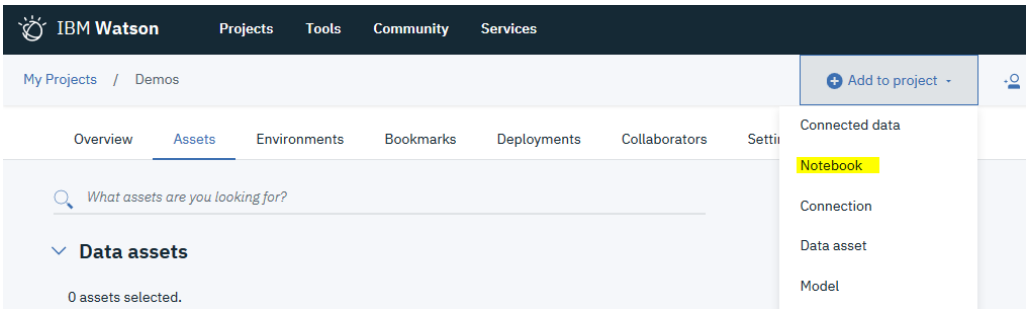
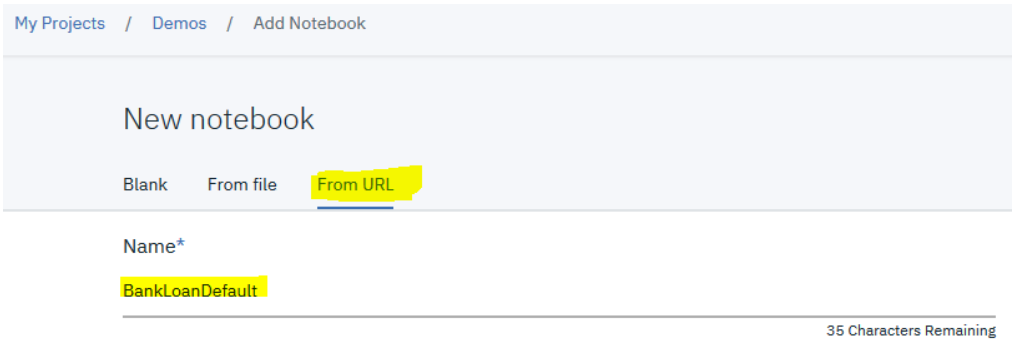
银行贷款违约预测 DSX 动手实践指导手册

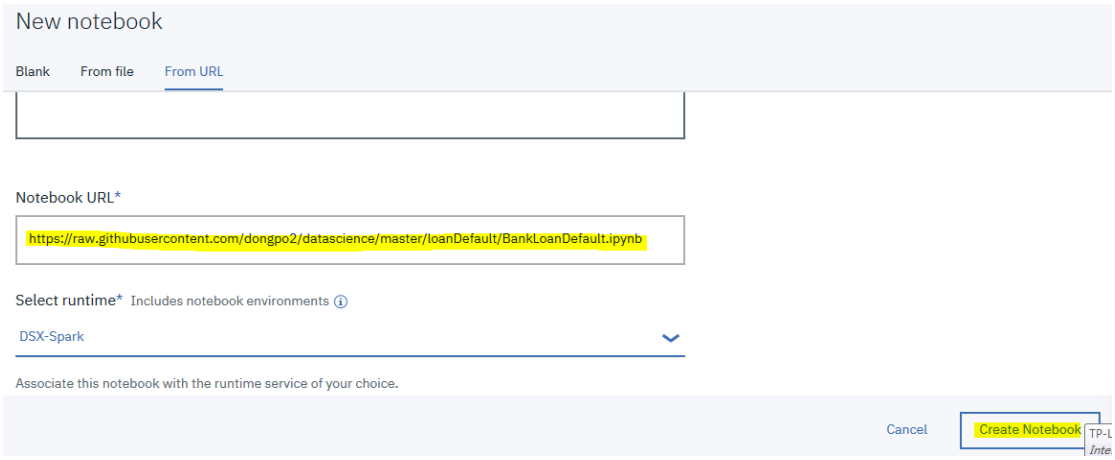
1. 登录 DSX，新创建一个 Project，如 Demos
   a) 点击右上方”Add to project”，选择”Notebook”



   b) 选择”From URL”，在名字处输入如”BankLoanDefault”



