**Пошаговое руководство по AutoML с H2O Flow**

**Вступление**

Разве не утомительно писать несколько строк кода каждый раз, когда мы хотим создать модель машинного обучения!

Вы когда-нибудь задумывались, насколько простым и эффективным стало бы, если бы мы могли построить модель машинного обучения, просто указав и щелкнув мышью? H2O Flow предлагает решение всех подобных проблем!

**H2O** - это платформа машинного обучения и искусственного интеллекта с открытым исходным кодом. Он поддерживает веб-интерфейс под названием Flow. **H2O Flow** можно использовать для создания различных типов моделей машинного обучения без написания кода. Мы можем просто указать и щелкнуть мышью, чтобы построить конвейеры машинного обучения. Имеет поддержку API для R, Python, Scala.

*В этой статье мы создаем модели в H2O Flow с помощью* ***AutoML,*** *но можно спросить, что же такое AutoML и что в нем такого шума?*

Чтобы ответить на этот вопрос, AutoML (автоматическое машинное обучение) просто автоматизирует процесс моделирования, что позволяет специалистам по обработке данных сосредоточиться на других важных аспектах конвейеров машинного обучения, таких как разработка функций и развертывание моделей.

**Установка h2O Flow**

Скачайте последнюю версию программы с официальной страницы. У вас должна быть предварительно установлена ​​64-битная Java, поскольку это предварительное условие для H2O Flow. После завершения загрузки мы можем использовать zip-файл с файлом jar H2O.

Запустите терминал, извлеките файл и следуйте приведенным ниже командам, чтобы запустить H2O Flow в вашем браузере.

cd ~ / Загрузки

разархивировать h2o-3.30.0.6.zip

cd h2o-3.30.0.6

java -jar h2o.jar

Переключитесь в свой браузер и перейдите по [адресу http: // localhost: 54321,](http://localhost:54321/) чтобы получить доступ к веб-интерфейсу Flow.

Первое, что может поразить вас, это дизайн Flow, очень похожий на ноутбуки Jupyter. Правая панель - это раздел справки, который может быть полезен новичкам.

Обратите внимание на эти сочетания клавиш, которые будут использоваться на протяжении всего проекта!

**Загрузка данных**

Мы будем использовать свободно доступный [набор данных](https://archive.ics.uci.edu/ml/machine-learning-databases/00222/) для нашего проекта. Эти данные относятся к прямым маркетинговым кампаниям банка, и нам необходимо спрогнозировать  
охват их клиентов. основанный на многих функциях. [Изучите набор данных](https://archive.ics.uci.edu/ml/datasets/Bank+Marketing), чтобы узнать больше об особенностях.

Теперь давайте начнем с создания нашей собственной записной книжки Flow. Обратите внимание на кнопку «+» на верхней панели инструментов. Мы можем использовать его для вставки ячеек. Как и в записных книжках Jupyter, мы можем включать ячейки разметки для любого текста, который мы хотим написать.

Нажмите на опцию «Импорт файлов», укажите местоположение файла данных и начните импорт. Мы также можем импортировать файлы из других источников, таких как HDFS и S3 bucket.

**Анализ данных**

Анализ данных относится к определению схемы. Схема автоматически обнаруживается “угадателем” синтаксического анализа. Мы можем изменить любой столбец по своему усмотрению в соответствии с нашими потребностями. Мы можем изменить тип данных с категориального на enum, здесь мы можем изменить столбец «день» на *enum,* поскольку в неделе всего 7 дней.

Мы знаем, что это проблема бинарной классификации, поскольку мы должны предсказать, будут ли клиенты зачислены на срочные вклады или нет.

Обычно мы применяем one hot encoding к категориальным данным перед построением модели машинного обучения, но Flow предоставляет нам функцию автоматического one hot encoding. Теперь давайте продолжим и нажмем Parse.

После завершения анализа мы можем просмотреть уточненные данные, включая размер датасета, количество столбцов и строк.

**Исследование данных**

Давайте изучим и визуализируем наши данные для лучшего понимания. Мы можем выбрать столбцы для их визуализации по отдельности. Мы можем получить распределение для числового столбца или частоту для категориальных столбцов.

