(https://analyticsindiamag.com/)

PUBLISHED ON JANUARY 14, 2021

IN CAREERS (HTTPS://ANALYTICSINDIAMAG.COM/CATEGORY/CAREERS/)

Top XGBoost Interview Questions For Data Scientists

By Ambika Choudhury(https://analyticsindiamag.com/author/ambika-choudhury/)

Introduced a few years ago by Tianqi Chen and his team of researchers at the University of Washington, eXtreme Gradient Boosting (https://xgboost.ai/) or XGBoost is a popular and efficient gradient boosting method (https://analyticsindiamag.com/guide-to-ensemble-methods-bagging-vs-boosting/). XGBoost is an optimised distributed gradient boosting library (https://analyticsindiamag.com/introduction-to-boosting-implementing-adaboost-in-python/), which is highly efficient, flexible and portable.

The method is used for supervised learning problems and has been widely applied by data scientists to get optimised results for various machine learning challenges. It implements ML algorithms under the Gradient Boosting framework

(https://analyticsindiamag.com/primer-ensemble-learning-bagging-boosting/) and helps in solving data science problems in a fast and accurate manner.

Here are the top ten interview questions on XGBoost that Data Scientists must know.

1| Is XGBoost faster than random forest?

Solution: XGBoost is usually used to train gradient-boosted decision trees (GBDT) and other gradient boosted models. Random forests (https://xgboost.readthedocs.io/en/latest/tutorials/rf.html) also use the same model representation and inference as gradient-boosted decision trees, but it is a different training algorithm. XGBoost can be used to train a standalone random forest. Also, random forest can be used as a base model for gradient boosting techniques.

Further, random forest (https://analyticsindiamag.com/random-forest-vs-xgboost-comparing-tree-based-algorithms-with-codes/) is an improvement over bagging that helps in reducing the variance. Random forest builds trees in parallel, while in boosting, trees are built sequentially. Meaning, each of the trees is grown using information from previously grown trees, unlike bagging, where multiple copies of original training data are created and fit separate decision tree on each. This is the reason why XGBoost generally performs better than random forest.

Know more here (https://kharshit.github.io/blog/2018/02/23/gradient-boosted-trees-better-than-random-forest#:~:text=That's%20why%20it%20generally%20performs%20better%20than%20random%20forest.&text=Random%20forest%20build%20treees%20in,2%20for%20its%20better%20results.).

2| What are the advantages and disadvantages of XGBoost?

Advantages:

- XGB consists of a number of hyper-parameters that can be tuned — a primary advantage over gradient boosting machines.
- XGBoost has an in-built capability to handle missing values.
- It provides various intuitive features, such as parallelisation, distributed computing, cache optimisation, and more.

Disadvantages:

- Like any other boosting method, XGB is sensitive to outliers.
- Unlike LightGBM, in XGB, one has to manually create dummy variable/ label encoding for categorical features before feeding them into the models.

Know more here (https://www.kaggle.com/questions-and-answers/77947).

3| How XGBoost Works?

Solution: When using gradient boosting for regression (https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost-HowItWorks.html), where the weak learners are considered to be regression trees, each of the regression trees maps an input data point to one of its leaves that includes a continuous score. XGB minimises a regularised objective function that merges a convex loss function, which is based on the variation between the target outputs and the predicted outputs. The training then proceeds iteratively, adding new trees with the capability to predict the residuals as well as errors of prior trees that are then coupled with the previous trees to make the final prediction.

Click here (https://analyticsindiamag.com/xgboost-internal-working-to-make-decision-trees-and-deduce-predictions/) to learn the step by step process of how XGB works.

4| What does the weight of XGB leaf nodes mean? How to calculate it?

Solution: The "leaf weight" can be said as the model's predicted output associated with each leaf (exit) node. Here is an instance of how to calculate the weights of the leaf nodes in XGB-

Consider a test data point, where age=10 and gender=female.To get the prediction for the data point, the tree is traversed from the top to bottom, performing a series of tests. At each of the intermediate nodes, a feature is needed to compare against a threshold.

