Table of Contents

[句子相似度匹配项目开题报告 1](#_Toc535861331)

[项目背景概述 1](#_Toc535861332)

[问题陈述 2](#_Toc535861333)

[数据 2](#_Toc535861334)

[数据分析 3](#_Toc535861335)

[评估标准 4](#_Toc535861336)

[基准模型 5](#_Toc535861337)

[项目设计 6](#_Toc535861338)

[引用 6](#_Toc535861339)

# 句子相似度匹配项目开题报告

## 项目背景概述

Quora Question Pairs是Quora于2017年公开的句子匹配数据集，其通过给定两个句子的一致性标签标注，从而来判断句子是否一致。

该问题来自于Quora问答网站，在该网站上，用户可以在上面提问，并有其他用户给出高质量的回答或是独特的想法。这能鼓励人们互相学习，了解更多的知识。

每个月，超过一亿的用户会访问Quora。在这么大的访问量下，经常会有用户提出相似的问题。相似的问题会导致用户花费更多时间去寻找最佳答案，来回答他们想问的问题。而且也会让回答问题的用户觉得，对于同一个问题，他们需要回答多次。因此在Quora上，一个问题的最典型形式是很有价值的。它能给提问者和回答这提供最好的用户体验。

目前，Quora使用的是一个随机森林模型来寻找类似的问题。而本次项目就是要探索使用更先进的技术来判断给定的两个问题是否是重复问题。高准确率的判断能够帮助网站找到每个问题最佳的回答，从而提高提问者，回答者以及网站浏览者的用户体验。这就是该项目的出发点。

而在技术层面上，该问题属于二分类问题。给定数据后，我们需要训练模型要将数据划分到“相同问题”和“不同问题”两个类别中。同时，给定的训练数据中包含了人工分类后的标签结果，所以该分类任务属于监督学习中的问题。最后，由于我们处理的数据属于自然语言，所以该问题也涉及到自然语言处理。

Quora 数据集训练集共包含40K的句子对，且其完全来自于Quora网站自身。句子对中每个句子以字符串的形式存储，每个字符串即句子的自然语言表达。

## 问题陈述

这个项目中要解决的问题是判断两个问题是否表达相同的意思。

使用的训练数据集主要包含三列数据：第一列为问题1，数据类型为字符串吗，包含了问题对中第一个问题的自然语言表示形式，使用语言为英语。第二列数据为问题2，数据类型和内容与第一列数据类似，包含的是问题对中第二个问题的自然语言表达。第三列是表示问题对中两个问题是否相同的一个标签，该标签为数值类型，只包含0,1两个值。0代表两个问题意思不同，1代表两个问题意思相同。数据样例如下：

|  | **question1** | **question2** | **is\_duplicate** |
| --- | --- | --- | --- |
| **0** | What is the step by step guide to invest in sh... | What is the step by step guide to invest in sh... | 0 |
| **1** | What is the story of Kohinoor (Koh-i-Noor) Dia... | What would happen if the Indian government sto... | 0 |
| **2** | How can I increase the speed of my internet co... | How can Internet speed be increased by hacking... | 0 |
| **3** | Why am I mentally very lonely? How can I solve... | Find the remainder when [math]23^{24}[/math] i... | 0 |
| **4** | Which one dissolve in water quikly sugar, salt... | Which fish would survive in salt water? | 0 |
| **5** | Astrology: I am a Capricorn Sun Cap moon and c... | I'm a triple Capricorn (Sun, Moon and ascendan... | 1 |
| **6** | Should I buy tiago? | What keeps childern active and far from phone ... | 0 |
| **7** | How can I be a good geologist? | What should I do to be a great geologist? | 1 |

## 数据

上一章节已经提及到处理的数据，具体的数据集为Quora Question Pairs，是Quora于2017年公开的句子匹配数据集，其通过给定两个句子的一致性标签标注，从而来判断句子是否一致。

Quora 数据集训练集共包含40K的句子对，且其完全来自于Quora网站自身。句子对中每个句子以字符串的形式存储，每个字符串即句子的自然语言表达。具体数据案例已经在上一节中展示。

数据量较大，在此仅提供kaggle网站上该数据集的说明，下载地址：<https://www.kaggle.com/c/quora-question-pairs/submissions?sortBy=date&group=all&page=1>

