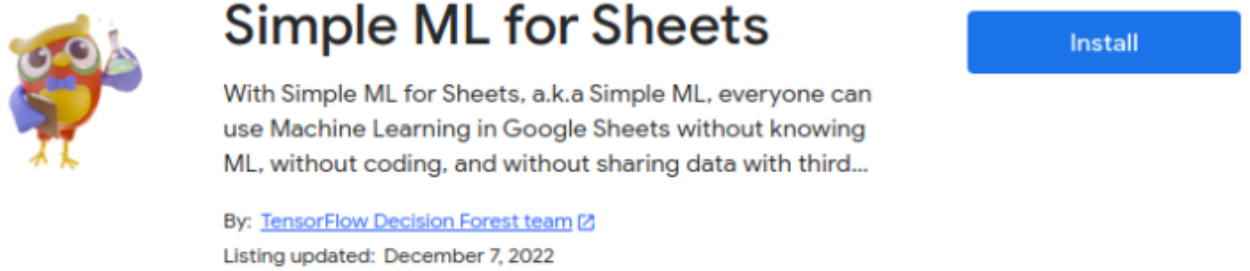
Применение машинного обучение к данным в Google Таблицах

Благодаря **Simple ML for Sheets** , или просто **Simple ML** , каждый может использовать машинное обучение (ML) в Google Sheets, не зная ML, не кодируя и не передавая данные третьим лицам. Simple ML for Sheets применяют для решения трех задач: прогнозирование отсутствующих значений (задача 1), выявление аномальных значений (задача 2) и обучение/оценка и оценка и объяснение модели (задача 3).

Установите Simple ML для Таблиц

Сначала установите Simple for Sheets.

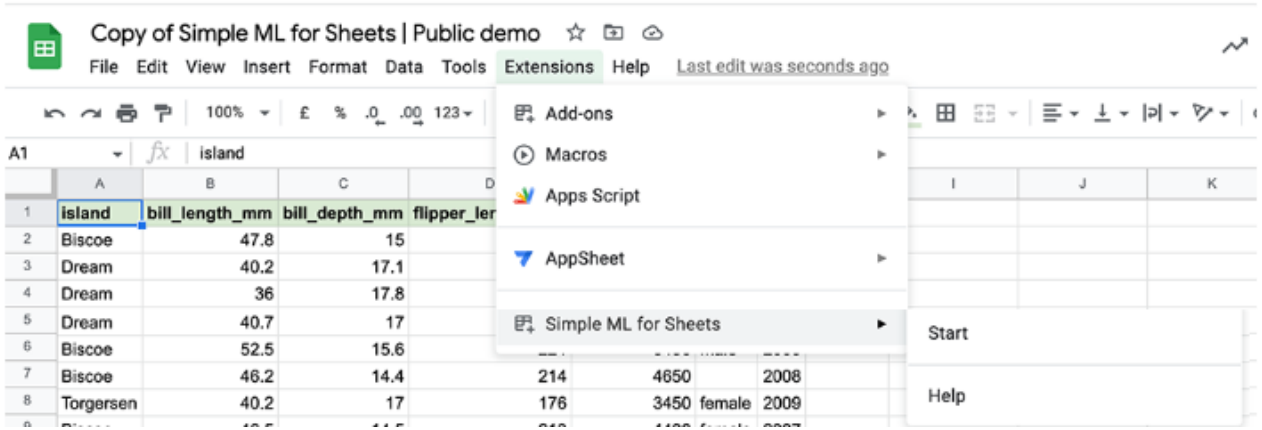
1. Перейдите к Simple ML for Sheets в Google Marketplace и нажмите кнопку **«Установить»** .



2. Предоставить разрешения.

   Почему?: Simple ML хранит ваши модели машинного обучения на вашем Google Диске в `simple_ml_for_sheets`каталоге.

3. Мы собрали электронную таблицу с примерными данными . Откройте и сделайте копию этой таблицы.
4. В электронной таблице Google Sheets убедитесь, что Simple ML отображается в меню **«Расширения»** . Если надстройка не видна, подождите несколько секунд и обновите страницу. Дополнение может появиться через минуту после установки.



Вы изучаете различия между этими видами. Ваш коллега собрал физиологические параметры (такие как размер и вес) примерно 300 пингвинов. Однако в процессе они отвлеклись и забыли отметить виды 30 пингвинов.

Вы хотите использовать Simple ML для восстановления видов этих 30 пингвинов.

