Sharing Experiences, Knowledge/Insight, and Informative Links

161 223

All Sections

This is where you post your findings and experiences on analytics related topics. For instance, if you have worked on an analytics project at your job, share your story and your insights here. If you come across a nice article or a web-post, tell us about it and post the link for further reading.

Best - D. Delen

This topic was locked May 1 at 11:59pm.

Search entries or author | Unread | Subscribed

1 | 2

Rudrakumar Ankaiyan (https://canvas.okstate.edu/courses/118118/users/190716)

Apr 9, 2022

Hello Everyone,

I have come across a very good article about how to overcome data shortage for ML Model Training with synthetic data.

As we know, the lack of data that reflects the full depth, granularity, and variety of real-life conditions is often the reason why a machine-learning model performs poorly. An enormous number of data sets are required to run an unbiased ML model that creates meaningful insights for all types of scenarios. Different model types have varying data requirements, but finding data is always a challenge.

Please go through the article where it discusses the data shortage, data requirements of different algorithms, causes of data shortage, and how creating synthetic data will be a realistic option.

https://tdwi.org/articles/2021/06/14/adv-all-overcome-data-shortages-for-ml-model-training-with-synthetic-data.aspx (https://tdwi.org/articles/2021/06/14/adv-all-overcome-data-shortages-for-ml-model-training-with-synthetic-data.aspx)

Rudrakumar Ankaiyan.

Edited by Rudrakumar Ankaiyan (https://canvas.okstate.edu/courses/118118/users/190716) on Apr 9 at 4:34pm

(http

Rithik Ponugoti (https://canvas.okstate.edu/courses/118118/users/189064)

Apr 14, 2022

Hi Rudra,

Thank you for sharing this article. The problem of data shortage is real due to the mentioned challenges like data privacy laws, finding enough users, customers, or employees who would agree to let their data be handed over for research purposes, small sample size due to the nature of the data itself, etc. The given article clearly explained the possible alternatives for the problems due to the data shortage. A notable recent advancement in creating synthetic data is Wasserstein GAN or WGAN. This critic neural network tries to find the minimal distance between the distribution observed in produced samples and the distribution of the data observed in the used training set. Then, WGAN trains the generator model to generate more realistic data.

Thanks to you I got to know the key difference between GAN and WGAN. Unlike GANs, WGAN does not pursue stability by looking for an equilibrium between two contrasting models. Instead, the WGAN seeks a junction between the models. As a result, it generates synthetic data with features more similar to real life.

I hope you share more informative articles like this in the future.

Thanks!

-Rithik.

(http

Moises Marin Martinez (https://canvas.okstate.edu/courses/118118/users/198182)

Apr 26, 2022

Hi Rudrakumar,

thanks for sharing this article from the datawarehouse institute.

I did a report to choose the best classification algorithm based on a dataset about obesity last year. The data set was gathered using surveys at universities. Surprisingly the data set had 23% of data gathered by surveys and 77% was generated, that's right, 77% was synthetic data. I think generating synthetic data is a great technique to use to test algorithms and visualizations, however it's not the data we would like to use primarily for a production level project.

This is a link to an article using the data set I mentioned.
https://www.sciencedirect.com/science/article/pii/S2352914820306225

Anudeep Nare (https://canvas.okstate.edu/courses/118118/users/188001)

May 1, 2022

Hello Rudra kumar.

I learned the essential distinction between GAN and WGAN thanks to you. Unlike GANs, WGAN does not seek stability by attempting to find a compromise between two opposing models. Instead, the WGAN looks for a point of convergence between the models. As a consequence, it creates synthetic data with more realistic characteristics.

Thank you

Rithik Ponugoti (https://canvas.okstate.edu/courses/118118/users/189064)

Apr 10, 2022

Hello class,

It is an absolutely vital step in any data science project to do Problem Framing. It is enacted in the way that you compile your data for analysis, answering questions related to potential features, defining the target, the level of granularity of the analysis, and the techniques to be used.

Most data science university programs, bootcamps, or online courses do not explicitly discuss problem framing, although it is an integral part of being an effective data scientist. These academic programs and what's expected of data scientists in the industry are very different. The disconnect between studying algorithms and learning about their applications is also a part of that.

Here is an excellent article by Slalom, a leading consulting company about the essence of Problem Framing. They have clearly articulated the tips for an effective framing of a business problem.

Link:

https://www.slalom.com/insight/problem-framing-data-scientists

Best,

Rithik Sai Ponugoti.

(http

Bhanu Teja Pulipalupula (https://canvas.okstate.edu/courses/118118/users/159029)

Apr 10, 2022

Hi Rithik,

Thanks for sharing the article by Slalom, completely agree with this. I've known Slalom as a top tech consulting firm for software development, CRM solutions but didn't know they're into Data Science too. I've checked the article you shared and here're a few things I found insightful, getting clarifications from the business team is the key to understanding the requirements and it is important to realize that there is no dumb question. Choosing a simple solution/algorithm over a complex solution might be the best option for many business problems.

Edited by **Bhanu Teja Pulipalupula (https://canvas.okstate.edu/courses/118118/users/159029)** on Apr 10 at 11:09pm

(http

Moises Marin Martinez (https://canvas.okstate.edu/courses/118118/users/198182)

Apr 26, 2022

Hi Rithik,

thanks for sharing this interesting article from slalom consulting company.

Having worked for few years in information technology and data warehousing for retail companies I find the Problem Framing idea very profound.

"A problem is not a problem unless you think it's a problem", I think starting with a clear understanding of a problem is key to solving it, sometimes I"ve been part of teams trying to find root cause of production issues, only to later found it wasn't really a problem. As the article and you summarize, starting with a well defined idea of what it is that we are trying to predict, model or solve will take us to the correct path; defining the level of granularity for the analysis will let us know right from the start how deep into the details we want to look at, too much detail may be overkill, too shallow and we may not obtain the results in the terms we need.

I don't go into a discussion about the rest of the points in the article, I think my point is clear. It is worth spending time at the beginning of a problem to define it and understand, from the point of view of the parties involved, what a solution will look like, in what terms and units, to not waste resources trying to predict, model or solve a different problem.

(http

Navya Mynedi (https://canvas.okstate.edu/courses/118118/users/157687)

May 1, 2022

Hi Rithik,

Thank you for sharing the article by slalom. Yes I agree that problem framing is a very important step they explained clearly about Setting out company challenge in a form that can be handled with data is known as problem framing. It's the process of turning an abstract aim, such as "we want to identify which customers are likely to churn," into what data will be utilized, how the data will appear, and what modeling tools will be employed. It's a mix of understanding the problem's underlying mechanics and matching your challenge to appropriate methodologies and approaches.

Anudeep Nare (https://canvas.okstate.edu/courses/118118/users/188001)

May 1, 2022

Hello, Rithik.

Thank you for sharing Slalom's essay; I absolutely agree. Slalom has long been renowned as a top tech consulting business for software development and CRM solutions, but I had no idea they were also into Data Science. I read the post you shared and found a few things helpful: seeking explanations from the business team is critical to understanding the requirements, and it's crucial to remember that there are no stupid questions. For many business situations, choosing a basic solution/algorithm over a complicated solution may be the best option.

Thank you

Sumanjali Etlam (https://canvas.okstate.edu/courses/118118/users/165323)

May 1, 2022

Hey Rithik!

I completely agree with Slalom's writing and thank you for sharing it with us. I had no idea Slalom was also into Data Science. Slalom has long been known as a top tech consulting firm for software development and CRM solutions, but I had no idea they were into it. I read your essay and found a few things to be useful: It's necessary to seek answers from the business team in order to fully comprehend the requirements, and it's also important to remember that there are no foolish questions. Choosing a simple solution/algorithm over a complicated solution may be the best option in many business scenarios.

Rudrakumar Ankaiyan (https://canvas.okstate.edu/courses/118118/users/190716)

Apr 10, 2022

Hello Everyone,

Please find the attached journal **Machine Learning Stock Market Prediction Studies: Review and Research Directions** which is about stock market predictions. It reviews the current literature and is trying to provide directions for future machine learning stock market prediction research. They have reviewed the current literature on stock market prediction by grouping the articles into the below four categories,

(1) Artificial Neural Network studies,

(2) Support Vector Machine studies,

(3) Studies using Genetic Algorithms with other techniques, and

(4) Studies using Hybrid or other Artificial Intelligence approaches.

Thank you,

Rudrakumar Ankaiyan.

Edited by Rudrakumar Ankaiyan (https://canvas.okstate.edu/courses/118118/users/190716) on Apr 10 at 2:36pm

Machine Learning Stock Market Prediction Studies.pdf (https://canvas.okstate.edu/files/14264198/download?download_frd=1&verifier=vRhT10D2tR2BrkuhTG9eO01WaChFG7iXsAvbO2Zf)

(http

Thirumala Krishna Kurakula (https://canvas.okstate.edu/courses/118118/users/161132)

Apr 14, 2022

Hi Rudra,

This journal was both fascinating and beneficial to me. It was clear that the journal's goal was to classify the studies with similar methods and situations. The team concluded that the Artificial Neural Network studies predict numerical stock market index values. They also stated that the Support Vector Machine studies are good at predicting categorization difficulties, Genetic Algorithms discover high-quality system inputs, and Hybrid ML techniques alleviate some of the drawbacks.

It is mentioned in the paper that the results were based on the Asian markets only and can be tested on different markets to get validation on this research. This research would have been an inviolable study if the models had distinguished between large and small firms.

Overall, this helped me in a lot of ways. If you find time, feel free to go through this website.

https://www.simplilearn.com/tutorials/machine-learning-tutorial/stock-price-prediction-using-machine-learning.

(http

Moises Marin Martinez (https://canvas.okstate.edu/courses/118118/users/198182)

Apr 26, 2022

Hi Rudrakumar,

thanks for sharing this pdf article about stock market prediction studies. This is a great example of the vast information available in the ML, AI and analytics domain. I read the first few pages and was quickly surprised about the scope of the article, they studied the last **20 years** of articles "where some form of machine learning was used to predict a stock market related outcome." Even though they only found 41 relevant articles, looking for information in the last 20 years is a titanic endeavor. It's interesting how, on the other side, the number of groups of algorithms used in the articles are just a handful. Will studying the next 20 years of articles result in more or less articles? Only time will tell.

(https:/

Bhanu Teja Pulipalupula (https://canvas.okstate.edu/courses/118118/users/159029)

Apr 10, 2022

Hello everyone!

Most job postings today in the Data Market (Data Analyst/Engineer/Scientist/ BI Engineer) have a requirement stating working knowledge of at least one of the major cloud services, i.e., AWS, Azure, or GCP. I feel overwhelmed to be aware of the different services offered by each of these cloud providers. I was going through the features of GCP (Google Cloud Platform) and found a cheat sheet that shows different services/features of GCP grouped by areas like Data Analytics, ML models, Compute, Storage, Database, etc.

Here's the link to the google cloud cheat sheet that you might find useful:
https://googlecloudcheatsheet.withgoogle.com/
(https://googlecloudcheatsheet.withgoogle.com/)

Bhanu Teja P

(http

Moises Marin Martinez (https://canvas.okstate.edu/courses/118118/users/198182)

Apr 26, 2022

Hi Bhanu,

thanks for sharing the cheat sheet for google cloud. It's helpful to have such cheat sheets because they can help to practice knowledge in more than one public cloud platform, due to the great similarities in services and functionalities among them.

I gave a presentation on the different cloud platforms last year, it's an interesting point of view about finding what cloud we need.

AWS, GCP, AZURE, IBM... What cloud do I need? | Cloud & DevSecOps Day
https://www.youtube.com/watch?v=Oiyj230rwpU (https://www.youtube.com/watch?v=Oiyj230rwpU)

(https://www.youtube.com/watch?v=Oiyj230rwpU)

(http **Jeya Subburaj (https://canvas.okstate.edu/courses/118118/users/167277)**
Apr 29, 2022

Hi *Bhanu Teja P,*

Thanks for Sharing the cheat sheet. I am on the learning journey of the cloud and this is really helpful and beneficial. There are plenty of cloud services to explore and learn.

Thanks

Jeya Subburaj

(http **Anudeep Nare (https://canvas.okstate.edu/courses/118118/users/188001)**
May 1, 2022

Hello, Bhanu.

Thank you for providing the Google Cloud cheat sheet. Cheat sheets like this are useful since they can be used to practice information across several public cloud platforms due to the many similarities in services and features. I'm learning about the cloud, and it's proving to be quite useful and valuable. There are several cloud services to investigate and learn about.

Thank you

(http

Sumanjali Etlam (https://canvas.okstate.edu/courses/118118/users/165323)

May 1, 2022

Hello there Bhanu.

Thank you for making the Google Cloud cheat sheet available to us. Because of the many similarities in services and capabilities, cheat sheets like this can be used to practice material across various public cloud platforms. I'm learning about the cloud, and it's proving to be really handy. There are a number of cloud services to look into and learn more about.

Thank you very much.

