Azure 机器学习实验

一、项目背景

【项目简介】

Azure Machine Learning（简称“AML”）是微软在其公有云 Azure 上推出的基于 Web 使用的一项机器学习服务，机器学习属人工智能的一个分支，它技术借助算法让电脑对大量流动数据集进行识别。这种方式能够通过历史数据来预测未来事件和行为，其实现方式明显优于传统的商业智能形式。微软的目标是简化使用机器学习的过程，以便于开发人员、业务分析师和数据科学家进行广泛、便捷地应用。这款服务的目的在于“将机器学习动力与云计算的简单性相结合”。AML 目前在微软的 Global Azure 云服务平台提供服务，用户可以通过站点：**https://studio.azureml.net/** 申请免费试用。

【项目涉及知识点】

- ✧ 下载、处理和上传收入普查的数据集；
- ✧ 创建一个新的 Azure 机器学习实验；
- ✧ 训练和评价一个预测模型；

二、项目基本需求及目的

【项目需求】

了解机器学习从数据到建模并最终评估预测的整个流程。

【项目目的】

根据人口普查数据预测不同人员收入情况

三、项目准备工作

【项目平台】

1，PC 机，如果你的电脑内存低于 512M，希望你不要安装虚拟机及项目所需的环境。
2，注册 Azure 平台并免费使用

四、项目实现步骤

【项目实现步骤】

1、数据集简介及准备

1.1 数据集简介

UCI 机器学习数据库的网址：**http://archive.ics.uci.edu/ml/**

该数据库是加州大学欧文分校(UniversityofCaliforniaIrvine)提出的用于机器学习的数据库，这个数据库目前共有 187 个数据集，其数目还在不断增加，UCI 数据集是一个常用的标准测试数据集。数据库不断更新，是所有学习人工智能、机器学习等都需要用到的数据库，是看文章、写论文、测试算法的必备数据集。数据库种类涉及生活、工程、科学各个领域，记录数也是从少到多，最多达几十万条。

Census Income Data Set

Download: Data Folder, Data Set Description

Abstract: Predict whether income exceeds $50K/yr based on census data. Also known as "Adult" dataset.

Data Set Characteristics:	Multivariate	**Number of Instances:**	48842	**Area:**	Social
Attribute Characteristics:	Categorical, Integer	**Number of Attributes:**	14	**Date Donated**	1996-05-01
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	324088

我们使用其中：美国人口普查数据集（**https://archive.ics.uci.edu/ml/datasets/census+income**）的数据，该数据从美国 1994 年人口普查数据库抽取而来，可以用来预测居民收入是否超过 50K/year。该数据集类变量为年收入是否超过 50k，属性变量包含年龄，工种，学历，职业，人种等重要信息，

值得一提的是，14 个属性变量中有 7 个类别型变量，数据集各属性:其中序号 0~13 是属性， 14 是类别

序号	字段名	含义	类型
0	age	年龄	Double
1	workclass	工作类型*	string
2	fnlwgt	序号	string
3	education	教育程度*	string
4	education_num	受教育时间	double
5	maritial_status	婚姻状况*	string
6	occupation	职业*	string
7	relationship	关系*	string
8	race	种族*	string
9	sex	性别*	string
10	capital_gain	资本收益	string
11	capital_loss	资本损失	string
12	hours_per_week	每周工作小时数	double
13	native_country	原籍*	string
14(label)	income	收入	string

数据集局部图如下图所示：

age	Workclass	Fnlwgt	Education	Education-num	Marital-status	Occupation	Relationship	Race	Sex	Capital-gain	Capital-loss	Hours-per-week	Native-country?	incom
39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
50	Self-emp-	83311	Bachelors	13	Married-civ-sp	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
38	Private	215646	HS-grad	9	Divorced	Handlers-clear	Not-in-family	White	Male	0	0	40	United-States	<=50K
53	Private	234721	11th	7	Married-civ-sp	Handlers-clear	Husband	Black	Male	0	0	40	United-States	<=50K
28	Private	338409	Bachelors	13	Married-civ-sp	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
37	Private	284582	Masters	14	Married-civ-sp	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
49	Private	160187	9th	5	Married-spouse	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K
52	Self-emp-	209642	HS-grad	9	Married-civ-sp	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K
31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	>50K
42	Private	159449	Bachelors	13	Married-civ-sp	Exec-managerial	Husband	White	Male	5178	0	40	United-States	>50K
37	Private	280464	Some-coll	10	Married-civ-sp	Exec-managerial	Husband	Black	Male	0	0	80	United-States	>50K
30	State-gov	141297	Bachelors	13	Married-civ-sp	Prof-specialty	Husband	Asian-Pac	Male	0	0	40	India	>50K
23	Private	122272	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0	0	30	United-States	<=50K
32	Private	205019	Assoc-acd	12	Never-married	Sales	Not-in-family	Black	Male	0	0	50	United-States	<=50K
40	Private	121772	Assoc-voc	11	Married-civ-sp	Craft-repair	Husband	Asian-Pac	Male	0	0	40	?	>50K
34	Private	245487	7th-8th	4	Married-civ-sp	Transport-movi	Husband	Amer-Indi	Male	0	0	45	Mexico	<=50K
25	Self-emp-	176756	HS-grad	9	Never-married	Farming-fishir	Own-child	White	Male	0	0	35	United-States	<=50K
32	Private	186824	HS-grad	9	Never-married	Machine-op-ins	Unmarried	White	Male	0	0	40	United-States	<=50K
38	Private	28887	11th	7	Married-civ-sp	Sales	Husband	White	Male	0	0	50	United-States	<=50K
43	Self-emp-	292175	Masters	14	Divorced	Exec-managerial	Unmarried	White	Female	0	0	45	United-States	>50K
40	Private	193524	Doctorate	16	Married-civ-sp	Prof-specialty	Husband	White	Male	0	0	60	United-States	>50K
54	Private	302146	HS-grad	9	Separated	Other-service	Unmarried	Black	Female	0	0	20	United-States	<=50K

注 1：

已清洗的数据仅供本课程学习使用，有一定的模拟性质。如需要更多的信息，则需要从原始数据按照相应的目的进行清洗。

注 2：

CSV 格式是数据分析工作中常见的一种数据格式。CSV 意为逗号分隔值（Comma-Separated Values），其文件以纯文本形式存储表格数据（数字和文本）。每行只有一条记录，每条记录被逗号分隔符分隔为字段，并且每条记录都有同样的字段序列。

CSV 格式能被大多数应用程序所支持，广泛用于在不同的系统之间转移数据，是一种容易被兼容的格式。实验楼中大量的数据分析类课程都使用了 CSV 格式的数据集，不仅如此，我们也推荐你在今后的数据分析工作中应用此格式来存储数据。

2、Azure 云平台的机器学习应用

2.1 观察数据集

现在，用 Microsoft Excel 或任何其他电子表格工具中打开 adult.data 文件，并为其添加网站中属性列表的详细信息，这些信息如下列出。注意，其中的一部分属性值为连续的，因为它们以数值的形式表现，另一部分则为离散的。

年龄（age），连续值

工作种类（Workclass）个人（Private），无限责任公司（Self-emp-not-inc），有限责任公司（Self-emp-inc），联邦政府（Federal-gov），地方政府（ Local-gov），州政府（State-gov），无薪人员（Without-pay），无工作经验人员（Never-worked）离散值

序列号（Fnlwgt）连续值

教育情况（Education） Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool ）离散值

受教育年限（Education-num），连续值

婚姻状况（Marital-status） 已婚（Married-civ-spouse），离婚（Divorced），未婚（Never-married），离异（Separated），丧偶（Widowed），已婚配偶缺席（Married-spouse-absent）、 再婚（Married-AF-spouse），离散值

职业情况（Occupation）技术支持（Tech-support），维修工艺（Craft-repair），服务行业（Other-service）、 销售（Sales）、执行管理（Exec-managerial）、 专业教授（Prof-specialty），清洁工（Handlers-cleaners），机床操控人员（Machine-op-inspct）、行政文员（Adm-clerical）、 养殖渔业（Farming-fishing）、 运输行业（Transport-moving），私人房屋服务（Priv-house-serv），保卫工作（Protective-serv）， 武装部队（Armed-Forces）职业情况，离散值

亲属情况（Relationship）妻子（Wife），子女（Own-child），丈夫（Husband），外来人员（Not-in-family）、 其他亲戚（Other-relative）、 未婚（Unmarried），离散值

种族肤色（Race）白人（White），亚洲太平洋岛民（Asian-Pac-Islander），阿米尔-印度-爱斯基摩人（Amer-Indian-Eskimo）、其他（Other），黑人（Black）离散值