### 数据分析

句子长度分析：

句子对中的句子长度和包含的词数等信息能够用来判断句子对是否能够表示相同含义。一般而言，在提问时，用户往往会用相似的词语和句子结构来描述相同的问题，而相似的句子结构和词语会使句子长度也比较相似。因此，在直觉上来说，相似长度的句子是相似句子的可能性更大。因此我们会分析句子长度相关的统计量。具体如下：

句子长度：句子字符串的长度

句子字符长度：句子中去除空格等空白符号后的字符长度

句子单词数量：用空格分割句子后得到的列表的长度。该长度为句子中单词和特殊符号的数量和。

进一步数据处理：直接获取到的长度特征为一个二元组，分别包含了第一个和第二个句子的长度特征。这个二元组不方便可视化，所以我将二元组的两个数据做比值并进一步处理来获取最终方便可视化的数据。以句子长度为例，处理过程如下：

句子长度的原始数据为二元组：（len\_q1, len\_q2）

将数据做比值，获得一个数据值：len\_q\_ratio = len\_q1 / len\_q2

为了反映该比值对判断的作用，我根据该比值将数据划分为较小的区间，并计算每个小区间中的数据中is\_duplicate段的平均值，平均值越大的区间表示该区间包含越多的正样本，而平均值越小，表示该区间包含的正样本越少。

但在可视化之前，发现len\_q\_ratio比值的分布空间为[0.006711409395973154, 117.0]。因为是比值，数据应该以1为中心分布，因此数据分布不均匀，不利于可视化，所以考虑使用对数操作调整数据分布。

比值对数：len\_q\_ratio\_ln = ln(len\_q1 / len\_q2)

该数值的分布大约在[-4，4]上。以0.04为小区间的划分范围，获得最终可视化分布如下：

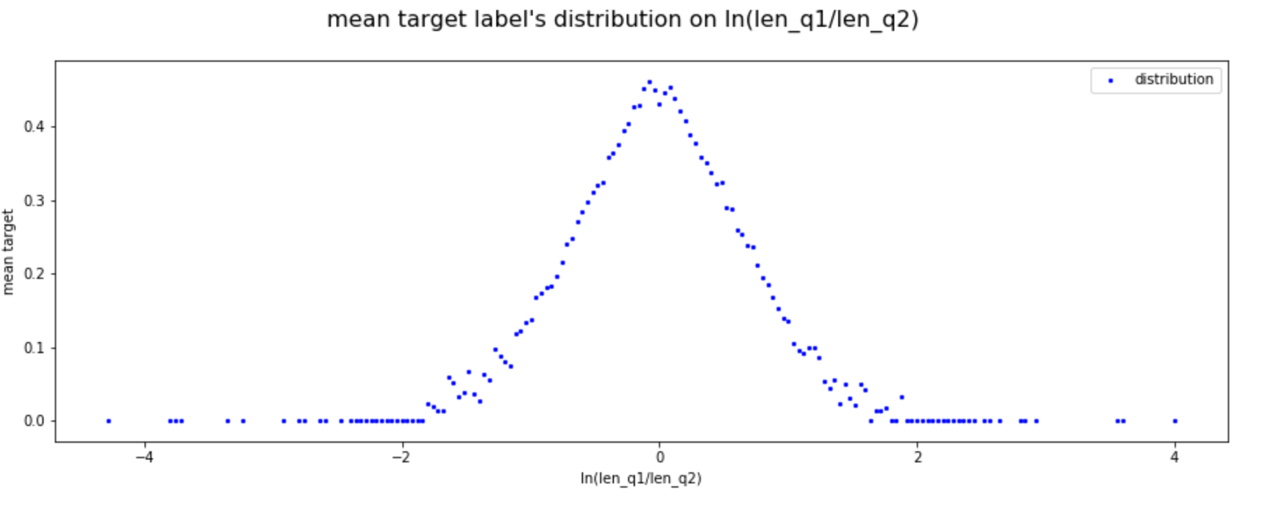
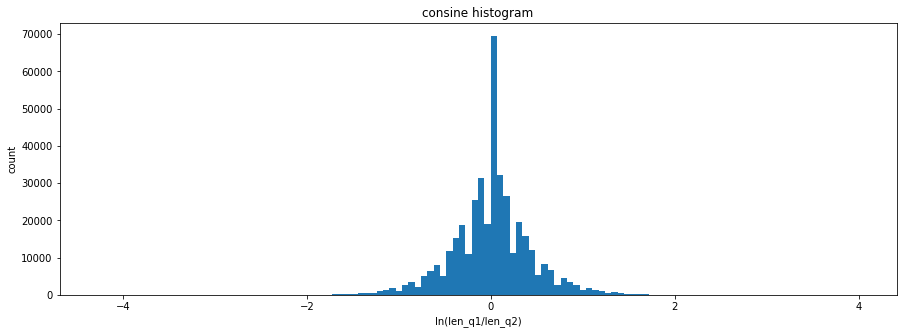