Мы можем видеть характеристики и сводку столбца «Возраст» вместе с частотным распределением.

Здесь мы можем увидеть распределение для столбца «y». Визуализируя столбец предикторов, мы делаем вывод о высоком уровне дисбаланса классов.

Точно так же мы можем проверить и другие столбцы.

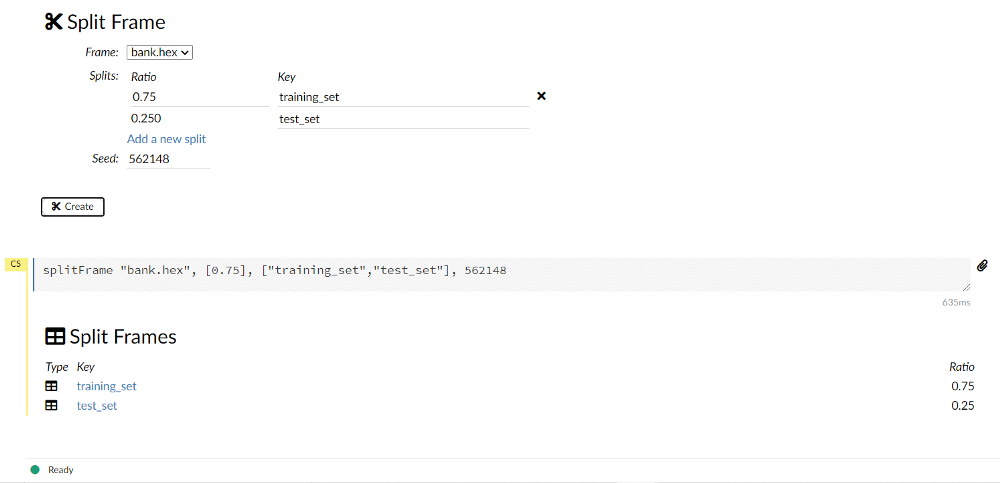
В Flow предусмотрена возможность вменения данных. Это может быть полезно в случае линейной модели, которая будет рассчитываться через пропущенные значения. Предусмотрен ряд методов условного исчисления. Метод по умолчанию установлен как *Среднее.*



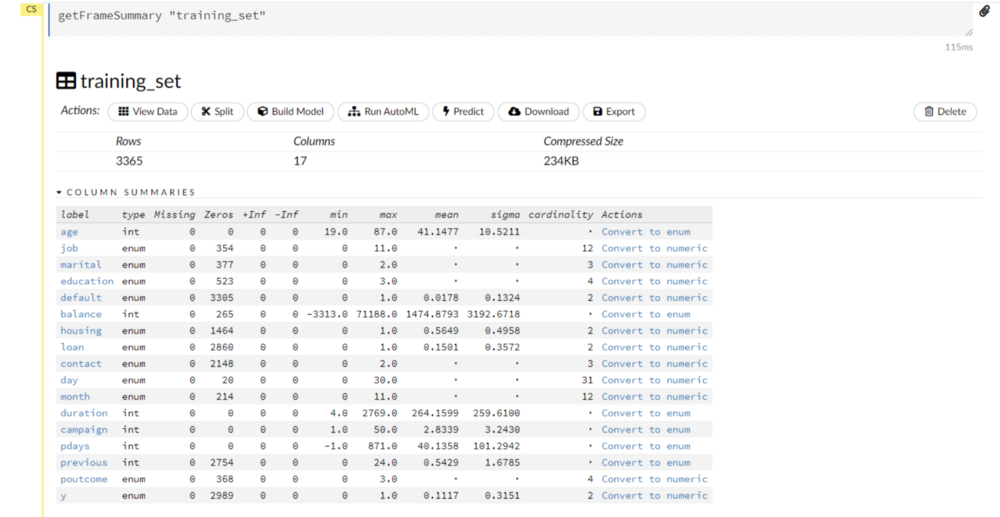
**Создание разделений для обучения и тестирования**

Прежде чем мы начнем обучение нашей модели, нам нужно разделить наши данные на обучающие и тестовые наборы. Мы можем добиться этого, перейдя к *данным -> разделить фрейм* на панели инструментов.

