|  |
| --- |
| * AutoML---salesforce/transmogrifai(google) |

**2022年9月9日**

目录

[1. 原始需求 3](#_Toc15842)

[2. 技术方案 3](#_Toc26236)

[2.1. 功能概述 3](#_Toc2154)

[2.2. 实现方案 3](#_Toc17565)

[2.2.1. 技术选型概述 3](#_Toc2685)

[2.2.2. 方案详述 4](#_Toc26660)

[2.3. 数据方案 4](#_Toc9542)

[2.3.1. 数据接口 4](#_Toc32715)

[2.3.2. 统计口径 6](#_Toc15855)

[2.3.3. 数据回刷 6](#_Toc20510)

[3. 对上下游影响 6](#_Toc21168)

[4. 相关附件 6](#_Toc715)

[5. 整体说明 6](#_Toc24947)

# 原始需求

* 需求编号:
* 需求名称:实现机器学习过程自动化

需求描述:

手动特征工程效率低下，可移植性低，所以需要机器学过程自动化

1. 对 raw data 进行预处理，特征提取和选择（top features）
2. 进行自动化特征工程（生成新的特征）
3. 进行自动化算法选择（确定具体下游任务，例如多分类任务，从该模块算法库中运行每一个相关算法，返回最优算法，并做预测）
4. 模型超参数优化，返回选取模型最优超参数数值

# 技术方案

## 功能概述

1. 超参数调节优化：超参数调节的方法为网格搜索，没有提供随机搜索算法和贝叶斯调参，这些预置模型中的每一个都带有一组预定义的超参数，将在确定最佳模型时进行测试。
2. 自动化特征推断与验证（feature inference and vaildation）：

可以删除几乎没有预测能力的特征，表现出零方差的特征，或者在训练样本中的分布与预测时的分布存在显着不同的特征

1. 自动化特征选择

选取与因变量的影响最大的top n 自变量

1. 自动化特征工程（生成新的特征）

TransmogrifAI为支持的所有特征类型提供了无数种编码技术，能做到不仅把数据转成算法可用的格式，还能优化转换，使机器学习算法更容易从数据中学习。例如，同样是年龄数字特征，它能根据特定问题（时尚行业、金融理财）把它们转成最合适的年龄段

1. 自动化算法（模型选择）

TransmogrifAI的模型选择器可以在数据上运行多种算法，并比较它们的平均验证错误，从中挑出最佳算法。除此之外，它还能通过适当地对数据进行采样并重新校准预测以匹配真实的先验，自动处理不平衡数据的问题，进一步提高模型性能。

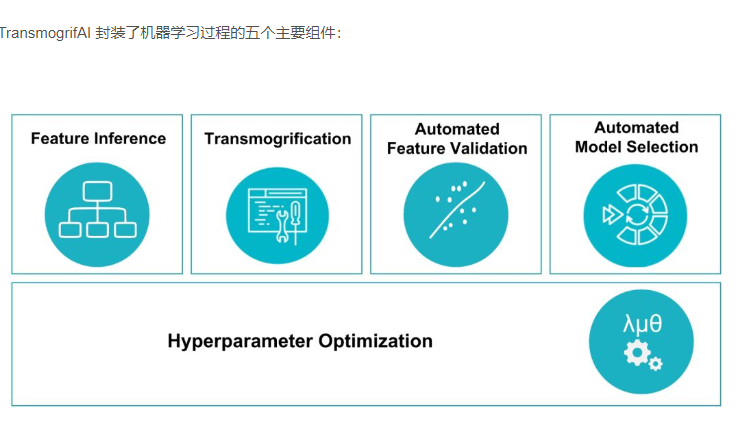
## 实现方案

### 技术选型概述

谷歌平台：AutoML Vision平台

TransmogrifAI 是一个基于 Scala 和 SparkML 构建的库，就是为了完成这项任务应运而生的，用于在Apache Spark上运行的Scala编写的结构化数据

