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）。

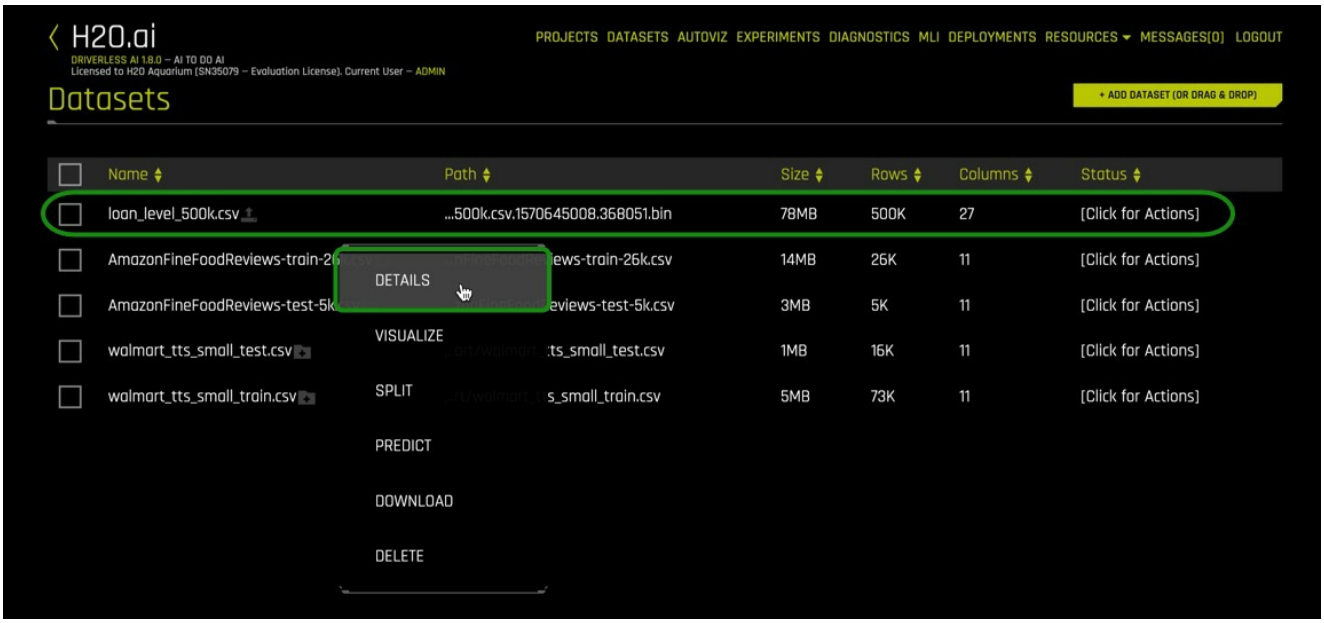
下载数据集

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**。



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：

Name	CREDIT_SCORE	FIRST_PAYMENT_DATE	FIRST_TIME_HOMEBUY...	MATURITY_DATE	METROPOLITAN_STATI...	MORTGAGE_INSURANC...	NUMBER_OF_UNITS	OCCUPANI
Logical Type	Auto-detect	Auto-detect	Auto-detect	Auto-detect	Auto-detect	Auto-detect	Auto-detect	Aut
Format	Auto-detect	Auto-detect	Auto-detect	Auto-detect	Auto-detect	Auto-detect	Auto-detect	Aut
Storage Type	int	int	str	int	int	int	int	
Count	497426	500137	369578	500137	429988	449089	500134	
Missing	2711	0	130559	0	70149	51048	3	
Mean	712.5362	200025.431	--	203023.1959	30777.8247	7.7445	1.0289	
Min	300	199901	--	202402	10180	0	1	
Max	839	201103	--	204101	49740	55	4	
Stdev	54.7913	109.8155	--	110.3842	11333.4011	12.0465	0.2184	
Freq	3881	72893	49160	72734	17051	309979	489352	
Unique	390	73	2	122	390	40	4	