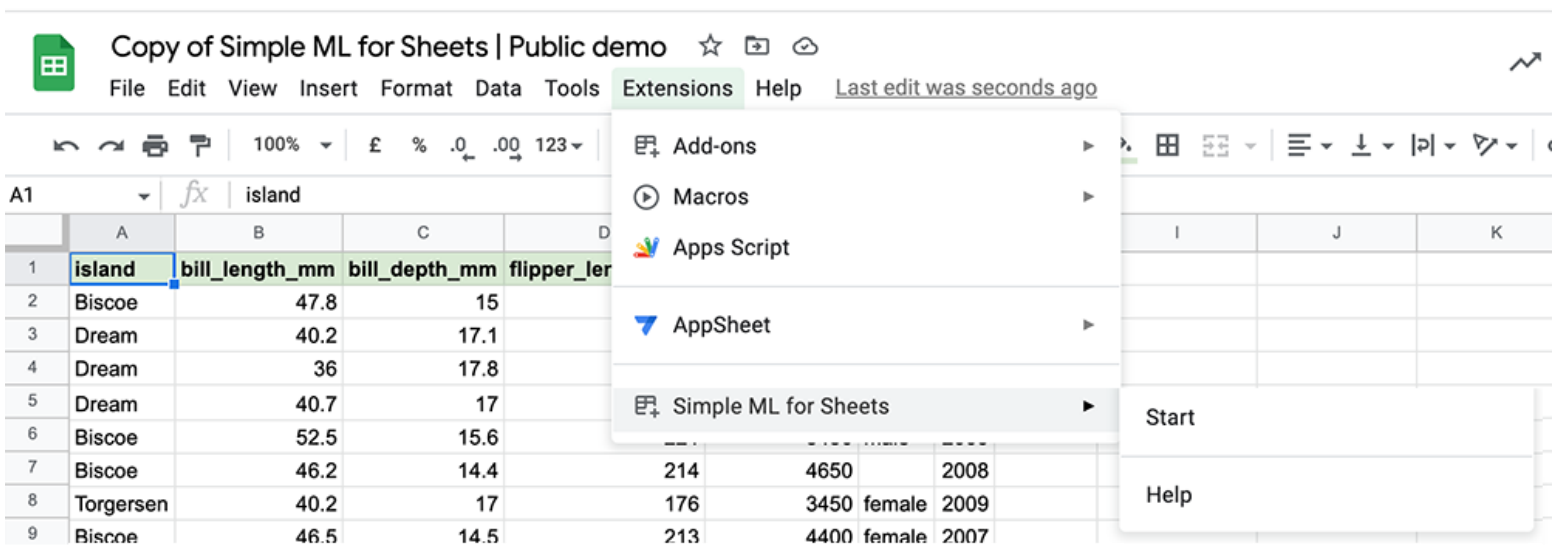
1. Прокручивайте данные вверх и вниз.

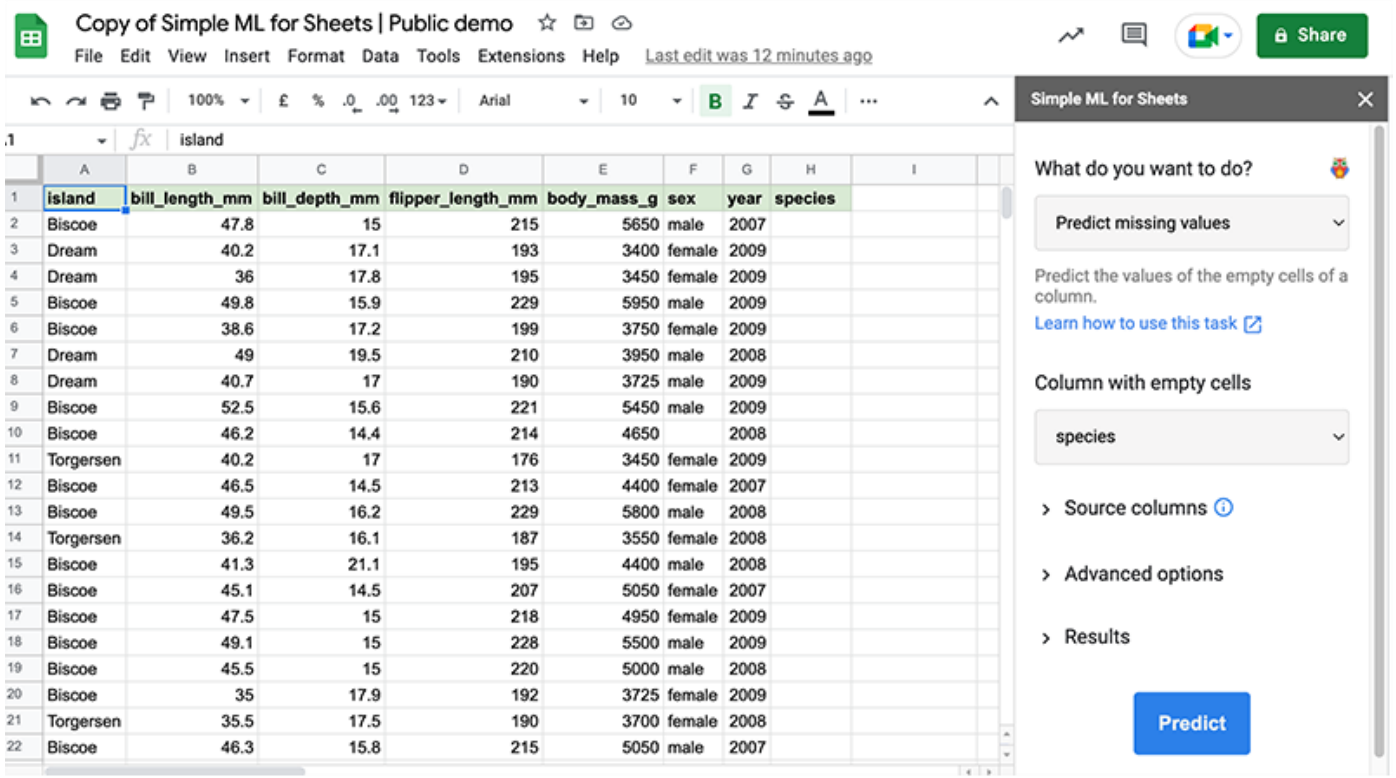
Этот лист содержит данные о пингвинах. Каждая строка представляет пингвина, а каждый столбец представляет физиологический аспект пингвина.