图 1

上图中和坐标为ln(len\_q1/len\_q2)，是我最终选定的比值对数特征。纵坐标为每个小区间中数据is\_duplicate字段的均值。

从图中可见，当比值对数在[-4, -2]和[2,4]之间时目标字段均值为0，说明该句子对长度差异过大时，基本不可能是相同含义的句子。而比值对数越接近0，目标字段均值越大，说明句子长度越相近，句子对的含义越肯能相同。因此该特征值对于模型的预测越有效果。

同时，我也对该比值的分布进行了可视化：

图2

同上，我也准备了句子字符长度和句子单词数量对应的比值对数分布，如下：

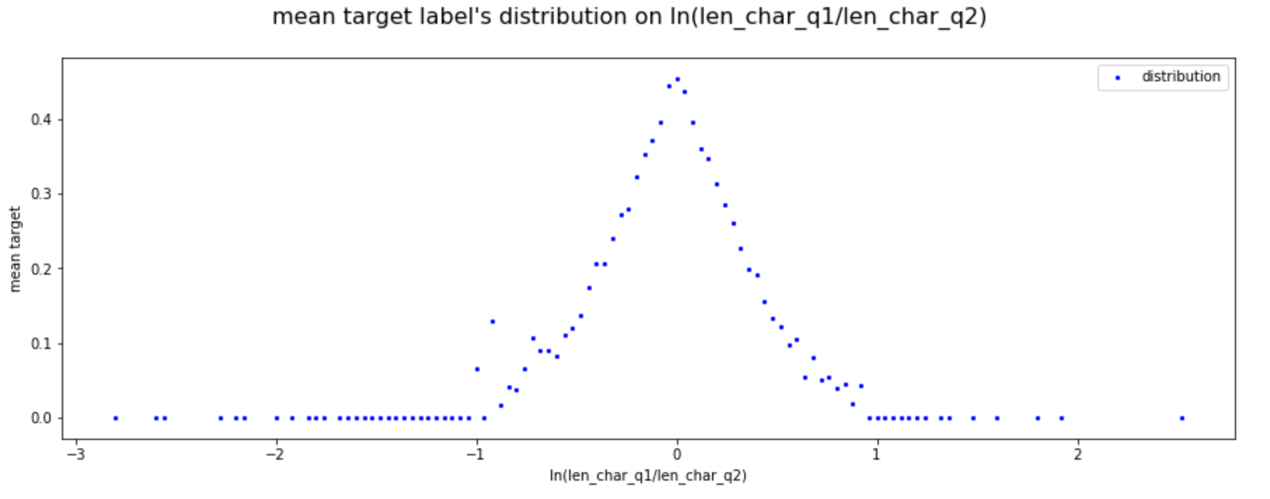


图3

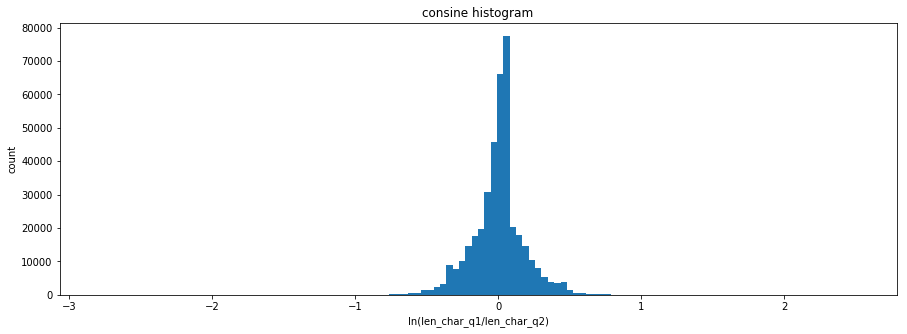


图4

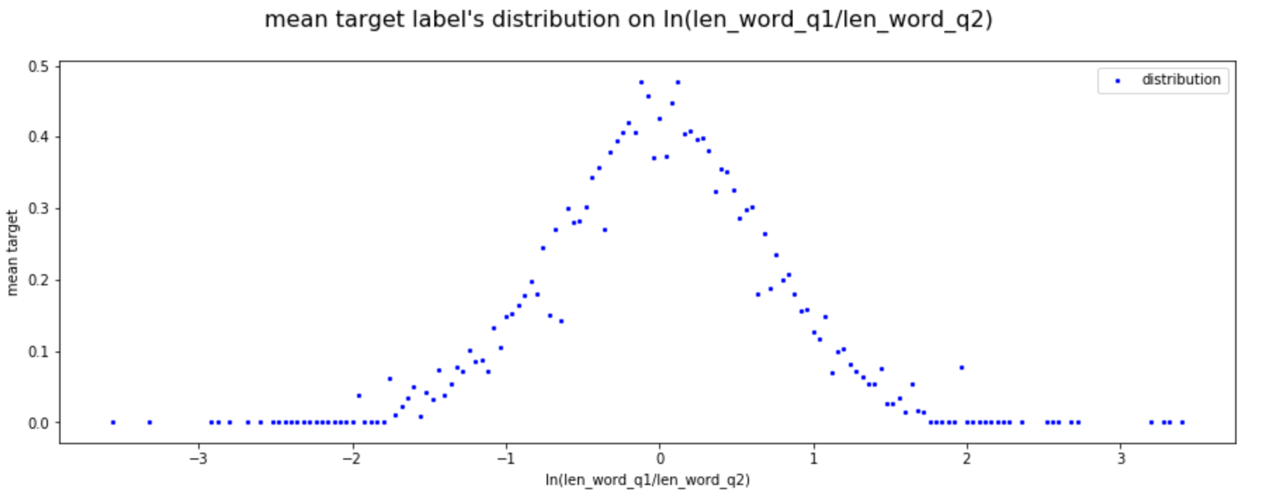


图5

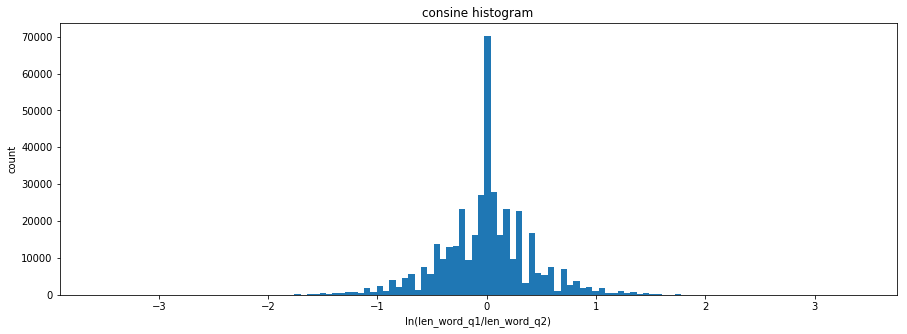


图6

发现分布与字符串长度的分布类似。

在此之后，再分析一下句子对中同时出现的单词数量与目标字段的关系。其中相同单词比例可定义为：

Commen = 2 \* commen\_count / (len1 + len2)