Now, depending on the result of the comparison, one must proceed to either the left or right child node of the tree. In case of (10, female), the test "age < 15" is to be performed first and then proceed to the left branch, because "age < 15" is true. Then, the second test "gender = male?" is performed, which evaluates to false, so we proceed to the right branch. We end up at the Leaf 2, whose output (leaf weight) is 0.1.

Click here (https://discuss.xgboost.ai/t/what-does-leaf-weight-mean/1587/2) to know more in detail.

5| What are the data pre-processing steps for XGB?

Solution: The data pre-processing steps for XGB include the following-

- Load the data
- Explore the data and remove the unneeded attributes
- Transform textual values to numeric
- Find and replace the missing values if needed
- Encoding the categorical data

- Break the dataset into training set as well as test set
- Perform feature scaling or data normalisation

Know more here (https://machinelearningmastery.com/data-preparation-gradient-boosting-xgboost-python/).

6| How does XGB calculate features?

Solution: XGB automatically provides the estimations of feature importance from a trained predictive model. After a boosting tree is constructed, it retrieves feature importance scores for each attribute. The feature importance contributes a score which indicates how much valuable each feature was in the construction of the boosted decision trees within the model.

Also, in terms of accuracy, XGB models show better performance (https://www.researchgate.net/publication/323570401_Comparison_of_Support_Vector_Machine_and_Extreme_Gradient_Boosting_for_predicting_daily_global_solar_radiation_using_temperature_and_precipitation_in_humid_subtropical_climates_A_case_study_in_China) for the training phase and comparable performance for the testing phase when compared to SVM models. Besides accuracy, XGB has higher computation speed than SVM.

Know more here (https://machinelearningmastery.com/feature-importance-and-feature-selection-with-xgboost-in-python/).

7| Why does XGBoost perform better than SVM?

Solution: In case of missing values, XGB is internally designed to handle missing values. The missing values are interpreted in such a way that if there endures any trend in the missing values, it is captured by the model. Users are required to supply a different value than other observations and pass that as a parameter.

XGBoost tries different things as it encounters a missing value on each node and learns which path to take for missing values in future. On the other hand, Support Vector Machine (SVM) does not perform well with the missing data and it is always a better option to impute the missing values before running SVM.

Know more here (https://medium.com/@taniyaghosh29/machine-learning-algorithms-what-are-the-differences-9b71df4f248f#:~:text=In%20fact%2C%20XGBoost%20is%20also,depending%20on%20the%20kernel%20used.).

8| Differences between XGBoost and LightGBM.

Solution: XGBoost and LightGBM are the packages belonging to the family of gradient boosting decision trees (GBDTs).

- Traditionally, XGBoost is slower than lightGBM but it achieves faster training through the Histogram binning process.
- LightGBM is a newer tool as compared to XGBoost. Hence, it has fewer users and thus a narrow user base than XGBoost and contains less documentation.

Know more here (https://analyticsindiamag.com/comparing-the-gradient-boosting-decision-tree-packages-xgboost-vs-lightgbm/).

9| How does XGB handle missing values?

Solution: XGBoost supports missing values by default. In tree algorithms, branch directions for missing values are learned during training. It is important to note that the gblinear booster treats missing values as zeros. During the training time XGB decides whether the missing values should fall into the right node or left node. This decision is taken to minimise the loss. If there are no missing values during the training time, the tree made a default decision to send any new missings to the right node.

Know more here (https://stats.stackexchange.com/questions/235489/xgboost-can-handle-missing-data-in-the-forecasting-phase#:~:text=1%20Answer&text=xgboost%20decides%20at%20training%20time,the%20right%20or%20left%20node.&text=If%20there%20are%20no%20missing,essentially%20fit%20by%20the%20model.).

10| What is the difference between AdaBoost and XGBoost?

Solution: XGBoost is flexible compared to AdaBoost as XGB is a generic algorithm to find approximate solutions to the additive modeling problem, while AdaBoost can be seen as a special case with a particular loss function.