性别（Sex ）男性（Female）,女性（ Male），离散值

资本盈利（Capital-gain ）连续值

资本损失（Capital-loss） ，连续值

每周工作时间（Hours-per-week ），连续值

国籍（Native-country ）美国（United-States）、 柬埔寨（Cambodia）、 英国（England），波多黎各（Puerto-Rico），加拿大（Canada），德国（Germany），美国周边地区（关岛-美属维尔京群岛等）（Outlying-US(Guam-USVI-etc)），印度（India）、日本（Japan）、 希腊（Greece）、 美国南部（South）、 中国（China）、 古巴（Cuba）、 伊朗（Iran）、 洪都拉斯（Honduras），菲律宾（Philippines）、 意大利（Italy）、 波兰（Poland）、 牙买加（Jamaica）、 越南（Vietnam）、 墨西哥（Mexico）、葡萄牙（Portugal）、 爱尔兰（Ireland）、 法国（France）、多米尼加共和国（Dominican-Republic）、 老挝（Laos）、 厄瓜多尔（Ecuador）、 台湾（Taiwan）、 海地（Haiti）、 哥伦比亚（Columbia）、 匈牙利（Hungary）、 危地马拉（Guatemala）、尼加拉瓜（Nicaragua）、苏格兰（Scotland）、 泰国（Thailand）、 南斯拉夫（Yugoslavia），萨尔瓦多（El-Salvador）、 特立尼达和多巴哥（Trinadad&Tobago）、 秘鲁（Peru），香港（Hong），荷兰（Holland-Netherlands）离散值

收入 （incom） >50K, <=50K ，离散值

注意，在插入这些列的标题后，一定要以 .csv 格式保存，且保存时将文件命名为 Adult.data.csv 。

2.2 导入数据

2.2.1 总结数据集

总括一下数据集的数据特征：

1，十四个与结果相关的唯一属性

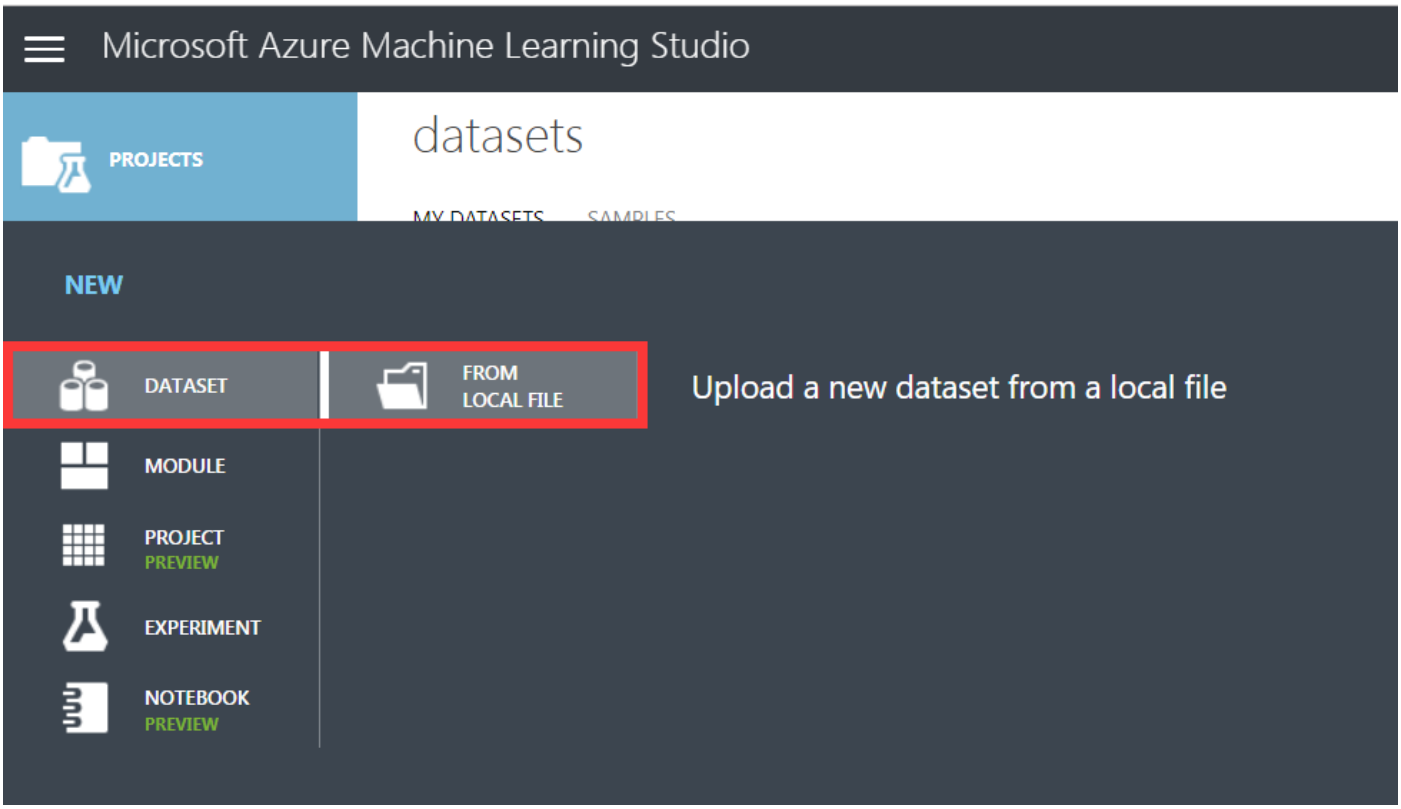
2，数据集的实例数为 48,842

3，预测任务是确定用户是否一年收入超过$50,000 美元。

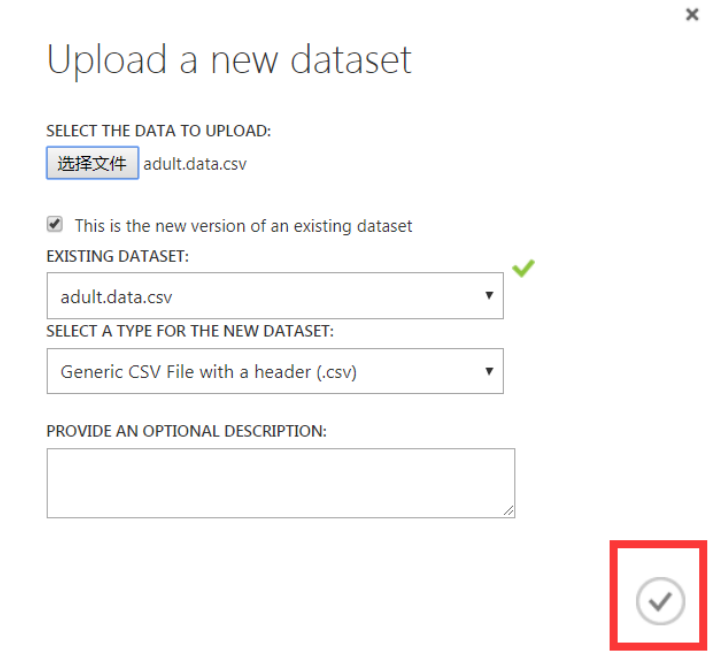
此人口收入的普查数据集以被微软作为一个样本数据提供出来了，在其成人普查收入的二元分类（Adult Census Income Binary Classification）数据集中便可以找到。以下我们将手动地一步步全面地介绍整个 Azure 机器学习工作流过程，很有可能，您的用于预测模型地真实数据集来自于其他外部资源，因此了解机器学习是怎么从开始至结束的全过程是很有必要的。

2.2.2 数据上载至 Azure 机器学习实验

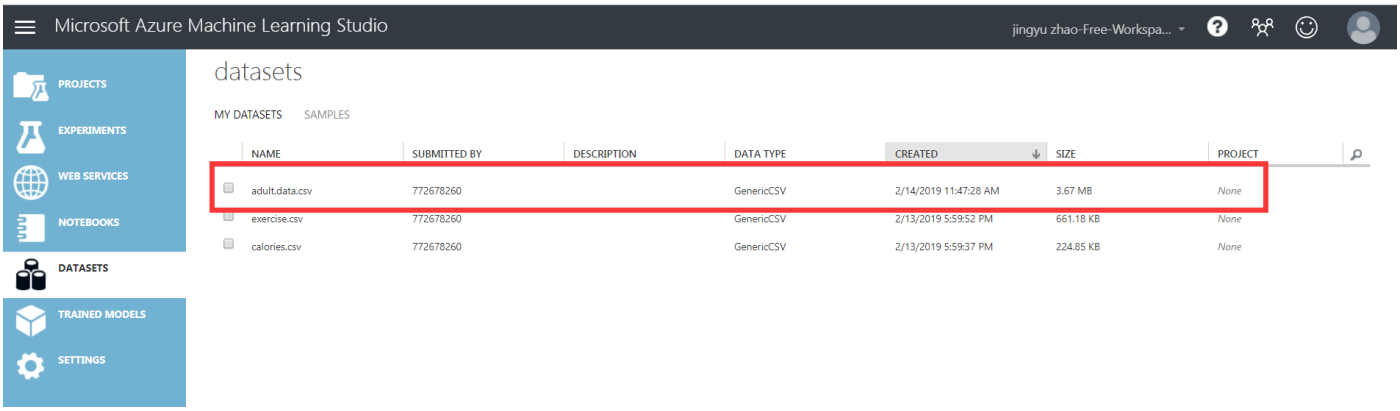
将人口收入普查数据集添加了列标题后，我们即可将数据上载至 Azure 机器学习工作区，并将其纳入预测模型。点击屏幕左下方的"+"，然后选择上传的数据集。下图显示上传本地数据文件的选项。



下一步，点击从本地文件选择即"FROM LOCAL FILE"，您可看见如下图所示的上载界面。在此界面您可指定上载文件的属性，比如文件的位置、名称（本例中我们使用 Adult.data.csv ）和类型（通常是 CSV 类型），以及新的数据集的可选说明。