Обратите внимание, что разделение по умолчанию составляет 75:25 для обучения и тестирования соответственно. Это может быть изменено в соответствии с потребностями. Переименуйте разбиения в «training\_set» и «test\_set».



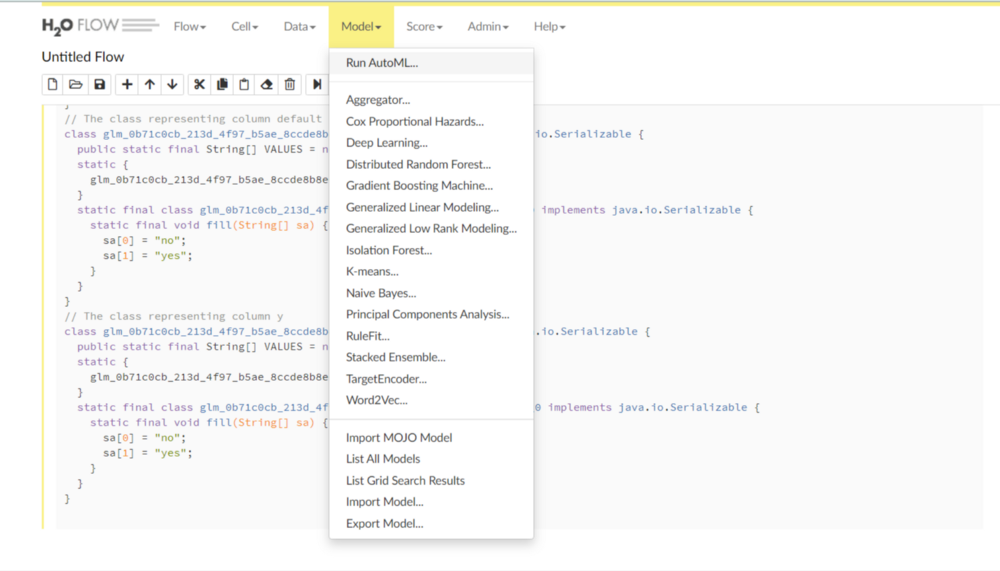
Теперь мы можем осмотреть каждый набор индивидуально, выбрав кадры.



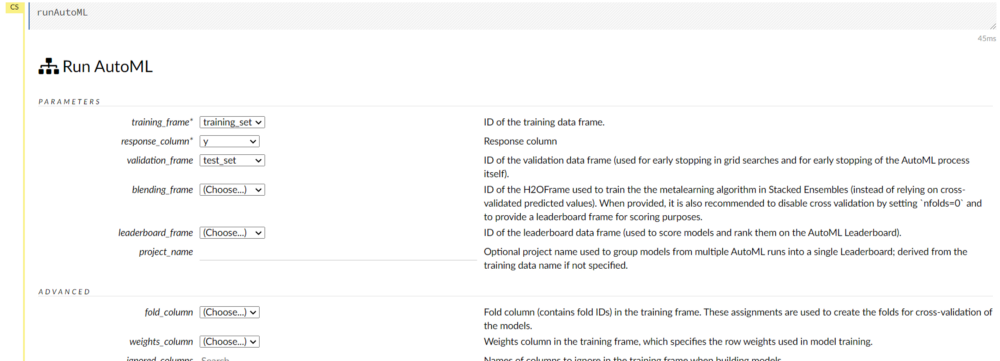
**Построение модели с помощью AutoML**

[AutoML](https://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html) обучает различные типы моделей, включая обобщённые линейные модели ([GLM](https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/glm.html)), распределенные случайные леса [DRF](https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/drf.html), [Gradient Boosting Machine (GBM)](https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/gbm.html), [XGBoost](https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/xgboost.html), [многослойный персептрон](https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/deep-learning.html) и составные ансамбли (стекинг [Stacked Ensembles](https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/stacked-ensembles.html)). Также представлена ​​таблица лидеров со всеми моделями, отсортированными по некоторым показателям.