其中commen\_count为两个句子中同时出现的单词的数量，len1和len2分别为句子1和句子2包含的单词数量。

然后显示目标字段的均值在该比例上的分布，如下图：

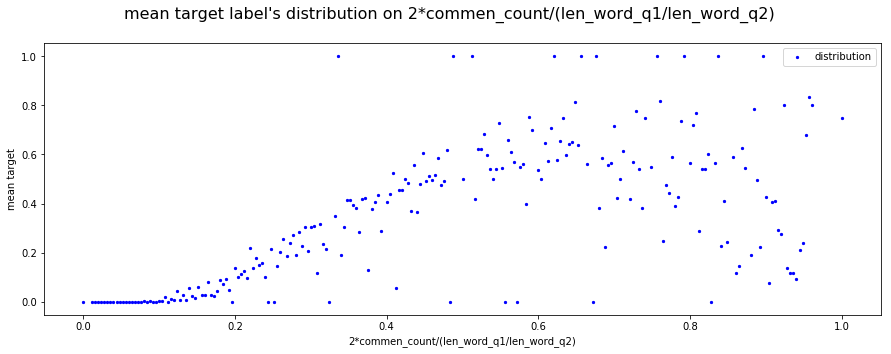


图7

从图7中可以看到，在比值较小时目标字段的均值较低，随着比值的升高，目标字段的均值分布开始变高。可见该比值对于句子对判断有作用。

而该比值的分布如下图：

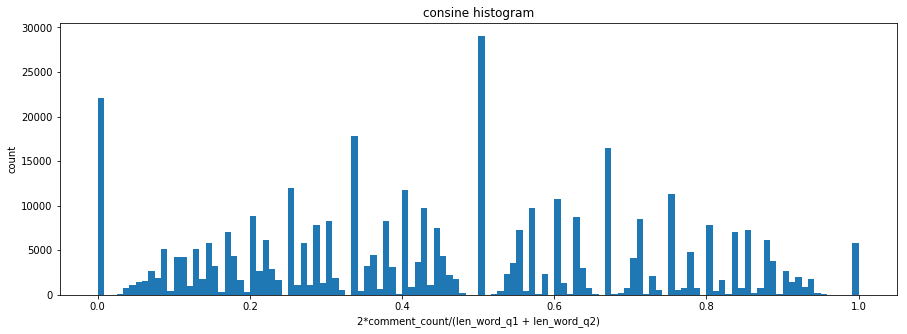


图8

从图8中可以看到，数据分布相对均匀，结合图7，可以发现，该比值能够帮助我们对大量数据进行划分。

## 评估标准

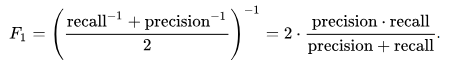
前面已经介绍过，该项目要解决的是一个分类问题，需要训练一个分类模型来解决该问题，因此可以考虑使用分类模型的评价标准。本次我们选取f1-score和accuracy score以及logloss来进行评价。F1-score和accuracy score评估指标是分类模型中非常典型，使用非常方便，而logloss是kaggle指定的评估标准，所以使用这几个评价标准。

accuracy score是计算模型预测准确性的指标，其定义为：

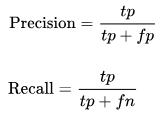


代表了预测模型中，准确预测的结果数量与总预测结果数量的比值。

f1-score的定义为：



其中precision和recall的定义为：



Precision代表了模型预测准确的阳性结果与所有预测为阳性结果的数量比，而Recall代表所有阳性结果中，被模型预测为阳性结果的比值。而f1-score则是Precision和Recall的调和平均数。能够综合反映模型预测的Precision和Recall。

这两个平均数能够反映出模型预测结果的准确性，因此用来做模型的评价标准。

Logloss, 即对数似然损失(Log-likelihood Loss), 也称逻辑斯谛回归损失(Logistic Loss), 是在概率估计上定义的.它常用于(multi-nominal, 多项)逻辑斯谛回归和神经网络,以及一些期望最大化算法的变体. 可用于评估分类器的概率输出.

对数损失通过惩罚错误的分类,实现对分类器的准确度(Accuracy)的量化. 最小化对数损失基本等价于最大化分类器的准确度.为了计算对数损失, 分类器必须提供对输入的所属的每个类别的概率值, 不只是最可能的类别. 对数损失函数的计算公式如下:



其中, Y 为输出变量, X为输入变量, L 为损失函数. N为输入样本量, M为可能的类别数, yij 是一个二值指标, 表示类别 j 是否是输入实例 xi 的真实类别。pij 为模型或分类器预测输入实例xi属于类别 j 的概率。Logloss越小表示其预测效果越好。

在实际模型训练过程中，我们将会使用sklearn中的train\_test\_split工具对训练数据集进行分割，以获取训练数据集和测试数据集。并用我们选取的指标，准确率， f1分数和logloss，测试模型在训练和测试数据集上的性能表现，以此作为评估标准。

## 基准模型

项目中我选取了RandomForest，GBDT，LogsitcRegression和SGD四个分类模型对该问题进行预测。基准模型为这四个模型的默认参数，并给定一个固定的random\_state以方便的进行后续的模型对比。

训练数据和测试数据生产过程如下：

Quora给定的训练数据集中包含约40万条数据。通过我们之前介绍的方法生成变量后，又对数据进行了清洗。

清洗过程主要去除了数据中的空值，NaN值以及无限大数值。最后对得到的数据集进行去重，获取到最终可以使用的数据。

最后对数据进行切割。以1：4的比例分割数据，获取到测试数据集和基本训练数据集，再对基本训练数据集以1：4的比例分割，得到最终的验证数据集合训练数据集。

各个数据集的尺寸如下:

|  |  |  |  |
| --- | --- | --- | --- |
| Train Set | Test Set | Validation Set | All |
| 251436 | 78575 | 62860 | 392871 |