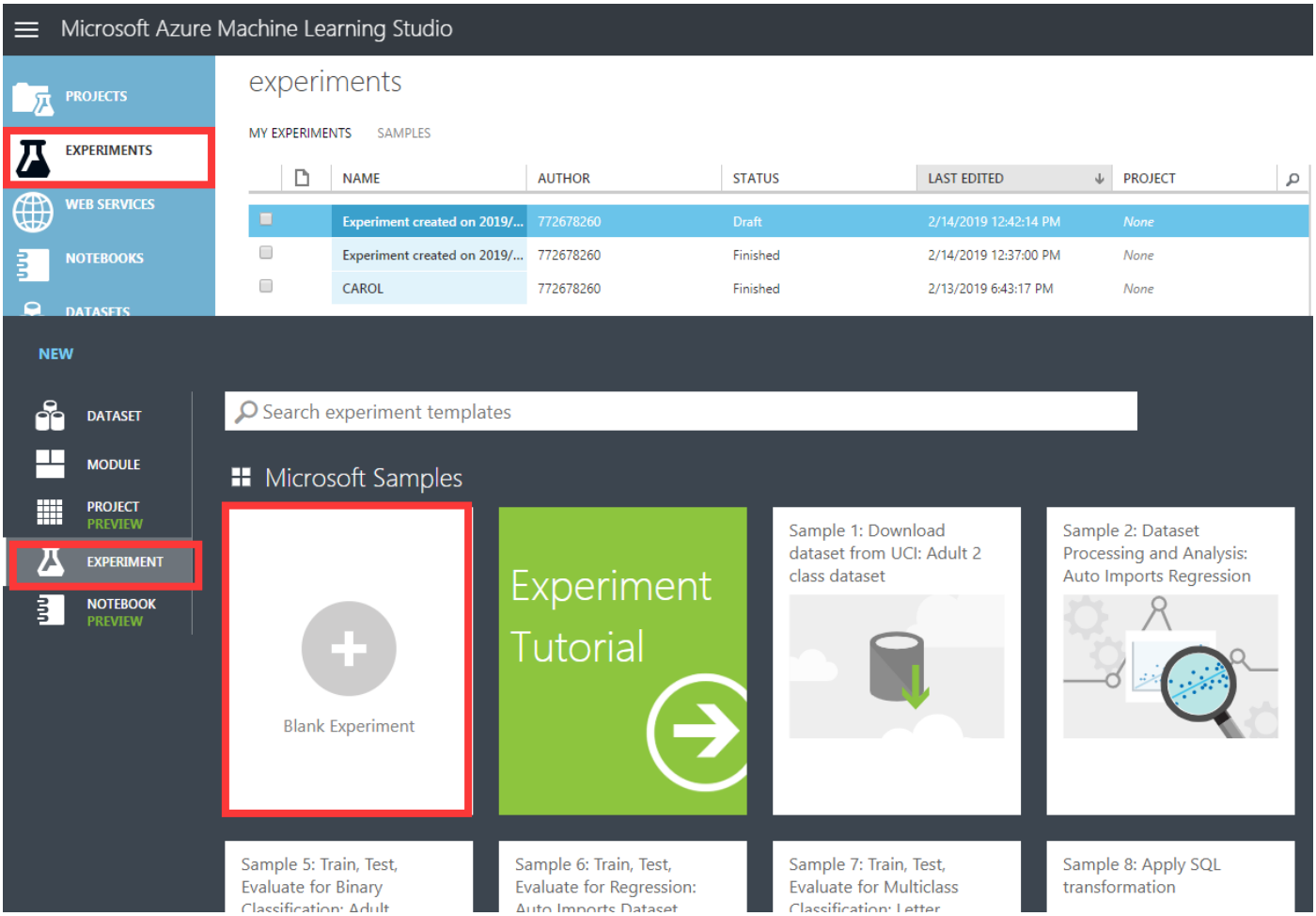


完成信息的输入并点击签入按钮后，您的数据集将异步加载至您的第一个 Azure 机器学习实验的工作区中：



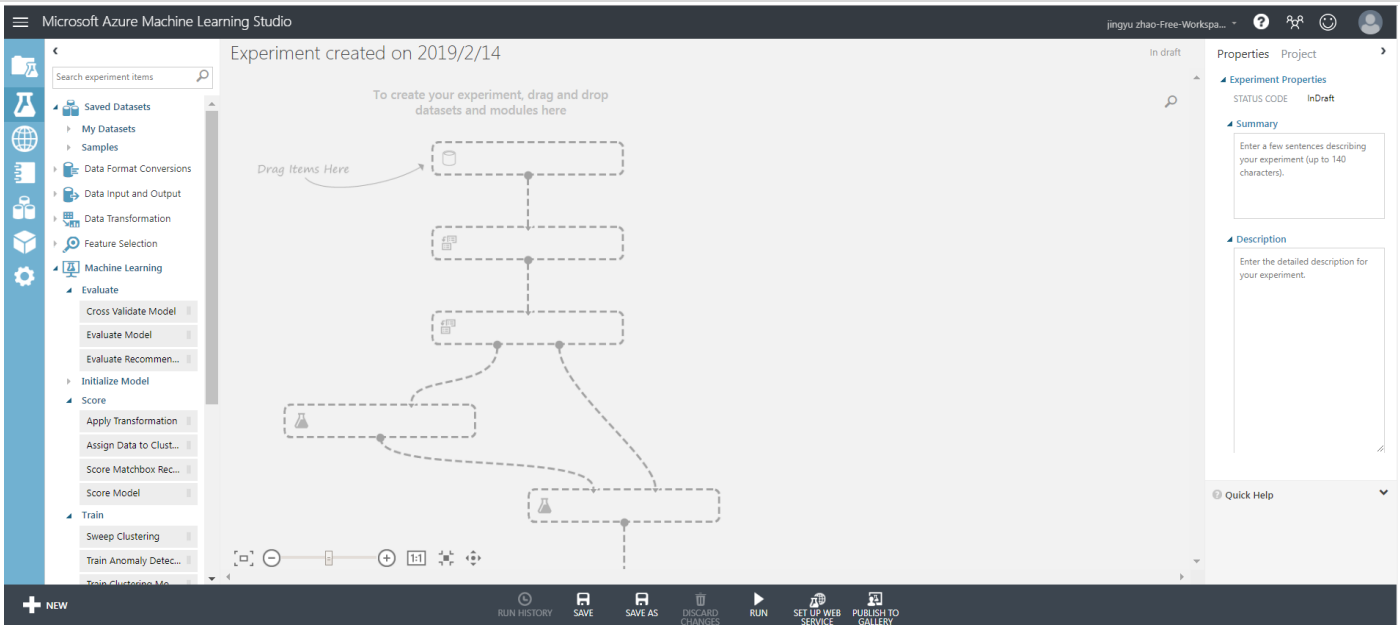
2.2.3 创建新的 Azure 机器学习实验

创建新的实验的方法是点击屏幕左下角的"+NEW"按钮，选择"实验"（EXPERIMENT）>"空白实验"（Blank Experiment）：



请注意，除了空白实验之外，还有许多示例实验模板可供您加载和修改，以便您快速掌握 Azure 机器学习的实践。

完成新的空白实验的加载后，您可见到如下图所示的 Azure ML Studio 可视化设计界面



可以看到设计器由**三个主要区域**构成：

左侧导航窗格 此区域包含 Azure 机器学习模块的可搜索列表，此模型可用于创建预测分析模型。

按功能区域分组的模块

数据集的读取和格式转换；

使用和训练机器学习算法；

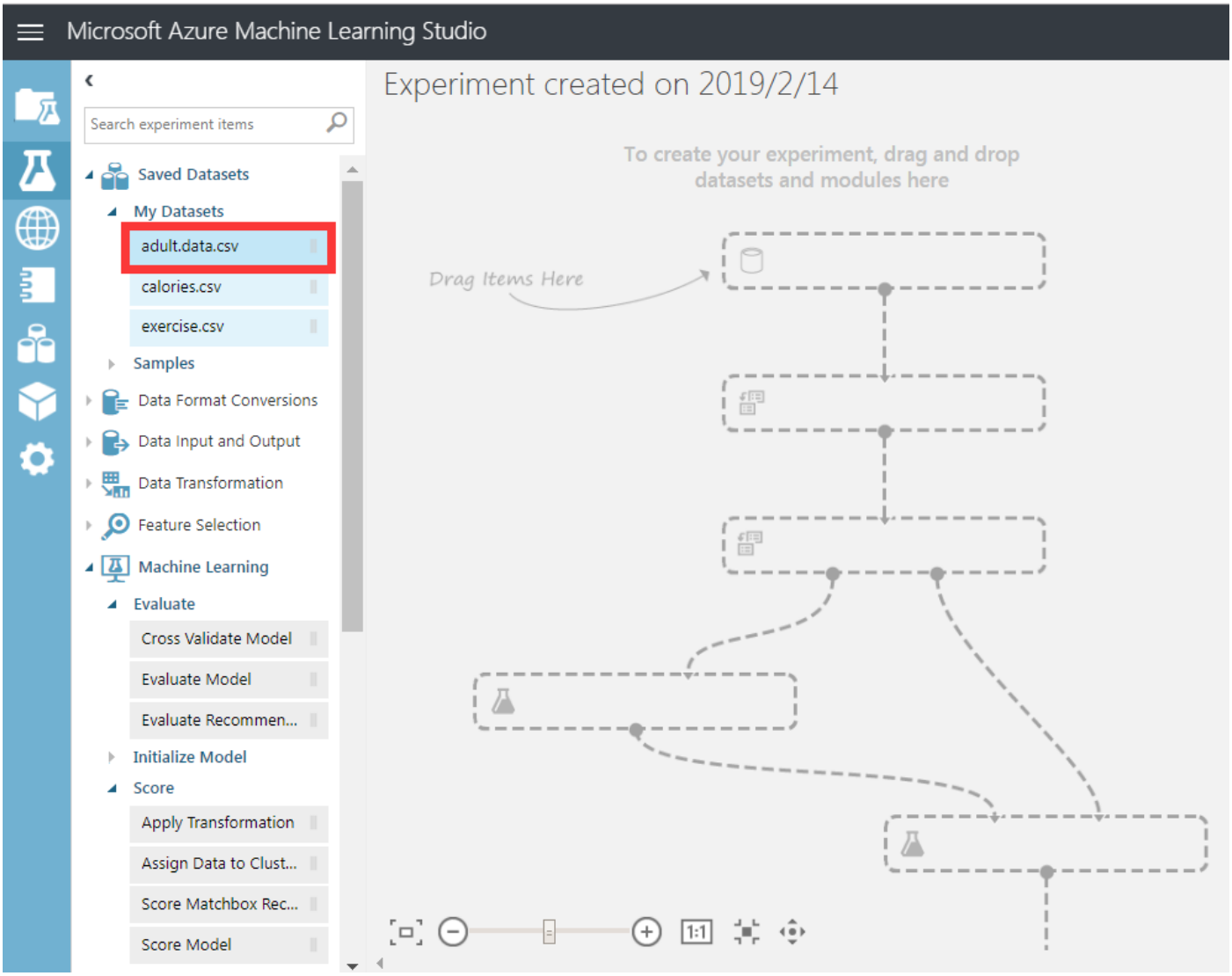
评估预测模型的结果。

中间窗格 在可视化设计器中，Azure 机器学习的实验类似于流程图的形式，可以通过拖拽左侧窗格中的功能模块至可视化设计器

的中间窗格组装成工作流。模块可以自由的被拖放在中间窗格的任意位置，模块之间通过输入和输出端口之间画线连接。