自动完成超参数优化、特征工程和模型选择，从而获得一个高性能的模型，并在后续进一步的探索和迭代。



自动模型选择

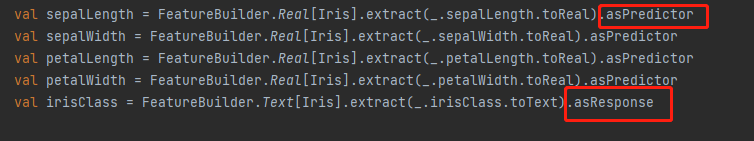
TransmogrifAI 模型选择器在数据上运行几种不同的机器学习算法，并基于平均验证错误自动选择最佳的算法。它还会对数据进行采样并重新校准预测来自动处理不平衡数据问题。数据科学家训练得出的最佳和最差模型的性能通常存在显著差距，而探索模型空间对于避免错过更好的模型来说是至关重要的

通过自动化，它实现了

1. 接近手动调整模型的精度，时间缩短了近100倍
2. 基于最新版的 Spark 构建, 利于使用和扩展

### 方案详述

1. 读取数据
2. 手动提取predictor，response特征



1. 自动化特征推断与验证（feature inference and vaildation）：



删除几乎没有预测能力的特征，表现出零方差的特征

SanityChecker是一种估计器，它可以在拟合模型之前分析特定数据集的明显问题。它基于特征类型对数据应用各种统计测试，并丢弃指示标签泄漏或显示很少或没有预测能力的预测值。除了标记和修复数据问题之外，SanityChecker还输出有关数据的统计信息，以便在ML管道中进一步进行诊断和洞察生成。

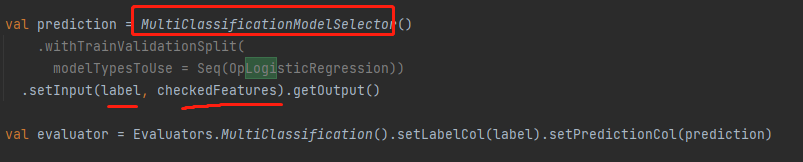
1. 选取想用的feature vector并实现自动化特征工程



自动构建新特征

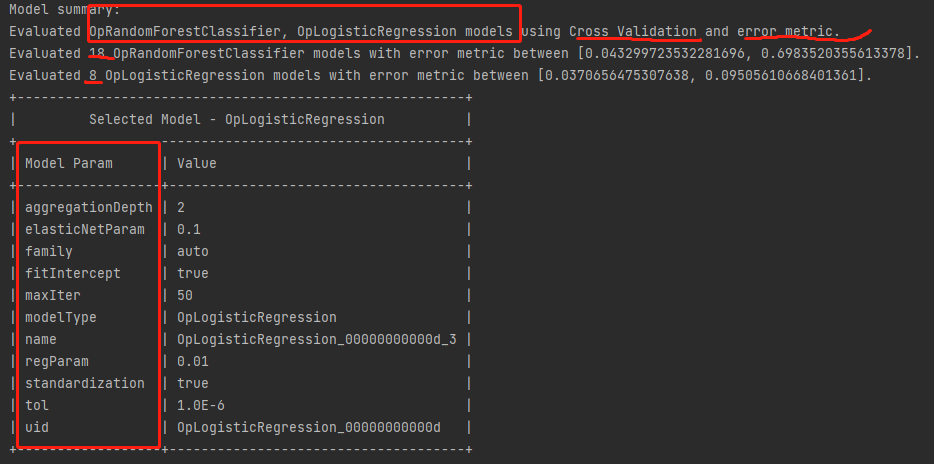
用户可以简单地对所有输入功能使用.transmogrify（）方法，而不是指定每个功能类型所需的所有操作。当然，特定功能工程也是可能的，可以与自动类型特定转换结合使用。

