

文章

[Jingwei Wang](#) · 五月 4, 2022 阅读大约需 8 分钟

IntegratedML 实践实验室

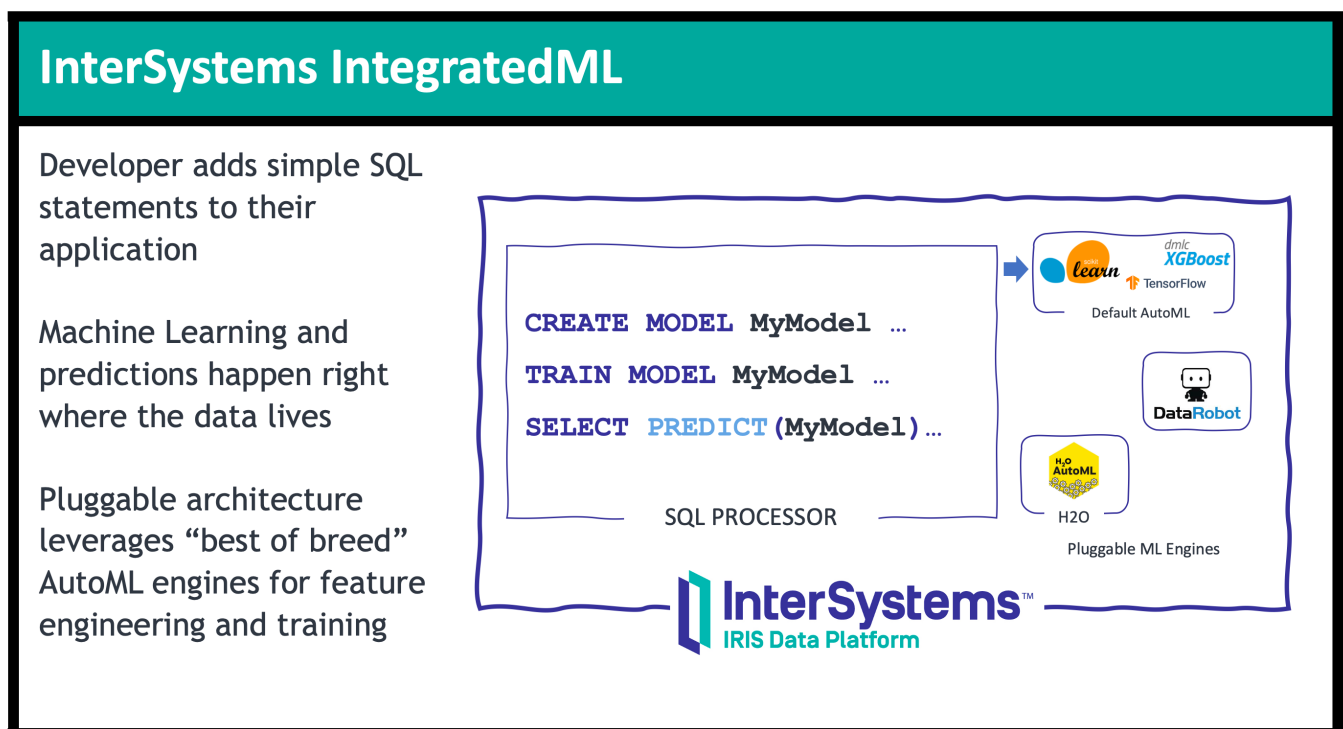
你是否尝试过[InterSystems IRIS IntegratedML学习平台](#)

？在这个平台中，你可以在再入院数据集上训练和测试一个模型，并能够预测一个病人何时会再入院，或计算其再入院的概率。

你不需要在你的系统上进行任何安装就可以尝试，你所要做的就是启动一个虚拟实验室环境（Zeppelin），然后玩一玩！

在这篇文章中，我们将利用这个实验室向你简要介绍IntegratedML，向你介绍要处理的问题，如何使用IntegratedML来创建一个再入院预测模型，以及如何分析其性能指标的一些见解。

