

机器学习实验评分和分析教程-财务重点

h2oai.github.io/tutorials/machine-learning-experiment-scoring-and-analysis-tutorial-financial-focus

3.任务1：启动实验

关于数据集

该数据集包含有关“房地美在1999年至2017年之间购买的部分完全摊销的固定利率抵押贷款的贷款水平信贷绩效数据的信息。功能包括人口统计学因素，每月贷款绩效，包括房地产处置的信贷绩效，自愿预付款，MI回收款，非MI回收款，费用，当前递延的UPB和最后一次付款的到期日。”

[1]

[1]我们的数据集是房地美单户贷款水平数据集的子集。它包含500,000行，大约80 MB。

本教程使用的数据集的子集共有27个要素（列）和500,137个借贷（行）。

下载数据集

将Freddie Mac单户贷款级别数据集的H2O子集下载到本地驱动器，并将其另存为csv文件。

loan_level_500k.csv

启动实验

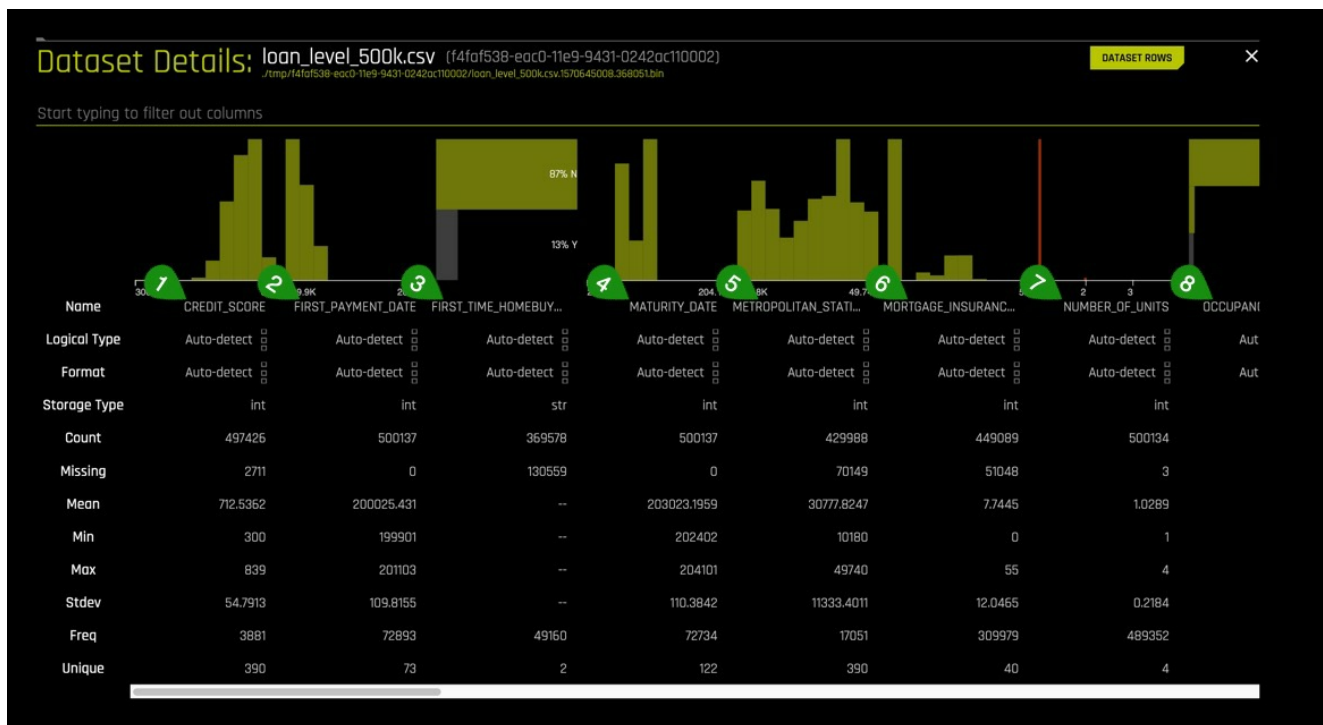
1.通过单击“数据集”概述页面上的“添加数据集”（或拖放），将 loan_level.csv加载到无人驾驶AI。点击上传文件，然后选择loan_level.csv文件。上传文件后，选择详细信息。

