Hey Reddit, I'm Anthony Goldbloom, founder of Kaggle. We recently teamed up with Google Cloud and NCAA® to apply machine learning to forecast the outcomes of March Madness®. AMA!

GoogleCloudOfficial [1] and r/Science AMAs[1]

[1]Affiliation not available

April 17, 2023

Abstract

Hi, I'm Anthony Goldbloom, co-founder and CEO of Kaggle. Kaggle is the world's largest community of data scientists and machine learners with over 1.4 million members. Data scientists come to Kaggle to compete in machine learning competitions, find and share open datasets and use Kaggle Kernels (Kaggle's cloud based data science workbench). Before starting Kaggle, I was a statistician at the Reserve Bank of Australia and the Australian Treasury, building models that forecast economic activity. The MIT Review has named me one of the top 35 innovators under 35 and Forbes has named me as one of the 30 under 30 in technology. For the first time, Kaggle, Google Cloud, and the NCAA ® will join together for the largest data-driven bracketology competition to date. As part of our continued collaboration, we've partnered with the NCAA to make 10 years (2008-2018) of historical NCAA Division I men's and women's basketball data available. This competition will be your chance to forecast the outcomes of March Madness® for both the Men's and Women's Basketball Championships. In my spare time I do kitefoil racing. I've written a bunch of kitefoiling related apps: two smart watch apps - a training app and a wind reporting app a Strava App used Kaggle Kernels to create a ranking system for kitefoilers (at last update, I was ranked 109). Proof I will be here to answer your questions at 1pm ET. EDIT: THANKS FOR THE QUESTIONS. THIS WAS MY FIRST REDDIT AMA. PLAN TO POP BACK LATER TODAY TO TRY TO ANSWER A FEW MORE QUESTIONS.

the WINNOWER

REDDIT

Hey Reddit, I'm Anthony Goldbloom, founder of Kaggle. We recently teamed up with Google Cloud and NCAA® to apply machine learning to forecast the outcomes of March Madness®. AMA!

GOOGLECLOUDOFFICIAL R/SCIENCE

Hi, I'm Anthony Goldbloom, co-founder and CEO of Kaggle. Kaggle is the world's largest community of data scientists and machine learners with over 1.4 million members. Data scientists come to Kaggle to compete in machine learning competitions, find and share open datasets and use Kaggle Kernels (Kaggle's cloud based data science workbench). Before starting Kaggle, I was a statistician at the Reserve Bank of Australia and the Australian Treasury, building models that forecast economic activity. The MIT Review has named me one of the top 35 innovators under 35 and Forbes has named me as one of the 30 under 30 in technology.
For the first time, Kaggle, Google Cloud, and the NCAA ® will join together for the largest data-driven bracketology competition to date. As part of our continued collaboration, we've partnered with the NCAA to make 10 years (2008-2018) of historical NCAA Division I men's and women's basketball data available. This competition will be your chance to forecast the outcomes of March Madness® for both the Men's and Women's Basketball Championships.
In my spare time I do kitefoil racing. I've written a bunch of kitefoiling related apps:
two smart watch apps - a training app and a wind reporting app
a Strava App
used Kaggle Kernels to create a ranking system for kitefoilers (at last update, I was ranked 109).
Proof
I will be here to answer your questions at 1pm ET.
EDIT: THANKS FOR THE QUESTIONS. THIS WAS MY FIRST REDDIT AMA. PLAN TO POP BACK LATER TODAY TO TRY TO ANSWER A FEW MORE QUESTIONS.

CORRESPONDENCE:

DATE RECEIVED:
March 01, 2018

DOI:
10.15200/winn.151984.40299

ARCHIVED:
February 28, 2018

CITATION:
GoogleCloudOfficial , r/Science , Hey Reddit, I'm Anthony Goldbloom, founder of Kaggle. We recently teamed up with Google Cloud and NCAA® to apply machine learning to

1) Why do you think machine learning will be effective at predicting the outcome of sports games?

2) Will you test the models on more than one year of March Madness to ensure that the winners aren't just due to luck?

DrFilbert

I actually think that tournaments like NCAA Marchness are not the sweet spot for machine learning. There are many fewer March Madness games than ad clicks/fraud events etc. One of the most interesting areas for machine learning research is figuring out how to make machine learning more powerful on smaller datasets.