Последний названный столбец speciesпредставляет виды животных. Вид отсутствует в первых 30 рядах.

2. Откройте Простой ML. В меню нажмите «Расширения» > «Простой ML для листов» > «Пуск» .

3. Подождите, пока появится боковая панель Simple ML (это может занять несколько секунд).





Предсказать пропущенные значения
После загрузки Simple ML вы можете использовать его для прогнозирования отсутствующих значений.

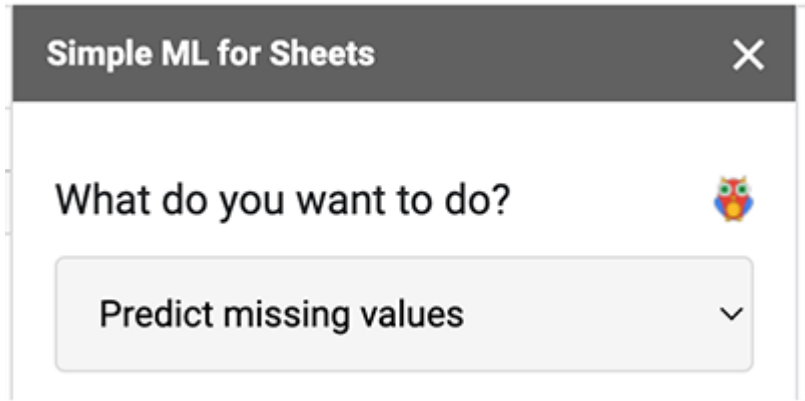
1. Во-первых, обратите внимание, что в некоторых строках отсутствуют значения в столбце H, species.

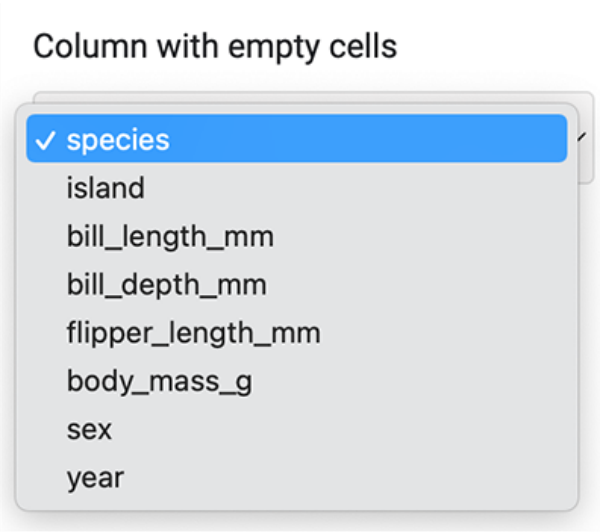
2. На боковой панели Simple ML for Sheets убедитесь, что поле Что вы хотите сделать? установлен на Прогнозирование отсутствующих значений .
3. В разделе «Столбец с пустыми ячейками» выберите столбец species. Это столбец, содержащий пропущенные значения, которые вы хотите предсказать.
4. Нажмите кнопку «Предсказать».

Через несколько секунд появятся два новых столбца:

Pred:species — это предсказанные значения для столбца видов .

Pred:Conf.species — достоверность предсказанного значения. Другими словами, этот столбец показывает, насколько модель уверена в предсказаниях в столбце Pred:species . Уверенность — это процент от 0% до 100%, где 100% означает, что модель уверена в своем прогнозе.