The screenshot shows the H2O.ai Datasets page. At the top, there's a navigation bar with links: PROJECTS, DATASETS, AUTOVIZ, EXPERIMENTS, DIAGNOSTICS, MLI, DEPLOYMENTS, RESOURCES, MESSAGES(0), and LOGOUT. Below the navigation bar, the page title is 'Datasets'. A yellow button labeled '+ ADD DATASET (OR DRAG & DROP)' is on the right. The main content is a table of datasets. The first row is highlighted with a green box. It has a checkbox, the name 'loan_level_500k.csv', a path, size '78MB', rows '500K', columns '27', and a status '[Click for Actions]'. Below the table, there are buttons for 'DETAILS', 'VISUALIZE', 'SPLIT', 'PREDICT', 'DOWNLOAD', and 'DELETE'.

<input type="checkbox"/>	Name	Path	Size	Rows	Columns	Status
<input type="checkbox"/>	loan_level_500k.csv	...500k.csv.1570645008.368051.bin	78MB	500K	27	[Click for Actions]
<input type="checkbox"/>	AmazonFineFoodReviews-train-26k	...AmazonFineFoodReviews-train-26k.csv	14MB	26K	11	[Click for Actions]
<input type="checkbox"/>	AmazonFineFoodReviews-test-5k	...AmazonFineFoodReviews-test-5k.csv	3MB	5K	11	[Click for Actions]
<input type="checkbox"/>	walmart_tts_small_test.csv	...walmart_tts_small_test.csv	1MB	16K	11	[Click for Actions]
<input type="checkbox"/>	walmart_tts_small_train.csv	...walmart_tts_small_train.csv	5MB	73K	11	[Click for Actions]

注意：您将看到另外四个数据集，但是您可以忽略它们，因为我们将使用该 `loan_level_500k.csv` 文件。

2. 让我们快速看一下这些列：



注意事项：

- C1-CREDIT_SCORE
- C2-FIRST_PAYMENT_DATE
- C3-FIRST_TIME_HOMEBUYER_FLAG
- C4-MATURITY_DATE
- C5-METROPOLITAN_STATISTICAL_AREA
- C6-MORTGAGE_INSURANCE_PERCENTAGE
- C7-NUMBER_OF_UNITS

3. 继续滚动浏览当前页面以查看更多列（不包括图像）

- C8-OCCUPANCY_STATUS
- C9-ORIGINAL_COMBINED_LOAN_TO_VALUE
- C10-ORIGINAL_DEBT_TO_INCOME_RATIO
- C11-ORIGINAL_UPB
- C12-ORIGINAL_LOAN_TO_VALUE
- C13-ORIGINAL_INTEREST_RATE
- C14-频道
- C15-PREPAYMENT_PENALTY_MORTGAGE_FLAG
- C16-PRODUCT_TYPE

- C17- PROPERTY_STATE
- C18-PROPERTY_TYPE
- C19-POSTAL_CODE
- C20-LOAN_SEQUENCE_NUMBER
- C21-贷款目的**
- C22-ORIGINAL_LOAN_TERM
- C23-NUMBER_OF_BORROWERS
- C24-SELLER_NAME
- C25-SERVICER_NAME
- C26-PREPAID掉落
- C27-DELINQUENT-此列是我们感兴趣的标签，用于预测False->未默认和True-> defaulted的位置

4.返回到数据集概述页面

5.单击loan_level_500k.csv文件，然后拆分

The screenshot shows the H2O.ai Datasets interface. At the top, there's a navigation bar with links like PROJECTS, DATASETS, AUTOVIZ, etc. Below the header, a table lists datasets. The first row, 'loan_level_500k.csv', is circled in green. A context menu is open for this dataset, and the 'SPLIT' option is highlighted with a green box and a mouse cursor. Other datasets listed include 'AmazonFineFoodReview', 'walmart_tts_small_test.csv', and 'walmart_tts_small_train.csv'.