(https:/

Jacob Wood (https://canvas.okstate.edu/courses/118118/users/214790)

Apr 11, 2022

Hello everyone,

As this course is my first experience with machine learning, I have learned a lot and am excited to continue to learn/practice. When our coursework switched from descriptive to predictive/prescriptive analytics , I found the following article very helpful to supplement our coursework and textbook readings. Studying this article in conjunction with our text book has helped me conceptualize the various ML algorithms we've employed and will employ further on our project/careers. Hopefully some of you all can benefit from this article as well.

https://towardsdatascience.com/11-most-common-machine-learning-algorithms-explained-in-a-nutshell-cc6e98df93be (https://towardsdatascience.com/11-most-common-machine-learning-algorithms-explained-in-a-nutshell-cc6e98df93be)

Regards,

Jacob Wood

(http

Sumanjali Etlam (https://canvas.okstate.edu/courses/118118/users/165323)

Apr 16, 2022

Hi Jacob,

Thanks for sharing the article. I found this article to be both intriguing and useful. Yes I agree continuous learning and practice makes us much more confident to grow in data science field. To practice I would suggest Codecademy which is a collaborative learning platform for

programming languages that has interactive platforms. In this article I came to know about DBSCAN Clustering which is my first time of knowing it in detail. I also like to share an article which tells us about machine learning from hype to real world applications. In this article I came to know about AI revolution, Deep learning, creativity and real-world applications of machine learning.

https://towardsdatascience.com/machine-learning-from-hype-to-real-world-applications-69de7afb56b6 (https://towardsdatascience.com/machine-learning-from-hype-to-real-world-applications-69de7afb56b6)

(http **Moises Marin Martinez (https://canvas.okstate.edu/courses/118118/users/198182)**

Apr 26, 2022

Hi Jacob!

wow, thanks so much for sharing that article, I find it very interesting and will be keeping it in my favorites for a quick reference. I'm doing a dual degree master, with some classes in a university in Mexico and other classes at OSU online. I've spend about two years so far in this program and, well, it's a lot of information. This summary is a good way to easily go thru the algorithms :)

(http **Srikanth Daruru (https://canvas.okstate.edu/courses/118118/users/193729)**

May 1, 2022

Hi jacob,

Thanks for sharing this link,

I found the following article to be really useful. It pleases my attention, and I have to definitely save this for my reference in further semester. This material helped me better understand the machine learning techniques we've used and will use in the future projects and careers.

(http **Ashish Kumar Pampana (https://canvas.okstate.edu/courses/118118/users/24369)**

May 1, 2022

Hi Jacob,

Thank you for this link.

I've been always the fan of towards data science articles. It was very helpful for me. I've been meaning to revise all my machine learning concepts. I'm working hard to be the

analytics specialist.

(http

Anudeep Nare (https://canvas.okstate.edu/courses/118118/users/188001)

May 1, 2022

Hi Jacob,

Thank you for this link.

You gave very good insights from your experience and i am a student from similar background . Your insights will be very helpful for us thank you

(https:/

Ben Lewis (https://canvas.okstate.edu/courses/118118/users/211833)

Apr 12, 2022

Hi all,

This is my first semester in the program. In addition to KNIME, I've been trying to learn Python this semester. There are a lot of great resources out there for the basics, but I've really enjoyed the following Youtube channels to understand some of the models better.

1. Codebasics - https://www.youtube.com/watch?v=gmvvaobm7eQ&list=PLeo1K3hjS3uvCeTYTeyfe0-rN5r8zn9rw (https://www.youtube.com/watch?v=gmvvaobm7eQ&list=PLeo1K3hjS3uvCeTYTeyfe0-rN5r8zn9rw)

(https://www.youtube.com/watch?v=gmvvaobm7eQ&list=PLeo1K3hjS3uvCeTYTeyfe0-rN5r8zn9rw) - This playlist has 10-20 minute videos about different predictive analytic models and a really simple demonstration. Then each video ends with an exercise you can practice on your own.

2. Statquest - https://www.youtube.com/c/joshstarmer (https://www.youtube.com/c/joshstarmer) - Statquest doesn't go into the code, but focuses on the math and statistics behind a lot of the popular models.

My method this semester has been to read about it in class, then watch these videos after getting it to work in KNIME. For more academic reading and an actual article , I really enjoy the site https://paperswithcode.com (https://paperswithcode.com) . This site is frequently

updated with machine learning models and articles about them, but also links to the GitHub! Here is a favorite of mine, where NLP and sentiment analysis tools were used to turn non-polite sentences into polite ones.

https://paperswithcode.com/paper/politeness-transfer-a-tag-and-generate (https://paperswithcode.com/paper/politeness-transfer-a-tag-and-generate)

(http

Jacob Wood (https://canvas.okstate.edu/courses/118118/users/214790)

Apr 13, 2022

Hi Ben,

Thanks for sharing these resources! I am in my first semester as well, and have been working on learning Python outside of class to help in future semesters. Within Python, I have not yet worked on coding predictive analytic models, but I've enjoyed a few of these videos which help reenforce lessons from this course and apply them within Python. During my free time I plan to go through this video series. I found another fun exercise which takes you through step by step in utilizing Python for predictive models (using CRISP-DM!).

https://towardsdatascience.com/end-to-end-python-framework-for-predictive-modeling-b8052bb96a78 (https://towardsdatascience.com/end-to-end-python-framework-for-predictive-modeling-b8052bb96a78)

Good luck with Python, I look forward to learning more Python in future courses.

(http

Thirumala Krishna Kurakula (https://canvas.okstate.edu/courses/118118/users/161132)

Apr 20, 2022

Hi Ben,

The YouTube channels and the coding website are helpful to me. My project in the last semester involved topic modeling, sentiment analysis, and named entity recognition. I can use the website that you mentioned "**https://paperswithcode.com/paper/politeness-transfer-a-tag-and-generat (https://paperswithcode.com/paper/politeness-transfer-a-tag-and-generate)** e" as a reference to extend my project.

(http

Ben Lewis (https://canvas.okstate.edu/courses/118118/users/211833)

Apr 25, 2022

Thank you Thirumala and Jacob for the responses! Very cool to hear how the website can be uses and to see the CRISP-DM model used elsewhere.

Moises Marin Martinez (https://canvas.okstate.edu/courses/118118/users/198182)

Apr 26, 2022

Hi Ben,

thanks so much for sharing these links, I hadn't come across them and they look like a perfect way to practice and learn the code.

I'd like to share with you a youtube channel with some tutorials for Knime that I have found useful.

https://www.youtube.com/channel/UCiXKwSPJLqjtfluopN1-4jg (https://www.youtube.com/channel/UCiXKwSPJLqjtfluopN1-4jg)

Anurag Budme (https://canvas.okstate.edu/courses/118118/users/155819)

Apr 12, 2022

Hello Everyone,

In recent years, the use of AI and machine learning in sports has grown more frequent. It has influenced audience involvement and has resulted in the development of a game strategy.

Artificial intelligence (AI) has a significant influence on how sports fans view them. AI systems may be used to automatically select the best camera viewpoint to display on the viewer's screen, provide real-time subtitles in several languages based on the viewer's location, and allow broadcasters to monetize their content.

Major League Baseball is a professional baseball company that is working closely with Google Cloud to find additional use cases for its huge volumes of data, considering the benefits of AI and machine learning deployment in sports.

Here’s the article on the new collaboration between Google Cloud and MLB for game-changing sports analytics

https://cloud.google.com/blog/products/data-analytics/mlb-pitches-new-data-uses-with-google-cloud-services (https://cloud.google.com/blog/products/data-analytics/mlb-pitches-new-data-uses-with-google-cloud-services)

(http) **Jacob Wood** (https://canvas.okstate.edu/courses/118118/users/214790)

Apr 13, 2022

Hi Anurag,

I'm a baseball lover and have enjoyed the collaboration between Google and MLB. The sport has always been rich in statistics and it's impressive to get a glimpse into the IT architecture utilized as well as the endless data points collected. In that video, it's too bad that the Director of Baseball systems for the Arizona Diamondbacks did not share how the team is using this data to inform in-game and front office strategies. It seems like there would be endless possibilities in this space (pitch selection, defensive shifts, best pinch hitter, etc.). Thank you for the share!

(http) **Nithya Satheneni** (https://canvas.okstate.edu/courses/118118/users/187996)

Apr 22, 2022

Hi Anurag,

I agree with you that AI and machine learning has grown in the sports in many ways. Many referees are using this technologies to make judgements on players, team statistics etc. Video Assistant Technology (VAR) is one type of technology used for judging decisions such as red cards, free kicks etc. Please check this article that has listed the 7 most used AI applications in sports.

https://www.v7labs.com/blog/ai-in-sports (https://www.v7labs.com/blog/ai-in-sports)

Thanks,

Nithya

(http) **Moises Marin Martinez** (https://canvas.okstate.edu/courses/118118/users/198182)

Apr 28, 2022

Hi Anurag,

thanks for sharing the link it is very interesting. My favorite reference for baseball and machine learning is **https://www.codebaseball.com/,** (https://www.codebaseball.com/,) if you like baseball and coding you'll find it interesting!

It is indeed very exciting how AI can be applied to complement the world of sports, specially during transmissions of sports, sports have a plethora of data.

(http

Jeya Subburaj (https://canvas.okstate.edu/courses/118118/users/167277)

Apr 29, 2022

Hey Anurag,

This is an interesting article about AI & ML in the sports world. For the last two decades, coaches have been using data science in sports to help improve the performance of their players. Can't wait to see lot more innovations in this sports world using AI before the unfolding of the AI regulation and compliance use of AI in the EU and US. Thanks for sharing the article.

Thanks,

Jeya Subburaj

(http

Anudeep Nare (https://canvas.okstate.edu/courses/118118/users/188001)

May 1, 2022

Hi Anurag ,

Thank you for this link.

I agree with you that AI and machine learning has grown in the sports in many ways. Many referees are using this technologies to make judgements on players, team statistics etc. You gave very good insights from your experience. Your insights will be very helpful for us thank you

(http

Sumanjali Etlam (https://canvas.okstate.edu/courses/118118/users/165323)

May 1, 2022

Hello, Anurag.

This is a fascinating piece regarding AI and machine learning in sports. Coaches have been employing data science in sports for the past two decades to help their athletes enhance their performance. Can't wait to see a lot more AI-powered advancements in the sports sector before the EU and US implement AI regulation and compliance. Thank you for sharing this content with us.

Jay West (https://canvas.okstate.edu/courses/118118/users/57886)

Apr 13, 2022

I found this article was interesting. In the hospitality and tourism industry, understanding customers is the key factor to success. By checking consumer reviews and opinions about properties and services on Yelp or TripAdvisor will help improve business performance and it will also help the business to better understand their demands. Therefore, it is not uncommon for the hospitality industry/hoteliers to adopt machine learning techniques to understand their customers. The early days of sentiment analysis, checking reviews and opinions were a daily task for managers, but today, automated text mining and sentiment analysis provides fast and efficient information which is a crucial tool for developing marketing strategies.

The webpage below explains how hospitality industry utilizes text mining and sentiment analysis, as well as how they can apply it to their businesses.

https://towardsdatascience.com/sentiment-analysis-for-hotel-reviews-3fa0c287d82e

Rithik Ponugoti (https://canvas.okstate.edu/courses/118118/users/189064)

Apr 22, 2022

Hello Jay,

Thank you for sharing this article. I had my hands-on Sentimental Analysis during my PDS-1 course here at OSU. We used Selenium web driver to scrape dynamic websites and it was a great learning experience. Thanks to you, I got to know more about the Octoparse in the article you shared.

Upon digging a bit, I got to know that it is a modern web data extraction software with a visual interface. Octoparse is simple to use for both expert and beginner users to bulk extract information from websites. For most scraping activities, no code is required! Octoparse makes getting data from the web easier and faster without requiring us to code.

Thanks!

Best,

Rithik Sai Ponugoti

Jay West (https://canvas.okstate.edu/courses/118118/users/57886)

Apr 23, 2022

Hello Rothik,

I am glad the article helped you discover extra knowledge! I am a new in this field, so I am not familiar with a number of terminology. But one thing that caught my attention was the statement that 'no code is required for scraping'. I might check into it in the future! Thank you for the explanations!

(http

Moises Marin Martinez (https://canvas.okstate.edu/courses/118118/users/198182)

Apr 28, 2022

Hi Jay!

thanks for sharing this interesting article, it blows my mind to think about the effort required by managers to obtain sentiment analysis before automation and programming. Can you imagine reading reviews from handwritten books ? or from log files? it must have been very time consuming. A domain that can greatly benefit from sentiment analysis is the hospitality area, it is thru feedback that they can get an idea of how well they are providing services and discover areas of opportunity. The link you share promotes octoparse, a low-code solution for web scraping, low-code solutions are very popular and allow non-programmers and non-formally trained experts to take advantage of powerful tools to obtain results. Citizen developers are bound to become more common in IT and data science because of the simplification of the tools and their cost reduction.

(https:/

Paul Dreyer (https://canvas.okstate.edu/courses/118118/users/183234)

Apr 13, 2022

I currently work in healthcare fraud detection, and the hot topics the last couple of years has been leveraging data science techniques to uncover a whole new slew of potential savings. I know this article is about 1.5 years old, but this is one I reference when presenting to clients and it's very helpful in providing a high-level overview of the techniques and processes we can leverage to discover new fraud, waste and abuse targets.

https://nycdatascience.edu/blog/student-works/healthcare-fraud-detecting-inconsistencies-in-provider-data/

(http

Moises Marin Martinez (https://canvas.okstate.edu/courses/118118/users/198182)

Apr 28, 2022

Hi Paul!

