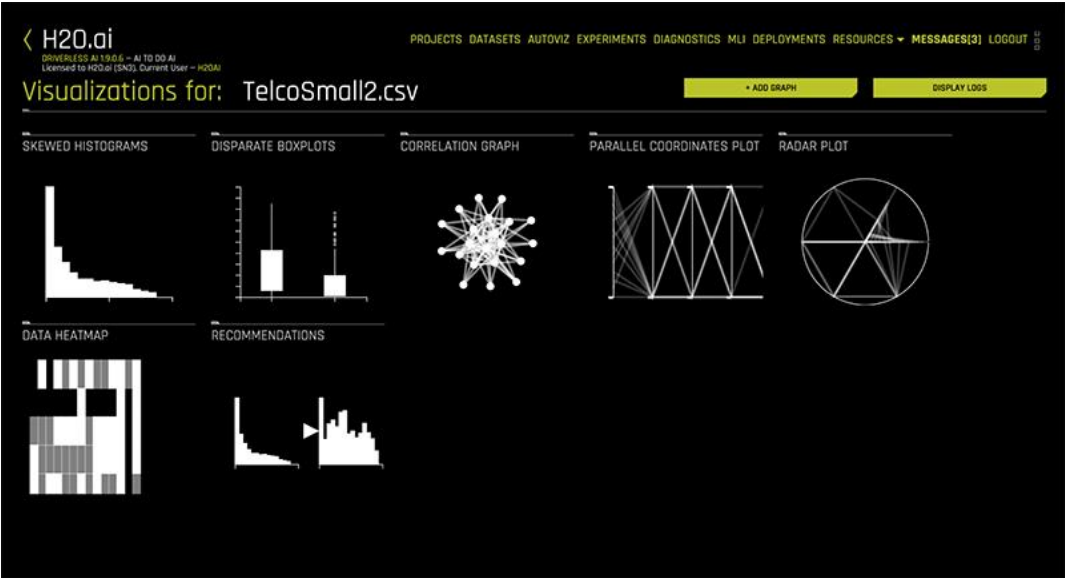
H2O предлагает выбор из 100+ предварительно настроенных преобразований данных, таких как одноразовое кодирование, вложение недостающих данных со средним или медианным значением и встраивание даты / времени, поэтому вы можете преобразовать свои данные в эффективные форматы машинного обучения одним щелчком мыши. Кроме того, платформа H2O является расширяемой и позволяет специалистам по обработке данных и инженерам использовать выбранные ими преобразователи.

| Colors | | | |
|--------|--|--|--|
| Black | | | |
| Yellow | | | |
| Gray | | | |
| Black | | | |

| Black | Yellow | Gray |
|-------|--------|------|
| 1 | 0 | 0 |
| 0 | 1 | 0 |
| 0 | 0 | 1 |
| 1 | 0 | 0 |

Automatic Data Visualization

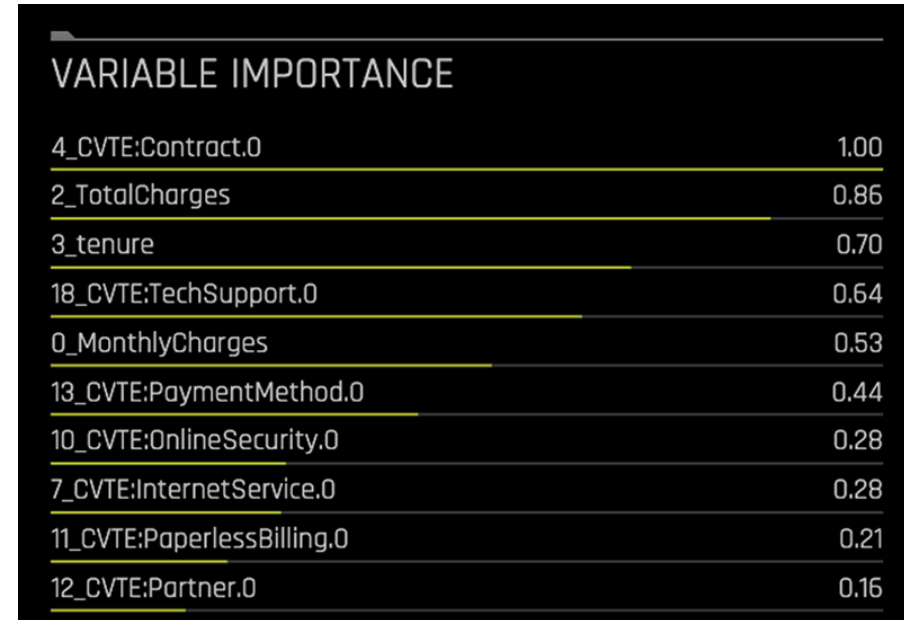
Автоматическая визуализация данных H2O (AutoViz) создает графики и графики данных, которые помогают специалистам по данным и инженерам исследовать и понимать свои данные перед построением модели. AutoViz автоматически определяет выбросы, необычные корреляции, кластеры, а также искаженные или ненормальные распределения, которые могут повлиять на точность моделей прогнозирования.