- Unlike XGB, AdaBoost can be implemented without the reference to gradients by reweighting the training samples based on classifications from previous learners
- now more here (https://datascience.stackexchange.com/questions/39193/adaboost-vs-gradient-boosting#:~:text=The%20main%20differences%20therefore%20are,with%20a%20particular%20loss%20function.&text=In%20Adaboost%2C%20'shortcomings'%20are,by%20high%2Dweight%20data%20points.).

More Great AIM Stories

All Major NVIDIA Announcements At Computex 2021 (https://analyticsindiamag.com/all-major-nvidia-announcements-at-computex-2021/)

DataOps Goes Mainstream As Atlan Lands Big (https://analyticsindiamag.com/dataops-goes-mainstream-as-atlan-lands-big/)

Stanford University Professor Maneesh Agrawala On Video Editing Tools, Deep Fakes & More (https://analyticsindiamag.com/stanford-university-professor-maneesh-agrawala-on-video-editing-tools-deep-fakes-more/)

Data Mesh: Moving Away From Monolithic & Centralised Data Lakes (https://analyticsindiamag.com/data-mesh-moving-away-from-monolithic-centralised-data-lakes/)

A Beginner's Guide To Intel oneAPI (https://analyticsindiamag.com/a-beginners-guide-to-intel-oneapi/)

An AI Tool To Assess Severity Of COVID-19 Cases (https://analyticsindiamag.com/an-

(https://analyticsindiamag.com/author/ambika-choudhury/)

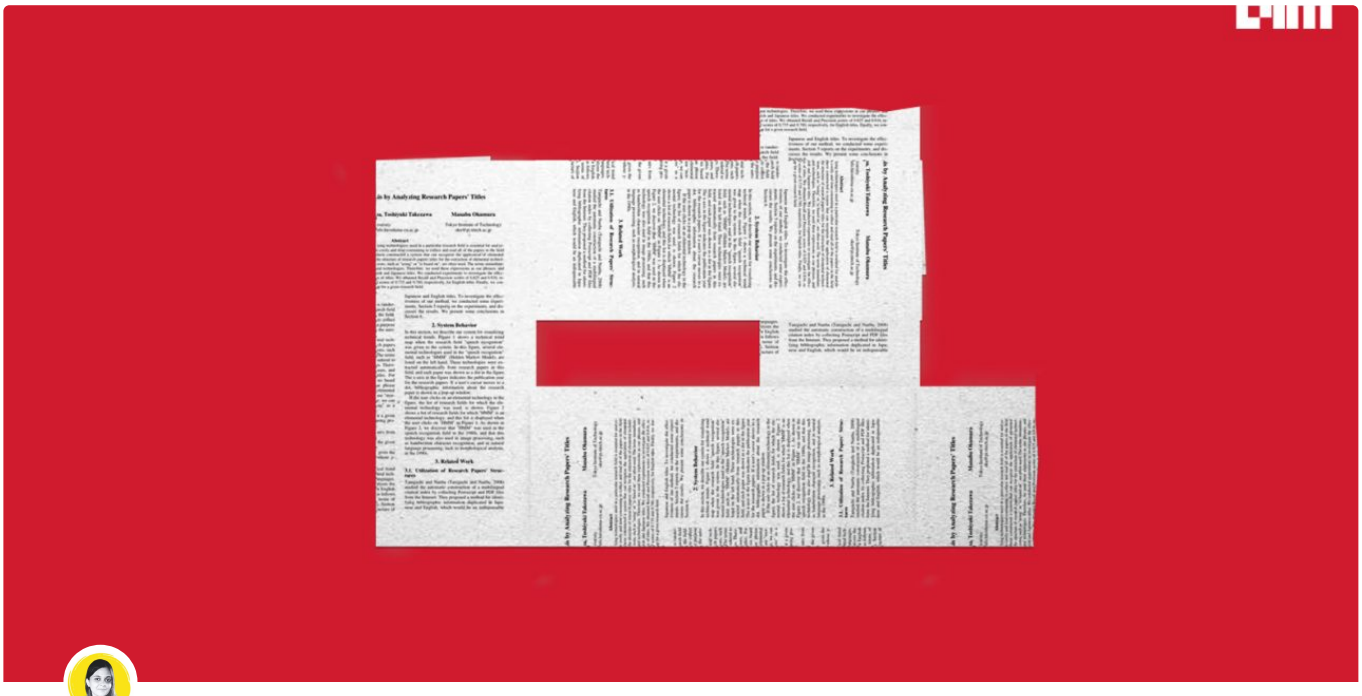
A Technical Journalist who loves writing about Machine Learning and Artificial Intelligence. A lover of music, writing and learning something out of the box.