Thanks for sharing this link, I have been working in IT for 10+ years and healthcare is one area where I have not had the opportunity to work in.

It's refreshing to read about how data science is being used in areas I am not familiar with, I looked at the article, the analysis is very interesting. Identifying potential fraudulent providers will result in savings, hopefully the tools will also help deter such attempts of dishonest claims in the future.

(http

Anurag Budme (https://canvas.okstate.edu/courses/118118/users/155819)

Apr 30, 2022

Hey Paul,

Thanks for sharing the link. I have an interest in the use cases leveraging data science techniques in the field of the health industry. It would be very helpful.

The article has some interesting analyses. Fraud identification is definitely helpful and adds value to cost savings.

(http

Seonwoo Ko (https://canvas.okstate.edu/courses/118118/users/194258)

May 1, 2022

Hello Paul,

Thank you for sharing your experience and interesting website! Even if I am not involved in healthcare industry, I think it would be helpful and needed in other industry as well. I have to look up more about information, but thank you for sharing the link!

Thanks,
Seonwoo

(https:/

Rithik Ponugoti (https://canvas.okstate.edu/courses/118118/users/189064)

Apr 14, 2022

Hello class,

What if I told you that the “measurement accuracy” parameter alone is insufficient to achieve a reliable result and that other factors must be considered. Yes! the sample space under consideration is maybe more essential than the accuracy. The “false positive paradox” is a concept used in statistics and data science to describe this argument. This paradox typically occurs when the probability of an event occurring is less than the error accuracy of the instrument used to measure the event.

Here is one beautiful article that helps you to dive into this topic.

https://medium.com/analytics-vidhya/when-the-measurements-accuracy-misleads-you-a9f0f1f7bb0d (https://medium.com/analytics-vidhya/when-the-measurements-accuracy-misleads-you-a9f0f1f7bb0d)

Thanks!

Rithik Sai Ponugoti.

(http

Ben Lewis (https://canvas.okstate.edu/courses/118118/users/211833)

Apr 26, 2022

Hi Rithik,

Great share! This is all new info to me. The example in the article with police and drunk drivers really brought the concept home for me. Going to have to make sure that the 'awareness test' is part of my process moving forward!

Thanks for sharing,

Ben

(http

Moises Marin Martinez (https://canvas.okstate.edu/courses/118118/users/198182)

Apr 28, 2022

Hi Rithik,

thanks for sharing the article. The topic of the article is very interesting. I totally fell for the first example, a detector that has 90% accuracy to find gold, and it has beeped on a given

stone, wow that has to be a good deal! Unfortunately being a 1% chance it is actually gold makes it a bad purchase.

Before reading this article I wasn't aware that paying attention to the sample space under consideration was essential to overcome the instrument’s inaccuracy, but not anymore! thanks!

(http

Anudeep Nare (https://canvas.okstate.edu/courses/118118/users/188001)

Apr 30, 2022

Hello, Rithik.

Interesting response! This is all new information to me. The article's example of cops and intoxicated drivers truly drove the point home for me. I'll have to make the 'awareness test' a part of my workflow going ahead!

Thank you for sharing.

(http

Sumanjali Etlam (https://canvas.okstate.edu/courses/118118/users/165323)

May 1, 2022

Hello, Rithik.

Thank you for sharing this content with us. The article's subject is quite intriguing.

I had no idea that paying attention to the sample space under consideration was necessary to overcome the instrument's inaccuracy before I read this post, but now I know! thanks!

(https:/

Jay West (https://canvas.okstate.edu/courses/118118/users/57886)

Apr 14, 2022

This is another interesting hospitality and tourism industry related article using text mining and structural topic modeling to help the industry understand their customers closely. This article demonstrates steps from data cleaning to sentiment analysis and topic modelling. By collecting data from Tripadvisor, the prevalent topics among customers’ negative reviews are extracted through topic modelling. As stated in the article, the machine learning approaches in the

hospitality and tourism literature have not been widely used yet, but are trending upwards. I am also new to this field and I see a lot of potential in learning about machine learning techniques.

Hu, N., Zhang, T., Gao, B., & Bose, I. (2019). What do hotel customers complain about? Text analysis using structural topic model. *Tourism Management*, *72*, 417-426.

1-s2.0-S0261517719300020-main(1).pdf (https://canvas.okstate.edu/files/14324302/download?download_frd=1&verifier=RhpVh1Aks0w14Hf00wdmONINHkZRuP41MHxfouYt)

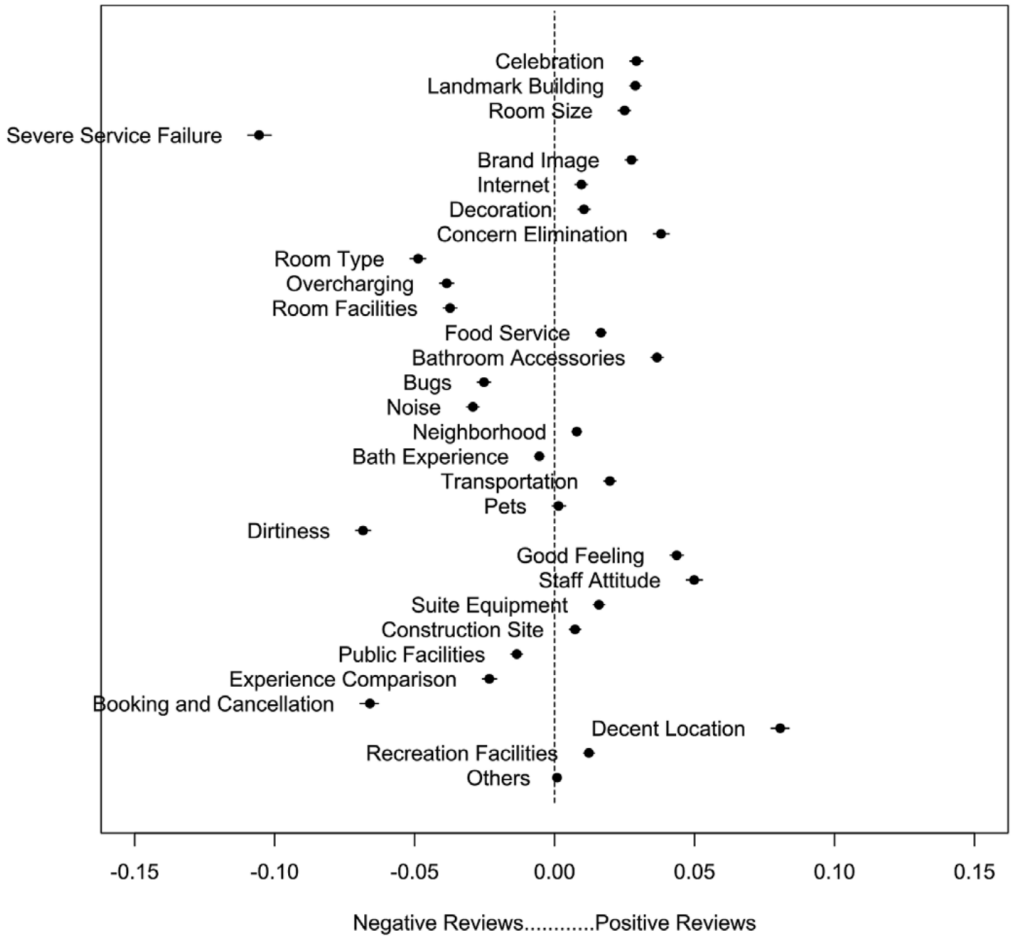
(http

Moises Marin Martinez (https://canvas.okstate.edu/courses/118118/users/198182)

Apr 28, 2022

Hi Jay,

thanks for sharing this pdf. What called my attention the most was the Figure 2, tha graphic showing the difference in the topic proportion (Negative vs. Positive). It shows in a single image topics that have been scored as positive and topics that have been scored as negative, I think it is a brilliant way to share reports of sentiments analysis with managers and other teams.



(https:/

Bhanu Teja Pulipalupula (https://canvas.okstate.edu/courses/118118/users/159029)

Apr 15, 2022

Hello everyone,

While going through chapters 2 and 3 of this course, i.e., Descriptive Analytics, I was checking for some online resources to learn Data Visualization with hands-on training. Here's a great website that I came across: **https://www.makeovermonday.co.uk/** **(https://www.makeovermonday.co.uk/)** . Makeover Monday is a weekly learning and development appointment with hundreds of passionate data people. Each week they post a link to a chart and its data, and then we can rework the chart. If you're enthusiastic about visualization, here're a few other resources you want to check out, **https://www.amazon.com/Storytelling-Data-Visualization-Business-Professionals/dp/1119002257** **(https://www.amazon.com/Storytelling-Data-Visualization-Business-Professionals/dp/1119002257)** , more than the tools used for data visualization, Storytelling with data is an important skill and this book is a go-to for that. If you're in the Tableau community, you may want to follow this person: **https://public.tableau.com/app/profile/pradeepkumar.g** **(https://public.tableau.com/app/profile/pradeepkumar.g)** and check out the dashboards on his Tableau public profile.

Thanks,

Bhanu Teja P

(http

Rithik Ponugoti (https://canvas.okstate.edu/courses/118118/users/189064)

Apr 28, 2022

Hi Bhanu,

Thank you for sharing these links! I am a big fan of Tableau and the wonders we can do with it. I do want to create a Tableau Public profile and I will definitely consider following the Pradeep Kumar. Also, good thought of sharing the Makeover Monday. They also host competitions and give the platform to shine. Hoping to enter into one soon!

Thanks,

Rithik Sai Ponugoti

(http

Rudrakumar Ankaiyan (https://canvas.okstate.edu/courses/118118/users/190716)

Apr 29, 2022

Hi Bhanu,

Thanks for sharing these useful links. Tableau is a interesting visualization tool to tell stories and make the underlying details easy for anyone to understand visually. I have taken Descriptive Analytics and Visualization course in my first semester and I learned a lot about Tableau. I have also done a Tableau project based on the Olympic stats and developed dashboards out of it.

Thanks,

Rudrakumar Ankaiyan

(http

Seonwoo Ko (https://canvas.okstate.edu/courses/118118/users/194258)

Apr 30, 2022

Hello Bhanu,

Thank you for your links!! I have used Tableau for this class and the other class from HTM. I think Tableau is really crucial and impactful for visualization tool. I am a visual person, so it is easier for me to understand visualization results rather than only reading written documents. I would use those links for my future research review! Thanks.

Best,
Seonwoo

(http

Anudeep Nare (https://canvas.okstate.edu/courses/118118/users/188001)

Apr 30, 2022

Hello Bhanu,

Tableau is the world's most popular data analytics application for visualizing and understanding data. While most businesses now recognize the value of data in driving growth and informing decision-making, it can also be quite difficult.

Transform your response to issues and easily generate actionable analysis results. With Tableau, it's time to create a data-driven culture.

Here's the link :

https://www.eidebailly.com/services/products-and-solutions/tableau?utm_source=google&utm_medium=paid_search&utm_campaign=dataanalytics&utm_term=datavisualization&utm_content=ad2&gclid=Cj0KCQjwma6TBhDIARIsAOKuANzwSIk5yM25-utsAHaKM9O7E4GMUhRILLALCrE5C-i4h6PnjkDIGPkaAnNHEALw_wcB (https://www.eidebailly.com/services/products-and-solutions/tableau?utm_source=google&utm_medium=paid_search&utm_campaign=dataanalytics&utm_term=data

visualization&utm_content=ad2&gclid=Cj0KCQjwma6TBhDIARIsAOKuANzwSIk5yM25-utsAHaKM9O7E4GMUhRILLALCrE5C-i4h6PnjkDIGPkaAnNHEALw_wcB)

Thank you ,

Anudeep

(http

Anurag Budme (https://canvas.okstate.edu/courses/118118/users/155819)

Apr 30, 2022

Hello, Bhanu.

Thank you for providing the links to Tableau visuals.

I've recently spent a significant amount of time using Tableau to create visuals. It's a lot of fun. I recently received my Tableau Desktop specialist certification.

If any of you are interested in visualizations & Tableau, I would recommend that you pursue the certification, which will add value to your professional profile.

(http

Sumanjali Etlam (https://canvas.okstate.edu/courses/118118/users/165323)

May 1, 2022

Hello there, Bhanu.

Thank you very much for the links!! Tableau is, in my opinion, a really important and significant visualization tool. Because I am a visual person, understanding visualization outcomes is easier for me than reading textual materials. Those are the links I'd use for my future research review! Thanks.

(http

Sumanjali Etlam (https://canvas.okstate.edu/courses/118118/users/165323)

May 1, 2022

Hi Anurag!

Congratulations for your Tableau Desktop specialist certification. I always had an interest to learn more about tableau. I would definitely reach you out for more tips and information to gain knowledge from you. Then I will give an exam for Tableau Desktop specialist certification after getting a good grasp.

(http

Anurag Budme (https://canvas.okstate.edu/courses/118118/users/155819)

May 1, 2022

Hi Suma,

Thank you. It's a great viz. tool and would always recommend one to learn Tableau.