什么是IntegratedML？



来源: <https://github.com/intersystems-community/integratedml-demo-template>

在开始本教程之前，让我们简单谈谈IRIS IntegratedML。这个工具使你能够直接在SQL语句中执行机器学习（ML）任务，抽象出复杂过程的实现，例如，选择哪些列和ML算法是对目标列进行分类或回归的最佳选择。

IntegratedML的另一个伟大功能是易于部署。一旦你的模型被训练并表现良好，你只需要运行SQL语句，以便让你的模型投入生产。

IntegratedML让你选择使用哪个 [ML提供者](#)。默认的提供者是 [AutoML](#)，这是InterSystems公司使用Python实现的ML库 [scikit-learn](#)。但是你也可以选择其他三个提供者: PMML, [H2O](#) and [DataRobot](#)。

在这篇文章中，我们将使用AutoML。

问题和解决方案

现在，让我们来介绍一下这个问题，以及我们如何努力提出一种方法来尽量减少其影响。我们将尝试减少由于再入院而产生的问题。

根据[维基百科](#)，医院再入院是指病人被批准离开医院（出院），但此人在一个意想不到的短时间间隔内再次回到医院（再入院）的事件。

来源: <https://news.yale.edu/2015/01/15/when-used-effectively-discharge-summar...>

再入院会造成病人护理质量的损失--出院和再入院之间的时间可能很关键，医院的资源也会优化。

克服这个问题的一个方法是尝试使用历史数据库来创建一个数据集，其中过去的再入院事件可以通过机器学习算法进行分析，创建一个ML模型。如果数据集足够丰富和干净，再入院模式可以被正确检测出来，新的事件可以通过这个模型计算出其概率。

因此，通过使用ML模型来预测再入院，能够避免错误的出院，肯定会成为医院提高服务质量和利润的一个有价值的选择。

使用IntegratedML创建和使用一个ML模型

模型创建

使用IntegratedML创建一个ML模型，就像执行一个SQL语句一样容易。你只需要定义数据所在的历史数据集，并为你的模型起一个名字。

```
CREATE MODEL Readmission PREDICTING (MxWillReAdmit) FROM EncountersHistory
```

After that statement, you have declared and created a model called Readmission intended to predict the values for a column called MxWillReAdmit, based on a dataset called EncountersHistory.

You can find more information about this statement [here](#). For example, in design phase, it ' s handy forces the same results on training, so you can use the USING argument with some arbitrary constant number, like this:

在该声明之后，你已经声明并创建了一个名为Readmission的模型，基于名为EncountersHistory的数据集，预测名为MxWillReAdmit的列的值。

你可以找到关于这个声明的更多信息，[点击这里](#)

。例如，在设计阶段，很方便地迫使训练中出现相同的结果，所以你可以用USING参数加上一些任意的常数，像这样。

```
CREATE MODEL Readmission PREDICTING (MxWillReAdmit) FROM EncountersHistory USING {"seed": 3}
```

模型训练

现在，通过使用TRAIN MODEL语句，你的模型已经准备好被训练。

```
TRAIN MODEL Readmission
```

在这里，IntegratedML为你做了很多工作（使用AutoML ML供应商）：

- [Feature engineering](#): 使用哪些列
- [Data encoding](#): 所选栏目中的数据必须如何呈现给ML算法
- [ML algorithm selection](#): 哪种算法能带来最好的结果
- [Model selection](#): 根据目标列的数据类型，应该选择一个分类模型或回归模型。在这个例子中，一旦目标列有布尔值，就会选择一个分类模型。

这个声明可能需要一些时间，取决于数据集的大小和你的数据的复杂性。

模型验证

到现在，我们应该急于知道我们的模型到底有多好。Integrated ML有一个计算性能指标的声明，基于另一个数据集，而不是用于训练的数据集

这个新的数据集被称为测试或验证数据集，这取决于你使用哪种策略进行验证。这个数据集通常是从用于创建训练数据集的同一个数据集中检索出来的。一个常见的方法是随机选择70%或80%的数据集用于训练，其余的用于测试/验证。