Our Upcoming Events

MORE FROM AIM

Alternative data analytics startup Synaptic raises USD 20 Mn in Series B (https://analyticsindiamag.com/alternative-data-analytics-startup-synaptic-raises-usd-20-mn-in-series-b/)

The funding is a significant step forward to harness the full potential of merging ML and analytics with alternative data to improve investing decisions.

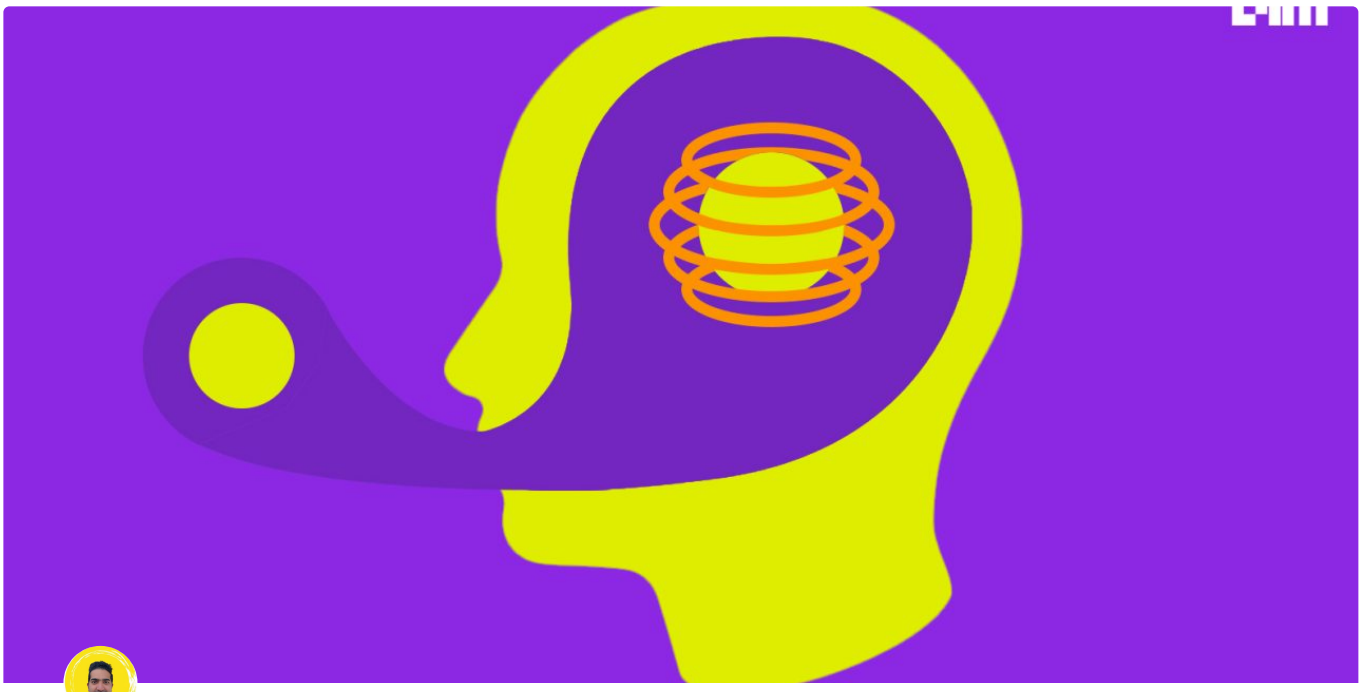


Inside ACL 2022 Test of Time Papers Awards (https://analyticsindiamag.com/inside-acl-2022-test-of-time-papers-awards/)

The ACL Test-of-Time Paper Award recognises up to four papers for their long-lasting impact in the field of Natural Language Processing and Computational Linguistics.