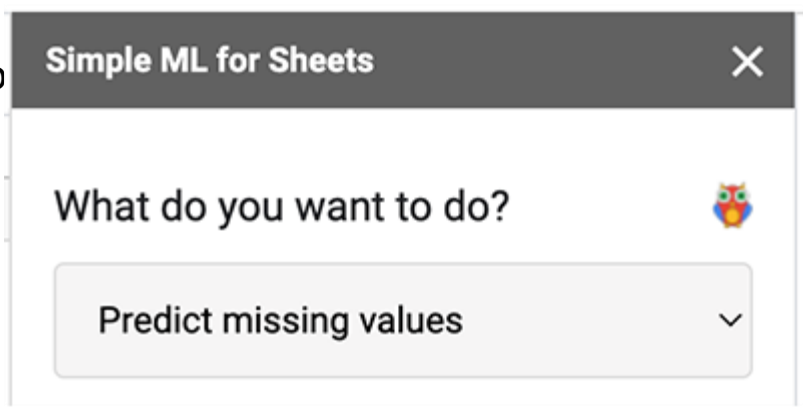


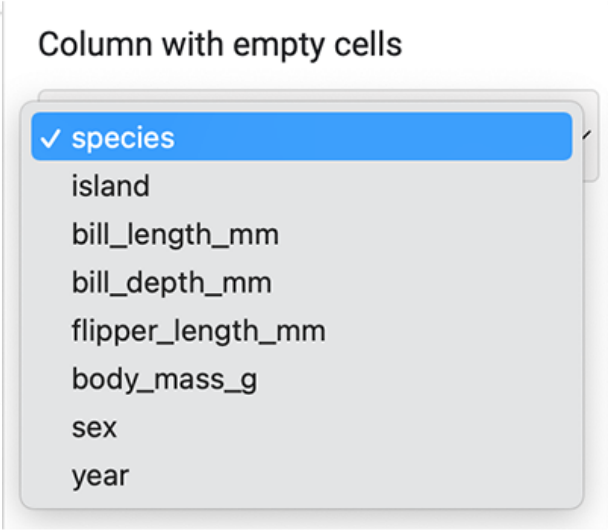
B	C	D	E	F	G	H	I	J
bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year	species	Pred:species	Pred:Conf.species
47.8	15	215	5650	male	2007		Gentoo	99.23 %
40.2	17.1	193	3400	female	2009		Adelie	99.06 %
36	17.8	195	3450	female	2009		Adelie	99.25 %
49.8	15.9	229	5950	male	2009		Gentoo	99.23 %
38.6	17.2	199	3750	female	2009		Adelie	99.25 %
49	19.5	210	3950	male	2008		Chinstrap	99.20 %
40.7	17	190	3725	male	2009		Adelie	99.10 %
52.5	15.6	221	5450	male	2009		Gentoo	99.23 %
46.2	14.4	214	4650		2008		Gentoo	99.23 %
40.2	17	176	3450	female	2009		Adelie	99.14 %
46.5	14.5	213	4400	female	2007		Gentoo	99.25 %
49.5	16.2	229	5800	male	2008		Gentoo	99.23 %
36.2	16.1	187	3550	female	2008		Adelie	98.69 %
41.3	21.1	195	4400	male	2008		Adelie	99.18 %
45.1	14.5	207	5050	female	2007		Gentoo	99.26 %
47.5	15	218	4950	female	2009		Gentoo	99.25 %
49.1	15	228	5500	male	2009		Gentoo	99.23 %
45.5	15	220	5000	male	2008		Gentoo	99.21 %
35	17.9	192	3725	female	2009		Adelie	99.25 %

Предсказать пропущенные значения
После загрузки Simple ML вы можете использовать его для прогнозирования отсутствующих значений.

1. Во-первых, обратите внимание, что в некоторых строках отсутствуют значения в столбце H, species.
2. На боковой панели Simple ML for Sheets убедитесь, что поле Что вы хотите сделать? установлен на Прогнозирование отсутствующих значений .
3. В разделе «Столбец с пустыми ячейками» выберите столбец species. Это столбец, содержащий пропущенные значения, которые вы хотите предсказать.
4. Нажмите кнопку «Предсказать».
Через несколько секунд появятся два новых столбца:
Pred:species — это предсказанные значения для столбца видов .
Pred:Conf.species — достоверность предсказанного значения. Другими словами, этот столбец показывает, насколько модель уверена в предсказаниях в столбце Pred:species . Уверенность — это процент от 0% до 100%, где 100% означает, что модель уверена в своем прогнозе.

Прогнозировать пропущенное значение собирает строки со species значениями и использует эти строки для обучения модели. Затем эта модель применяется к строкам с отсутствующими species значениями. Модель доступна в задаче Управление моделями и в папке simple_ml_for_sheetsвашего Google Диска.





B	C	D	E	F	G	H	I	J
bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year	species	Pred:species	Pred:Conf.species
47.8	15	215	5650	male	2007		Gentoo	99.23 %
40.2	17.1	193	3400	female	2009		Adelie	99.06 %
36	17.8	195	3450	female	2009		Adelie	99.25 %
49.8	15.9	229	5950	male	2009		Gentoo	99.23 %
38.6	17.2	199	3750	female	2009		Adelie	99.25 %
49	19.5	210	3950	male	2008		Chinstrap	99.20 %
40.7	17	190	3725	male	2009		Adelie	99.10 %
52.5	15.6	221	5450	male	2009		Gentoo	99.23 %
46.2	14.4	214	4650		2008		Gentoo	99.23 %
40.2	17	176	3450	female	2009		Adelie	99.14 %
46.5	14.5	213	4400	female	2007		Gentoo	99.25 %
49.5	16.2	229	5800	male	2008		Gentoo	99.23 %
36.2	16.1	187	3550	female	2008		Adelie	98.69 %
41.3	21.1	195	4400	male	2008		Adelie	99.18 %
45.1	14.5	207	5050	female	2007		Gentoo	99.26 %
47.5	15	218	4950	female	2009		Gentoo	99.25 %
49.1	15	228	5500	male	2009		Gentoo	99.23 %
45.5	15	220	5000	male	2008		Gentoo	99.21 %
35	17.9	192	3725	female	2009		Adelie	99.25 %

Задача 2: Найдите аномальные значения
Откройте первую вкладку под названием «Случай № 2: обнаружение аномальных значений».