   c) 在下面的”Notebook URL”输入 https://raw.githubusercontent.com/dongpo2/datascience/master/loanDefault/BankLoanDefault.ipynb
   然后，点击”Create Notebook”



   d) 可以直接在 Notebook 中下载数据文件，如果直接展现，可以把内容保存为一个叫做 bankloan.csv 的文件

银行客户贷款违约预测

这里是利用IBM DSX来对银行客户贷款的数据集进行建模预测未来违约的可能性。

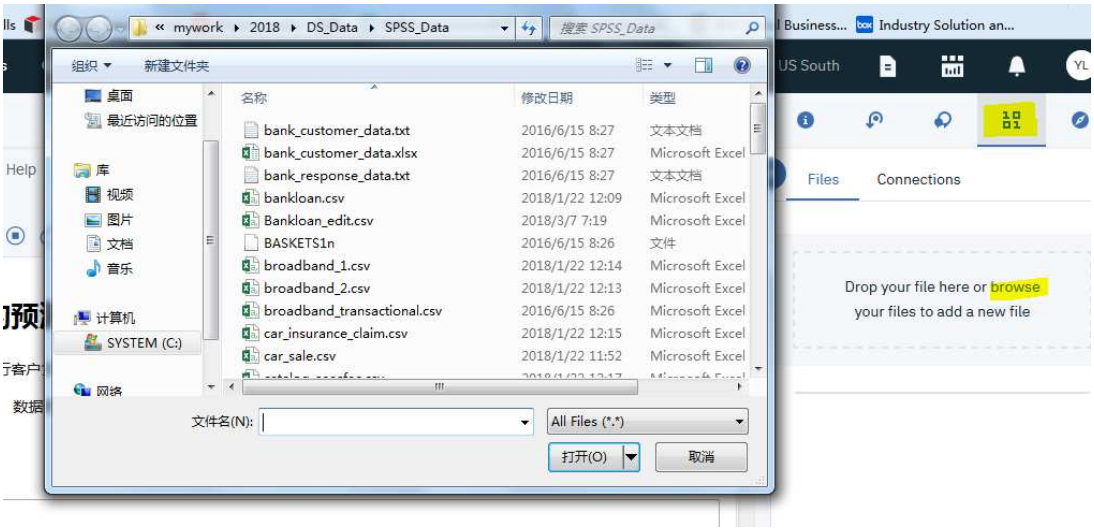
这个文档包括了从数据加载、数据理解、数据处理准备、建模、模型部署与API测试。

1. 加载数据

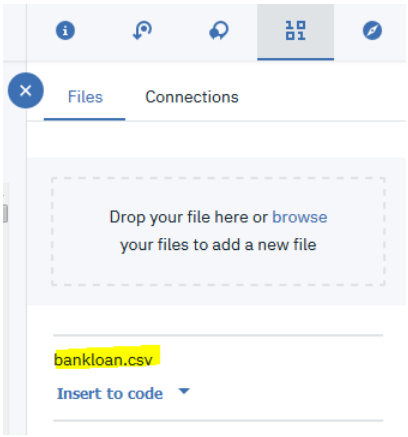
从github下载数据文件，可以采用2种方式(二选一，不同时运行)：

1. 下载数据文件到本地，然后直接从本地装载文件到DSX
   - URL:https://raw.githubusercontent.com/dongpo2/datascience/master/loanDefault/bankloan.csv
2. 或者，使用wget直接从github下载到DSX，然后保存到cloud-object-storage