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

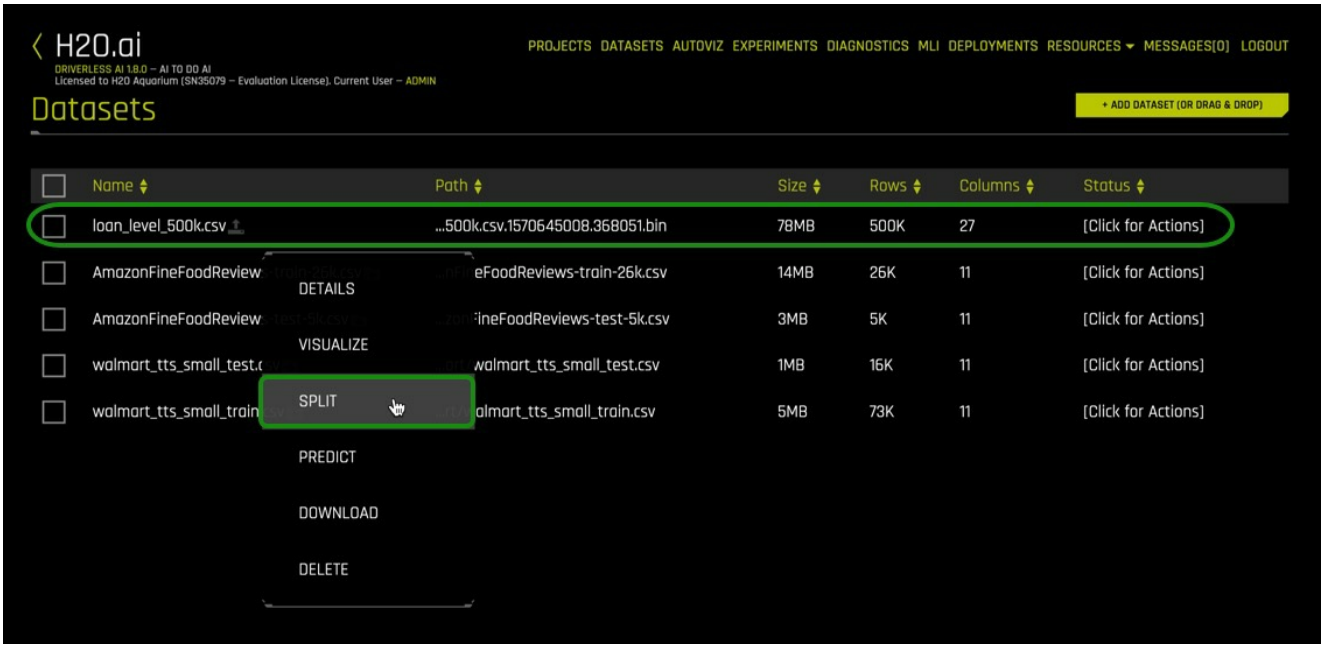
3.继续滚动浏览当前页面以查看更多列（不包括图像）

- 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

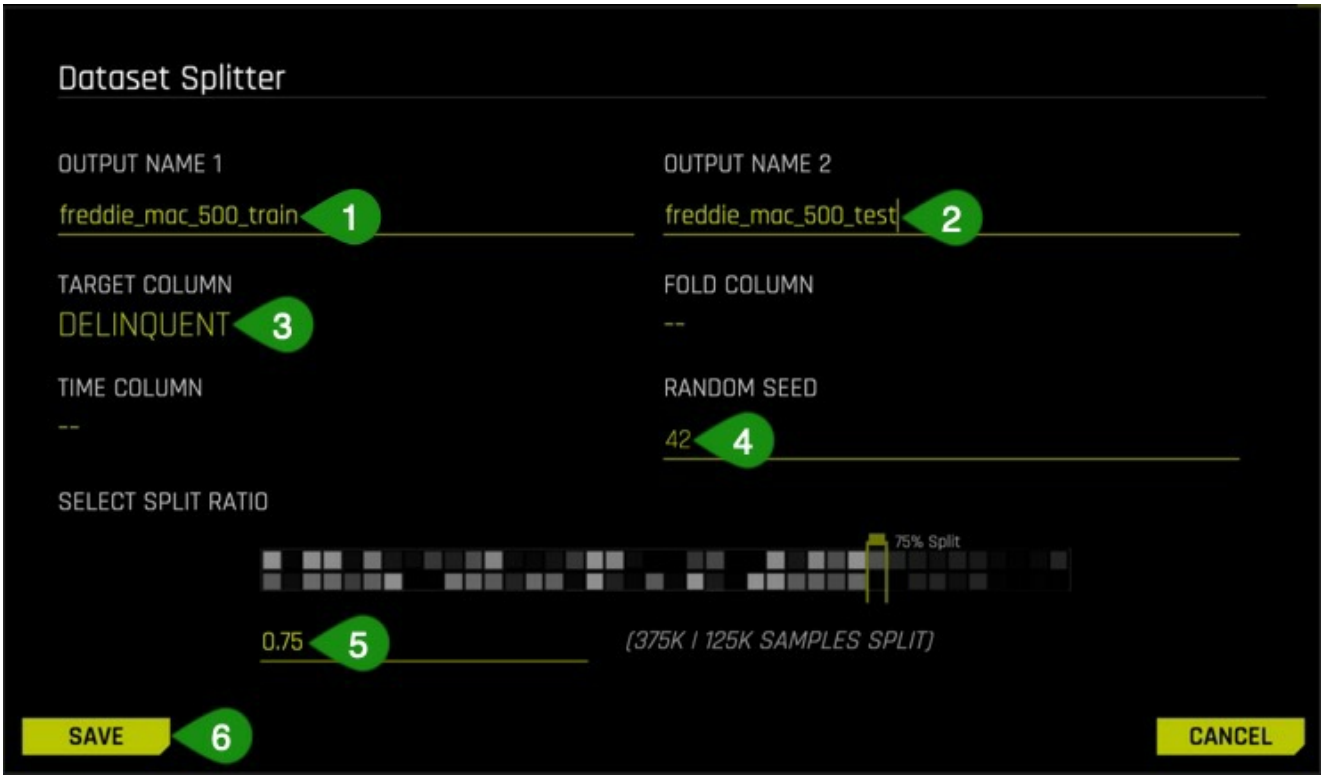
- C17- PROPERTY_STATE
- C18-PROPERTY_TYPE
- C19-POSTAL_CODE
- C20-LOAN_SEQUENCE_NUMBER
- C21-**贷款目的****
- C22-ORIGINAL_LOAN_TERM
- C23-NUMBER_OF_BORROWERS
- C24-SELLER_NAME
- C25-SERVICER_NAME