右侧窗体 在属性视图中，可在右侧窗体查看和设置被选择模块的属性。

在左侧窗体展开"已保存的数据集（Saved Datasets）"选项，便可以看到我们上载的用于 Azure 机器学习的 Adult.data.csv 数据文件出现在数据集的列表中，如图显示 Adult.data.csv 将被拖放至可视化设计器的中间窗体:



2.3 分割数据集

通常，创建 Azure 机器学习实验后，我们都会将数据集分割为两个分组即**训练数据**和**验证数据**，这样做有两个特定目的：

1，训练数据通常用来创建预测模型，基于机器学习算法发现历史数据中的固有模式。

2，验证数据的分组用来测试训练数据创建的预测模型对于已知结果预测的精度和概率。

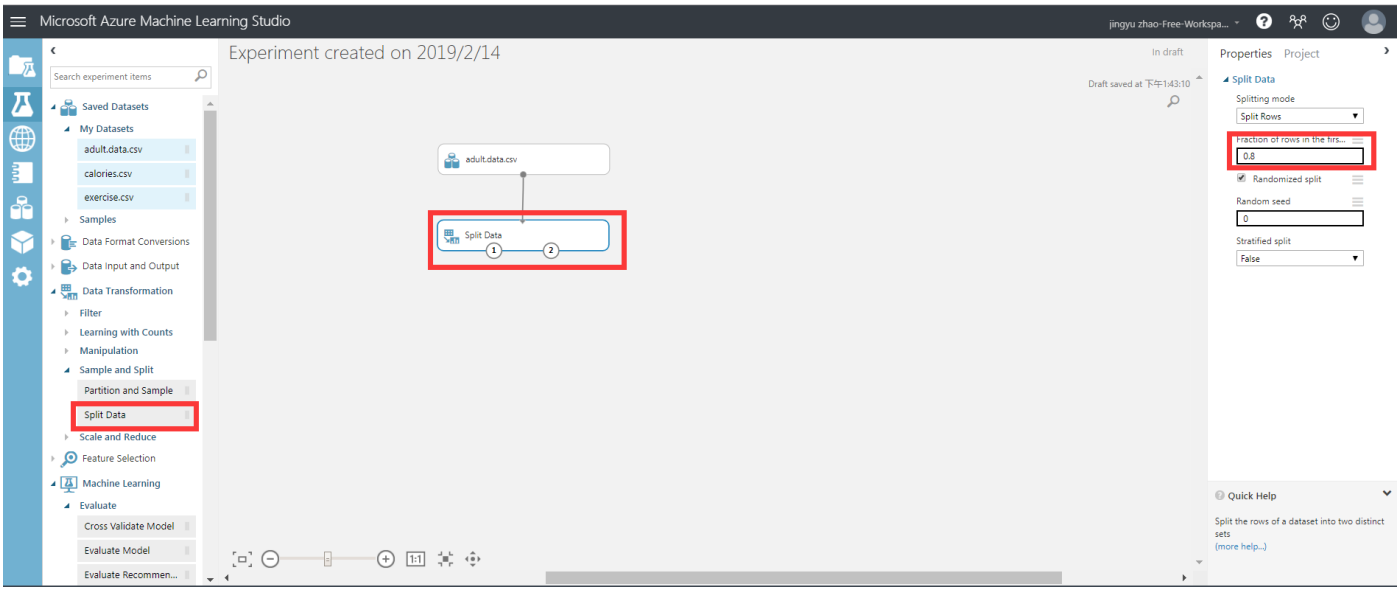
执行以下的步骤将数据集分割成两部分。

1，在左侧窗体中展开"Data Transformation"即数据转换模块。

2，拖动"Split"即分割模块至 Azure 机器学习设计器。

3，连接"Split"模块与 Adult.data.csv 数据集。

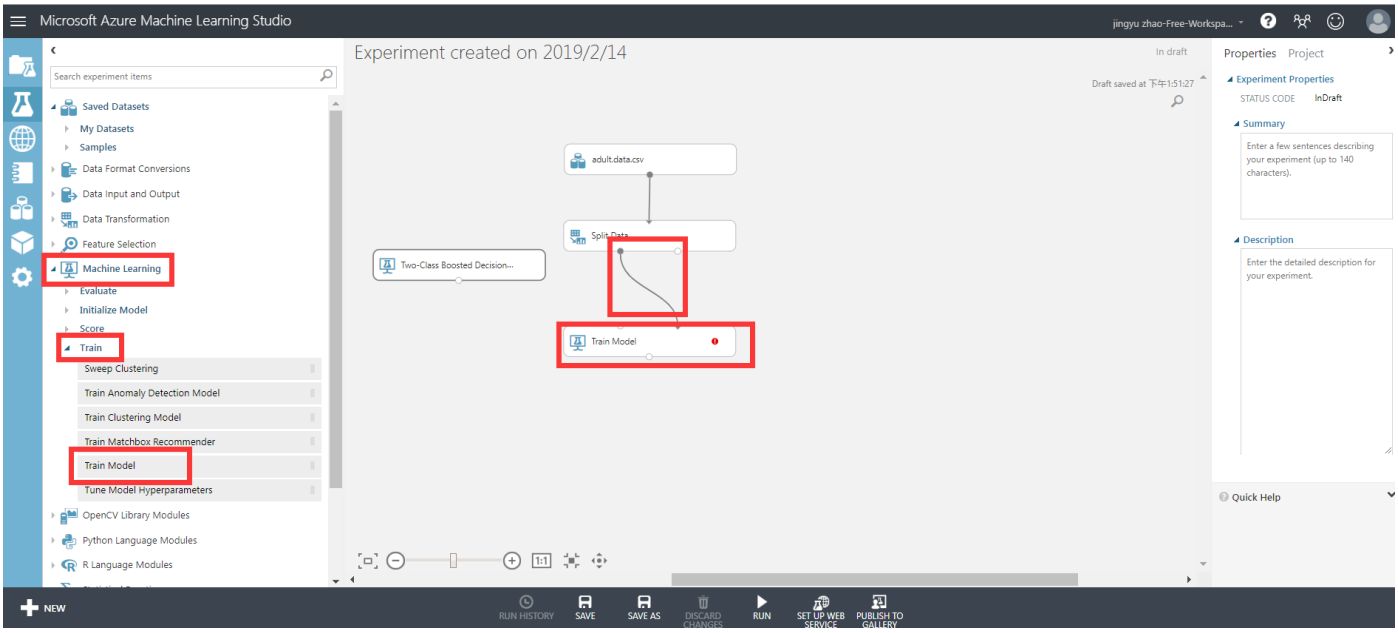
4，点击分割模块并设置"Fraction of rows in the first output dataset"为 0.8。这将 80%的数据分割至训练数据集中。