e) 点击右上侧的”Data”按钮，出现右侧操作栏的”Files”页签下，点击”browse”选择下载的文件 bankloan.csv



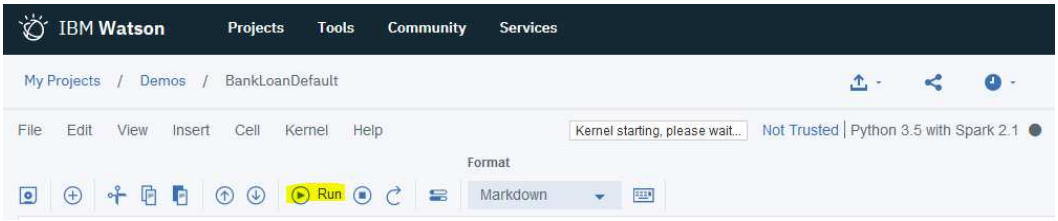
f) 上传完成后，出现以下



2. 运行程序

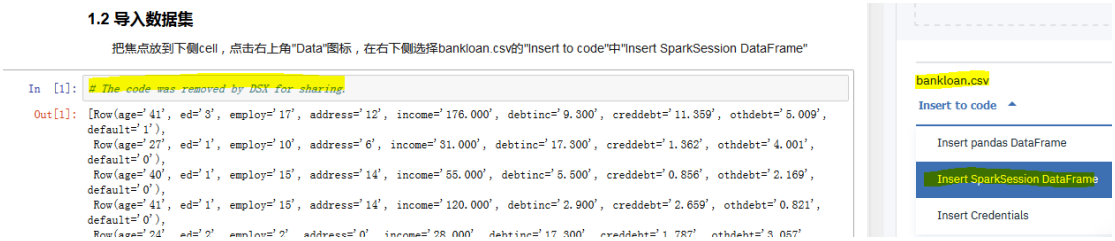
a) 运行程序可以使用鼠标点击工具栏上的”Run”图标，或者菜单 Cell 中各种运行命令。

一般交互式只运行当前焦点停留的 Cell。也可以使用快捷键 Ctl+Enter 运行。



b) 导入数据集

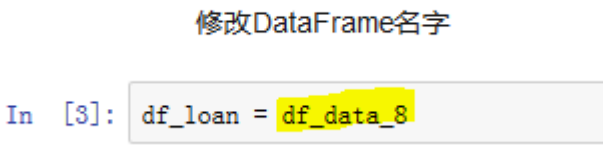
焦点选择到 1.2 节第一个 Cell，并在右侧”Data”操作栏，选择 bankloan.csv Insert to code 下 Insert SparkSession DataFrame



会自动插入导入数据集代码，然后运行，查看结果。

在运行状态 Cell 左侧会显示”*”，运行完毕会显示数字。

继续运行下一个 Cell 前，需要根据插入代码的变量名修改



c) 根据 Notebook 中的解释理解并运行，一直到 2.4 节完成。自己练习时，可以根据示例程序修改或者添加自己希望实现的内容

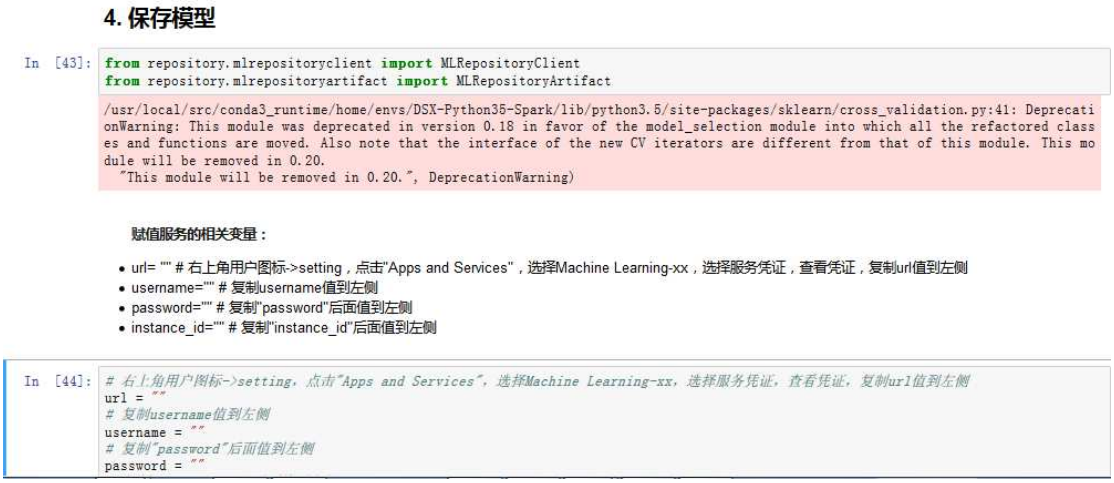
d) 保存数据集，在 2.5 节第一个 Cell，选中，然后在右侧”Data”操作栏，插入 pandas 代码 Insert pandas DataFrame

然后根据 notebook 中的提示，分别修改再下一个 Cell 中的 BUCKET 变量值，以及写入 csv 文件的函数名(原来是 client_manual)。