H2O AI Hybrid Cloud

Automatic Feature Engineering

H2O автоматизирует весь процесс разработки функций, значительно ускоряя выполнение одной из самых трудоемких задач по науке о данных. H2O обнаруживает соответствующие функции, находит взаимодействия в этих функциях и извлекает новые функции из данных. Благодаря недавно полученным функциям технология H2O пересчитывает соответствующие функции и продолжает повторять до тех пор, пока не будут созданы лучшие функции и ранжированы по важности.



| VARIABLE IMPORTANCE | |
|----------------------------|------|
| 4_CVTE:Contract.0 | 1.00 |
| 2_TotalCharges | 0.86 |
| 3_tenure | 0.70 |
| 18_CVTE:TechSupport.0 | 0.64 |
| 0_MonthlyCharges | 0.53 |
| 13_CVTE:PaymentMethod.0 | 0.44 |
| 10_CVTE:OnlineSecurity.0 | 0.28 |
| 7_CVTE:InternetService.0 | 0.28 |
| 11_CVTE:PaperlessBilling.0 | 0.21 |
| 12_CVTE:Partner.0 | 0.16 |

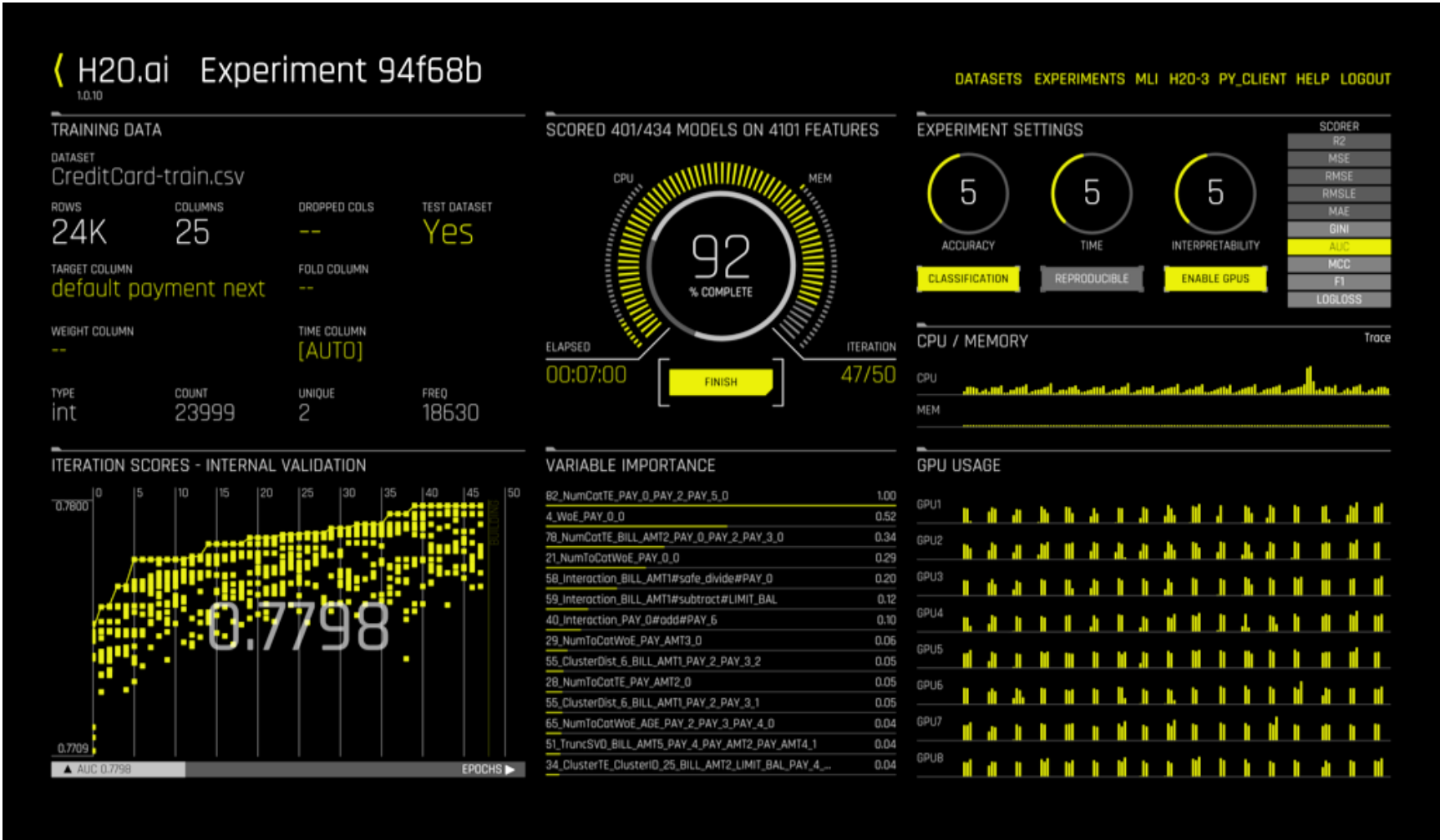
Быстрое создание моделей искусственного интеллекта мирового класса