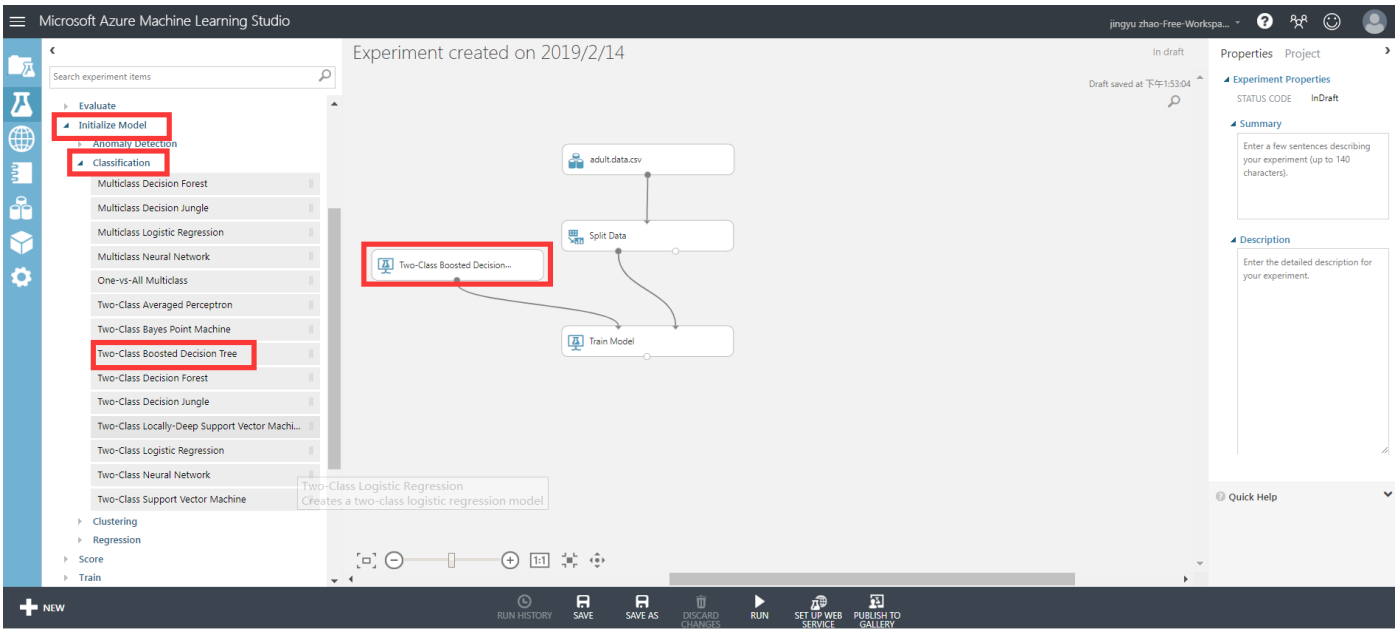
以上操作就将数据集中的 80%的数据用于训练模型，我们可使用剩余的 20%数据验证模型的精度。

2.4 模型训练

借助 Azure 机器学习算法"教"模型如何评估数据:在左侧窗体中展开"Machine Learning"即机器学习模块,然后展开"Train"子模块,将"Train Model"拖放至设计器中，最后在设计器中连接"Train Model"和"Split"图形。



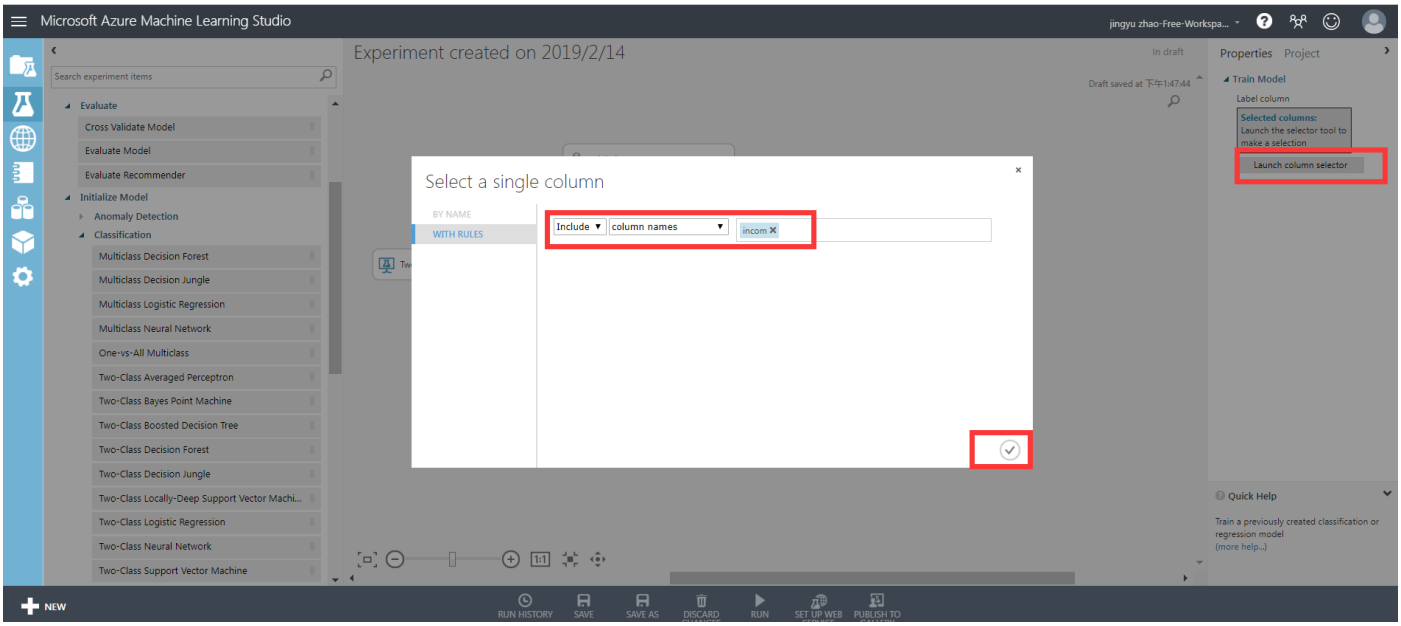
然后，我们展开"Machine Learning"即机器学习模块下的"Initialize Model"即初始化模型，展开"Classfication"即分类子模块。在此实验中，我们使用"Two-Class Boosted Decision Tree"即双类提升的决策树算法。在左侧窗体中选中该算法模块并将其拖放至设计器中，至此您的实验应该如下图所示。



2.5 选择预测项

要完成算法的配置，我们需要指定数据集中的哪一列数据作为输出或者预测列，数据集中的任意列将基于其他列的数据做预测。若要执行此操作，在设计器中点击"Train Model"，属性窗体将在 Azure ML Studio 的右侧窗体中显示， 若您在设计器中设置，请选择"Launch column selector"即启动列选择器，选择"Include"和列名称为"income"即收入的列。

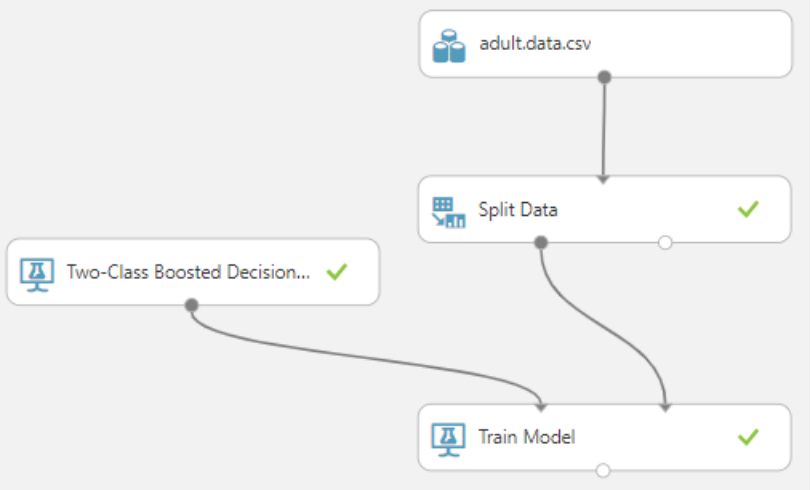
下图所示的列选择器将数据集中的收入列作为预测列，即要预测的是用户收入。如下图所示。



按照这种方式，Azure 机器学习算法从每行数据中的其他列训练模型，以预测收入。我们使用数据集中的 80%基于已知的输入和输出数据训练训练模型。

至此，我们已经做好训练模型的准备，选择屏幕底端的"RUN"即运行选项，然后静待 Azure 机器学习训练我们的模型。您会注意到，实验每个阶段完成的时候，绿色的复选框就出现在每个操作的右侧，如下图所示。