Name	Path	Size	Rows	Columns	Status
loan_level_500k.csv	...500k.csv.1570645008.368051.bin	78MB	500K	27	[Click for Actions]
AmazonFineFoodReview	...AmazonFineFoodReviews-train-26k.csv	14MB	26K	11	[Click for Actions]
AmazonFineFoodReview	...AmazonFineFoodReviews-test-5k.csv	3MB	5K	11	[Click for Actions]
walmart_tts_small_test.csv	...walmart_tts_small_test.csv	1MB	16K	11	[Click for Actions]
walmart_tts_small_train.csv	...walmart_tts_small_train.csv	5MB	73K	11	[Click for Actions]

6.将数据分为两组：**freddie_mac_500_train**和**freddie_mac_500_test**。使用下面的图片作为指导：

The screenshot shows the 'Dataset Splitter' window. It has two columns for input fields. The first column contains 'OUTPUT NAME 1' with the value 'freddie_mac_500_train' (callout 1), 'TARGET COLUMN' with 'DELINQUENT' (callout 3), 'TIME COLUMN' with '--', and 'SELECT SPLIT RATIO' with a slider set to 0.75 (callout 5). The second column contains 'OUTPUT NAME 2' with 'freddie_mac_500_test' (callout 2), 'FOLD COLUMN' with '--', 'RANDOM SEED' with '42' (callout 4), and a '75% Split' label above the slider. At the bottom are 'SAVE' (callout 6) and 'CANCEL' buttons. A status bar at the bottom indicates '(375K | 125K SAMPLES SPLIT)'.

注意事项：

1. 为输出名称1键入“freddie_mac_500_train”，它将用作训练集
2. 为输出名称2输入“freddie_mac_500_test”，它将用作测试集
3. 对于目标列，选择拖欠
4. 您可以将“随机种子”设置为任意数字，我们选择42，通过选择随机种子，我们将获得一致的分割
5. 通过将滑块调整为75%或在显示“火车/有效拆分比率”的部分中输入.75，将拆分值更改为.75
6. 保存

训练集包含37.5万行，每行代表一笔贷款，而27列代表每笔贷款的属性，包括具有我们要尝试预测的标签的列。

注意：训练和测试拆分中的实际数据因用户而异，因为数据是随机拆分的。测试集包含125k行，每行代表一个借贷，以及27个属性列，代表每个借贷的属性。

7.验证是否存在三个数据

集，freddie_mac_500_test，freddie_mac_500_train和loan_level_500k.csv：

H2O.ai
DRIVERLESS AI 1.8.0 - AI TO DO AI
Licensed to H2O Aquarium (SN35079 - Evaluation License). Current User - ADMIN

PROJECTS DATASETS AUTOVIZ EXPERIMENTS DIAGNOSTICS MLI DEPLOYMENTS RESOURCES ▾ MESSAGES(0) LOGOUT

Datasets + ADD DATASET (OR DRAG & DROP)

<input type="checkbox"/>	Name ↕	Path ↕	Size ↕	Rows ↕	Columns ↕	Status ↕
<input type="checkbox"/>	freddie_mac_500_test	...c_500_test.1570645820.6505.bin	20MB	125K	27	[Click for Actions]
<input type="checkbox"/>	freddie_mac_500_train	...0_train.1570645821.1046062.bin	59MB	375K	27	[Click for Actions]
<input type="checkbox"/>	loan_level_500k.csv	...500k.csv.1570645008.368051.bin	78MB	500K	27	[Click for Actions]
<input type="checkbox"/>	AmazonFineFoodReviews-train-26k.csv	...nFineFoodReviews-train-26k.csv	14MB	26K	11	[Click for Actions]
<input type="checkbox"/>	AmazonFineFoodReviews-test-5k.csv	...zonFineFoodReviews-test-5k.csv	3MB	5K	11	[Click for Actions]
<input type="checkbox"/>	walmart_tts_small_test.csv	...art/walmart_tts_small_test.csv	1MB	16K	11	[Click for Actions]
<input type="checkbox"/>	walmart_tts_small_train.csv	...rt/walmart_tts_small_train.csv	5MB	73K	11	[Click for Actions]