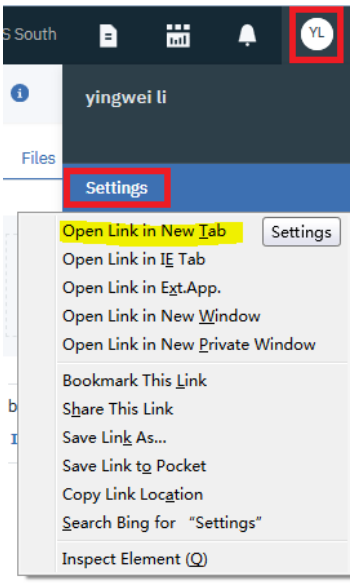
根据 notebook 中的提示，删除不需要的代码，然后继续运行 Cell。

e) 建模阶段，根据提示理解建模过程，逐个运行 Cell，一直到 3 章节结束

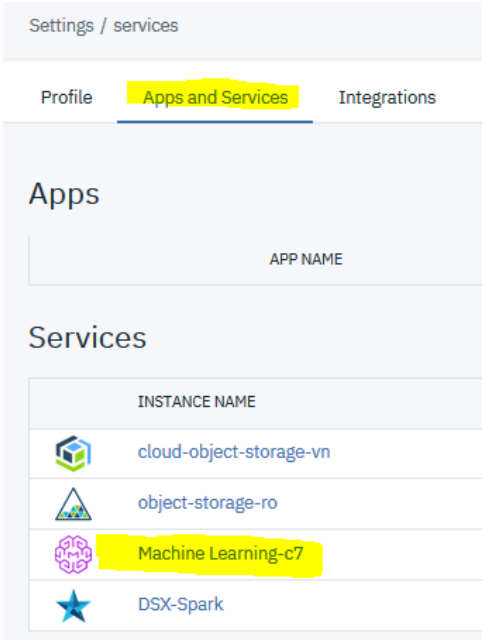
f) 保存模型，运行完 4 章节的第一个 Cell，我们需要对 4 变量赋值



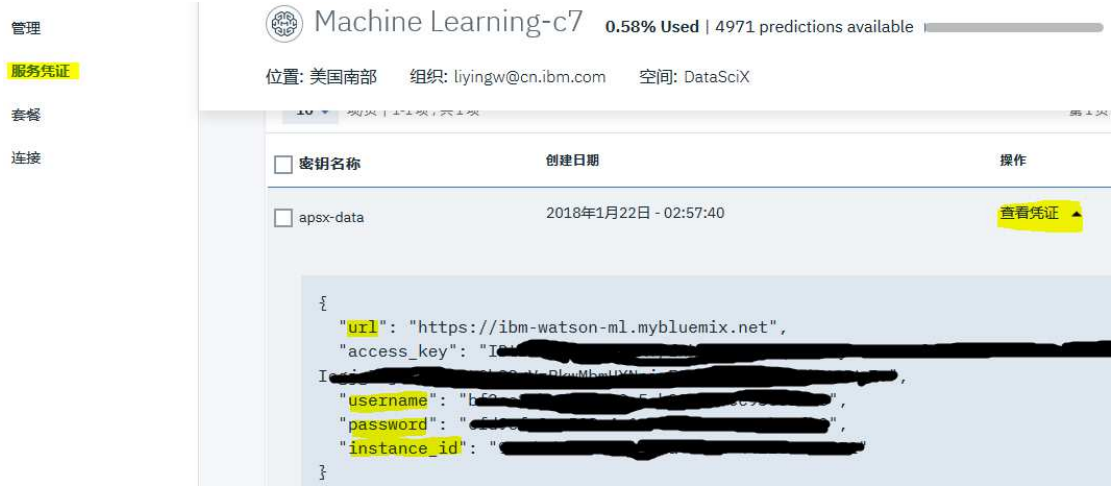
点击右上方用户图标，鼠标移动到下拉菜单的”Setting”，右键选择”Open Link in New Tab”，新打开一个标签页，管理