、

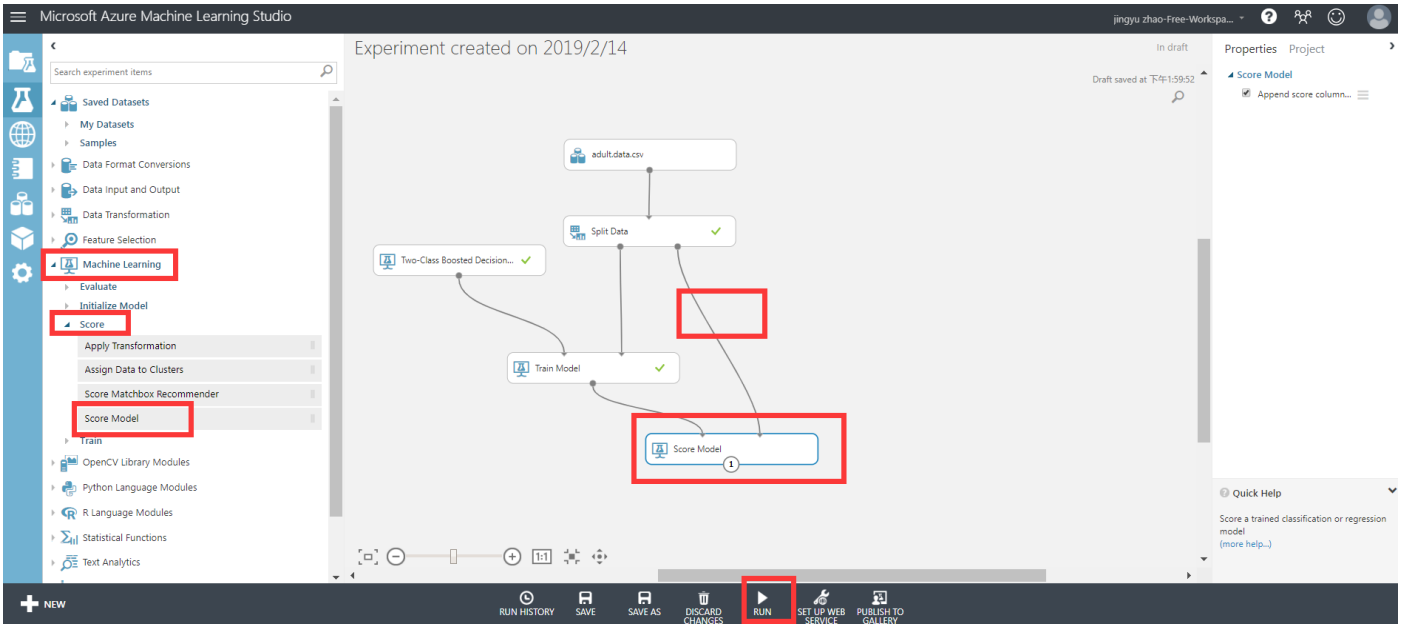


2.6 模型评分

现在我们已经训练完成新的 Azure 机器学习预测模型，下一步我们从解决方案的适用性的角度评估预测结果的正确性，以确定模型的精度。请牢记，Azure 机器学习解决方案伟大之处在于迭代开发，最终成功的关键是快速试错。

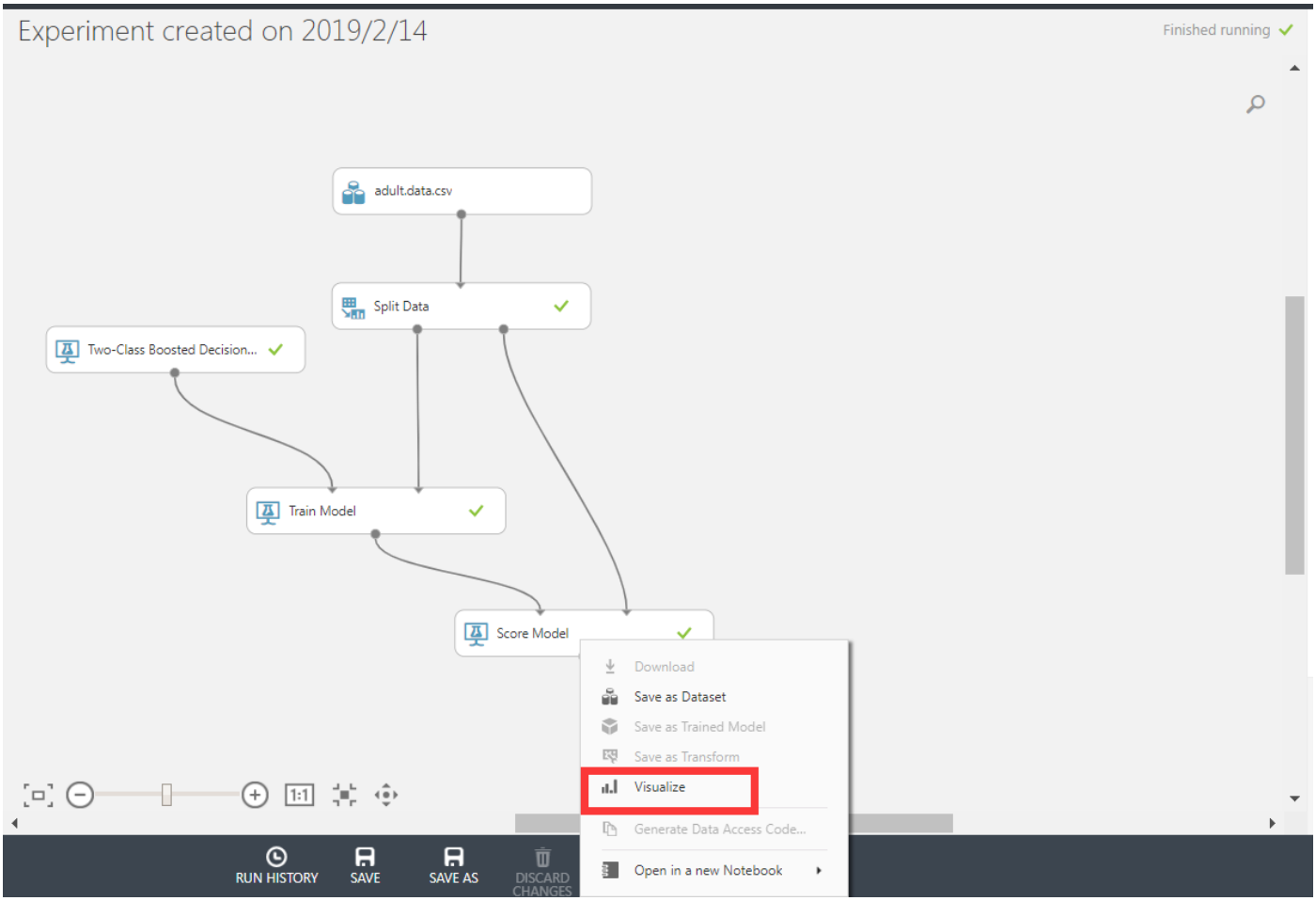
如要实现对模型的评价，首先展开 Azure ML Studio 左侧的"Machine Learning"即机器学习模块，然后展开"Score Model"即评分模型子模块，将"Score Model"拖放至设计器中，下一步连接"Score Model"和"Train Model"，最后链接"Score Model"和"Split"模块。至此，基本上就完成了利用数据集中 20%的数据评估预测模型的准确性。