Мы можем выбрать опцию ***Run AutoML*** из *раскрывающегося меню:*

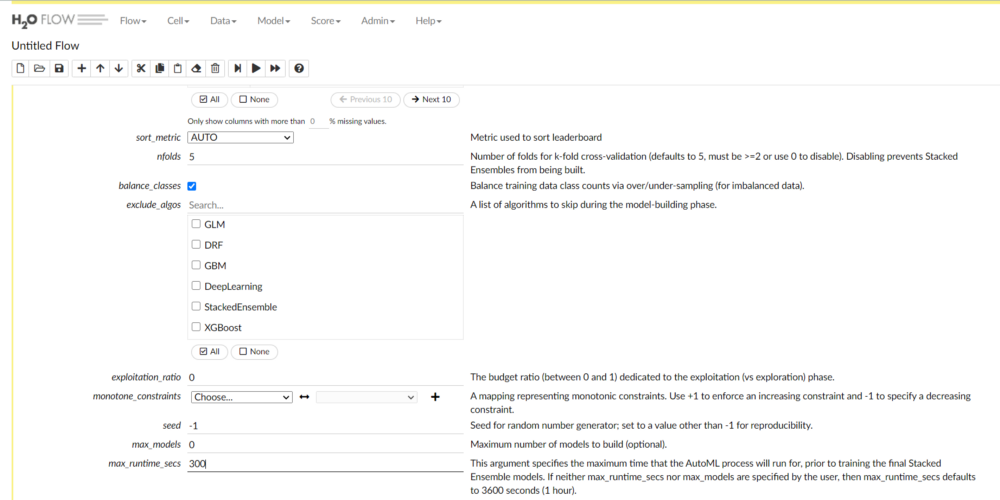


Теперь мы можем выбрать обучающий кадр, кадр ответа и кадр проверки как «training\_set», «y» и «test\_set» соответственно. Мы можем игнорировать другие параметры, поскольку они используются для добавления дополнительных функций.



По умолчанию число перекрестной проверки равно пяти. Поскольку у нас есть ситуация дисбаланса классов, мы можем выбрать опцию классов баланса.

Мы также можем исключить определенные модели, если знаем, что они не актуальны. Измените *max\_runtime\_seconds* на 300 *секунд. AutoML* обучает модели до *max\_runtime\_seconds, после чего он остановится.* По умолчанию установлено значение 3600 секунд.



Наконец, мы можем начать построение моделей, выбрав опцию «построить модели». Есть возможность просмотра обновлений моделей в реальном времени во время обучения. Мы также можем видеть историю оценок в режиме реального времени с визуализацией обучения (графическими представлениями).

