Kaggle Introdution

2019年5月30日　　下午 01:50

資料來源：https://zhuanlan.zhihu.com/p/25686876

1. 甚麼是Kaggle

   Kaggle成立于2010年，是一個進行資料發掘和預測競賽的線上平臺。 從公司的角度來講，可以提供一些資料，進而提出一個實際需要解決的問題；從參賽者的角度來講，他們將組隊參與專案，針對其中一個問題提出解決方案，最終由公司選出的最佳方案可以獲得5K-10K美金的獎金。

   除此之外，Kaggle官方每年還會舉辦一次大規模的競賽，獎金高達一百萬美金，吸引了廣大的資料科學愛好者參與其中。 從某種角度來講，大家可以把它理解為一個眾包平臺。 但是不同于傳統的低層次工作力需求，Kaggle一直致力於解決業界難題，因此也創造了一種全新的工作力市場——不再以學歷和工作經驗作為唯一的人才評判標準，而是著眼于個人技能，為頂尖人才和公司之間搭建了一座橋樑。

2. Kaggle 的競賽模式

   Kaggle上的競賽有各種分類，例如獎金極高競爭激烈的的「Featured」，相對平民化的「Research」等等。 但他們整體的專案模式是一樣的，就是通過出題方給予的訓練集建立模型，再利用測試集算出結果用來評比。

   同時，每個進行中的競賽專案都會顯示剩餘時間、參與的隊伍數量以及獎金金額，並且還會即時更新選手排位。 在截止日期之前，所有隊伍都可以自由加入競賽，或者對已經提交的方案進行完善，因此排名也會不斷變動，不到最後一刻誰都不知道花落誰家。

   由於這類問題並沒有標準答案，只有無限逼近最優解，所以這樣的模式可以激勵參與者提出更好的方案，甚至推動整個行業的發展。

   Kaggle競賽另一個有趣的地方在於每個人都有自己的Profile，上面會顯示所有自己參與過的專案、活躍度、即時排位、歷史最佳排位等，不僅看上去非常有成就感，更能在求職和申請的時候起到Certificate的作用。

3. Kaggle的意義

   Kaggle提供了一個非常好的學習平臺，在這裡你可以接觸到真正的業界案例，收穫實際的專案經驗，在每一個專案中不斷挑戰自己，甚至在Kaggle榜上佔據一席之位，提高自己在業內的知名度， 優秀的排位甚至可能帶來的非常好的工作機會。 同時，也可以認識一群志同道合的人，擴展自己的professional network，與業內最頂尖的高手互動，尤其是很多隊伍在比賽結束後都會公開自己的解法，如果這個專案恰好你參與過，為之投入過無數個日日夜夜， 此時就是不可多得的學習機會。

   對於剛剛進入這個行業的菜鳥而言，參加Kaggle的專案是非常「長見識」的，可能初期的嘗試會非常吃力，畢竟都是非常前沿的問題，但是如果能堅持完整的把一個專案做下來，且不說coding能力會有一個很大的提高， 在實際案例中解決問題的能力也會得到極

大的鍛煉，為自己的職業生涯打下一個良好的基礎。 如果能在Kaggle這種高手雲集的比賽中獲得一個還不錯的成績，寫在簡歷上足以打動你今後的Boss，跳槽就翻倍的高薪工作指日可待！ 值得一提的是，雖然是彙集精英的平台，但是Kaggle的論壇氛圍很好，對新人非常友好，大家一定要多看Script多請教！

4. 新人該如何上手

理論上來講，Kaggle歡迎任何資料科學的愛好者，不過實際上，要想真的參與其中，還是有一定門檻的。 一般來講，參賽者最好具有統計、電腦或數學相關背景，有一定的coding技能，對機器學習和深度學習有基本的瞭解。 Kaggle任務雖然不限制程式設計語言，但絕大多數隊伍會選用Python和R，所以你應該至少熟悉其中一種。 此外，對於那些對成績有追求的人，Feature Engineering也是必不可少的。 但對於Data Science的入門者來說，這樣的要求實在是有些過分了。 對於這一塊想要進一步瞭解的同學可以看這個問題：特徵工程到底是什麼？

> 特征工程到底是什么？ - 城东的回答 - 知乎
> https://www.zhihu.com/question/29316149/answer/110159647

當然，如果你從未獨立做過一個專案，還是要從練習賽開始熟悉。 因為競賽模式中的任務是公司懸賞發佈的實際案例，並沒有標準的答案；而練習賽不僅專案難度低，而且是有官方給出的參考方案的，大家可以用來對比改善自己的測試結果，從中進行提高。 所以呢，建議感興趣的同學先去獨立做一下101和playground的訓練賽，至於做多少個案例才能上道，就要看個人素質啦。 這裡為大家推薦幾篇非常好的文章，裡面手把手的教了大家入門級的三個經典練習專案，供大家學習。

> 1. Titanic（泰坦尼克之灾）
> 中文教程： 逻辑回归应用之Kaggle泰坦尼克之灾
> 英文教程：An Interactive Data Science Tutorial
> 2. House Prices: Advanced Regression Techniques（房价预测）
> 中文教程：Kaggle竞赛 — 2017年房价预测
> 英文教程：How to get to TOP 25% with Simple Model using sklearn
> 3. Digital Recognition（数字识别）
> 中文教程：大数据竞赛平台—Kaggle 入门
> 英文教程：Interactive Intro to Dimensionality Reduction

5. 如何提升排名

• 選擇合適的隊友：

由於Kaggle的專案是由公司提供的，涉及各個行業，所以一般都是不同背景的人組隊參加（如統計、CS、DS，專案相關領域如生物等）。 因此對於新手來講，很重要的一點就是要抱好大腿，不僅可以蹭到好的排名，還有機會近距離向大牛學習，技能值必然漲。 而自己可以從力所能及的工作做起，如清洗資料等等，積累專案經驗。

• 選擇「正確」的專案；

首先，選擇資料量小的專案，這樣不管使用什麼演算法都不會耗時太久，對機器性能要求也不高，出結果也比較快；其次，選擇難度低獎金少的專案，一方面競爭小，另一方

面也適合新手；最後，選擇參與人數多的專案，畢竟有那麼多「僵屍號」撐著。 這樣下來，基本上認認真真做下來排名都不會太難看。

• 選擇恰當的工具：

我們都知道循序漸進的道理，因此對於剛剛涉獵Kaggle，只是希望從中學習，而不追求高排名的同學，可以先從學習Machine Learning中常用的模型開始，比如Logistic Regression和Random Forest，這兩個模型對於大部分問題就夠了；基礎好的還可以學習一下Gradient Boosting，雖然難度高一點，但是視覺化效果會好很多。

當然，說到底，想獲得更好的名次，提高自己的Skills才是終極解決方案！