在新的浏览器页面上，选择”App and Services”，下面会列出使用的服务，选择点击“Machine-Learning-xx”进入服务管理页。



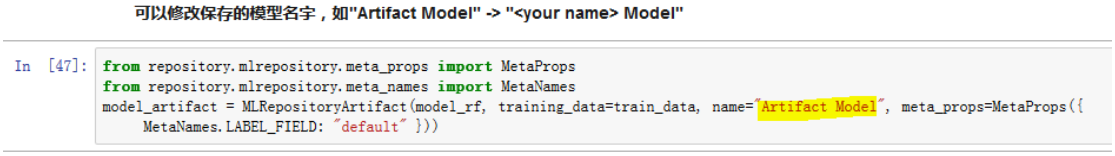
点击左侧的“服务凭证”，然后再点击“查看凭证”展开内容。



分别复制四项值到 Notebook 中对应的变量值。

运行后续 Cell 完成模型保存。

可以在保存时，修改为自己希望的名字



g) 模型部署

运行模型部署相关 Cell，其中可以修改为自己希望的名字

可以修改部署模型名称

```
In [57]: payload_online = {"name": "Bank Loan Default Prediction", "description": "My Deployment", "type": "online"}
response_online = requests.post(endpoint_deployments, json=payload_online, headers=header)

print(response_online)
#print(response_online.text)
```

h) 测试，可以在 Notebook 中完成测试，

6. 测试模型API

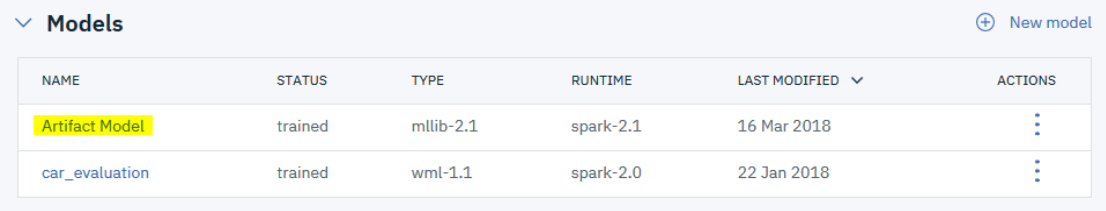
```
[81]: payload_scoring = {"fields": ["age","ed", "address","employ","income","debtinc","creddebt","othdebt"],"values": [[33,1,5,8,
response_scoring = requests.post(scoring_url, json=payload_scoring, headers=header)

print(response_scoring.text)
```


```
{
  "fields": ["age", "ed", "address", "employ", "income", "debtinc", "creddebt", "othdebt", "features", "rawPrediction",
"probability", "prediction", "predictedLabel"],
  "values": [[33, 1, 5, 8, 116.0, 15.0, 5.0, 10.0, [33.0, 1.0, 8.0, 5.0, 116.0, 15.0, 5.0, 10.0], [10.587353053077383, 9
.412646946922617], [0.5293676526538691, 0.47063234734613085], 0.0, "not default"], [31, 3, 3, 3, 52.0, 20.0, 15.0, 5.0,
[31.0, 3.0, 3.0, 3.0, 52.0, 20.0, 15.0, 5.0], [8.641658162473622, 11.358341837526378], [0.4320829081236811, 0.5679170918
763189], 1.0, "default"]]
}
```