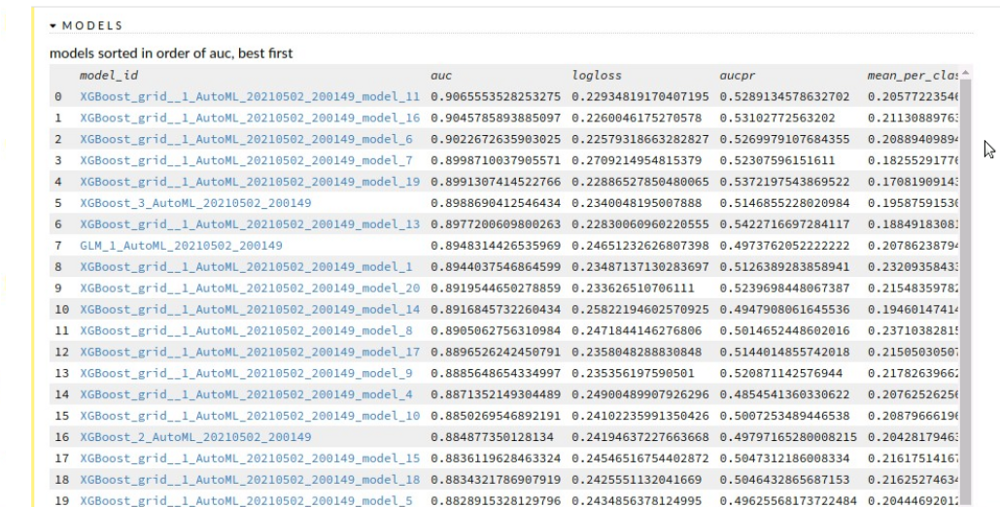


**Модель исследования**

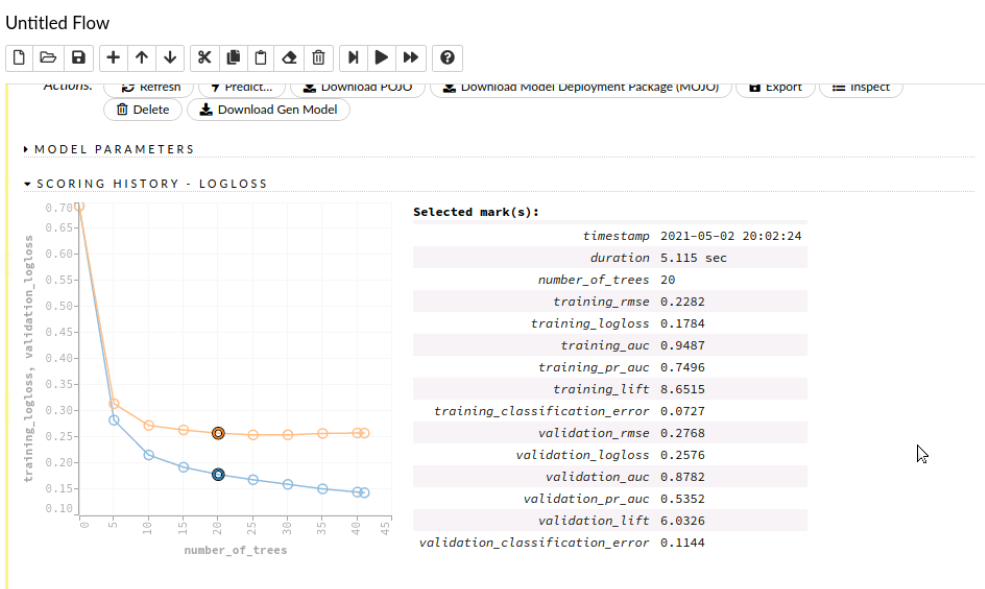
Процесс обучения займет пять минут, так как мы указали *max\_runtime\_seconds.* После завершения работы мы можем перейти к таблице лидеров модели.

Все модели, обученные AutoML, отображаются в отсортированном порядке в соответствии с производительностью.

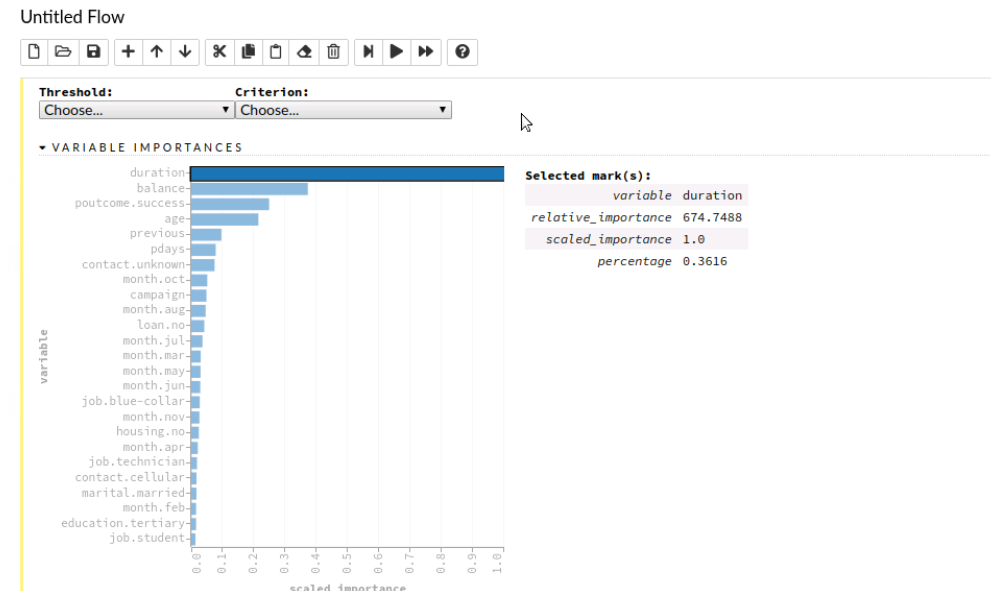
В этом случае модель **XGboost** является лучшей моделью.