I am happy to provide the tips, information, and resource links that will help you to complete Tableau Desktop Certification.

Thanks,

Anurag

Thirumala Krishna Kurakula (https://canvas.okstate.edu/courses/118118/users/161132)

Apr 16, 2022

Predictive analytics uses modeling and forecasting techniques to determine the probability of an event occurring in the future. Clinicians, scientists, health professional organizations, pharmaceutical manufacturers, and everyone involved in healthcare could use those estimates to achieve the best quality care for specific patients.

The involvement of predictive analytics in the healthcare sector has many advantages. Some of them are:

1. Clinical prognoses, outbreak prediction, hospital readmissions, and overstay estimation

2. Allocation of resources and identifying patients at high risk.

3. Preventing newborn diseases with genetic screening.

4. Engagement and conduct of patients' choice for the most effective consumer treatments.

5. Preventing breakdowns of equipment.

6. Using sensors for monitoring.

This is an interesting research paper that talks about predictive analytics in the healthcare industry

https://www.researchgate.net/publication/303480761_Predictive_Analytics_in_Healthcare_System_Using_Data_Mining_Techniques

Jay West (https://canvas.okstate.edu/courses/118118/users/57886)

Apr 23, 2022

Thank you for sharing the article! I also came across a similar online article that mentioned about predictive analytics which discussed the probability of an event occurring in the future. As you mentioned above, predictive analytics is a great tool in health industry but it also a popular strategy in casino industry as well. So, the predictive analytics are popular for the following industry. 1. Finance: Forecasting Future Cash Flow 2. Entertainment & Hospitality: Determining Staffing Needs 3. Marketing: Behavioral Targeting 4. Manufacturing: Preventing Malfunction 5. Health Care: Early Detection of Allergic Reactions.

https://online.hbs.edu/blog/post/predictive-analytics

(http

Seonwoo Ko (https://canvas.okstate.edu/courses/118118/users/194258)

Apr 30, 2022

Hello Thirumala,

Thank you for sharing your article! I agree with this article that predictive analytics will be helpful and crucial in healthcare industry. They have to track patients' results and predict the treatments. I am attaching one article related to predictive analytics in healthcare sector. I hope it helps. Thanks!

https://scholar.google.com/scholar?hl=en&as_sdt=0%2C37&q=predictive+analytics+in+healthcare&oq=predictive+analytic#d=gs_qabs&t=1651296404367&u=%23p%3DI5HvuAD9lxAJ (https://scholar.google.com/scholar?hl=en&as_sdt=0%2C37&q=predictive+analytics+in+healthcare&oq=predictive+analytic#d=gs_qabs&t=1651296404367&u=%23p%3DI5HvuAD9lxAJ)

Best,
Seonwoo

(https:/

Sumanjali Etlam (https://canvas.okstate.edu/courses/118118/users/165323)

Apr 16, 2022

Hi Everyone,

I would like to share my experience and insights about working on analytics project.

As a Data Science enthusiast, I've had the opportunity to work on a few intriguing Data Analysis projects. But the one that interested me the most was the one where I completed the entire thing on my own. I used the New York City Dataset for the research, which was about Uber Data Analysis. This project appealed to me because it elegantly demonstrates the art of storytelling,

which every Data Science enthusiast should be familiar with. In addition, data storytelling is important in a variety of Machine Learning tasks. As a result, there was another additional motive to take on this endeavor. Although Python appears to be the language of choice for the vast majority of Data Science projects, I chose R Programming for my project. This is due to the fact that R programming is ideal for data analysis tasks. This project is more of a data visualization than a data analysis project, and it is constructed using the ggplot2 library to acquire an intuition for understanding the users who book trips. Companies may benefit from visualization by better comprehending complicated data and gaining insights that will help them make better decisions. I learned how to produce data visualizations at the end of the Uber data analysis R project. Using ggplot2, which allowed me to create a variety of visuals for various time periods throughout the year. As a result, I were able to deduce how time influenced customer travels. Finally, I created a geo map of New York that showed us how different people traveled from different locations.

Looking forward to seeing your thoughts and responses on this!!

(http

Navya Mynedi (https://canvas.okstate.edu/courses/118118/users/157687)

Apr 21, 2022

Hi Sumanjali,

Thank you for sharing your project's insights. They are very interesting. I can completely relate to you. When I was working on my project about electrical cars, I became nervous after viewing the enormous dataset and was concerned about how to build models for that dataset. However, after examining the data, I, too, believe that this is more close to like story telling because I have to demonstrated which features influence car sales. When it comes to coding Initially, I felt more at ease with Python; but, for some aspects, I believe R is more comfortable than Python. I realized that having better results is more important than the language we use to get them.

(http

Anudeep Nare (https://canvas.okstate.edu/courses/118118/users/188001)

Apr 26, 2022

Hello, Sumanjali.

Thank you for taking the time to share your project's findings. They're quite intriguing. However, after reviewing the statistics, I, too, think that this is more akin to story telling because I must explain which characteristics impact automobile sales.

Thank you for a valuable information sumanjali.

(https:/

Sumanjali Etlam (https://canvas.okstate.edu/courses/118118/users/165323)

Apr 16, 2022

Hi Everyone,

Data science is employed in a wide range of applications, including digital advertising and internet searches. Data Science is important in the development of machine learning and AI. Then they create an algorithm with the help of data analysts. I want to land my career in data scientist position. From my research and knowledge here I want to recommend some key areas to be strong for the data scientist aspirants to achieve their dream job role.

Broadly speaking, a data scientist should be aware of the following:

- Applied Statistics, Data Mining, and computational methods such as neural networks and machine learning are also required.
- Knowledge in database systems such as MySQL, Hive, and others are necessary.
- Statistics may be a useful tool (DS). In a broad sense, statistics is the application of mathematics to the technical study of data. A simple visualization, such as a bar chart, may provide some high-level information, but statistics allows us to work on the data in a far more targeted and information-driven manner. Rather than guesstimating, the math lets us develop specific conclusions about our facts.

To put it another way, a data scientist works on both prediction and perspective analytics, which should aid you in determining the knowledge base necessary.

Hope this helps!!

(http

Anudeep Nare (https://canvas.okstate.edu/courses/118118/users/188001)

Apr 30, 2022

Hello Sumanjali,

Interesting response! This is all new information to me. Applied Statistics, Data Mining, and computational methods such as neural networks are very important to a data analyst and you have given a valuable insights.

Thank you for sharing.

Thirumala Krishna Kurakula (https://canvas.okstate.edu/courses/118118/users/161132)

Apr 16, 2022

Predictive analytics models analyze historical data, identify patterns, spot trends, and use that data to make predictions about future events.

Classification models:

A classification model takes some data and produces an output that categorizes it into one of several categories.

Time series models:

A collection of quantities assembled over even periods and ordered chronologically is referred to as time-series data. The time series frequency refers to the frequency at which data is collected over a period.

Outliers models:

An outlier is an observation that differs so significantly from the rest of the data in a database. This description is directly adapted by the outlier identification approach, which specifies a hypothesis that an outlier is distributed differently from all other occurrences in a database.

Clustering models:

Clustering models are concerned with locating groups of similar data and categorizing them according to their membership in that group.

Forecast models:

Forecasting models are one of the many techniques used by businesses to forecast sales and consumption patterns. In the marketing domain, these models are extremely useful. Businesses use a variety of forecasting methodologies that provide differing degrees of knowledge.

This website briefly explains these models.

https://seleritysas.com/blog/2019/12/12/types-of-predictive-analytics-models-and-how-they-work/

Anudeep Nare (https://canvas.okstate.edu/courses/118118/users/188001)

Apr 30, 2022

Hello Thirumala Krishna,

It was a very Interesting response! This is all new information to me. **Classification model** ,**Clustering models, Clustering models** very important in predictive analytics models to analyze historical data up to my knowledge and you have given a valuable insights.

Thanks,

Anudeep Reddy Nare

Anudeep Nare (https://canvas.okstate.edu/courses/118118/users/188001)

Apr 17, 2022

Hello class,

I'd want to tell my ideas and experiences from working on an analytics project. Data science is used in a variety of applications, such as digital advertising and internet searches. Data Science is critical to the advancement of machine learning and AI. Then, with the assistance of data analysts, they develop an algorithm. AI and machine learning in sports are becoming increasingly common. It affected audience participation and culminated in the creation of a game strategy. Artificial intelligence (AI) is having a huge impact on how sports fans perceive them. AI systems may be used to automatically determine the optimum camera angle to display on the viewer's screen, give real-time subtitles in many languages based on the viewer's location, and enable broadcasters to monetise their content.

Sumanjali Etlam (https://canvas.okstate.edu/courses/118118/users/165323)

May 1, 2022

Hi Anudeep,

Thanks for sharing your experience on analytics project. I agree Data science is very crucial now-a-days.

Again thanks for sharing!

Thanks,

Sumanjali

Seonwoo Ko (https://canvas.okstate.edu/courses/118118/users/194258)

Apr 18, 2022

Hello y'all,

For my first poster presentation at the hospitality conference, I have been working on collecting Reddit API. I have been utilizing R studio, but I have found some issues using R studio. I would like to try Python soon to figure out gathering and analyzing data. Since I am a pretty new researcher in hospitality and data analytics, I usually get a lot of resources from Github. The website below has helped me a lot to process the Reddit API. If anyone has utilized Reddit for analyzing data, please share your experience with me!

https://github.com/reddit-archive/reddit/wiki/OAuth2 (https://github.com/reddit-archive/reddit/wiki/OAuth2)

(https:/

Seonwoo Ko (https://canvas.okstate.edu/courses/118118/users/194258)

Apr 18, 2022

Hello,

I would like to share my research project with our department! I am a research assistant in hospitality and tourism management. My advisor bought a service robot by collaborating with other professors and we are gathering data using a service robot at Taylor's restaurant. We are using the Nao robot from Softbank company in Japan. He is a programmable robot and interacts with customers. We have used their own developing program, but we also can use the Python program. After this semester, we will develop the robot more! If you are interested in the robot, you are more than welcome to come to Taylor's next week!

https://www.softbankrobotics.com/emea/en/nao (https://www.softbankrobotics.com/emea/en/nao)

(http

Ben Lewis (https://canvas.okstate.edu/courses/118118/users/211833)

Apr 26, 2022

Whoa! I so wish I was on campus to see this. I had to look up Nao but the Youtube videos alone have me excited. What all tasks are you having the robot do?

(http

Seonwoo Ko (https://canvas.okstate.edu/courses/118118/users/194258)

May 1, 2022

Hello Ben,

I am glad you are interested in Nao! He can recognize your voice and face as long as you interact with Nao. He can speak more than 20 languages, including English, French, Chinese, Japanese, etc. He also can sing, dance, and answer your questions! (like How are you, Is robot dangerous?) I think robot in hospitality and tourism industry is in initial stages, but it'll get popular at some point! If you are interested, you can come Taylor's restaurant next semester! I hope you see there. Thank you!

Rithik Ponugoti (https://canvas.okstate.edu/courses/118118/users/189064)

Apr 19, 2022

Hello Class,

The Curse of Dimension.

When we have a larger dataset in terms of features, the computing work required for processing or analysis grows exponentially. In principle, adding more dimensions to data might provide a lot of information, but in practice, it adds more noise and redundancy to the data. Higher-dimensional data won't function well with distance-based machine learning techniques.

Techniques in Dimensionality Reduction

1. Feature Selection Methods.

2. Manifold Learning (e.g. t-sne, MDS, etc..)

3. Matrix Factorization (e.g PCA, Kernal PCA, etc..)

Principal Component Analysis (PCA) is a frequently used dimensionality reduction approach that falls under the category of unsupervised machine learning because it does not need the input of a label. PCA can be used to reduce dimensionality or to analyze higher-dimensional data in a lower-dimensional space. The aim of the PCA method is to discover new axes or basis vectors that maintain a larger variance for data in lower dimensions.

To learn more about PCA, please use the link below.

https://www.analyticsvidhya.com/blog/2022/03/learn-about-principal-component-analysis-in-details/ (https://www.analyticsvidhya.com/blog/2022/03/learn-about-principal-component-analysis-in-details/)

Navya Mynedi (https://canvas.okstate.edu/courses/118118/users/157687)

Apr 21, 2022

Hi Rithik,

Thank you for sharing your knowledge about dimensionality techniques. This appears to be a whole new topic for me. I learned a lot about the PCA after looking at the link you gave. Now I have a better understanding of how to deal with higher-dimensional data in a lower-dimensional. They described the steps involved in the PCA in details. Even though the mathematical calculations they presented were difficult for me to follow, I can state that I understood the main concepts of PCA.

Bhanu Teja Pulipalupula (https://canvas.okstate.edu/courses/118118/users/159029)

Apr 19, 2022

Hello everyone,

I worked as a web/analytics developer for OSU's Fire Protection publications (FPP) department as a student employee last semester and here're some notes from that experience. For an e-commerce website where OSU FPP sells training material for firefighters, we used Google Analytics, Search Console, and Google Ads for digital marketing purposes. I made a few user interface changes to the product category pages in the e-commerce website, and we used Google Analytics reporting to compare the KPI metrics like pageviews, avg time on page, bounce rate, etc to that of the period before the UI changes were made so that we can decide if we should keep the UI changes made or revert them back. Tools like Google Analytics come really come in handy for tracking end-user reactions to changes like these.