- C26-PREPAID**掉落**
- C27-DELINQUENT-**此列是我们感兴趣的标签，用于预测**False->**未默认和**True-> defaulted **的位置**

4.**返回到数据集概述页面**

5.**单击loan_level_500k.csv文件，然后拆分**



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：**



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**：

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：

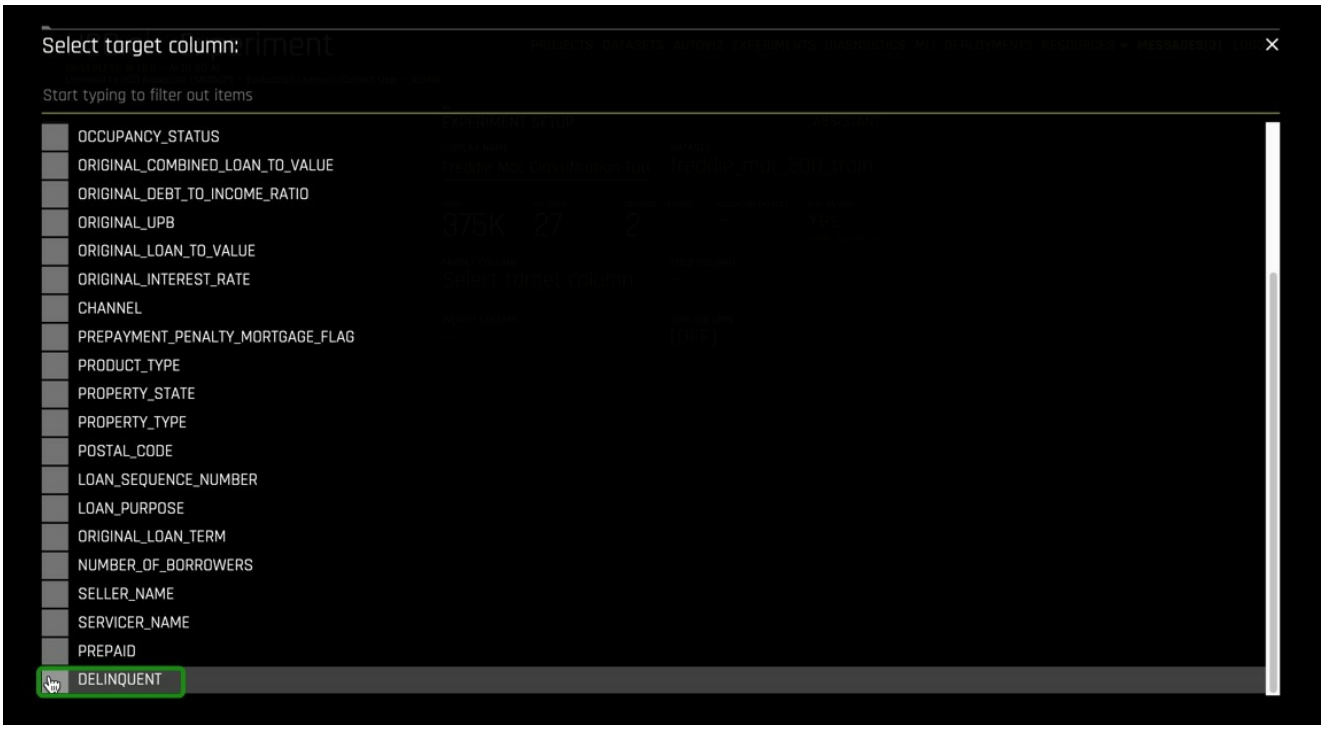
为实验命名 Freddie Mac Classification Tutorial