下一步，单击屏幕底部的"Run"即运行按钮等待处理的结果（每个模块右侧出现绿色的复选标记表示运行完毕）。下图是机器学习实验预测收入的运算过程截图。

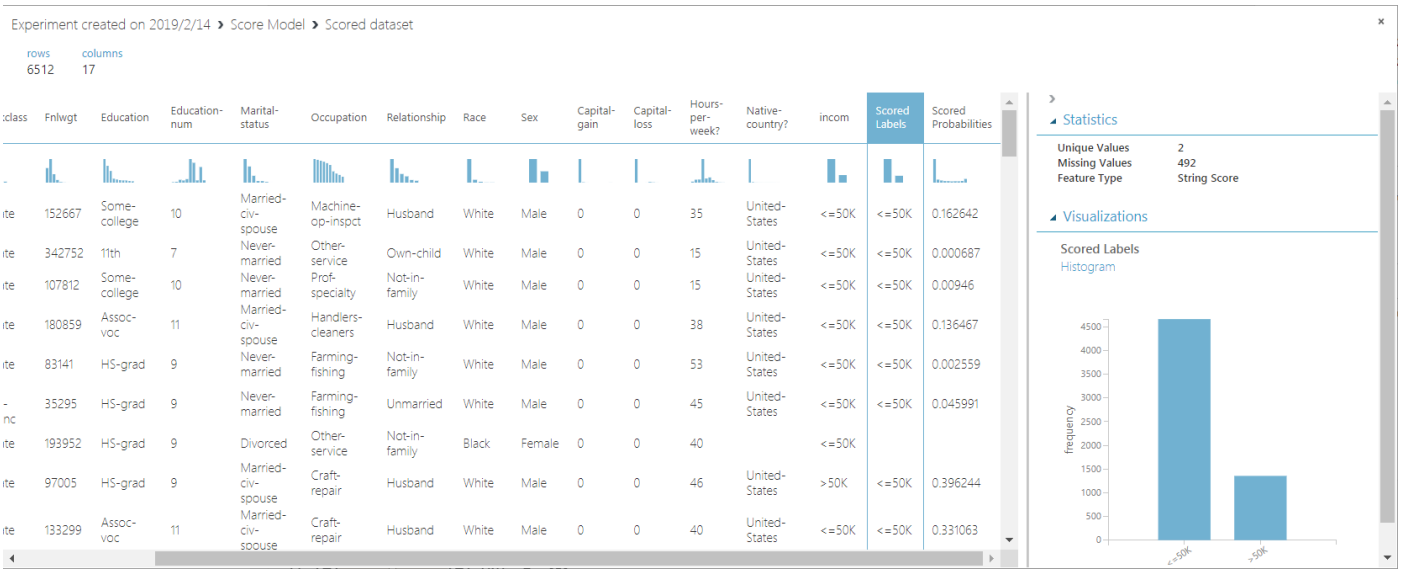


2.7 模型计算结果的可视化

当所有的模型运算结束后，将鼠标悬停在"Score Model"即评分模型上点击右键，从快捷菜单中选择"Visualize"即可视化，如下图所示。



当您选择可视化新训练的模型数据选项后，会生成一个新的页面。在可视化的界面中滑动滚动条至最右端，您会发现两个额外的列显示在数据集中，如下图所示。



可以看到现在有两个额外的列添加到了我们的数据集中：

1，**"Scored Lables"**即评分标签表示数据集中此行数据的预测结果

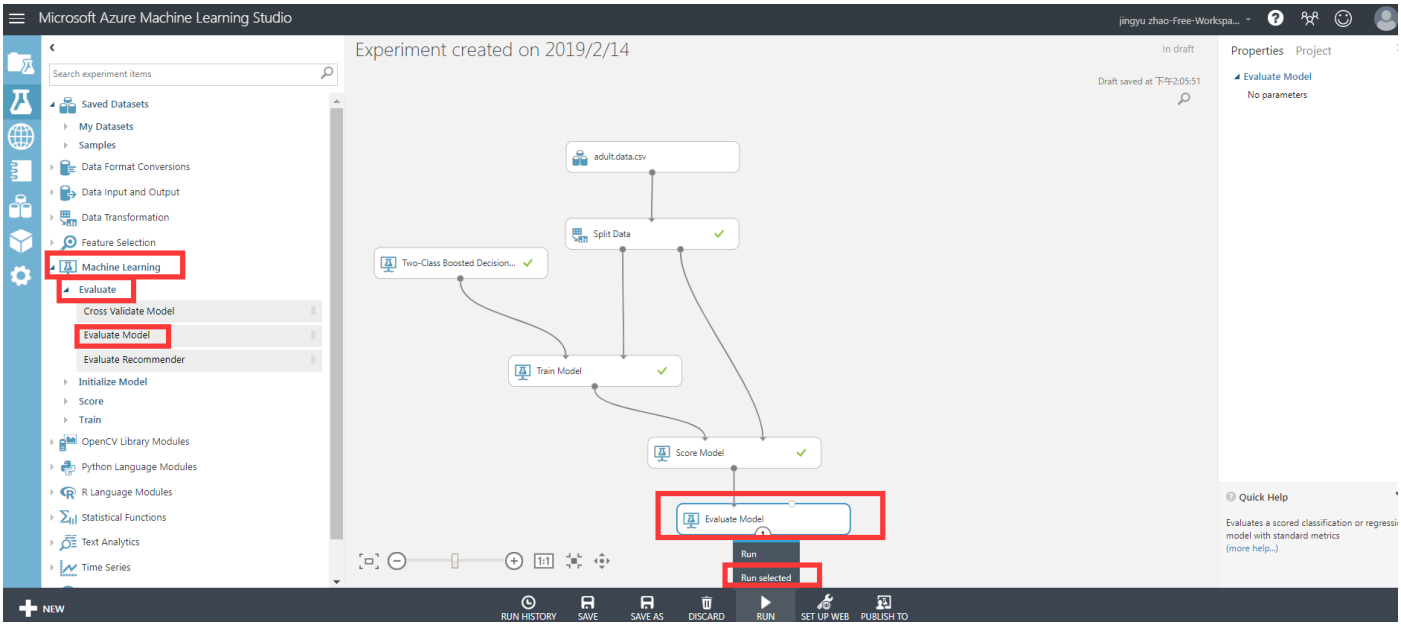
2，**"Scored Probabilities"**即评分概率表示收入水平超过 $50000 的概率 （或可能性）。

在我们数据集中新增的列提供了算法针对每行数据计算的预测结果和概率因子。概率因子是模型基于数据集中其他列数据预测结果的准确度的概率估计。通常情况下，预测分析是一个多轮迭代的过程。可能您会尝试许多不同的算法，或者将他们联合使用（在高级的机器学习主题文章中被称为集成）以证明预测模型的有效性。

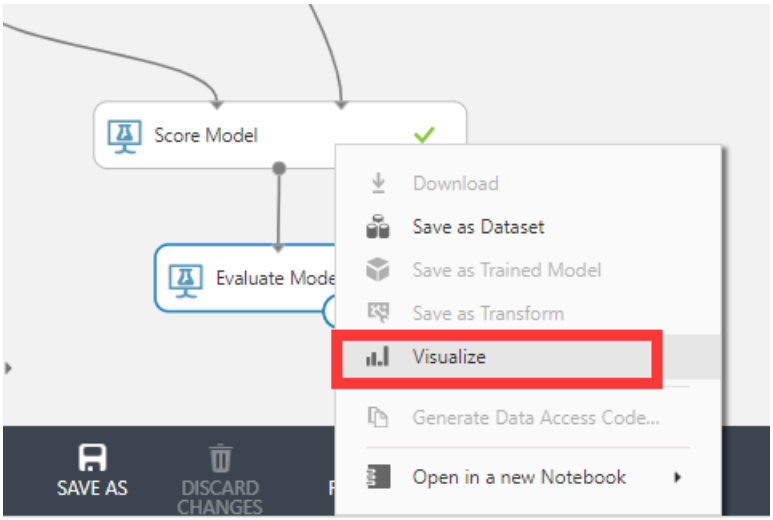
2.8 模型评估

Azure 机器学习最引入注目的功能之一就是它能够快速评估不同的算法，只要轻点鼠标就可完成这些功能，这一切都归功于评估模型。确定模型的精准度的方法很简单，我们只要使用 Azure ML Studio 内置的评估模型就轻松完成模型的评价。

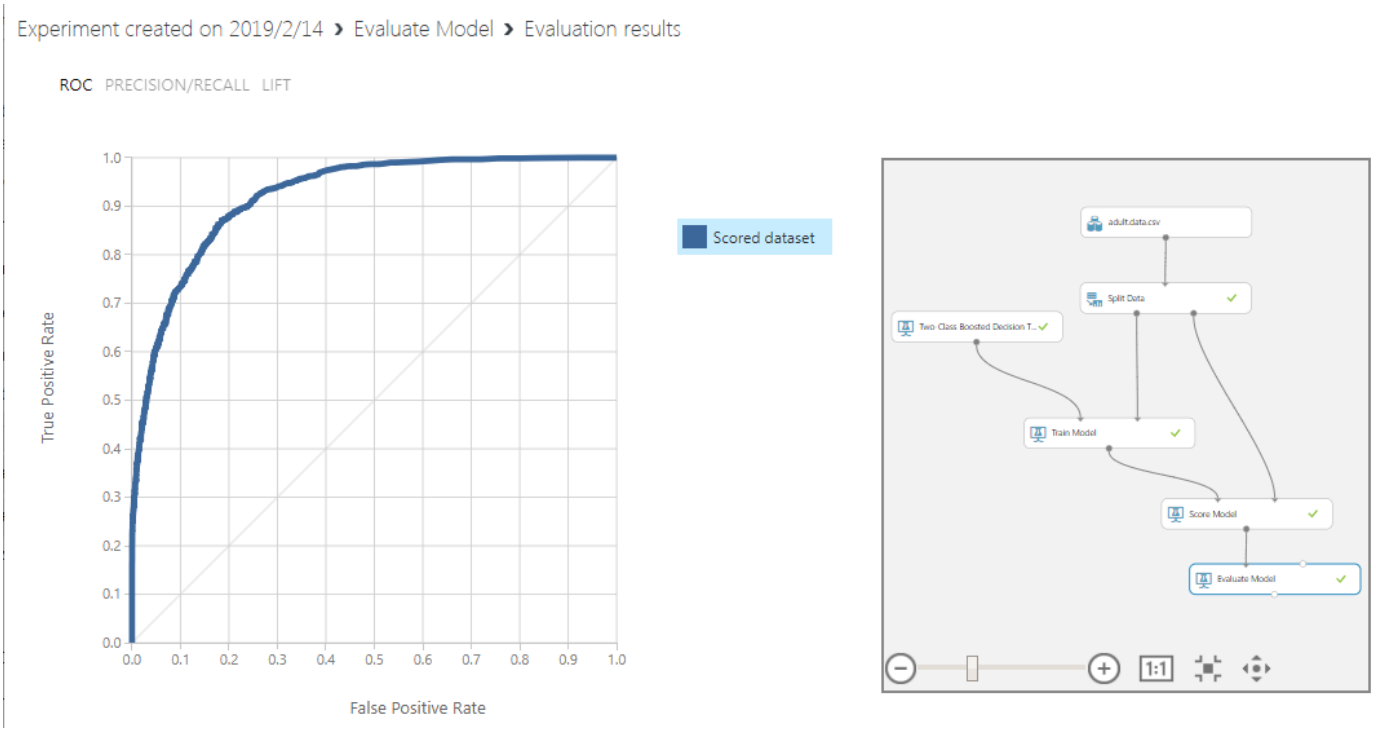
若要执行此操作，在 Azure ML Studio 的左侧导航窗格中点击"Machine Learning"即机器学习模块，选择"Evaluate"即评估子模块，最后选择"Evaluate Model"即评估模型的模块，将其拖至可视化设计器页面中的"Score Model"模块下方。连接"Split Model"和"Score Model"即分割模型和评分模型，以及"Evaluate Model"和"Score Model"即评价模型和评分模型，如下图所示。