H2O AI Hybrid Cloud помогает специалистам по обработке данных ускорить процесс построения модели с помощью расширенного автоматического проектирования функций, автоматической настройки гиперпараметров, автоматического выбора алгоритма и автоматической проверки модели. Автоматизированное машинное обучение H2O было разработано ведущими специалистами в области данных в мире и позволяет создавать высокоточные и надежные модели для **структурированных данных, текста, изображений, видео и данных временных рядов.**

H2O AI Hybrid Cloud

AutoML для любых данных

H2O AI Hybrid Cloud позволяет пользователям быстро создавать модели мирового класса не только для табличных данных, но также для текста, изображений, видео и данных временных рядов. H2O автоматически выбирает лучший алгоритм, настраивает модель и предоставляет специалистам по данным и разработчикам таблицу лидеров, чтобы постоянно проверять и тестировать модели чемпионов.



H2O AI Hybrid Cloud

Структурированные / табличные данные

Табличные и структурированные данные часто являются наиболее распространенными и ценными данными для таких организаций, как ERP и CRM. Эта информация может использоваться для прогнозирования сбоев обслуживания, обнаружения мошенничества, выявления клиентов, которые могут отказаться от обслуживания, и многого другого. AutoML H2O работает с табличными наборами данных так же хорошо, как и ведущие доступные методы глубокого обучения.

Обработка естественного языка (NLP)

Компании настойчиво ищут текстовые данные для решения таких бизнес-задач, как оптимизация обслуживания клиентов, персонализация маркетинга, отслеживание настроений в отношении бренда или продукта, классификация документов и добавление тегов к контенту. H2O AI Hybrid Cloud помогает специалистам по данным и разработчикам быстро создавать модели NLP, автоматически преобразовывая текстовые строки в функции с помощью таких методов, как BERT, TFIDF, CNN и GRU. Подобно автоматическому проектированию функций, H2O автоматизирует практически всю предварительную обработку текста, необходимую для решения самых сложных задач НЛП.

H2O AI Hybrid Cloud также включает в себя современные трансформаторы PyTorch BERT. Используя передовые методы НЛП, пользователи могут также обрабатывать более крупные текстовые блоки и строить модели, используя все доступные данные и текст.

H2O AI Hybrid Cloud

Временные последовательности

Прогнозирование временных рядов - одна из самых распространенных проблем аналитики, с которыми сегодня сталкиваются компании, от прогнозирования доходов до FP&A и динамического ценообразования. H2O AI Hybrid Cloud поддерживает обширный набор методов моделирования временных рядов, а также уникальную автоматическую разработку функций специально для задач временных рядов, позволяя пользователям быстро определять, какие временные особенности имеют значение в данных.

Зрение

Организации стремительно и все чаще стремятся использовать данные изображений и видео для повышения безопасности, выявления производственных ошибок, оптимизации макетов розничной торговли и многого другого. H2O AI Hybrid Cloud автоматически создает модели изображений и видео, используя более 30 предварительно обученных преобразователей и моделей изображений, включая (SE) -ResNe (X) ts, DenseNets, MobileNets, EffientNets и Inceptions. Изображения можно обрабатывать отдельно или как часть наборов данных, которые включают другие типы данных, например табличные и текстовые данные.

Быстрое обучение больших наборов данных

H2O AI Hybrid Cloud оптимизирован для работы с новейшими графическими процессорами Nvidia, IBM Power 9 и процессорами Intel x86. Он поддерживает распределенные вычисления и быстро обучает модели на основе ТБ данных.

H2O AI Hybrid Cloud

Объяснимость (Explain). Комплексный набор лучших в своем классе объяснимых методов искусственного интеллекта

H2O.ai уже несколько лет является пионером в области объяснимого ИИ и интерпретируемости машинного обучения. Помимо публикации нескольких ведущих книг о том, как создать более ответственный ИИ, H2O.ai также разработал передовые методы в отрасли, такие как K-Lime.

Надежный набор инструментов постфактум для объяснения

H2O.ai создал один из самых обширных наборов возможностей для анализа ваших моделей машинного обучения после того, как они были разработаны, такие как: Shapley Values, K-Lime, суррогатные деревья решений, Reason Codes, Partial Dependency Plots, Disparate Impact Analysis, Exportable Rules Based Systems и многое другое.

AutoML с передовыми методами моделирования белого ящика

Ведущие методы, такие как Explainable Neural Networks (XNNs), GA2Ms, Light GBM и XGBoost, позволяют пользователям автоматически создавать более прозрачные модели без ущерба для точности.