Этот лист содержит физиологические данные об морских ушках (разновидность морских улиток). Как и в задании № 1, каждая строка представляет животное, а каждый столбец представляет физиологический аспект животных.

Ученые могут определить возраст морского ушка по количеству колец в его раковине, как вы можете определить возраст дерева по количеству колец в его стволе. Этот набор данных содержит записи примерно о 4000 морских ушек. В этом задании цель состоит в том, чтобы найти аномалии в количестве колец в морских ушках.

Вычислите показатель аномалии

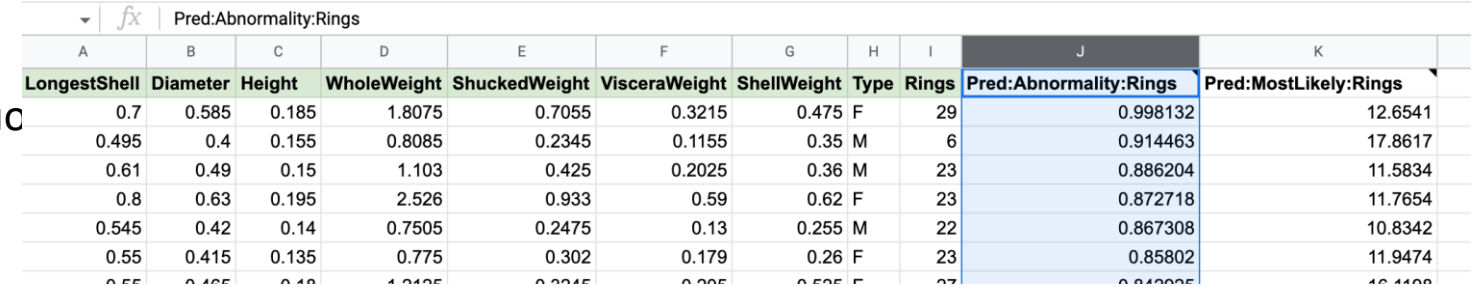
1. Выберите вкладку **Дело № 2: Выявить аномальные значения.**
2. Если Simple ML еще не открыт, откройте его **Extension** > **Simple ML** > **Start**
3. В разделе **Что вы хотите сделать?** выберите **Выявить аномальные значения.**
4. Выберите столбец, который вы хотите проанализировать. В данном случае это столбец **«Кольца»** .
5. Щелкните **Выявить аномальные значения** .

Через несколько секунд справа от ваших данных будут созданы два новых столбца: **Pred:Abnormality:Rings** и **Pred:Mostlikely:Rings** .

- **Pred:Abnormality:Rings** — это «показатели отклонения» от 0 (наиболее нормальное) до 1 (наиболее ненормальное). Они говорят, насколько похожа на другие значения каждая строка.
- **Pred:MostLikely** — это наиболее вероятные прогнозируемые значения для выбранного столбца. В некоторых случаях это будет то же самое, что и существующее значение, в других случаях прогнозируемое значение будет отличаться от существующего значения.

A	B	C	D	E	F	G	H	I	J	K
LongestShell	Diameter	Height	WholeWeight	ShuckedWeight	VisceraWeight	ShellWeight	Type	Rings	Pred:Abnormality:Rings	Pred:MostLikely:Rings
0.455	0.365	0.095	0.514	0.2245	0.101	0.15	M	15	0.00452901	8.66309
0.35	0.265	0.09	0.2255	0.0995	0.0485	0.07	M	7	0	8.4635
0.53	0.42	0.135	0.677	0.2565	0.1415	0.21	F	9	0	11.2651
0.44	0.365	0.125	0.516	0.2155	0.114	0.155	M	10	0	9.5178
0.33	0.255	0.08	0.205	0.0895	0.0395	0.055	I	7	0	6.65391
0.425	0.3	0.095	0.3515	0.141	0.0775	0.12	I	8	0	7.97577

Ищите самые ненормальные строки
Иногда интересно посмотреть на самые «ненормальные» значения, отсортировав строки по столбцу Pred:Abnormality. Аномальные значения иногда указывают на ошибку в данных или нормальный аспект набора данных.