An advertisement or product page showing up on the top few search results in a search engine like Google for most search terms in the domain is the key for any business. I worked on setting up campaigns in the Google Ads console, for this to work, each campaign is associated with a set of Search terms (of our choice) that we can configure and must pay for. Google Ads also provides us a lot of search term data with word frequencies, search term number of occurrences, clicks, impressions and Click Through Rate (CTR), Cost Per Click (CPC) etc. With this data, I used to analyze the top search terms that can be added to the campaign to ensure a better return on investment (ROI). Apart from this, from analyzing the existing search term data that we're already paying for, based on various metrics, for example, if the Cost Per Click for a search term that we're paying for is comparatively high, I used to present it to the marketing team, and we removed those search terms for which there was low ROI. Google Ads console is a great tool for digital marketing and helps us to get our product pages in the top few search results in a search engine.

I hope you find this insightful, you may want to google KPIs like CTR, and CPC to learn more.

Thanks,

Bhanu Teja P

(http

Anurag Budme (https://canvas.okstate.edu/courses/118118/users/155819)

Apr 21, 2022

Hi Bhanu Teja,

Thank you for sharing your real-time UI and marketing experience with Google Analytics. For every organization to increase revenue compared to previous periods, UI/UX has become a critical factor to consider. Using the UI metrics to address user difficulties will result in a significant increase in revenue for the company. I've looked at the CTR and CPC numbers, and they're really useful.

Thanks,

Budme

(http

Navya Mynedi (https://canvas.okstate.edu/courses/118118/users/157687)

Apr 23, 2022

Hi Bhanu Teja,

Thank you for sharing your experience. I believe you did an excellent job with your work. I used to be surprised at how these advertisements appear in the Google search engine when we type in similar terms. I know that advertising is very vital in any business. You presented a thorough overview of the technology involved, and I looked into the KPIs you indicated. That helped me a lot to understand better.

Thanks,

Navya

(http

Seonwoo Ko (https://canvas.okstate.edu/courses/118118/users/194258)

Apr 23, 2022

Hello Bhanu Teja,

Thank you for sharing your real experience! I was always wondering which programs they use in the real industries. I think Google Analytics is reliable and effective because almost all people use google in many ways, such as google maps, search engines, PowerPoint, and advertising. I have never used Google Analytics, but I will definitely look it up and see if I can use these analytics in the Hospitality and Tourism Management field! Thank you again for your information.

Thanks,

Seonwoo

Thirumala Krishna Kurakula (https://canvas.okstate.edu/courses/118118/users/161132)

Apr 19, 2022

Variable selection is an essential step in building classification models. We must consider many factors to narrow down the variables list.

1. Choosing candidate variables is a technique to narrow down the list of potential variables. Candidate variables exhibit predictive accuracy with the outcome. Candidate variables for a certain topic might be chosen based on specialist knowledge. We can also accomplish this by researching relevant material.

2. Variables with a high positive or negative correlation can be eliminated.

3. Variables can be selected based on the variable importance chart/values.

4. If a variable is critical, we can expect the model's performance to deteriorate after permuting the values of the variable. The more significant the difference in performance, the more significant the variable.

5. By using variable selection techniques:

i) Backward elimination: "A variable selection procedure in which all variables are entered into the equation and then sequentially removed. The variable with the smallest partial correlation with the dependent variable is considered first for removal."

ii) Forward Selection: "A stepwise variable selection procedure in which variables are sequentially entered into the model. The first variable considered for entry into the equation is the one with the largest positive or negative correlation with the dependent variable."

It took the explanation for backward elimination and forward selection from the website "https://www.ibm.com/docs/el/spss-statistics/beta?topic=regression-linear-variable-selection-methods". There are many other important techniques for variable selection.

Nithya Satheneni (https://canvas.okstate.edu/courses/118118/users/187996)

Apr 22, 2022

Hi Krishna,

Thanks for sharing this important topic that is very crucial in the building models. We will have lot of unwanted variables in the dataset which we usually work on and do not understand which one is useful for modeling and prediction. Variable selection is an important step which helps us in cleaning our data. This is the major part of Data Cleaning process. The two techniques which you mentioned are very good. Based on the model and data, we can choose which one to use.

Thanks,

Nithya

Bhanu Teja Pulipalupula (https://canvas.okstate.edu/courses/118118/users/159029)

Apr 23, 2022

Hi Krishna, .We end up spending most time in variable selection while working on prediction problems. Variable importance is one of the powerful techniques that helps us in the variable selection, the more a model relies on a variable to make predictions, the more important it is for the model. The interesting thing about variable importance is that we can check this for different models and check how each model ranks the important variables, check for common variables and remove the least important variables from our consideration. I read about the other techniques you shared and I find them very insightful, thanks for sharing these links!

Navya Mynedi (https://canvas.okstate.edu/courses/118118/users/157687)

Apr 23, 2022

Hi Krishna,

Thank you for sharing your insights on variable selection. When I first started working on the project, I was unsure how to choose the important variables and eliminate the others. When

we have a little amount of data with a few columns, we can manually look through the data and grasp the relevance of each one. However, when we have a huge amount of data with many columns, it is very difficult to comprehend the importance of each one. The methods you described are very helpful in understanding the various strategies involved in variable selection.

Thanks,

Navya

(https:/

Nithya Satheneni (https://canvas.okstate.edu/courses/118118/users/187996)

Apr 20, 2022

Greetings all,

I came across an interesting article about how block chain makes predictive analytics more accessible

One of the major challenges of predictive analytics is the volume of the data necessary to train the algorithms, Access to such large volumes of data is not always feasible and also requires high computational power for processing hence because very expensive.

This article provides a brief overview on how blockchain provides tenable solutions for some of the challenges of predictive analytics and mentions some of the potential applications of predictive analytics through block chain.

It would be worthwhile if you take a quick glance on the below article

https://towardsdatascience.com/how-will-blockchain-make-predictive-analytics-accessible-d256d543081d (https://towardsdatascience.com/how-will-blockchain-make-predictive-analytics-accessible-d256d543081d)

Thanks,

Nithya Satheneni

(http

Anurag Budme (https://canvas.okstate.edu/courses/118118/users/155819)

Apr 21, 2022

Hi Nithya,

Thanks for sharing the knowledge and informative link. Statement in the article – “For years blockchain was synonymous with Bitcoin since it was the underlying technology. This is not the case anymore” is completely agreed.

More organizations will utilize blockchain technology and address business challenges as a result of the shift in approach from centralized to decentralized approaches offered by blockchain.

Thanks,

Budme

Andrea Zerman (https://canvas.okstate.edu/courses/118118/users/214524)

Apr 20, 2022

Hello,

I know everyone was not able to attend the KNIME data connect back in February. I wanted to share a node I learned about as it has played an important part for me in the last homework and group project. It is called the Parameter Optimization Loop. After finding a model that produces great accuracy, you can potentially optimize it further. For a selected parameter in your training model, it will loop through the specified values and find the best value to achieve your goal. Your goal can be to maximize or minimize any of the scorer accuracy statistics such as accuracy, sensitivity, etc. Here’s a link to a youtube video that walks through an example: https://youtu.be/IlqepyIba6Y

Jay West (https://canvas.okstate.edu/courses/118118/users/57886)

Apr 21, 2022

Thank you for sharing your knowledge! I also attended the KNIME conference but it was not easy to understand the use of nodes and the concept. I might look into the Parameter Optimization Loop and hope this node helps our group project improve accuracy. Thank you!

Anurag Budme (https://canvas.okstate.edu/courses/118118/users/155819)

Apr 21, 2022

Hi Zerman,

Thank you for sharing the information.

After watching the video, I decided to try out the node in my project, and it turned out to be fantastic. In the project, trial and error with various sets of parameters is a time-consuming process. This node allowed me to adjust all of the settings at once. Aside from that, it allows you to choose the computational workload, which is useful when dealing with computationally intensive models.

The option of minimizing and maximizing the specified metric saved a lot of time in the model selection process when it came to the final metrics related to accuracy and sensitivity.

Thanks,

Budme

Edited by **Anurag Budme (https://canvas.okstate.edu/courses/118118/users/155819)** on Apr 21 at 1:09pm

---

Grant Lackey (https://canvas.okstate.edu/courses/118118/users/87846)

Apr 20, 2022

Hi everyone, I'm Grant Lackey. I'm a senior in the 4 + 1 BAnDS program graduating this semester with a bachelor's degree in Marketing Research and Analytics and aiming for my master's degree shortly after. I have always had a passion for video games. This past week has been a huge week for the game Valorant. It was the first international tournament for this season, which means plenty of data to go through for player statistics and rankings! In my free time, I try to give back worthwhile visualizations to the gaming community for big events. I started with scraping the data from the website into an excel sheet. I have been editing an equation to find the Player Value Rating (my created KPI for this tournament) measuring the different weights/importance of every statistic given by Valorant professionals during their games. Think about player ratings in the NFL, MLB, or NBA for fantasy leagues, that is what I am trying to replicate but with Valorant. After creating the PVR, I try to find the correlation between this statistic to other data points such as character selection, map choice, team composition, etc. Once I find an interesting find, then I begin to visualize the data through Excel or Tableau. This allows the audience to comprehend the great amount of information at a glance! The finals are coming up this weekend, so I will continue to find new ways to analyze this data, but I will be focusing on my finals first, haha!

Here is a link to the data I initially started with: **https://www.vlr.gg/event/stats/926/valorant-champions-tour-stage-1-masters-reykjav-k (https://www.vlr.gg/event/stats/926/valorant-champions-tour-stage-1-masters-reykjav-k)**

Is this informative? Other than about myself, not so much. But maybe this will inspire someone here to start a small project like this in their free time! Thanks for reading! - Grant Lackey

(http

Rithik Ponugoti (https://canvas.okstate.edu/courses/118118/users/189064)

Apr 23, 2022

Hi Grant,

I am quite inspired by your vision of giving back to the gaming community! I am also interested in the games and I gave a thought to starting a project in the same. The link you have provided gives perfect data for detail-oriented visualizations. I think the BeautifulSoup package in Python can help to scrape the data or maybe the web crawler. I will surely get my hands on this project.

Thanks!

Rithik Sai Ponugoti

(http

Seonwoo Ko (https://canvas.okstate.edu/courses/118118/users/194258)

Apr 23, 2022

Hello Grant,

Thank you for your information! I have recently started League of Legends and realized that there are such big game communities. I agree with you that there is plenty of data in the game industry and we will definitely be able to find some topics or data that we can use for research or find really interesting. I will definitely look the website up soon! Thank you for your interesting information! It was really informative. :)

Best,

Seonwoo

(http

Fayaz Shaik (https://canvas.okstate.edu/courses/118118/users/190988)

May 1, 2022

Hey Grant,

I am also another gaming fan, who loves Valorant. I have been playing it for quite a while now and I try to follow up with all the tournaments happening. I am aware about the Valorant Champions Tour 2022: Stage 1 that happened from April 10 to April 24, the pool prize size was the biggest which was for $675,000. The winner was the deserving OpTic Gaming

(OPTC) who won $200,000 and 750 points followed by Loud (LLL). I completely agree with you on how much amount of data we can play with for visualizations in such tournaments.

And, Yes, Seonwoo gaming communities are under-rated, imagine working on real-time visualizations for such tournaments using various software tools to predict data using various variables, it would be so much fun and also very informative.

Grant, I am amazed by how you spend your free time. I would love to see more of it and do share your interesting find using your PVR and the variables you used for capturing your data points.

P.S. - All the best on your finals!

Thank you Grant, for sharing the source link, where you scraped your data from. In the below link you can find the statistics of the subject tournament and other details which I followed up with:

https://liquipedia.net/valorant/VCT/2022/Stage_1/Masters (https://liquipedia.net/valorant/VCT/2022/Stage_1/Masters)

Best Regards,

Shaik Fayaz

MSIS Fall 2021

Anurag Budme (https://canvas.okstate.edu/courses/118118/users/155819)

Apr 21, 2022

Hello Everyone,

I'd want to discuss a key issue that came up during the Data Science project's execution, particularly model deployment. Model deployment is the process of integrating a machine learning model into an existing production environment so that it may be used to make data-driven business decisions.

According to the report by VentureBeat, 90% of the models never make it into production. The lack of cross-language and framework compatibility is one of the causes of the failure to commercialize machine learning models. When it comes to different languages and frameworks, there are significant disparities.

Furthermore, certain pipelines may employ Docker and Kubernetes for containerization, while others may not. Specific APIs will be deployed by some pipelines, but not by others. The list

goes on and on.

To address this void, tools like TFX, Mlflow, and Kubeflow are starting to emerge. However, these tools are still in their infancy, and knowledge about them is now scarce. Aside from that, when it comes to machine learning models, versioning and reproducibility remain a challenge.