Marathon: An unusual marketing strategy for TCS (https://analyticsindiamag.com/marathon-an-unusual-marketing-strategy-for-tcs/)

TCS' association with marathons began in 2008 when they became the junior sponsors for the Mumbai Marathon.



NLP gets a quantum boost (https://analyticsindiamag.com/nlp-gets-a-quantum-boost/)

QSANN is effective and scalable on larger data sets and can be deployed on near-term quantum

devices.

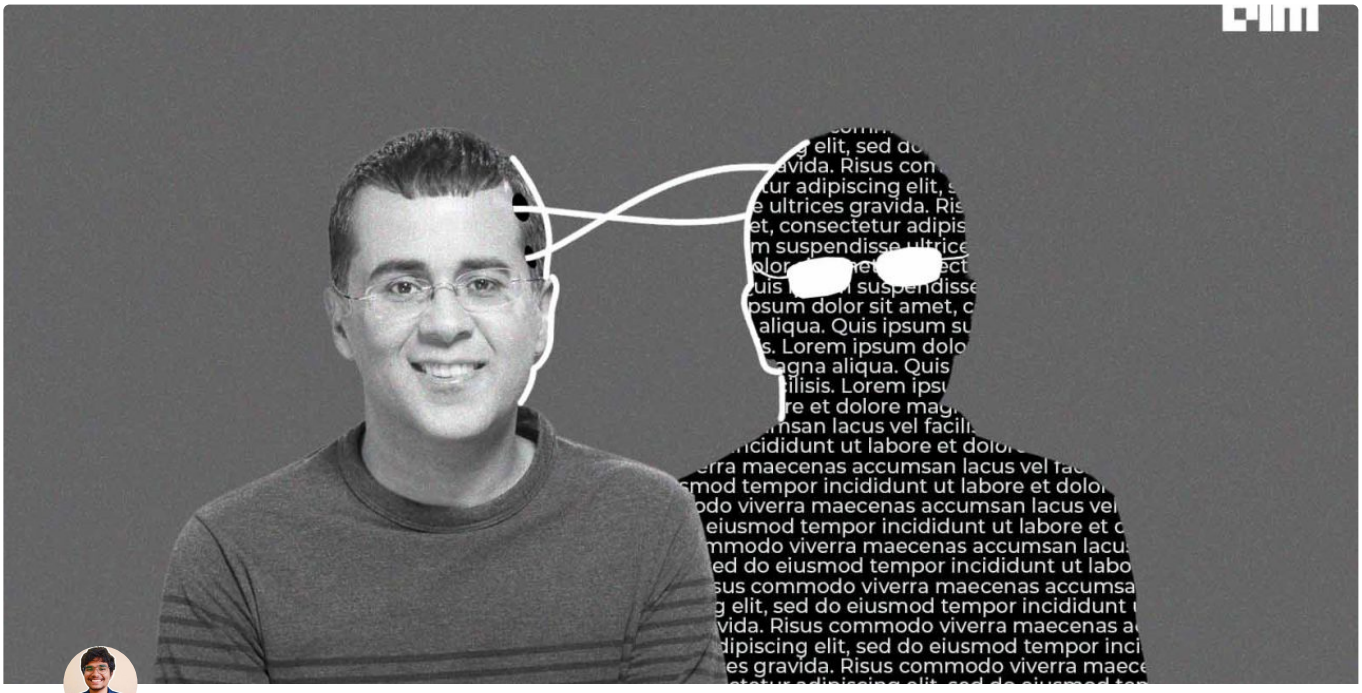
IIT Roorkee partners with Deloitte to bridge AI skill gap in India (https://analyticsindiamag.com/iit-roorkee-partners-with-deloitte-to-bridge-ai-skill-gap-in-india/)

This partnership has the potential to strengthen the AI roadmap of India.

Key announcements from NVIDIA at ISC 2022 (https://analyticsindiamag.com/key-announcements-from-nvidia-

at-isc-2022/)

Venado will be the first system in the US to feature a mix of Grace CPU Superchip nodes and Grace Hopper Superchip nodes.