Pred:Abnormality:Rings

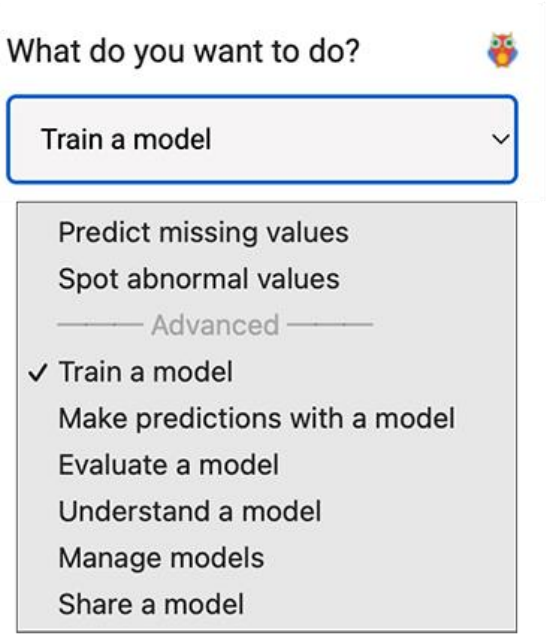
A	B	C	D	E	F	G	H	I	J	K
LongestShell	Diameter	Height	WholeWeight	ShuckedWeight	VisceraWeight	ShellWeight	Type	Rings	Pred:Abnormality:Rings	Pred:MostLikely:Rings
0.7	0.585	0.185	1.8075	0.7055	0.3215	0.475	F	29	0.998132	12.6541
0.495	0.4	0.155	0.8085	0.2345	0.1155	0.35	M	6	0.914463	17.8617
0.61	0.49	0.15	1.103	0.425	0.2025	0.36	M	23	0.886204	11.5834
0.8	0.63	0.195	2.526	0.933	0.59	0.62	F	23	0.872718	11.7654
0.545	0.42	0.14	0.7505	0.2475	0.13	0.255	M	22	0.867308	10.8342
0.55	0.415	0.135	0.775	0.302	0.179	0.26	F	23	0.85802	11.9474

Хотя это и противоречит здравому смыслу, во многих ситуациях нормальным является наличие аномальных значений, и ненормально, чтобы все строки вели себя одинаково. Это означает, что значение с высокой оценкой аномальности не обязательно означает, что значение содержит ошибки или было неверно сообщено.
Выберите заголовок столбца Pred:Abnormality:Rings, щелкнув текст «Pred:Abnormality:Rings».
В меню «Данные» выберите «Сортировать лист» > «Сортировать лист по столбцу J (от Z до A)» .
Обязательно выберите от Z до A , иначе строки со значениями больше 0 в этом столбце будут внизу листа.
Как только лист отсортирован с самыми высокими значениями аномалий вверху, посмотрите на первую строку/животное: у первого морского ушка 29 колец, что очень много. Но модель считает, что у этого морского ушка должно быть только 12 или 13 колец. Другими словами, это морское ушко очень старое, но выглядит гораздо моложе.

Посмотрите на другие столбцы
Продолжайте исследовать аномальные значения для других столбцов (и измените некоторые значения вручную, если хотите).

Под капотом десять разных моделей обучаются с использованием 10-кратного протокола перекрестной проверки. Затем каждое значение в целевом столбце сравнивается с прогнозом соответствующей модели при перекрестной проверке. Если существующее значение и прогнозируемое значение не совпадают, строка считается ненормальной. Подробности смотрите в документации по задаче .

Вы также можете удалить некоторые функции из исходных столбцов или изменить алгоритм обучения: по умолчанию Simple ML использует алгоритм обучения Gradient Boosted Trees.



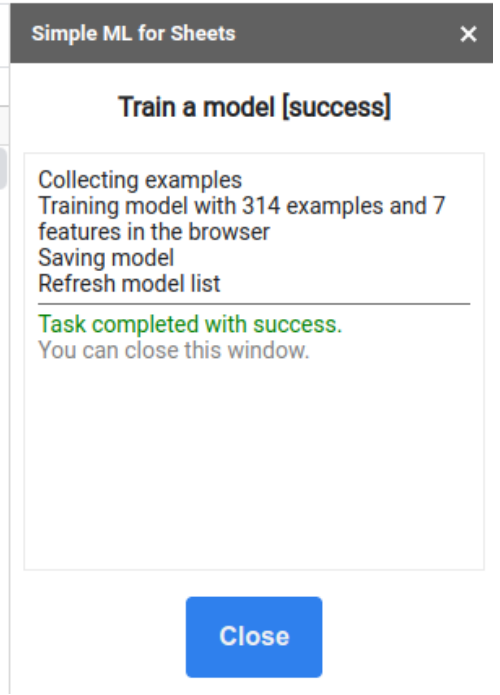
Задача 3: Как обучать/оценивать/интерпретировать и производить модель