点击 Azure ML Studio 屏幕底部的"Run"即运行按钮，在执行过程中您可以查看实验中每个模块的运行情况，如果模块运行完毕会在模块的右侧显示绿色的复选标记。整个过程运行完毕后，右键单击评估模型的模块底部连接器，在快捷菜单中选择"Visualize"即可视化：



评估的结果就会如下图显示：



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
1022	485	0.864	0.754	0.5	0.922

False Positive	True Negative	Recall	F1 Score
334	4179	0.678	0.714

Positive Label	Negative Label
>50K	<=50K

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	506	37	0.090	0.828	0.494	0.932	0.336	0.817	0.992	0.002
(0.800,0.900]	180	65	0.131	0.847	0.598	0.871	0.455	0.843	0.977	0.008
(0.700,0.800]	132	62	0.163	0.858	0.657	0.833	0.543	0.863	0.964	0.015
(0.600,0.700]	114	83	0.196	0.863	0.694	0.791	0.618	0.881	0.945	0.026
(0.500,0.600]	90	87	0.225	0.864	0.714	0.754	0.678	0.896	0.926	0.038
(0.400,0.500]	93	129	0.262	0.858	0.723	0.707	0.740	0.912	0.897	0.059
(0.300,0.400]	82	148	0.300	0.847	0.722	0.662	0.794	0.926	0.865	0.084
(0.200,0.300]	103	208	0.352	0.830	0.717	0.613	0.863	0.947	0.819	0.122
(0.100,0.200]	89	351	0.425	0.786	0.683	0.543	0.922	0.966	0.741	0.192
(0.000,0.100]	118	3343	1.000	0.250	0.400	0.250	1.000	1.000	0.000	0.922

五、总结

5.1 曲线和度量指标

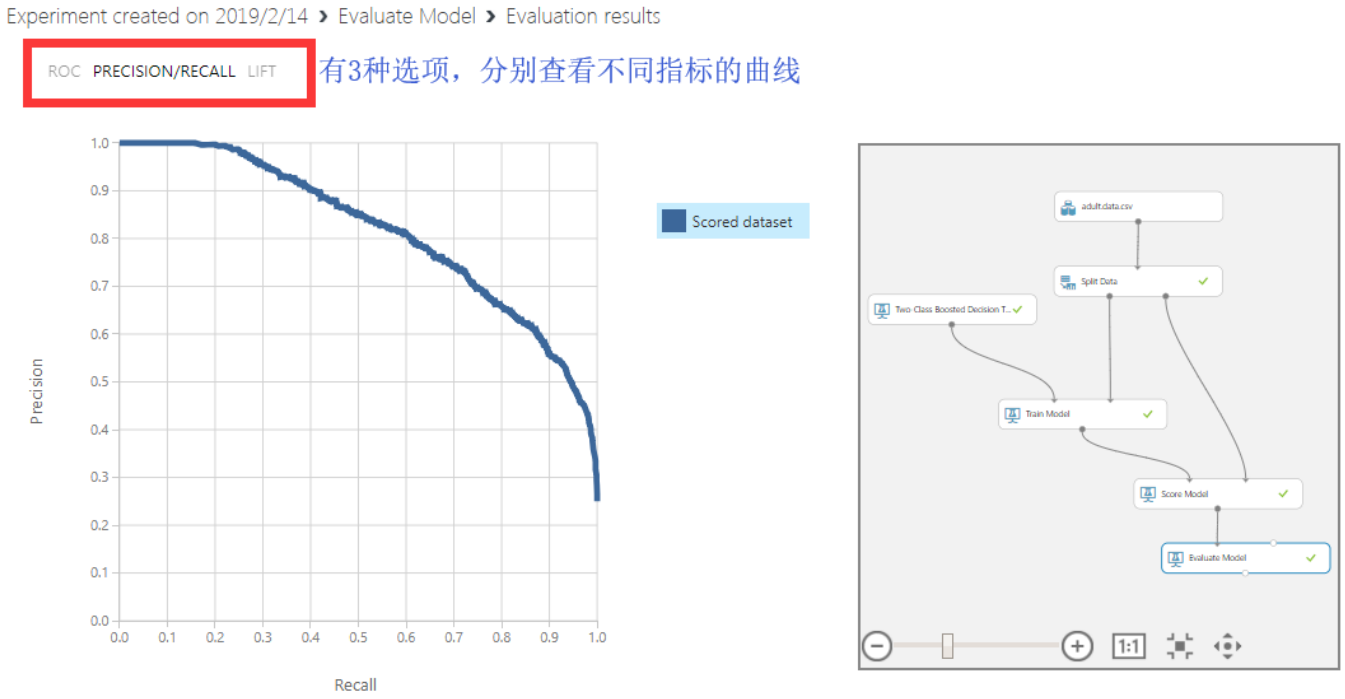
评估模型模块会产生一套曲线和度量指标，让您对于评分模型的结果或者两个评分模型的对比情况一目了然。评分结果以以下三种形式展示：

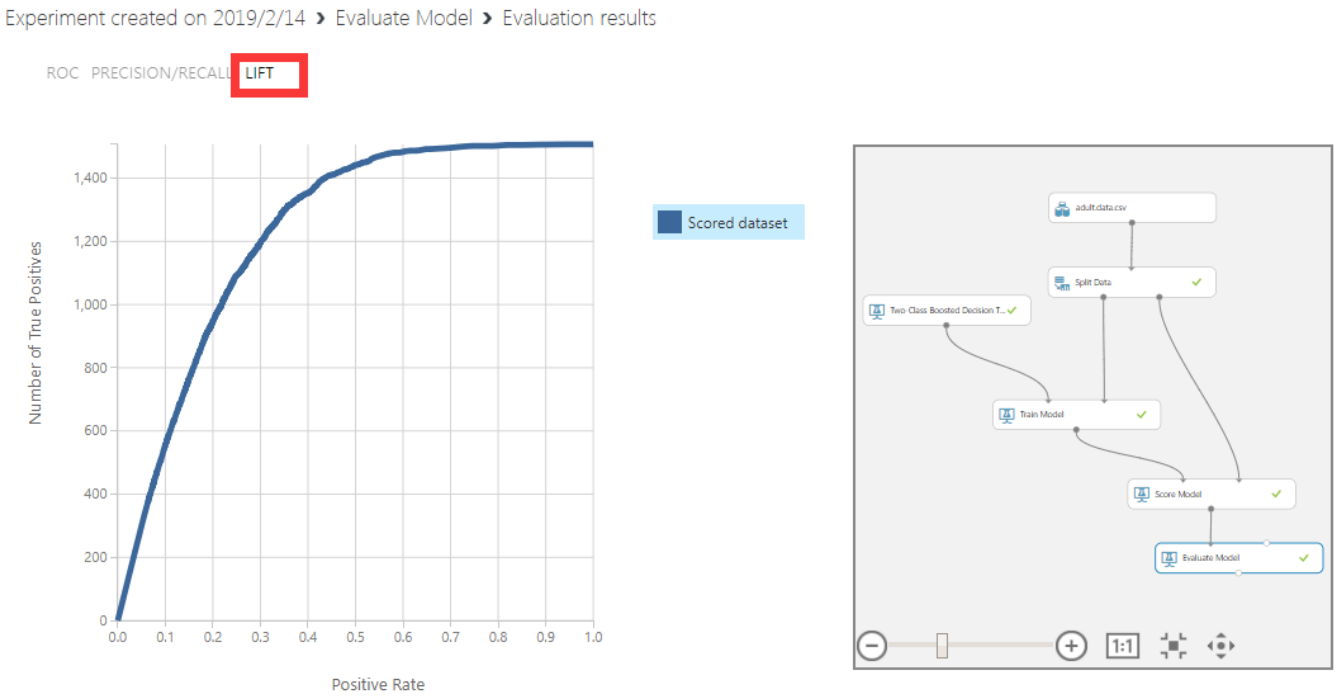
ROC 曲线（Receiver Operator Characteristic）即受试者工作特征曲线反映的是真阳性占总的实际阳性的比例。将它与在各种阈值设置情况下假阳性占总的实际阴性的比例进行对比。对角连线表示 50%预测的准确性，并可作为评价的基准以便后续提高。曲线位于左边高出对角线的部分表示模型的精准度高，当然您也会希望实验的结果曲线出现在此区域。

准确率和召回率是衡量信息检索系统性能的重要指标。准确率是指检索到相关文档数占检索到的文档总数的比例，而召回率是指检索到相关文档数占所有相关文档总数的比例。

lift 曲线是数据挖掘分类器最常用的方式之一，与 ROC 曲线不同的是 lift 考虑分类器的准确性，也就是使用分类器获得的正类数量和不使用分类器随机获取正类数量的比例。

可视化结果中，您可看到两个数据集（"训练"数据集和"验证"数据集）几乎完全相同，即红色和蓝色曲线几乎完全重合，这表明我们的预测模型相当准确。Azure 机器学习入门的初衷就是构建合理准确的预测模型，并在下一个阶段中进行应用。





5.2 保存实验

在此步骤中，我们将要保存实验的副本。在屏幕的底部点击"Save As"另存为按钮。在后面的实验中，你可能将实验的核心功能做出重大的修改，所以要先将实验另存，保存的名称建议具有描述性的说明，比如 Azure 机器学习的收入预测——训练模型试验（Azure ML Income Prediction – Train Model Experiment）。