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**

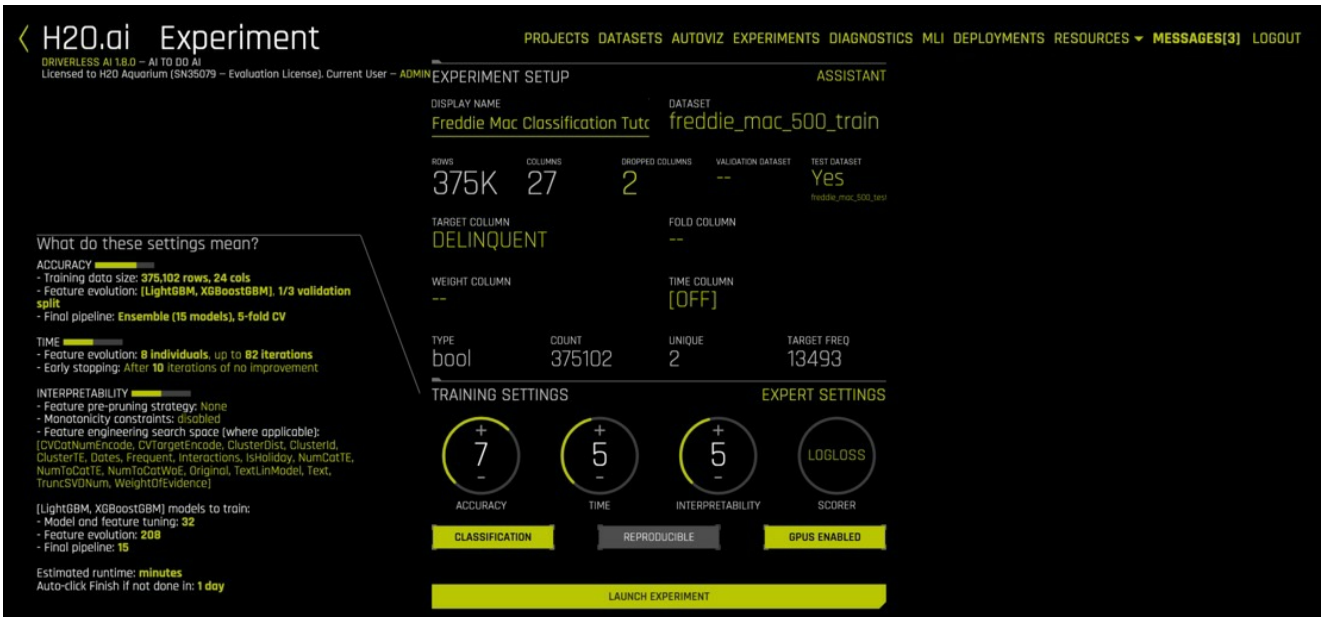
删除这两个列是因为它们都是贷款的明显指标，它们将拖欠还款并导致数据泄漏。

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