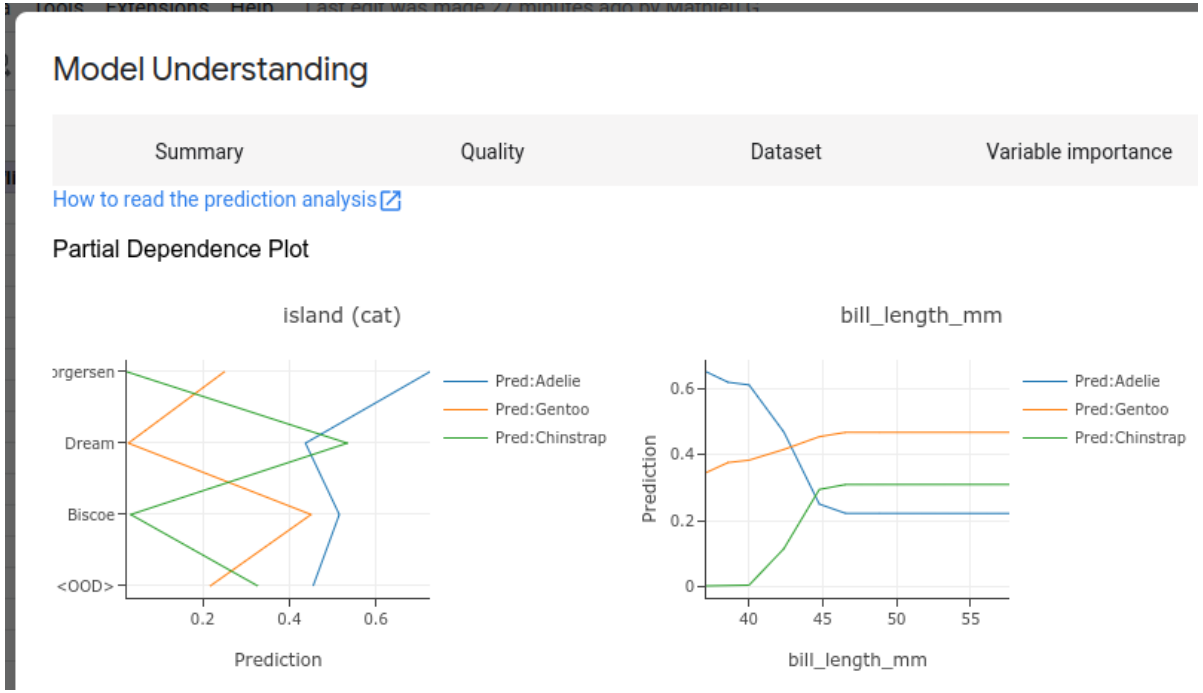
Вы узнали, как использовать Simple ML для прогнозирования отсутствующих значений и выявления отклонений в данных.

Ниже Simple ML создал несколько моделей машинного обучения: одна модель была обучена в задаче № 1, а десять моделей были обучены, но не сохранены в задаче № 2. Однако иногда необходимо создавать, оценивать и использовать модели машинного обучения вручную. В задаче № 3 вы увидите, как обучать, оценивать, анализировать и интерпретировать и, наконец, экспортировать модель машинного обучения в Colab.

Colab — популярная платформа для написания программ с использованием языка программирования Python для обучения машинному обучению. В этом примере модель будет обучена с помощью Simple ML for Sheets, а затем экспортирована в совместную работу для выполнения вывода.

Обучение модели
Первый шаг — обучить модель машинного обучения. Вкладки «Случай № 3: набор данных для обучения» и «Случай № 3: набор данных для оценки» содержат соответственно наборы данных для обучения и тестирования, которые вы будете использовать для обучения и оценки модели.
Выберите вкладку «Случай № 3: набор обучающих данных».
В разделе Что вы хотите сделать? выберите Обучить модель .
Назовите свою модель «Моя первая модель».
В разделе Label выберите виды . Модель попытается предсказать столбец видов.
Пока не меняйте исходные столбцы или дополнительные параметры .
Нажмите Train.
Через несколько секунд модель обучена и готова к использованию. Нажмите кнопку **Закрыть** .





Анализ и интерпретация модели

Иногда интересно понять, что находится внутри модели. Давайте взглянем.

В разделе Что вы хотите сделать? выберите Понять модель .

В разделе «Модели» выберите модель, которую вы только что обучили, под названием «Моя модель».

Установите флажок Включить данные листа .

Нажмите «Понять» .

Через несколько секунд появится окно понимания модели.

На вкладке «Сводка» вы можете увидеть информацию о входных функциях модели.