表2

训练基本模型时，使用训练数据集新型模型训练，使用测试数据集对模型进行测试。验证数据集将会在后面的模型调优过程中使用。

因为是基准模型，使用默认值参数进行训练。得到模型的效果如下结果如下：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| model | LogLoss | | F1-Score | | Accuracy | |
| Train | Test | Train | Test | Train | Test |
| RandomForest | 0.337508 | 9.314371 | 0.985816 | 0.618121 | 0.989503 | 0.732165 |
| GBDT | 9.205588 | 9.339426 | 0.674664 | 0.669399 | 0.734828 | 0.729203 |
| LogsitcRegression | 10.972802 | 11.009331 | 0.577043 | 0.578439 | 0.681540 | 0.681716 |
| SGD | 20.730326 | 20.711837 | 0.000000 | 0.000000 | 0.626821 | 0.626817 |

表3

从表3可以看出，除SGD外，其他几个模型都有一定的预测效果。其中效果最好的是GBDT模型，训练和测试的准确率都很高。而RandomForest明显出现过拟合，而SGD则明显欠拟合。考虑是没有进行对应参数调节造成的。

由于最终的评价标准为logloss，我们在选择基准模型时将以logloss为主要选择标准。而从logloss数据上也能得到上面的结论。RandomForest明显过拟合，而SGD明显欠拟合。再比较logloss数值，GBDT是最好的模型，将被定为基准模型。

## 项目设计

这个问题是分类问题，所以使用分类模型来解决该问题。方案的基本思路按照分类模型的训练过程进行：首先从原始数据中提取特征，然后选择并训练模型，并进行模型调参。最后，选择效果最好的模型，并用该模型解决问题。具体细节如下

1. 特征提取
   1. 原始数据为自然语言，无法直接作为分类模型的输入，因此要在自然语言的基础上生成一些特征，例如字符串的长度，包含单词数量等。
   2. 由于原始数据为自然语言，因此可以使用自然语言处理的技术来生成特征。例如使用词袋模型，tf-idf模型，句子模糊匹配等技术对自然语言处理，并生成与自然语言相关的特征。
2. 模型训练
   1. 数据清洗：通过特征提取获得了特征数据，但该数据中往往包含一些异常数据。例如空值，或数据范围过大的数据。因此在训练模型前需要对数据进行清理
   2. 分割训练数据集和测试数据集。由于后续涉及到模型选择，还需要检验数据集（validation data set）
   3. 对选取的模型进行调参
3. 模型选择
   1. 模型训练过程中会选取多个分类模型进行训练。在最后比较模型间性能差异，通过测试数据集选择最好的模型来解决问题

以上为解决问题的过程。最终期望的结果是能够获得效果明显的模型。通过该模型判断给定的两个句子是否含有相同得到意思。

## 引用

Quora Question Pairs背景信息：<https://www.kaggle.com/c/quora-question-pairs>

Matplotlib visualization：

Matplotlib: A 2D Graphics Environment：

<http://aip.scitation.org/doi/pdf/10.1109/MCSE.2007.55?class=pdf>

Beginning Python Visualization: Crafting Visual Transformation Scripts：

<https://books.google.com/books?hl=zh-CN&lr=&id=N1InCgAAQBAJ&oi=fnd&pg=PP3&dq=python+visualization+matplotlib&ots=9GN8i96vNs&sig=wT6Oa4_ehjizHz2hnTpM0teGM-w#v=onepage&q=python%20visualization%20matplotlib&f=false>

Python matplotlib 数据可视化：

<https://absentm.github.io/2017/03/18/Python-matplotlib-%E6%95%B0%E6%8D%AE%E5%8F%AF%E8%A7%86%E5%8C%96/>

评价指标：

Log Loss： <http://wiki.fast.ai/index.php/Log_Loss>

对数损失函数(Logarithmic Loss Function)的原理和 Python 实现：

<https://www.cnblogs.com/klchang/p/9217551.html>

Accuracy, Precision, Recall or F1：

<https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9>

机器学习算法：

Understanding Logistic Regression in Python：<https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python>

The Random Forest Algorithm：

<https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd>

GBDT：梯度提升决策树

https://www.jianshu.com/p/005a4e6ac775