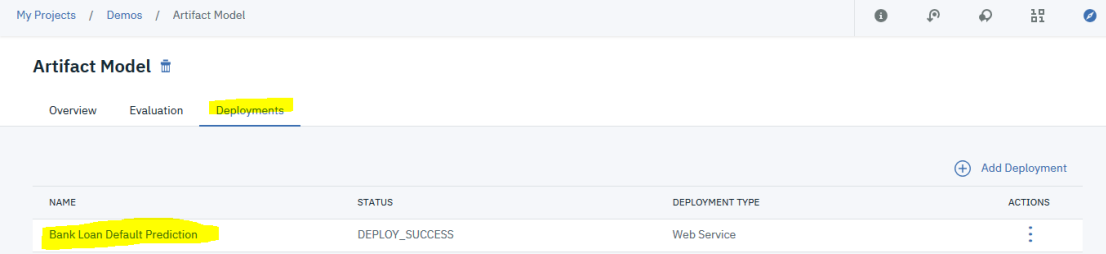

修改数据，观察返回结果。

i) 也可以在 DSX 中进行测试

在项目首页，可以看到自己保存的模型



点击自己的模型，进入管理页，选择”Deployment”页签，出现已经部署的模型。

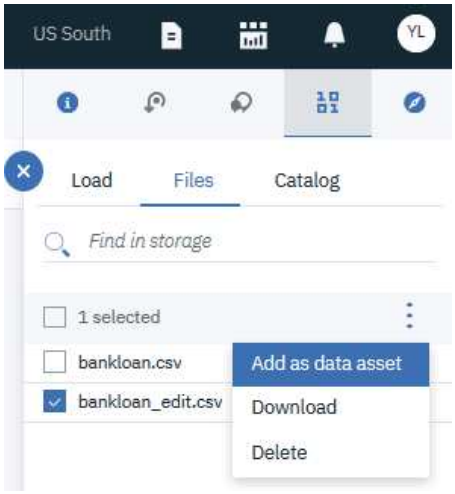


点击部署名，选择”Test”页签，输入数据，点击”Predict”，可以得到预测结果，可以显示为图形和文本两种方式。

3. 使用自动化建模工具进行建模

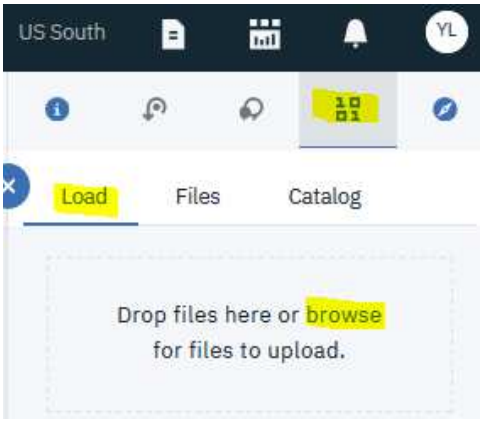
a) DSX 支持不编程模式的建模方式，如果已经完成 Notebook，或至少完成其中 2.5 章节部分的保存数据集

i. 在项目首页，右侧 Data 操作栏中，选择”bankloan_edit.csv”点击右上角三个点的图标，选择”Add as asset”



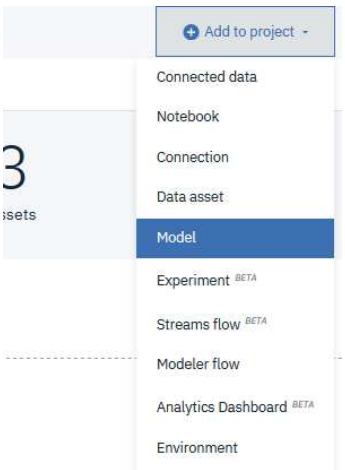
ii. 完成后，Data Assets 会出现该数据集

b) 如果没有完成 Notebook 的 2.5 章节，可以直接下载数据集文件
https://raw.githubusercontent.com/dongpo2/datascience/master/loanDefault/bankloan_edit.csv