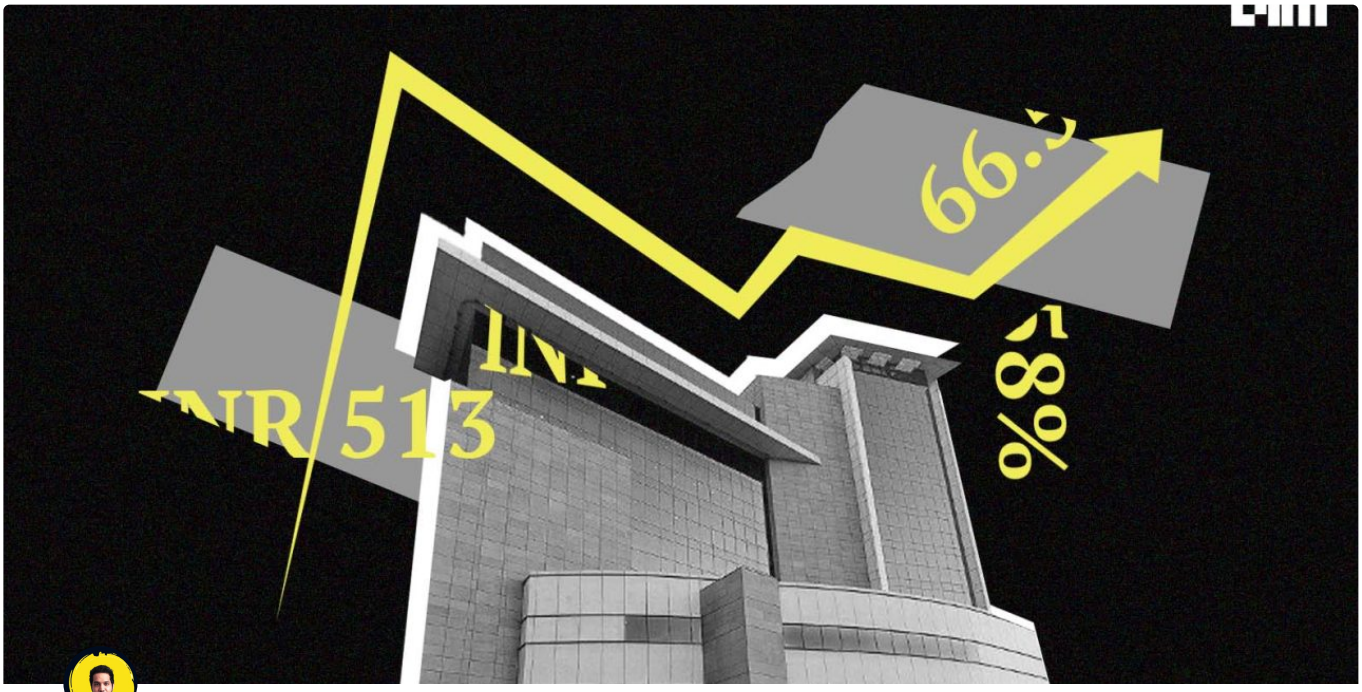


How to build an AI model that writes like Chetan Bhagat (https://analyticsindiamag.com/how-to-build-an-ai-model-that-writes-like-chetan-bhagat/)

I used TextBlob for sentiment analysis.