This competition is only judged based on the 2018 tournament. We have Kagglers predict the probability a team will win. So if a model is very confident and correct, it gets more points than a model that is less confident and correct. And on the flipside, if a model is very confident and incorrect it loses more points. Using that approach makes it harder to win the competition with dumb luck.

forecast the outcomes of March Madness®. AMA!, *The Winnower* 5:e151984.40299 , 2018 , DOI: 10.15200/winn.151984.40299

Love that name. I hope you guys really flex your data and build a tight organization. Who came up with/ decided on the name ?

TheSwearPolice

It's my fault :). I wrote an algorithm that iterated over phonetic domain names and printed out a list of those that were available. I then sent around a vote to friends and family.

I'm from Australia originally and so had never heard of "kegel" exercises. And the Australian pronounciation of Kaggle is very different from kegel. Once we moved to the US and realized the (unfortunate) pronunciation we considered changing the name...

Has it been harder to get Kaggle competitions since Google acquired Kaggle? I ask because there are only two paid non-Google competitions available at: https://www.kaggle.com/competitions

I haven't participated in many, but I've kept an eye on them and it feels like there are fewer competitions these days.

JoshVarty

No. Quite the opposite. We've been swamped by competition demand since the acquisition. My sense is that big customers are more willing to work with us now that we're not just a small standalone company.

The reason you're seeing so many Google competitions is because we're currently understaffed and unable to handle the demand. We've prioritized Google problems because we can outsource the work behind setting those competitions up to other Google teams, which allows us to launch more competitions given the size of our current team.

And by the way... we're hiring for our competition team if anybody wants to help us clear our launch backlog.

Hi, I just wanted to say I love Kaggle. I do full time Data Strategy consulting and if I ever need inspiration for portfolio work I browse Kaggle looking through all the great data sets that you have.

I have a question - In what direction do you think governments will trend regarding Open Data in the next 5 - 10 years. Right now it seems to vary by country, but wondering if you had any thought on how it could play out globally.

OffTheChartsC

My view is that potential of open data has not been reached. There are fewer success stories than I would have hoped.

Open data is actually a big focus for Kaggle. We have apublic data platform that allows our community to share public datasets. We also allow our community to share their analysis on that data using our cloud-based workbench called Kaggle Kernels. Our hope is that by having our open data platform be more than just a data catalogue, that it attracts more engagement and helps the open data movement reach its potential.

What advice do you have for new college graduates who are looking to get started in the data science field? My younger brother just finished his degree in engineering, but has an interest in data/machine learning and doesn't know where to begin looking.

up_um0p

I'm pretty biased, but I think Kaggle is a great place to start ;).

Shared some other learning resources in a previous answer

Hi Anthony! Thank you for taking the time to join us for this AMA. One of our favorite questions to ask every person who holds an AMA here is: Can you remember a time where the use of statistics dramatically changed your opinion on something? A scenario where the stats disproved many of your preconceived notions about a topic?

rhiever

Great question... I'm thinking about it.

Help me win my March Madness bracket! What tips do you have for this year? Also, does your strategy change based on how many people your competing against?

Bonobo42

Read the forums and look at other people's kernels. There's some great stuff in there.

For example, there's an awesome thread in the forum for the women's competition pointing out that upsets are less common in the female tournament. That probably means that you need to make sure your model is predicting with more conviction for the female tournament.

Of course, you're going to have to come up with unique ideas to win...

A lot of scientists still use MATLAB. Are there any plans for Kaggle to add first-class support for MATLAB? How about on Google Cloud?

crossmirage

We don't have plans to add first call support for MATLAB. We have prioritized Python and R for Kaggle Kernels. They're open source, have a vibrant package ecosystem and are by far the most popular choices in our community.

Lightly related but hopefully interesting: R was most dominant in Kaggle's early days but it's now being eclipsed by Python.

Will you be assessing the accuracy of your forecasting for the tournament as compared to other ratings systems like Kenpom, T-rank, fivethirtyeight, etc?

I'd love to see a comparison at the end.

INeedMoreCreativity

We won't be. But it'd be a great forum post or kernel for somebody to share. I suspect other Kagglers

would also be very interested in this.