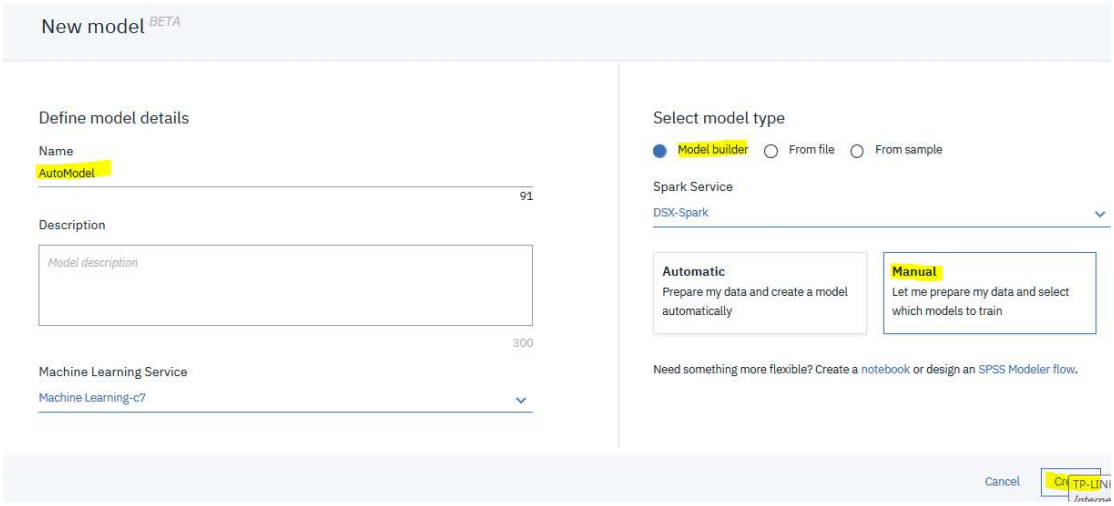


在项目首页，点击右侧”Data”图标，选择”Load”页签，点击”browse”，选择下载的文件，上传为数据集。

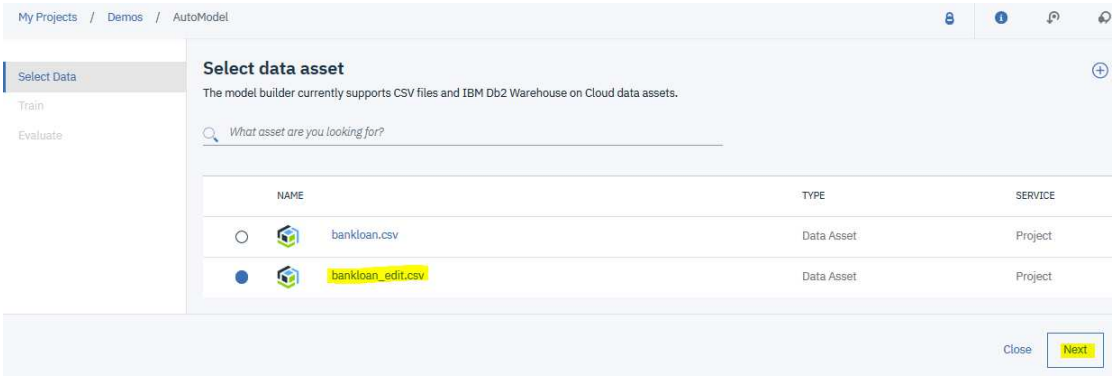
c) 项目首页，点击”Add to project”，选择 Model



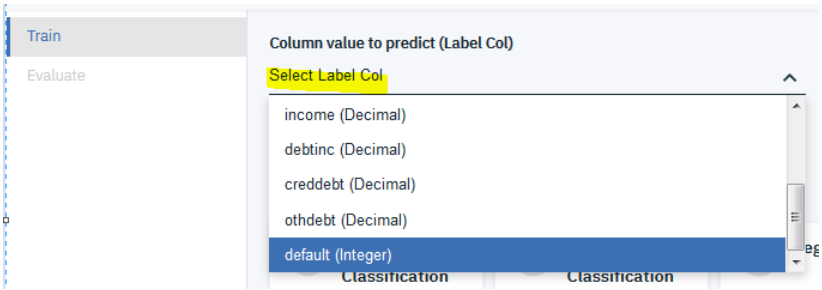
在新建 Model 页，输入名字，选择 Model Builder，选择 Manual，开始创建