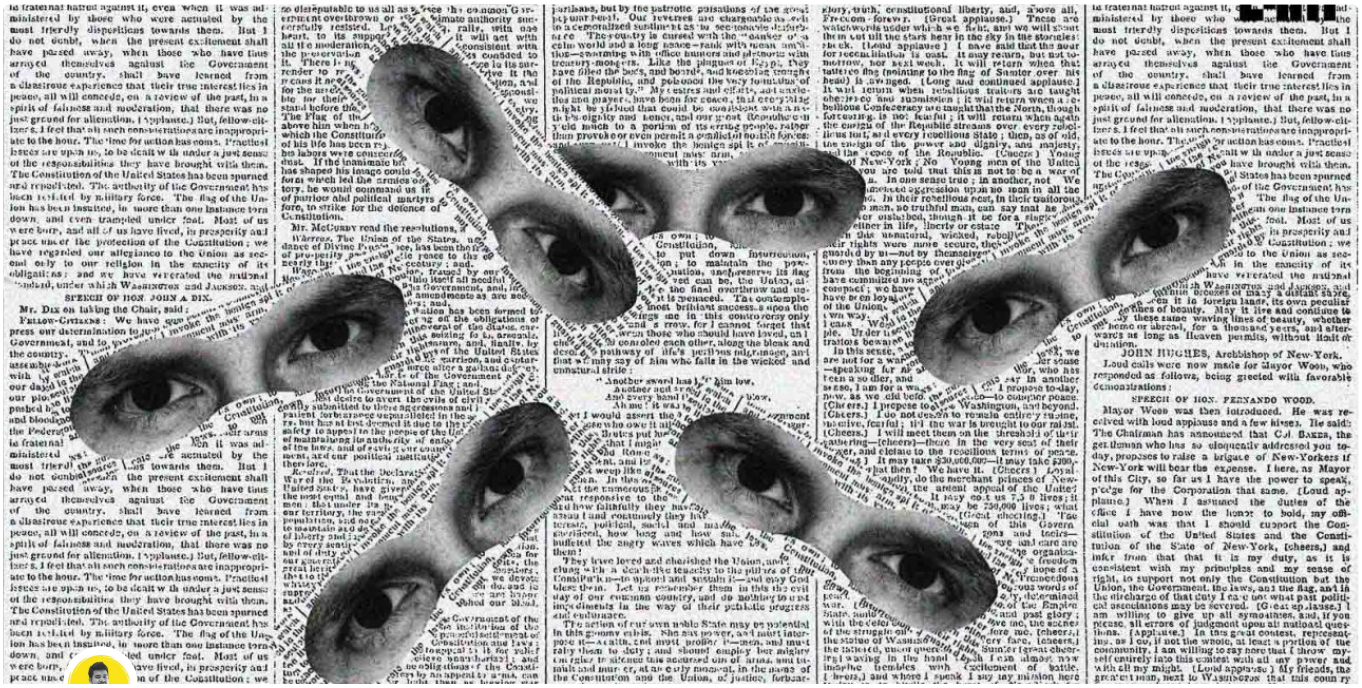
Run your first data analysis program in the browser with PyScript (https://analyticsindiamag.com/run-your-first-data-analysis-program-in-the-browser-with-pyscript/)

PyScript allows users to build Python applications on the web using an HTML interface.



Why Info Edge succeeds (https://analyticsindiamag.com/why-info-edge-succeeds/)

The numbers indicate that the IT hiring is going through the roof and is providing additional revenues for the platform.



‘I don’t really trust papers out of top AI labs anymore’ (https://analyticsindiamag.com/i-dont-really-trust-papers-out-of-top-ai-labs-anymore/)

Findings that can’t be replicated are intrinsically less reliable.

(https://analyticsindiamag.com)

(https://www.linkedin.com/company/analytics-india-magazine/)

About Us

Advertise

Weekly Newsletter

Write for us

Careers

Contact Us

RANKINGS & LISTS

Academic Rankings

Best Firms To Work For

PeMa Quadrant

OUR CONFERENCES

Cypher

The MachineCon

Machine Learning Developers Summit

The Rising

Data Engineering Summit

OUR BRANDS

MachineHack

AIM Recruits

AIM Leaders Council

Best Firm Certification

AIM Research

VIDEOS

Documentary – The Transition Cost

Web Series – The Dating Scientists

Podcasts – Simulated Reality

Analytics India Guru

The Pretentious Geek

Deeper Insights with Leaders

Curiosum – AI Storytelling

AWARDS

AI50

40 under 40 Data Scientists

Women in AI Leadership

EVENTS

AIM Custom Events

AIM Virtual

FOR ML DEVELOPERS

Hackathons

Discussion Forum

Job Portal

Mock Assessments

Practice ML

Free AI Courses

NEWSLETTER

Stay up to date with our latest news, receive exclusive deals, and more.

Enter Your Email Address

SUBSCRIBE ⟶