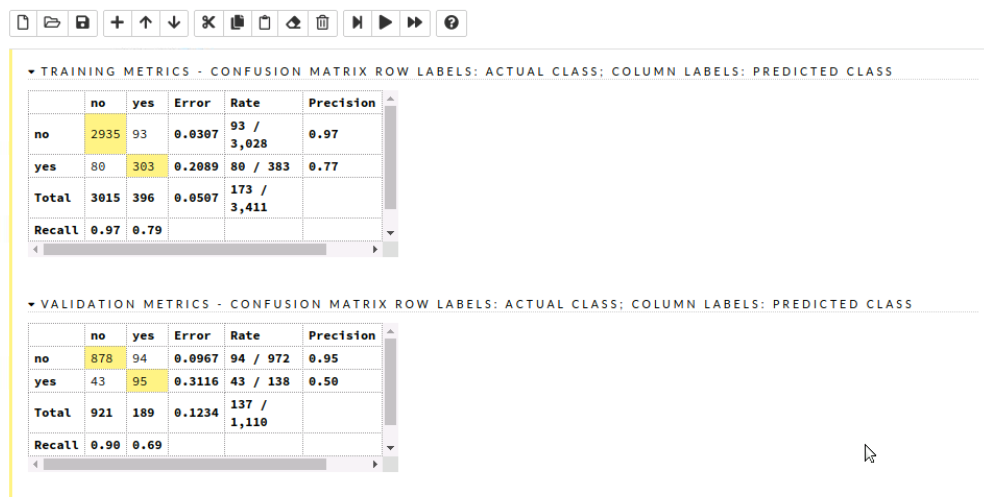
AutoML обеспечивает визуализацию различных показателей, которые можно использовать для исследования модели. Мы можем щелкнуть по кривой, чтобы узнать о ней более подробную информацию.



Мы также можем проверить важность переменных. Мы видим, что переменная продолжительность контакта очень предсказуема и используется в этой модели.



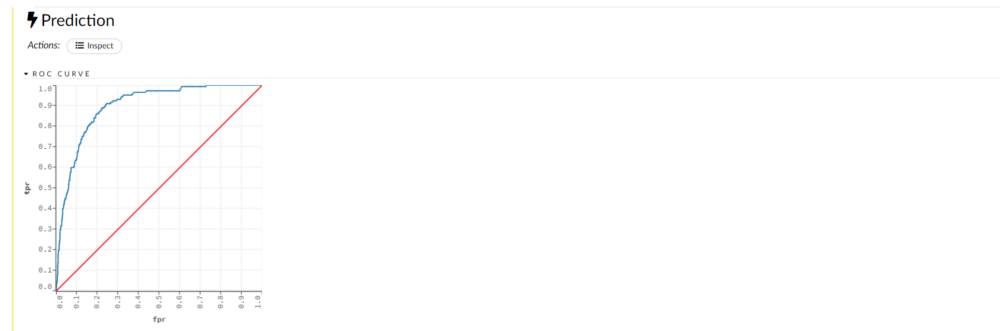
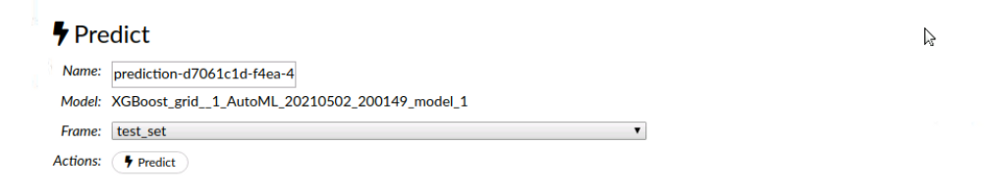
[Матрица ошибок](https://en.wikipedia.org/wiki/Precision_and_recall)  также имеет решающее значение, поскольку она дает различную информацию, такую ​​как корреляция. Мы также можем выделить определенные переменные и просмотреть их.