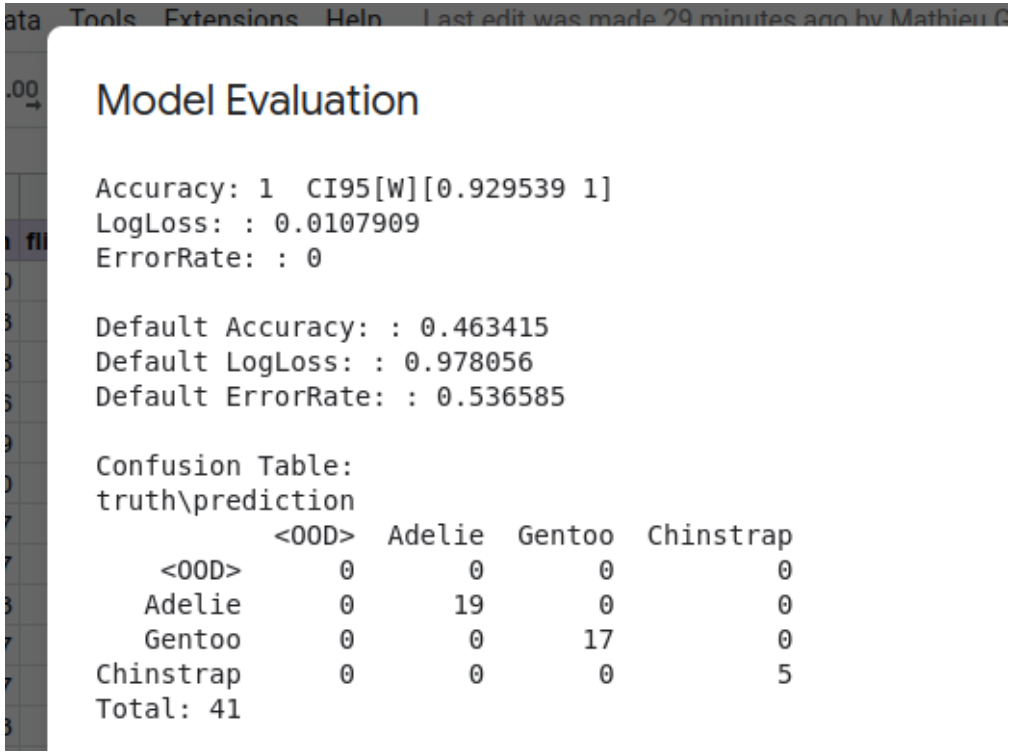
На вкладке «Качество» вы можете увидеть метрики оценки модели. В этом случае оценка была рассчитана на наборе данных проверки, автоматически извлеченном из набора обучающих данных. Различные алгоритмы обучения (в разделе «дополнительные параметры») могут вести себя по-разному.

На вкладке «Набор данных» вы можете увидеть подробную информацию о входных функциях модели.

На вкладке Важность переменных вы можете увидеть, как каждая функция влияет на модель. Например, важность переменной признака MEAN_DECREASE_IN_ACCURACY указывает, насколько «упадет» качество модели, если признак будет удален. Функции с наивысшей важностью являются наиболее важными для модели.

На вкладке Прогнозы вы можете увидеть, как на прогноз модели влияют различные значения признаков.

Наконец, на вкладке Модель графика вы можете увидеть представление модели. Обратите внимание, что на графике отображаются только модели «дерева решений» (которые необходимо выбрать в «дополнительных параметрах» при обучении модели).



Оцените модель
Измерение качества модели на тестовом наборе (также называемом «удерживаемым набором») имеет решающее значение для измерения переобучения модели.

Во время обучения набор проверочных данных автоматически извлекается из набора обучающих данных для управления обучением. Оценка набора данных для проверки представлена в окне «Понимание модели» , показанном на предыдущем шаге. На текущем этапе модель будет оцениваться вручную на новом наборе данных.

Выберите вкладку «Случай № 3: набор данных для оценки».
В разделе Что вы хотите сделать? выберите Оценить модель .
В разделе «Модели» выберите модель, которую вы только что обучили, под названием «Моя модель».
Щелкните Оценить .
Через несколько секунд появится окно оценки модели. Не стесняйтесь прокручивать, чтобы увидеть оценочные графики.

Sharing model

Copy and execute the following code in a Colab to run the model.

```
!pip install tensorflow tensorflow_decision_forests -U -qq

# Transfer the model from Google Drive to Colab
from google.colab import drive
drive.mount("/content/gdrive")
!cp "/content/gdrive/My Drive/simple_ml_for_sheets/My first model" ydf_model

# Prepare and load the model with TensorFlow
import tensorflow as tf
import tensorflow_decision_forests as tfdf

tfdf.keras.yggdrasil_model_to_keras_model("ydf_model", "tfdf_model")
model = tf.keras.models.load_model("tfdf_model")

# Make predictions with the model
examples = {
  "sex" : ["male", "male", "female"],
  "year" : [2009, 2009, 2007],
  "flipper_length_mm" : [190, 228, 182],
  "island" : ["Biscoe", "Biscoe", "Dream"],
  "body_mass_g" : [3900, 5600, 3150],
  "bill_length_mm" : [38.2, 50.8, 36.5],
  "bill_depth_mm" : [20, 17.3, 18],
}
model.predict_step(examples)
```

Экспорт модели

На последнем шаге этого руководства вы экспортируете модель в Colab.
Выберите вкладку «Случай № 3: набор данных для оценки».
В разделе Что вы хотите сделать? выберите Экспорт модели .
В разделе «Модели» выберите модель, которую вы только что обучили, под названием «Моя модель».
В разделе «Назначение» выберите «Colab (внешний)» .

Щелкните Экспорт.

Через несколько секунд появится окно с фрагментом кода Python. Вы можете вставить и запустить этот код в новом документе Colab .
Вы увидите предсказания модели.

Прогнозы модели также можно вычислить в Таблицах с помощью Simple ML с помощью задачи Сделать прогнозы с помощью модели .