8.单击**freddie_mac_500_train**文件，然后选择**Predict**。

9. 在“**第一次无人驾驶AI**”上选择“**不立即**”，单击“**是**”进行浏览！。将会出现类似的图像：

H2O.ai Experiment
DRIVERLESS AI 1.8.0 - AI TO DO AI
Licensed to H2O Aquarium (SN35079 - Evaluation License). Current User - ADMIN

PROJECTS DATASETS AUTOVIZ EXPERIMENTS DIAGNOSTICS MLI DEPLOYMENTS RESOURCES ▾ MESSAGES(3) LOGOUT

EXPERIMENT SETUP

ASSISTANT

DISPLAY NAME
 Display name

DATASET
 freddie_mac_500_train

ROWS
 375K

COLUMNS
 27

DROPPED COLUMNS
 --

VALIDATION DATASET
 --

TEST DATASET
 --

TARGET COLUMN
 Select target column

FOLD COLUMN
 --

WEIGHT COLUMN
 --

TIME COLUMN
 [OFF]

为实验命名 **Freddie Mac Classification Tutorial**

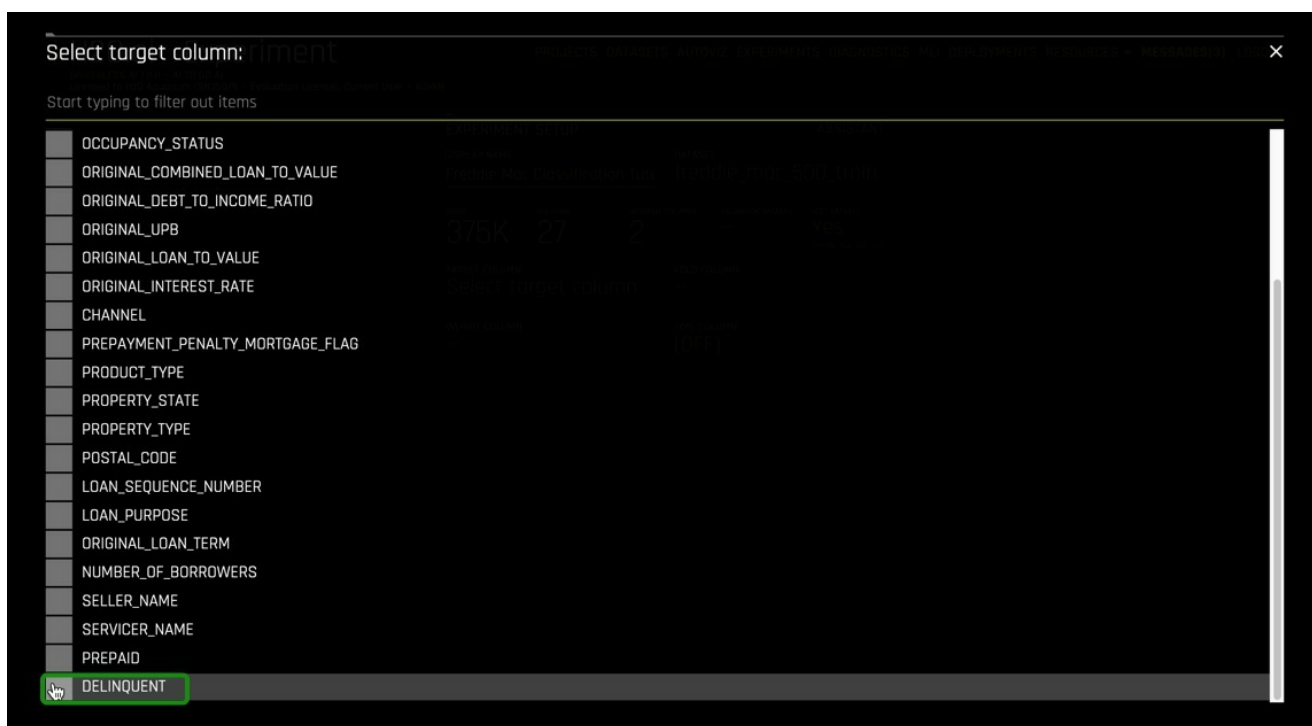
10.选择**Dropped Cols**，拖放以下两列：

- 预付款_罚金_抵押_标志
- 预付
- 选择**完成**

删除这两个列是因为它们都是贷款的明显指标，它们将拖欠还款并导致数据泄漏。



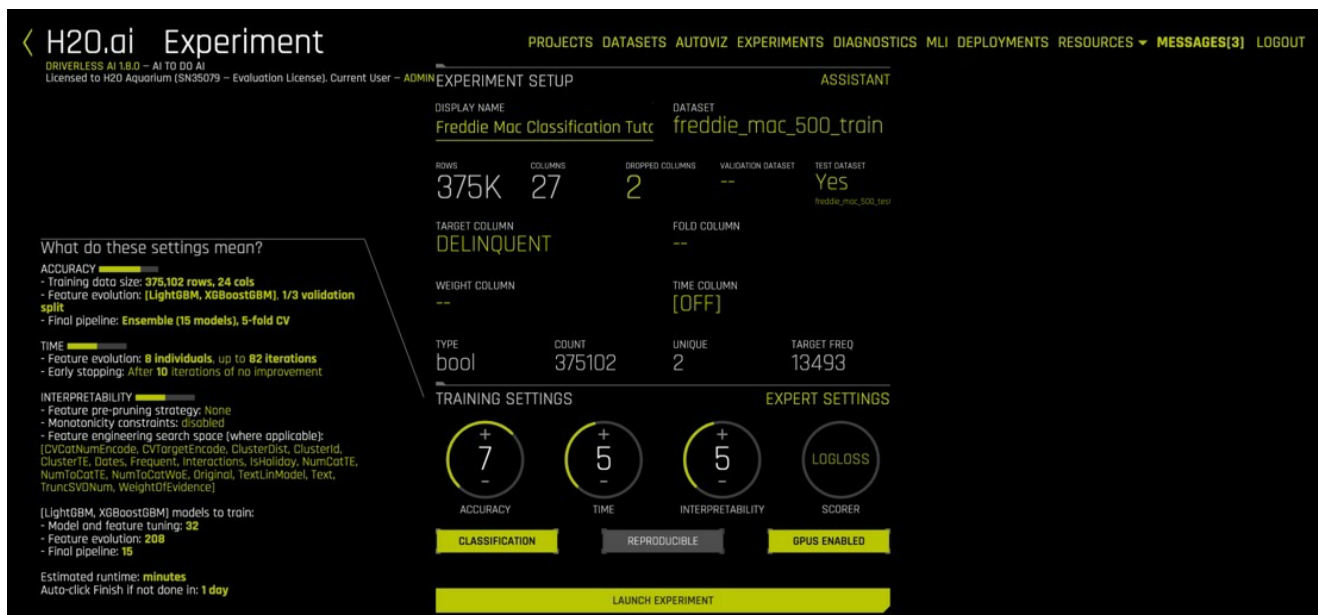
11.选择Target Column，然后选择Delinquent



12.选择Test Dataset，然后选择freddie_mac_500_test



13.将会出现类似的实验页面：



在任务2中，我们将探索和更新“实验设置”。

背部下一个