Checking out your Kaggle profile, it looks like you don't compete often, or at least recently. Is that a conflict of interest thing?

Jhinra

Or just that I'm insecure about my abilities ;).

With my job and a new baby, I find it hard to find time to put significant effort into a competition. I find it easier to play with Kaggle Kernels and the public data platform. There's no deadline so I can drop in when I have time.

I imagine Kaggle gets a lot of *interesting* competition proposals. Are there any weird (or terrible) proposals that come to mind that Kaggle had to reject?

Jhinra

Most times we reject a competition it's because it's not a good fit for machine learning (not enough data or the problem is too vague).

We have run competitions before that didn't find any signal. for example trying to identifypeople's personality type from their Tweets.

We have recently learned that Bill Gates uses tabs instead of spaces, what about you?

Are you a space guy or a tab guy?

Amlo12345678

The Kaggle convention is spaces.

I personally like to mix it up ;)

Welcome Anthony! Were you a basketball fan before this project? Has it changed the way you look at the game?

GarretHobart

As an Australian, I never really had much exposure to college basketball. March Madness competitions are my co-worker, willis77's idea. He played basketball but I suspect he was a better data scientist than basketball player... which is why he proposed it.

How do you see machine learning playing a role in other areas besides sports?

Paris720

I'd say sports is actually a relatively minor (albeit fun) area for machine learning.

I'm proud of the competitions we've hosted around automated essay grading and medical diagnosis for example.

Hi Anthony,

Does your algorithm account for random spoiler effects (upsets) or is strictly probability of winning each game?

pipsdontsqueak

To clarify, Kaggle doesn't build a March Madness model. We host a competition where data scientists can submit their models and we judge their performance.

Basic models will take account of factors like win-loss %. To win, a data scientist is likely to have to use much more sophisticated features. There's been nice discussion in our forums about the altitude of Denver making it harder for visiting teams for eg

Do you see super computers or quantum computers as a major leap forward for ML and if yes in which areas of predicting models (apart from the weather;))

incommunicado2000

GPUs have been crucial to the deep learning revolution. TPUs are the next promising hardware leap that's imminent. Quantum Computing is still a way out.

Impossible to predict what new uses cases will be unlocked by the next big leap in hardware....

How accurate do you predict your model to be? Do you foresee your model being an issue in terms of sports betting?

artvol11

One top Kaggler is saying he expects his model to score a log-loss of 0.52..

You'd have to do some mapping to betting market odds to know whether that's accurate enough to have an advantage in the betting markets.

What kind of pizza do you like?

cfatt

Dislike pie charts ;)

Where and when will we get to see the published results of this project?

DragonSwagin

Final results posted here:

Mens: https://www.kaggle.com/c/mens-machine-learning-competition-2018/leaderboard

Womens: https://www.kaggle.com/c/womens-machine-learning-competition-2018/leaderboard

When can we expect skynet?

paramach

In my view, machine learning is a useful tool and we are nowhere near achieving artificial general intelligence.

How do you plan on monetizing your incredible user base of data science competitors and the algorithms they create?

Also, any chance you guys are hiring?

itshuey88

We charge for hosting competitions. And we also have ajobs board.

Yes we are hiring!

Did you run the analysis for previous years? How accurately did your models predict those outcomes? What were the strongest correlations tired to winning individual games?

GwnHobby

There's a forum thread on who are the perennial top performers in the March Madness competitions. These are the people you want to ask about what relationships really work.

As mentioned above, you should read and ask questions in the forums. There are lots of interesting discussions on topics ranging from the altitude of home stadiums, the importance of derived statistics (e.g. possession % measures) and about college teams that have more recent NBA drafts attracting better new talent.

What's your favourite coding environment? I'm partial to vim

zu7iv

Kaggle Kernels of course!

I also like vim.

What is your favorite kind of data?

palmp3

Weather data

What technology do you see standing out and having the biggest impact on Data Analysis over the next 5 years?

I-DrawLines

To quote the hackneyed William Gibson quote: "the future is already here, it's just notwidely

distributed".

Deep learning is a major breakthrough but it's not really used much outside of companies like Google. We'll see big leaps in everything from medical diagnosis to insurance claim processing as a result of what we can do with image data thanks to deep learning. We've barely scratched the surface...