To handle the difficulties of model deployment. I found some articles that will assist you to comprehend the various ways of deploying machine learning models. Apart from that, I've included a link to the instructions for deploying a machine learning model in Google Cloud.

https://cloud.google.com/vertex-ai/docs/predictions/deploy-model-console (https://cloud.google.com/vertex-ai/docs/predictions/deploy-model-console)

https://towardsdatascience.com/3-ways-to-deploy-machine-learning-models-in-production-cdba15b00e (https://towardsdatascience.com/3-ways-to-deploy-machine-learning-models-in-production-cdba15b00e)

(http

Navya Mynedi (https://canvas.okstate.edu/courses/118118/users/157687)

Apr 21, 2022

Hi Anurag,

Thank you for sharing your deployment knowledge. I've always had doubts about the deployment. I don't have much experience dealing with these issues because I haven't done any real-world projects with machine learning models. I learned a lot from the links you supplied. Now I know which tools are the most effective.

Furthermore, the link you provided for deploying a machine learning model in Google Cloud is really precise. That made it easier for me to understand since they detailed everything with step-by-step instructions.

Thanks,

Navya

(http

Rithik Ponugoti (https://canvas.okstate.edu/courses/118118/users/189064)

Apr 21, 2022

Hi Anurag,

Thank you for sharing this information! I went through those articles and learned quite a few things one of the simpler ways to deploy a machine learning model is to create a web

service for prediction. I gave a quick reading about the three steps involved in creating a web service including building a model, persisting the model, and serving the persisted model using a web framework.

Thanks!

Rithik.

(http

Bhanu Teja Pulipalupula (https://canvas.okstate.edu/courses/118118/users/159029)

Apr 22, 2022

Hi Anurag,

Thanks for sharing this, I think deployment is the most important skill which we do not experience in academics. From the GCP article you shared, I read a little more about model deployment using APIs. In the API led model deployment, it must be deployed to an endpoint; deploying a model associates physical resources with the model, allowing it to deliver online predictions with minimal latency.

Here's a link that shows sample scripts to create and hit endpoints, **https://cloud.google.com/vertex-ai/docs/predictions/deploy-model-api#aiplatform_create_endpoint_sample-python** (https://cloud.google.com/vertex-ai/docs/predictions/deploy-model-api#aiplatform_create_endpoint_sample-python) .

Thanks,

Bhanu Teja P

(https:/

Rithik Ponugoti (https://canvas.okstate.edu/courses/118118/users/189064)

Apr 21, 2022

Hello Class,

One can always expect a question on Variance/Bias in any Data Science interview. It is absolutely vital to understand the concepts behind these two terms and the key differences.

To explain these in simple terms

- **Bias** is the simplifying assumptions made by the model to make the target function easier to approximate.
- **Variance** is the amount that the estimate of the target function will change, given different training data.

- **Bias-variance trade-off** is the sweet spot where our machine model performs between the errors introduced by the bias and the variance.

If you want to revise these critical concepts in less than 3 minutes, I would suggest diving deep into this article.

https://towardsdatascience.com/understanding-bias-variance-trade-off-in-3-minutes-c516cb013513 (https://towardsdatascience.com/understanding-bias-variance-trade-off-in-3-minutes-c516cb013513)

Thanks,

Rithik Sai Ponugoti

○

Rudrakumar Ankaiyan (https://canvas.okstate.edu/courses/118118/users/190716)

Apr 24, 2022

Hi Rithik,

Thanks for sharing the information about Variance/Bias. I guess it will be very helpful for the rest of us as most of the organizations are expecting us to be strong and clear about the fundamentals of Data Science. Keep Sharing!

Thanks,

Rudrakumar Ankaiyan

○

Navya Mynedi (https://canvas.okstate.edu/courses/118118/users/157687)

Apr 21, 2022

Hello all,

As we all know that Preparing data for a machine learning (ML) system is time consuming, difficult, and mistake prone. Data normalization is an important part of this data preparation process. Every time I come across this step, I'm always unsure what it does and what are the various normalizing approaches and how they differ. I read a few articles to gain a better understanding of normalization. Normalization, in simple words according to my understanding, is each variable is given equal weight, such that no single variable leads model performance in one direction just because they are larger numbers.

I found these three normalization techniques that are more useful.

Rescaling: also known as "min-max normalization," is a process in which the data's minimum value is converted to a 0, the highest value is converted to a 1, and all other values are converted to a decimal between 0 and 1.

Mean normalization: this method uses the mean of the observations.

Z-score normalization: This technique, also known as standardization, employs the Z-score or "standard score." SVM and logistic regression are two examples of machine learning algorithms that utilize it

These two articles are more interesting because they go into great detail about the differences and applications of these techniques.

https://visualstudiomagazine.com/articles/2020/08/04/ml-data-prep-normalization.aspx (https://visualstudiomagazine.com/articles/2020/08/04/ml-data-prep-normalization.aspx)

https://towardsdatascience.com/data-normalization-in-machine-learning-395fdec69d02 (https://towardsdatascience.com/data-normalization-in-machine-learning-395fdec69d02)

(http

Nithya Satheneni (https://canvas.okstate.edu/courses/118118/users/187996)

Apr 22, 2022

Hi Navya,

Thanks for sharing this topic about the types of normalization and how it's used. I used to wonder what this step actually does in the model that is very useful for data preparation. The article you provided has good insights about when this technique is useful in the Data Preparation Pipeline and what are the different types we can use. Thanks for sharing!

Thanks,

Nithya

(http

Anudeep Nare (https://canvas.okstate.edu/courses/118118/users/188001)

Apr 26, 2022

Hello, Navya.

Thank you for discussing the many forms of normalization and how they are utilized. The article you linked to provides important information regarding when this strategy is useful in the Data Preparation Pipeline. Thank you for your contribution.

(https:/

Rithik Ponugoti (https://canvas.okstate.edu/courses/118118/users/189064)

Apr 22, 2022

Hello Class,

I recently started learning about NoSQL as part of my Data Warehousing course here at OSU. I wanted to quickly share about it and why we need it.

NoSQL databases (aka "not only SQL") are non-tabular databases and store data differently than relational tables. NoSQL databases come in a variety of types based on their data model. The main types are document, key-value, wide-column, and graph. They provide flexible schemas and scale easily with large amounts of data and high user loads. They have become incredibly popular in the industry and here are a few reasons responsible:

1. The pace of development with NoSQL databases can be much faster than with a SQL database.

2. The structure of many different forms of data is more easily handled and evolved with a NoSQL database.

3. The amount of data in many applications cannot be served affordably by a SQL database.

4. The scale of traffic and the need for zero downtime cannot be handled by SQL.

5. New application paradigms can be more easily supported.

If you are curious about knowing more about NoSQL and MongoDB here are the links you can refer to:

https://www.mongodb.com/nosql-explained (https://www.mongodb.com/nosql-explained)

https://www.mongodb.com/nosql-explained/when-to-use-nosql#:~:text=The%20pace%20of%20development%20with,iterations%2C%20and%20frequent%20code%20pushes. (https://www.mongodb.com/nosql-explained/when-to-use-nosql#:~:text=The%20pace%20of%20development%20with,iterations%2C%20and%20frequent%20code%20pushes.)

(http

Bhanu Teja Pulipalupula (https://canvas.okstate.edu/courses/118118/users/159029)

Apr 23, 2022

Hi Rithik, Thanks for sharing these links. It was good to learn more about the NoSQL schema that has a different structure than the traditional row-and-column table model used with relational database management systems (RDBMS) and the four main types of NoSQL

databases, i.e., document databases (stores data in JSON/XML docs), key-value stores (key-value pairs), column-oriented databases, graph databases (relationships between data elements). Here's the link to the webpage that describes each of these NoSQL database types, **https://www.mongodb.com/scale/types-of-nosql-databases (https://www.mongodb.com/scale/types-of-nosql-databases)**

Anudeep Nare (https://canvas.okstate.edu/courses/118118/users/188001)

Apr 26, 2022

Hello, Rithik. Thank you for providing these resources. It was interesting to learn more about the NoSQL schema, which differs from the traditional row-and-column table model used with relational database management systems (RDBMS), as well as the four main types of NoSQL databases.

Edited by **Anudeep Nare (https://canvas.okstate.edu/courses/118118/users/188001)** on Apr 26 at 7:31pm

Navya Mynedi (https://canvas.okstate.edu/courses/118118/users/157687)

Apr 29, 2022

Hi Rithik,

Thank you for sharing the articles about the NoSQL database. It was interesting to learn about how they differ from relational database systems. The articles you provided helped me to learn more about MongoDB and the different types of NoSQL databases.

Thanks,

Navya

Bhanu Teja Pulipalupula (https://canvas.okstate.edu/courses/118118/users/159029)

Apr 23, 2022

Hello everyone,

Most organizations use a variety of analytics together to make smart decisions that help a business. Here's a short article comparing Descriptive, Predictive, Prescriptive, and Diagnostic

Analytics with a healthcare setting example: https://insightsoftware.com/blog/comparing-descriptive-predictive-prescriptive-and-diagnostic-analytics/ (https://insightsoftware.com/blog/comparing-descriptive-predictive-prescriptive-and-diagnostic-analytics/) . At this point in time of this coursework, we all know this already but I hope you find the healthcare example insightful.

Thanks,

Bhanu Teja P

Jacob Wood (https://canvas.okstate.edu/courses/118118/users/214790)

Apr 28, 2022

Hi Bhanu Teja P,

Thanks for sharing this brief article. Although we've been exposed to all of this information already, I really enjoy reading about use cases for different analytics types. Sometimes, it's valuable to build your knowledge base of potential business insights needed and how various analytics types can provide these insights. I found a paper that might be of interest to you. It describes a dashboard used in an Emergency Department to reflect the status of the Emergency Department at any time. This reminds me of the descriptive analytics example described in your article.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6284143/ (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6284143/)

Hope you find some use out of this literature.

Bhanu Teja Pulipalupula (https://canvas.okstate.edu/courses/118118/users/159029)

Apr 23, 2022

Hello All, If you plan to start learning the basics of machine learning from scratch, you may want to check out the Machine Learning course by Andrew Ng on Coursera. Professor Andrew Ng talks clearly and the way he transfers knowledge is very simple and easy to understand. This course provides a lot of basic knowledge for anyone who doesn't know anything about machine learning. Here's the link to the course: https://www.coursera.org/learn/machine-learning (https://www.coursera.org/learn/machine-learning)

Thanks,

Bhanu Teja P

(http

Rudrakumar Ankaiyan (https://canvas.okstate.edu/courses/118118/users/190716)

Apr 24, 2022

Hi Bhanu,

Thanks for sharing the information. Andrew Ng is a pioneer in the field of AI. His courses on machine learning teach us about the fundamental concepts required for a beginner. His courses on Neural Networks are also very good. Please find the below link for his course about the foundational concepts of neural networks and deep learning.

https://www.coursera.org/learn/neural-networks-deep-learning?specialization=deep-learning (https://www.coursera.org/learn/neural-networks-deep-learning?specialization=deep-learning)

Thanks,

Rudrakumar Ankaiyan.

(http

Bhanu Teja Pulipalupula (https://canvas.okstate.edu/courses/118118/users/159029)

Apr 26, 2022

Hey Rudra, Thanks for sharing this link. Andrew Ng is launching a newly updated course collaborating with DeepLearning.AI and Stanford Online in June 2022, and here's the link: **https://www.deeplearning.ai/program/machine-learning-specialization/** (https://www.deeplearning.ai/program/machine-learning-specialization/) . Follow him on Twitter for more updates: **https://twitter.com/AndrewYNg** (https://twitter.com/AndrewYNg)

(http

Ashish Kumar Pampana (https://canvas.okstate.edu/courses/118118/users/24369)

Apr 29, 2022

Hi Bhanu,

Thank you for sharing the links. I have been doing a lot of research about learning the basics of machine learning techniques. I'm looking forward to learn for these links.

Thanks,

Ashish

Edited by **Ashish Kumar Pampana** (https://canvas.okstate.edu/courses/118118/users/24369) on Apr 29 at 3:43pm

(http

Fayaz Shaik (https://canvas.okstate.edu/courses/118118/users/190988)

May 1, 2022

Hello Guys,

Thank you Bhanu Teja and Rudra for the learning links of

- ML
- Neural Networks
- Deep Learning

These are very useful and will definitely help me out. Also, these are some great reads and courses to cover up this summer.

Best Regards,

Shaik Fayaz

MSIS Fall 2021

Edited by Fayaz Shaik (https://canvas.okstate.edu/courses/118118/users/190988) on May 1 at 12:03pm

Navya Mynedi (https://canvas.okstate.edu/courses/118118/users/157687)

Apr 23, 2022

Hello all,

While I was working on the project. when I saw the many columns with the categorical values. I was worried about how to handle them. To have a better understanding, I read a few articles. And I found that categorical data were divided into two types.

1. **Nominal Data:** In this, we can change the order of categories, changing the order doesn't affect its value, for example, Gender(male/female)
2. **Ordinal Data:** In this, we are not allowed to change the order of the categories. For example, Ranks: 1st/2nd/3rd, Education: (High School/Undergrads/Postgrads/Doctorate), etc.

Some ways to handle these categorical values.

Frequency Encoding: Replacing each category with the number of times that occurred in that column.

Ordinal Number Encoding: When we found that the data is ordinal we can replace the category with some ordinal number based on the ranks.

Mean Encoding: Here we replace the category with the mean value with respect to the target column.