在实验室里，之前已经为这项任务准备了一个新的可使用的数据集：EncountersNew数据集。

现在，我们可以找出我们的模型有多好（或多坏）。

```
VALIDATE MODEL Readmission FROM EncountersNew
```

为了得到结果，你应该查询 INFORMATION_SCHEMA.ML_VALIDATION_METRICS 表。

```
SELECT * FROM INFORMATION_SCHEMA.ML_VALIDATION_METRICS
```

模型的性能是用 [IntegratedML的四个指标](#):

- Accuracy: 预测正确率（数值接近1表示正确率高）。
- Precision: 对模型所做的所有阳性预测的正确率（接近1的值意味着很少有假阳性预测）。
- Recall: 关于数据集中所有实际阳性值的正确阳性预测率（接近1的值意味着很少有假阴性预测）。
- F-Measure:
另一种衡量准确率的方法，当准确率表现不佳时使用，一般用于不平衡问题（数值接近1意味着高正确率）。

关于性能指标的一些讨论

在这里，我们可以看到，该模型的准确率为82%。但是，这个指标不应该被单独分析，其他的指标如精确度和召回率也必须被评估。

让我们思考一下错误预测的影响--假阳性和假阴性。对于我们的模型来说，假阳性意味着预测的再入院并不是实际的再入院。而假阴性，意味着一个病人已经再入院，而模型预测这个病人不会再入院。

这两种情况都会导致错误的决定。一个假阳性的预测可能会导致一个超过必要的病人留在医院的决定；而一个假阴性的预测可能会导致一个提前出院的决定，然后再让这个病人重新入院。

请注意，在我们的模型中，假阴性会导致再入院病例，这正是我们要避免的。所以，我们必须选择能减少假阴性病例的模型，即使假阳性病例的数量增加。

因此，为了选择具有低假阴性率的模型，我们需要选择具有高召回率的模型。这是因为召回率越高，假阴性就越低。

由于我们的模型有84%的召回率，让我们暂时假定它是一个合理的值。

预测

一旦你训练好了你的模型，而且它的性能可以接受，现在你就可以执行这个模型，以便预测结果。

你应该使用 [PREDICT](#) 函数，以便用户对模型进行预测。

```
SELECT TOP 100
    ID,
    PREDICT(Readmission) AS PredictedReadmission,
    MxWillReAdmit AS ActualReadmission
FROM
    EncountersNew
```

这个函数根据内部阈值，评估模型的目标列的行是真值还是假值的概率。

你可以看到我们的模型做了一些错误的预测--这很正常。

另一个可以用来进行预测的函数，是 [PROBABILITY](#) 函数。

```
SELECT TOP 10
    ID,
    PROBABILITY(Readmission FOR '1') AS ReadmissionProbability,
    PREDICT(Readmission) AS PredictedReadmission,
    MxWillReAdmit AS ActualReadmission
FROM
    EncountersNew
```

与PREDICTION相同，PROBABILITY使用一行的列作为指定模型的输入，但在这里我们需要定义我们想要计算的概率是哪一类。

如果我们想定制用于预测的阈值，这个函数可能很有用。

查询训练好的模型

在创建和训练你的模型后，你可以使用 INFORMATION_SCHEMA.ML_TRAINED_MODELS 表来查询它们。

```
SELECT * FROM INFORMATION_SCHEMA.ML_TRAINED_MODELS
```

结论

在这篇文章中，我们对历史上的病人再入院数据集进行了创建、训练、验证和预测。这些预测可以作为寻求更好决策的另一个工具，从而改善对病人的护理，降低医院服务的成本。

所有这些步骤都是通过InterSystems IntegratedML使用SQL语言完成的，它为广大的开发者带来了机器学习的力量。

[#IntegratedML](#) [#InterSystems IRIS](#)

源

URL:

<https://cn.community.intersystems.com/post/integratedml-%E5%AE%9E%E8%B7%B5%E5%AE%9E%E9%AA%8C%E5%AE%A4>