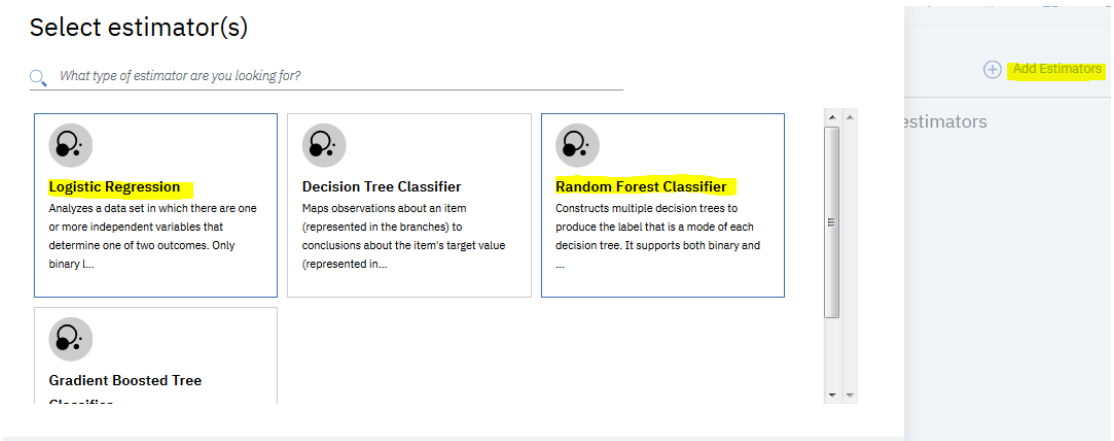
在”Select Data”，选择数据集，Next



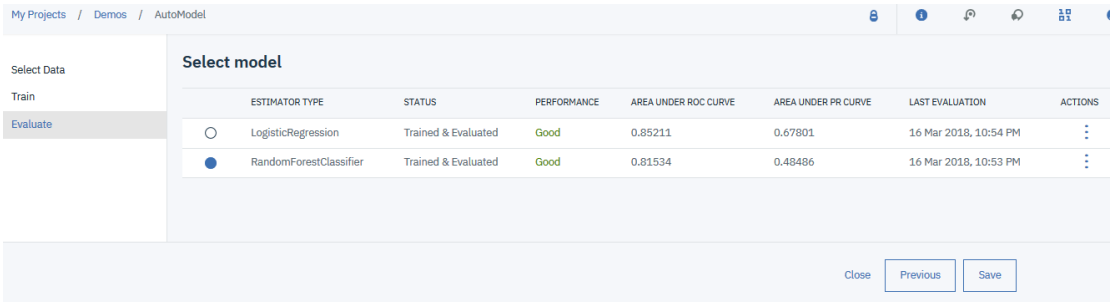
在”Train”页，先选择标注列名称”default”



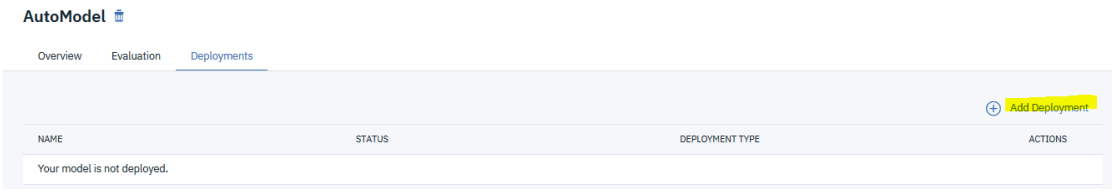
DSX 将自动推荐任务类型(二分类)，在右上角的“Add Estimators”，按 Ctl+鼠标可以选择多项算法，这里选择随机森林和逻辑回归。



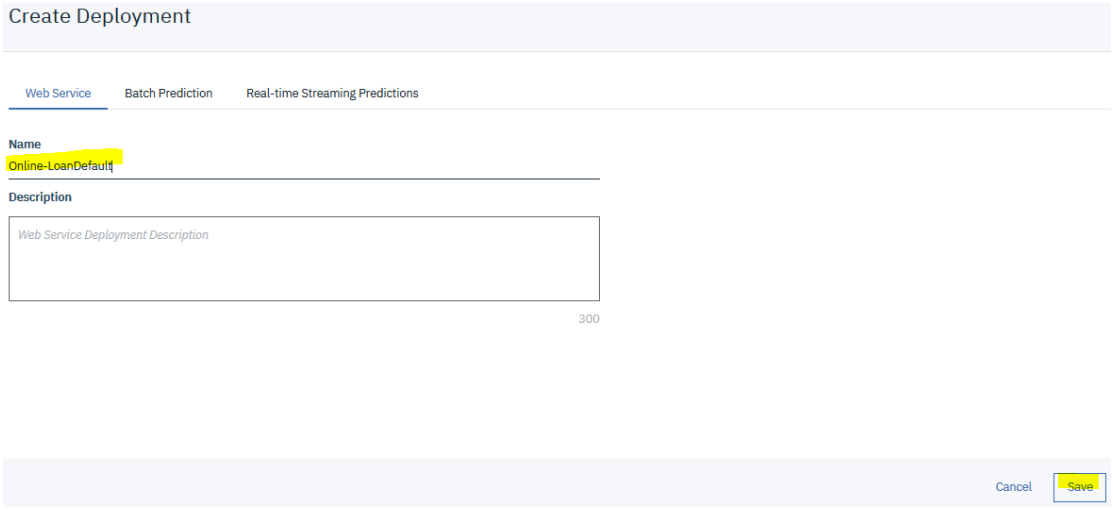
然后 Next 继续。在“Evaluate”页将给出不同算法的性能指标，选择合适的保存。



保存成功后，进入模型页，选择”Deployments”页。点击右上角图标“Add Deployment”



在新建部署页，输入名称，保存



d) 测试，参见 2.i 章节测试