Probability Ratio Encoding: Here the category column will be replaced with the probability ratio based on the target variable.

Look at this article to understand more about the various encoding methods.

https://www.kdnuggets.com/2021/05/deal-with-categorical-data-machine-learning.html (https://www.kdnuggets.com/2021/05/deal-with-categorical-data-machine-learning.html)

Edited by **Navya Mynedi** (https://canvas.okstate.edu/courses/118118/users/157687) on Apr 23 at 7:08am

---

Rudrakumar Ankaiyan (https://canvas.okstate.edu/courses/118118/users/190716)

Apr 24, 2022

Hi Navya,

Thanks for the information on different types of data attributes. It's very important for us to first understand the type of variables. As we know, each algorithm works well with a particular set of attributes type. So, we need to encode data as per our requirements. Keep sharing!

Thanks,

Rudrakumar Ankaiyan

Thirumala Krishna Kurakula (https://canvas.okstate.edu/courses/118118/users/161132)

May 1, 2022

Hi Navya,

Thanks for sharing this information. It was challenging for me to work on categorical attributes. If the category variable is concealed, decoding its meaning becomes difficult. There will only be one level that occurs for most possible samples. Due to the low variance, these factors have little effect on simulation results. No categorical parameters are allowed in ML libraries. So we make them numerical parameters.

Paul Davis (https://canvas.okstate.edu/courses/118118/users/194957)

Apr 23, 2022

Hey everyone - I've been taking this class as a fun elective for my MBA and it really has been fun for me, while also being pretty challenging. I work at an education non-profit and I have been

thinking quite a bit about the different ways we should be thinking about applying predictive modeling at work.

One of the key ways is actually addressed in detail in this article here: **https://www.mdrc.org/sites/default/files/Predictive_Modeling_of_K-12_Academic_Outcomes.pdf (https://www.mdrc.org/sites/default/files/Predictive_Modeling_of_K-12_Academic_Outcomes.pdf)**

Education uses a core set of risk factors named the A, B, Cs (Academics, Behavior and Course Performance) that have been consistently identified as being predictive markers of likelihood to graduate and can be used as early as the third grade. However they are binary markers, and today most public school students in large urban school districts have at least one risk factor. What that means is it is hard to pristine intervention time and understand which student is at most need. This paper is interesting because it is seeking to use the rich granular data that schools have to translate the 'binary' A, B, Cs into risk scores. It really can be a great tool to target student intervention and understand how best to apply risk factors to work.

It is also a great stepping stone into thinking about how to build a model with the data we have around students here in Tulsa.

(http **Nithya Satheneni (https://canvas.okstate.edu/courses/118118/users/187996)**

Apr 24, 2022

Hi Paul,

Thank you for sharing this useful information with us. I agree with this approach since any school must assess the risk to each student's academic career and take appropriate precautions. Many schools have already started using these methods. This article was very useful and fun to read.

Thanks,

Nithya

(https:/ **Jay West (https://canvas.okstate.edu/courses/118118/users/57886)**

Apr 23, 2022

It looks like big data in hospitality and tourism literature is not new anymore. Sentiment analysis appears to be widely used in our field to understand customers and employees. The attached

article discovers employees job satisfaction by using sentiment analysis. I found it is interesting as it helps visualizing the prominent words and feelings among employees!

10-1108_IHR-08-2018-0007.pdf (https://canvas.okstate.edu/files/14431891/download?download_frd=1&verifier=5NMr297ucQ5LOjcfHRJoYVyF8b0Ta2LOYDnZLYYf)

(http

Rithik Ponugoti (https://canvas.okstate.edu/courses/118118/users/189064)

Apr 28, 2022

Hi Jay,

Thank you for sharing the article. Sentimental Analysis is one of the hot techs that is being leveraged by many companies. For one's who are interested in knowing about the Sentimental Analysis, I would recommend them to take the Programming for Data Science I course here at Spears. The course deeply dives into the foundations of Sentimental Analysis in both Python and R. I would like to reminisce about the Text Blob. It is a Python (2 and 3) library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more

To know more about Sentimental Analysis you can also dive into this article:

https://monkeylearn.com/sentiment-analysis/#:~:text=Sentiment%20analysis%20(or%20opinion%20mining,feedback%2C%20and%20understand%20customer%20needs. (https://monkeylearn.com/sentiment-analysis/#:~:text=Sentiment%20analysis%20(or%20opinion%20mining,feedback%2C%20and%20understand%20customer%20needs.)

Thanks,

Rithik Sai Ponugoti

(https:/

Thirumala Krishna Kurakula (https://canvas.okstate.edu/courses/118118/users/161132)

Apr 23, 2022

Equal size sampling:

It removes entries from the input set to equalize the distribution of values in a category column. This node is useful if a learning technique is prone to asymmetrical class distributions and you want to decrease the data set such that the class attributes appear equally frequently.
The node will drop rows from the majority classes at random. This node's rows will include all records from the minority classes and a random sample from each of the majority classes, with each sample containing the same number of objects as the minority class.

Smote (Synthetic Minority Over-sampling Technique) :

To enrich the training data, smote oversamples the input data. To have an excellent classification performance, several supervised methods, like decision trees, require an equal class distribution. When the input data is uneven, such as when there are few items of the "active" class but many of the "inactive" class, this node changes the class distribution by creating synthetic rows.

Rudrakumar Ankaiyan (https://canvas.okstate.edu/courses/118118/users/190716)

Apr 24, 2022

Hi Thirumala,

Thanks for sharing information on Sampling. As we know, Class Imbalance is one of the major problems that we will come across while performing modeling, and Sampling techniques are very helpful to reduce the imbalance in the distribution of data between the classes.

Thanks,

Rudrakumar Ankaiyan.

Bhanu Teja Pulipalupula (https://canvas.okstate.edu/courses/118118/users/159029)

Apr 26, 2022

Hi Krishna, Thanks for bringing up this Class imbalance topic.
If there is an imbalance in the data set, the prediction for the majority class will have a high degree of accuracy, but we miss out on the minority class. Apart from the techniques you mentioned, here's a link that describes a few other techniques to deal with Imbalanced Classes: **https://www.analyticsvidhya.com/blog/2020/07/10-techniques-to-deal-with-class-imbalance-in-machine-learning/** (https://www.analyticsvidhya.com/blog/2020/07/10-techniques-to-deal-with-class-imbalance-in-machine-learning/) .
Thanks,

Bhanu Teja P

(http

Jacob Wood (https://canvas.okstate.edu/courses/118118/users/214790)

Apr 28, 2022

Hi all,

Thanks for sharing this information. When working with our team through iterations of our predictive models for the group project, we had to make some decisions about how to best balance the data. Initially we tried to model without extensive balancing and our minority class was under represented, so accuracy was high but we had sensitivity/specificity concerns. We were able to improve the models through different data balancing techniques. While we didn't try all the 10 techniques in the above article, I'll make a note of those techniques for future models.

(https:/

Nithya Satheneni (https://canvas.okstate.edu/courses/118118/users/187996)

Apr 24, 2022

Hello,

I'd want to talk about how predictive analytics and machine learning may assist discern between AI hype and reality. Today's market has been employing these tactics, which have proven to be extremely beneficial in keeping them current. Predictive analytics and machine learning offer a wealth of intelligence to the process of answering these questions, and here is where AI is generating value today. This aids in the management of both sales and marketing efficiency.

Please read the article below to learn more about this subject.

https://irelandstechnologyblog.com/predictive-analytics-ai-separating-hype-from-reality-a57f65a4f786 (https://irelandstechnologyblog.com/predictive-analytics-ai-separating-hype-from-reality-a57f65a4f786)

Thanks,

Nithya

(http

Rithik Ponugoti (https://canvas.okstate.edu/courses/118118/users/189064)

Apr 28, 2022

Hi Nithya,

Thank you for bringing this story to our attention. It's critical to completely comprehend the challenges we're attempting to tackle in order to truly appreciate what a new technology like AI can bring to the table. When it comes to AI solutions for marketing and sales, the present reality is less about future robots or automating every marketing procedure, and more about how data can answer one key go-to-market question: who to sell and/or advertise to. Predictive analytics and machine learning offer a lot of knowledge to the process of answering these questions, and here is where AI is generating value today.

Best,

Rithik Sai Ponugoti

Josh Basquez (https://canvas.okstate.edu/courses/118118/users/158078)

Apr 30, 2022

Rithik and Nithya -- There does appear to be some hype in the AI and analytics markets so it is a worthy discussion. I believe the value in the customer data does have another aspect however, besides the potential customers that might be interested in your product or service, and that is the matching of certain types of products and services with a customer's aligned interests. For example, if a company offers both electrical wiring as well as security or alarm systems, there may be a subset of customers that would be interested in the electrician services, and another subset (with some overlap potential) that may be more interested in the security services. So another value of the analytics could be in processing of the potential clients for both targeted as well as relevancy-based advertising.

Another dimension to this might be the time-dimension predictions, that is, predicting not only who might be interested in a product/service but also when over a certain timeline those potential clients might be most open to engaging in a transaction with the source marketing.

I appreciate this discussion and believe it is necessary and responsible for the data science community to try to self-regulate and call out potential issues that might give the industry a bad name.

Nithya Satheneni (https://canvas.okstate.edu/courses/118118/users/187996)

Apr 24, 2022

Hi Everyone,

I wanted to know what all the different types of machine learning methods are and when each of these methods shine out based on datasets because we're all working on different ML strategies

for prediction in our assignments/projects. Unsupervised and supervised machine learning are the two main types of machine learning. When we know what we want to teach the machine, we employ the supervised technique. This strategy uses previous data to forecast the company's future hazards. Unsupervised learning is the process of forming patterns from data and grouping the data. It is simple to implement and deploy because it does not require vast datasets. Two alternative strategies are discussed in the below article, along with the techniques that each method employs.

https://mobidev.biz/blog/5-essential-machine-learning-techniques (https://mobidev.biz/blog/5-essential-machine-learning-techniques)

Thanks,

Nithya

(http

Rudrakumar Ankaiyan (https://canvas.okstate.edu/courses/118118/users/190716)

Apr 24, 2022

Hi Nithya,

Thanks for sharing the information about the different types of machine learning techniques. I guess it's very important for a Data Scientist to know how to choose a particular machine learning technique that produces the expected results after studying and examining the dataset.

Thanks,

Rudrakumar Ankaiyan

(http

Rithik Ponugoti (https://canvas.okstate.edu/courses/118118/users/189064)

Apr 25, 2022

Hi Nithya,

Thank you for sharing the article. I went through the link you have provided and it is really a good source to brush up on the basics. I was able to go through different methods in both Supervised (Classification, Regression) and Unsupervised learning(Clustering). I would definitely recommend this!

Hi Rudra,

Yes! It is very important for a Data Scientist to know the context of the model before using it. I guess having domain knowledge might aid the knowledge.

Best,

Rithik Sai Ponugoti

Rithik Ponugoti (https://canvas.okstate.edu/courses/118118/users/189064)

Apr 25, 2022

Hello Class,

I would like to share about Reinforcement Learning. I quite hear less about this type, so I wanted to read a bit of it and share my insights.

Reinforcement learning is all about taking the right steps to maximize your benefit in a given circumstance. It is used by a variety of software and computers to determine the best feasible action or path in a given scenario. Reinforcement learning differs from supervised learning in that supervised learning includes the answer key, allowing the model to be trained with the correct answer, whereas reinforcement learning does not include an answer and instead relies on the reinforcement agent to decide what to do to complete the task. It is obligated to learn from its experience in the absence of a training dataset.

To get to know more about Reinforcement learning, please feel free to dive into the below link

https://www.geeksforgeeks.org/what-is-reinforcement-learning/
(https://www.geeksforgeeks.org/what-is-reinforcement-learning/)

Best Regards,

Rithik Sai Ponugoti.

Nithya Satheneni (https://canvas.okstate.edu/courses/118118/users/187996)

Apr 26, 2022

Hi Rithik,

Thanks for sharing this area of machine learning. We do hear about very less, but I feel that this is a good technique that can be used for maximizing the results. The example of the robot and the diamond was very well explained in the article that gives us a good base on this reinforcement learning.

Thanks,

Nithya Satheneni

(https:/

Anudeep Nare (https://canvas.okstate.edu/courses/118118/users/188001)

Apr 26, 2022

Hello class,

Analytics has been impacting the bottom line for organizations. Now that more businesses have mastered the use of analytics, they are diving deeper into their data in order to improve productivity, acquire a competitive edge, and boost their bottom lines even more. That's why businesses are aiming to integrate machine learning (ML) and artificial intelligence (AI) as part of a broader analytics strategy to meet their objectives. The first step is to understand how to incorporate contemporary machine learning algorithms into their data architecture. Many people are turning to firms who have already started the implementation process and have had success.

Here I have attached a link of complete article about Data analytics:

https://callminer.com/blog/smart-implementation-machine-learning-ai-data-analysis-50-examples-use-cases-insights-leveraging-ai-ml-data-analytics
(https://callminer.com/blog/smart-implementation-machine-learning-ai-data-analysis-50-examples-use-cases-insights-leveraging-ai-ml-data-analytics)

Anudeep.