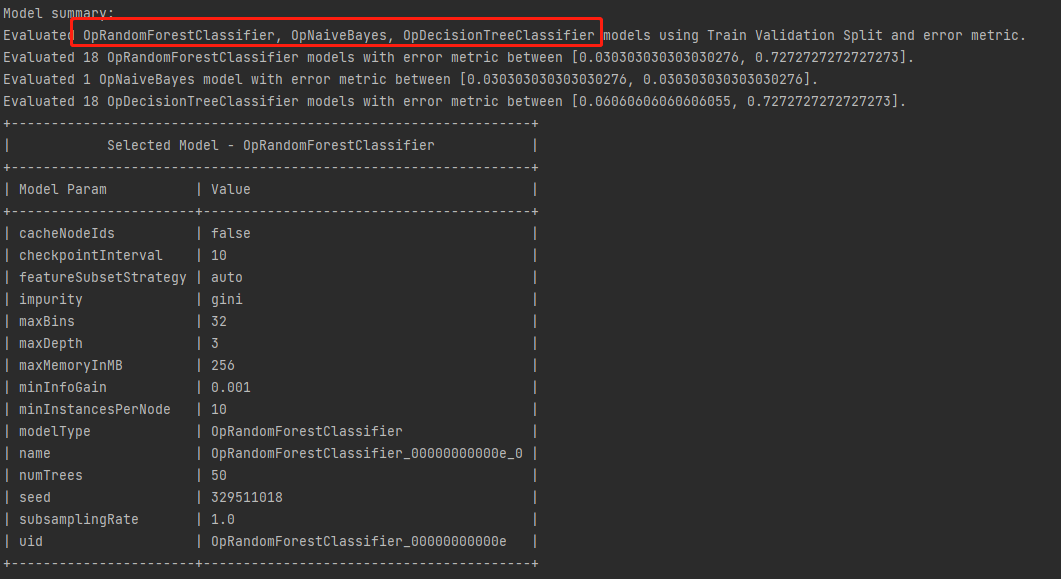
1. 构建下游算法种类模型（例如多分类）和评估器



1. 模型训练得到最优模型和最优超参数

下图分别为为两个，三个算法的最优选取，数量和种类可以自己设置





例如：

分类任务中可供选择的是GBTCLassifier, LinearSVC, LogisticRegression, DecisionTrees, RandomForest  NaiveBayes

回归任务可供选择的是：

GeneralizedLinearRegression, LinearRegression, DecisionTrees, RandomForest  GBTreeRegressor

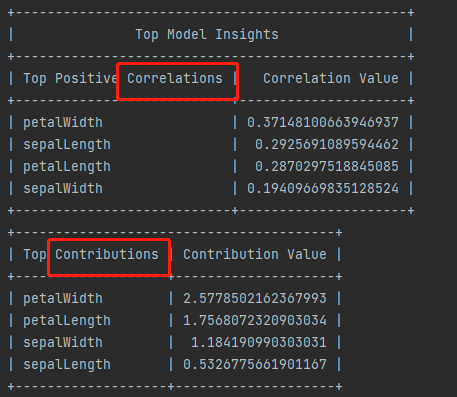
The best model is selected via a CrossValidation and error metric

1. 根据model insights 选取最优特征

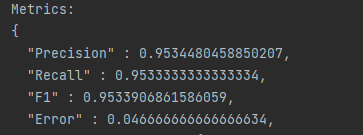


结果：

与因变量最相关的Top 4 features



7.最终所选取的模型验证最终结果



## 数据方案

### 数据接口

1. 接口格式约定
2. 使用标准http协议post方式进行请求，请求参数放在body中；返回包格式为json；
3. json数据采用标准json格式，即：所有字段用双引号括起来。
4. 请求url和json请求包返回包节点名称全部小写，统一utf-8格式；
5. 内容允许为空的含义：表示在特定情况下字段没有返回，而不是字段的值为空；
6. 输入数据格式

1）接口地址

2）功能说明

**3）请求参数说明**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **字段名称** | **字段说明** | **类型** | **是否允许为空** | **备注** |
| csvFilePath | 文件路径 | 字符串 | 否 |  |

1. 输出数据格式

|  |  |  |  |
| --- | --- | --- | --- |
| **字段名称** | **字段说明** | **是否允许为空** | **备注** |
| status | 响应状态 | 否 | Fail或者success。 |
| desc | 响应描述信息 | 是 | 如：成功 |
| **ret\_time** | 接口返回时间 | 否 | 时间戳 |
| **args.save\_model\_path** | 保存模型路径 | 否 |  |

### 统计口径

无

### 数据回刷

无

# 对上下游影响

无

# 相关附件

# 整体说明