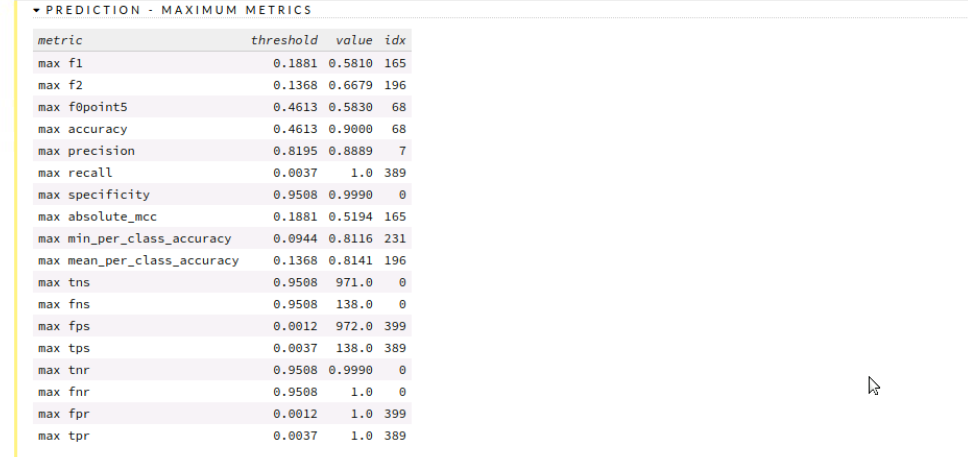
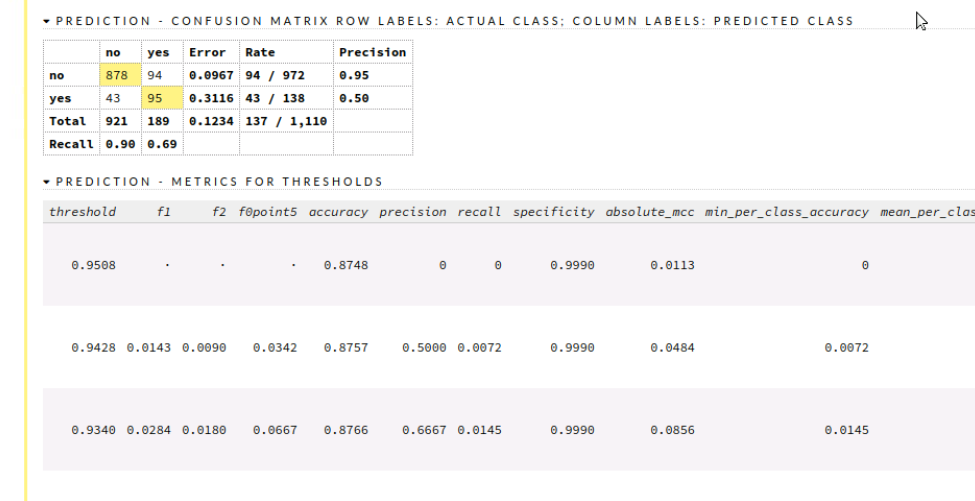
**Прогноз**

После того, как мы будем удовлетворены выбранной моделью, можно переходить к прогнозам.

Начните с выбора параметра прогнозирования на панели инструментов и выберите модель, которую вы хотите использовать, а затем выберите рамку проверки. Теперь, просто нажав кнопку прогнозирования, мы можем просмотреть прогнозируемые значения вместе с различными метриками оценки, такими как средний квадрат.



Мы можем видеть матрицу путаницы и анализировать наш результат вместе с различными показателями и представлениями графиков.



**Конечные заметки**

Мы успешно научились использовать H20 из веб-интерфейса, называемого H2O Flow, и обучили и визуализировали все этапы создания и обучения моделей, не написав ни единой строчки кода! Из этого урока вы получили представление о том, насколько простым становится создание и обучение моделей машинного обучения с использованием AutoML.

Для получения дополнительной информации рекомендуется изучить официальную [документацию](https://docs.h2o.ai/) .