(http

Rithik Ponugoti (https://canvas.okstate.edu/courses/118118/users/189064)

Apr 28, 2022

Hi Anudeep,

Indeed! It is critical to understand the impact of Analytics on businesses. It was great to know that Twitter uses machine learning technology and AI to evaluate tweets in real-time and score them using various metrics to display tweets that have the potential to drive the most engagement. Also, other companies like Edgecase uses machine learning to analyze customer behaviors and actions to provide a better experience for shoppers who may not know what they want to buy, in an effort to make casual online browsing more similar to a traditional retail experience.

Keep sharing the links that provide the business aspect of the analytics!

Thanks,

Rithik Sai Ponugoti.

(https:/

Anudeep Nare (https://canvas.okstate.edu/courses/118118/users/188001)

Apr 26, 2022

Hello Classs,

I would like share my experience TIBCO Spotfire. TIBCO Spotfire the most comprehensive analytics solution on the market, TIBCO Spotfire software allows anybody to explore and visualize new data discoveries through immersive dashboards and sophisticated analytics. Predictive analytics, geolocation analytics, and streaming analytics are just a few of the features Spotfire® analytics offers at scale. You can also construct bespoke analytic apps quickly, regularly, and at scale using Spotfire Mods. You receive a smooth, single-pane-of-glass experience for visual analytics, data discovery, and point-and-click insights with the Spotfire analytics platform and the TIBCO Hyperconverged Analytics advantage. Immerse yourself with interactive historical and real-time data: With fully brush-linked, dynamic visualizations, drill down or across multi-layer, diverse data sources.

Please go through the article to know more about TIBCO Spotfire:

https://www.tibco.com/products/tibco-spotfire?utm_medium=cpc&utm_source=google&utm_content=s&utm_campaign=ggl_s_en_nam_SPT_nonbrand_beta&utm_term=%2Bdata%20%2Bvisualization&_bt=583764788349&_bm=b&_bn=g&gclid=CjwKCAjwsJ6TBhAIEiwAfl4TWK94N5PqkwNxvv8LPLBX5vk8zn9BW9ZNZQuvoFtvrlwvUygeepif_hoC8XUQAvD_BwE (https://www.tibco.com/products/tibco-spotfire?utm_medium=cpc&utm_source=google&utm_content=s&utm_campaign=ggl_s_en_nam_SPT_nonbrand_beta&utm_term=%2Bdata%20%2Bvisualization&_bt=583764788349&_bm=b&_bn=g&gclid=CjwKCAjwsJ6TBhAIEiwAfl4TWK94N5PqkwNxvv8LPLBX5vk8zn9BW9ZNZQuvoFtvrlwvUygeepif_hoC8XUQAvD_BwE)

Thank you,

Anudeep

(http

Ben Lewis (https://canvas.okstate.edu/courses/118118/users/211833)

Apr 29, 2022

Very cool Anudeep! I'm curious, have you used this tool much? If so, how intense is the development time compared to other analytics solutions like SAS? Are visualizations easy to share with others once published?

Nithya Satheneni (https://canvas.okstate.edu/courses/118118/users/187996)

Apr 26, 2022

Hi Everyone,

I've been going through the CRISP-DM methodology and the benefits of utilizing it because we're all using it for our data mining project. This post explains why we will be adopting CRISP-DM for Data Science Projects, which I have put below. CRISP-DM is cost-effective, and it may be used in any project, regardless of domain. Please take a look at the article below.

https://analyticsindiamag.com/why-is-crisp-dm-gaining-grounds/ (https://analyticsindiamag.com/why-is-crisp-dm-gaining-grounds/)

Thanks,

Nithya Satheneni

Jay West (https://canvas.okstate.edu/courses/118118/users/57886)

Apr 26, 2022

Thank you for sharing the website! As stated in the article, CRISP-DM is a great methodology for researchers and data analytics. Due to the demand from data science, CRIPS-DM provides top down approach or solution oriented approach to solve problems. On the other hand, CRISP-DM does not address the application scenario in which an ML model is maintained as an application. Also, quality assurance needs to be addressed in CRISP-DM.

Again, thank you for sharing your knowledge and website! I learned a lot from the website!

Rithik Ponugoti (https://canvas.okstate.edu/courses/118118/users/189064)

Apr 28, 2022

Hi Nithya,

By now we actually know what the CRISP-DM is and its impact of it on the business world. Thanks for sharing the article! For anyone who wants to brush up on the concept of CRISP-DM can actually visit the website to know more. It beautifully interprets the steps of Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment.

Thanks again!

Best,

Rithik Sai Ponugoti

Anurag Budme (https://canvas.okstate.edu/courses/118118/users/155819)

Apr 30, 2022

Hi Nithya,

Thanks for recommending the article. It would be helpful for people looking to brush up on the CRISP-DM concepts.

It explains each stage of the CRISP-DM process effectively.

Thanks,

Anurag

Thirumala Krishna Kurakula (https://canvas.okstate.edu/courses/118118/users/161132)

May 1, 2022

Hi Nithya Sathineni,

Thanks for sharing this information. Crisp-DM allows us to specify all project's goals, especially deadlines, outcomes, understandability and trustworthiness, information safety concerns, and regulatory challenges. It also allows us to develop budgeting for the projects that evaluate the project's expenses to the possible advantages of the firm.

Bhanu Teja Pulipalupula (https://canvas.okstate.edu/courses/118118/users/159029)

Apr 26, 2022

Hello everyone, I'd like to share about the k-fold cross-validation technique, the data is separated into k subgroups, One of the k subsets is utilized as the test/validation set each time, while the remaining k-1 subsets are combined to form a training set. This considerably minimizes bias because the majority of the data is utilized for fitting, as well as variance because the majority of the data is also used in the validation set. Interchanging the training and test sets improves the method's effectiveness. K = 5 or 10 is often chosen as a general rule, although nothing is fixed, and it can take any value.

Bhanu Teja Pulipalupula (https://canvas.okstate.edu/courses/118118/users/159029)

Apr 26, 2022

In continuation to the above post, here's something about Stratified K-Fold Cross-Validation. The K Fold cross-validation technique is tweaked slightly such that each fold contains around the same percentage of samples from each target class as the entire set, or in the case of prediction problems, the mean response value is roughly identical in all folds.

If you'd like to read more about other cross-validation techniques, check out:
(https://towardsdatascience.com/cross-validation-in-machine-learning-72924a69872f)
https://www.analyticsvidhya.com/blog/2021/05/4-ways-to-evaluate-your-machine-learning-model-cross-validation-techniques-with-python-code/
(https://www.analyticsvidhya.com/blog/2021/05/4-ways-to-evaluate-your-machine-learning-model-cross-validation-techniques-with-python-code/)

https://towardsdatascience.com/cross-validation-in-machine-learning-72924a69872f
(https://towardsdatascience.com/cross-validation-in-machine-learning-72924a69872f)

Fayaz Shaik (https://canvas.okstate.edu/courses/118118/users/190988)

May 1, 2022

Thank you for the info on this ML technique Bhanu. The links you provided have a very detailed insight into the cross-validation techniques. I would definitely recommend everyone to read this to get a better understanding of this machine learning technique.

Best Regards,

Shaik Fayaz

MSIS Fall 2021

(https:/

Anudeep Nare (https://canvas.okstate.edu/courses/118118/users/188001)

Apr 26, 2022

Hello class,

I have come across a very good article about Analyzing Process Data Using Data Mining Techniques,

Many researchers have been interested by a new sort of data, process data, created through computer-based assessment, or new sources of data, such as keystroke or eye tracking data, as technology has advanced in educational assessment. Most of the time, such material, referred to as a "data ocean," is quite huge in volume and has few ready-to-use properties. It's been difficult to figure out how to explore, uncover, and extract relevant information from such a vast ocean.

Please go through the article :

https://www.frontiersin.org/articles/10.3389/fpsyg.2018.02231/full (https://www.frontiersin.org/articles/10.3389/fpsyg.2018.02231/full)

Thank you,

Anudeep Reddy

(http

Rithik Ponugoti (https://canvas.okstate.edu/courses/118118/users/189064)

Apr 28, 2022

Hi Anudeep,

Data mining is crucial and also most overlooked by the upcoming Data Science aspirants. They focus more on building complex models and less on the quality and relevance of the Data! The techniques you mentioned are quite new to me and the link you provided gave me a kind of Insight into it

Thanks and Keep sharing!

Best

Rithik Sai Ponugoti

(https:/

Moises Marin Martinez (https://canvas.okstate.edu/courses/118118/users/198182)

Apr 26, 2022

Hello!

There's a topic that I like when it comes to data. The so called "dark data". Dark data is defined by the Gartner Inc as the "information assets organizations collect, process and store during regular business activities, but generally fail to use for other purposes" [1]. In other words, it is data that a company collects but is not used to obtain insights for decision making, or for any other use. It is a problem because of the size, it's usually data bigger in size than the data being used for driving decisions. Storing it adds to the overall storage cost of data in a company, what is worst of this situation is that dark data could be hiding valuable insights.

Up to 90 percent of big data is dark data [2].

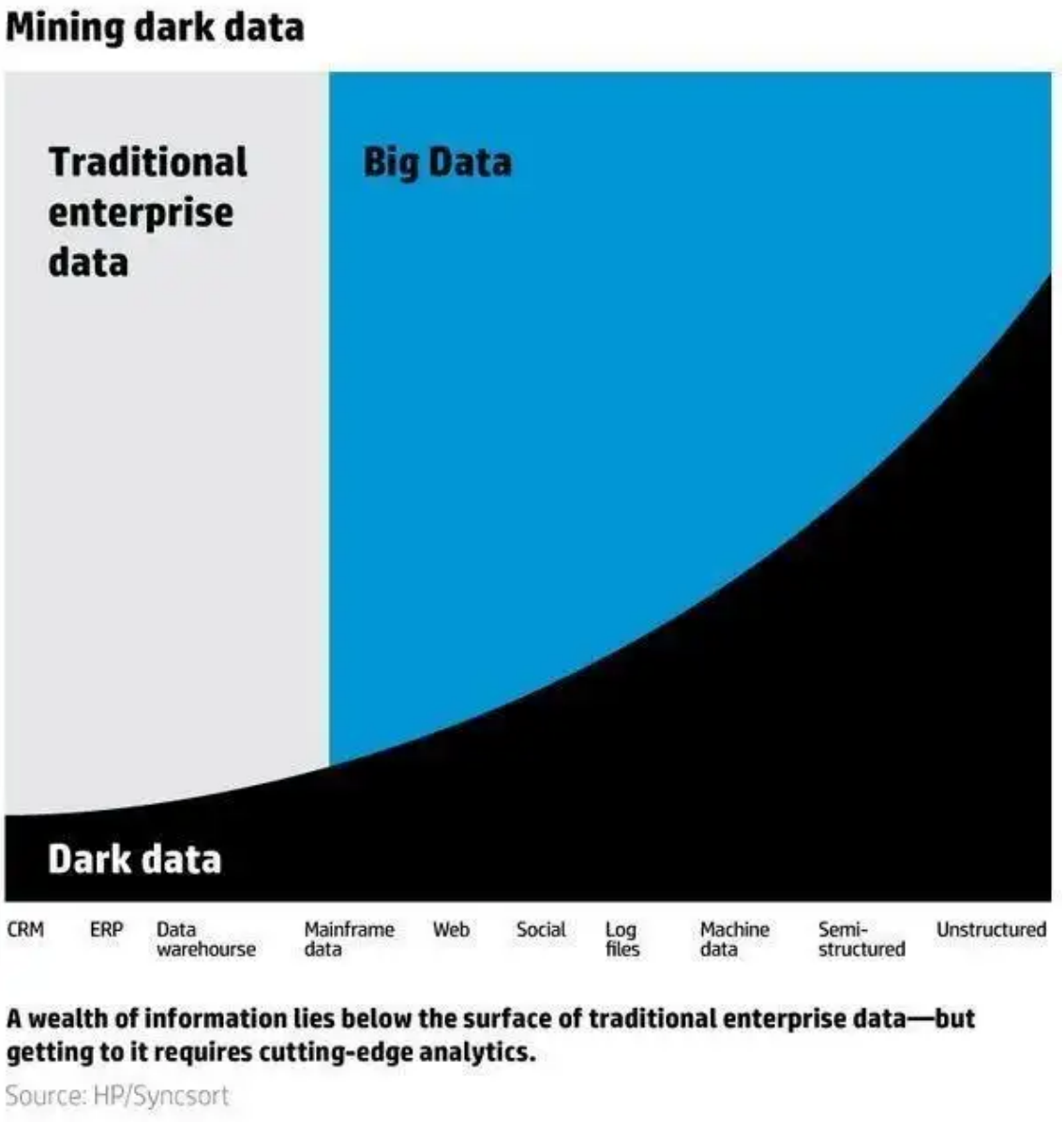


Image taken from reference [2].

[1]
Gartner. (2022). *Dark Data*. Gartner, Inc.
Web Site: https://www.gartner.com/en/information-technology/glossary/dark-data (https://www.gartner.com/en/information-technology/glossary/dark-data)

[2]
Banafa, A. (2021). Dark Data Explained. SemiWiki.com
Web Site: https://semiwiki.com/general/298187-dark-data-explained/ (https://semiwiki.com/general/298187-dark-data-explained/)

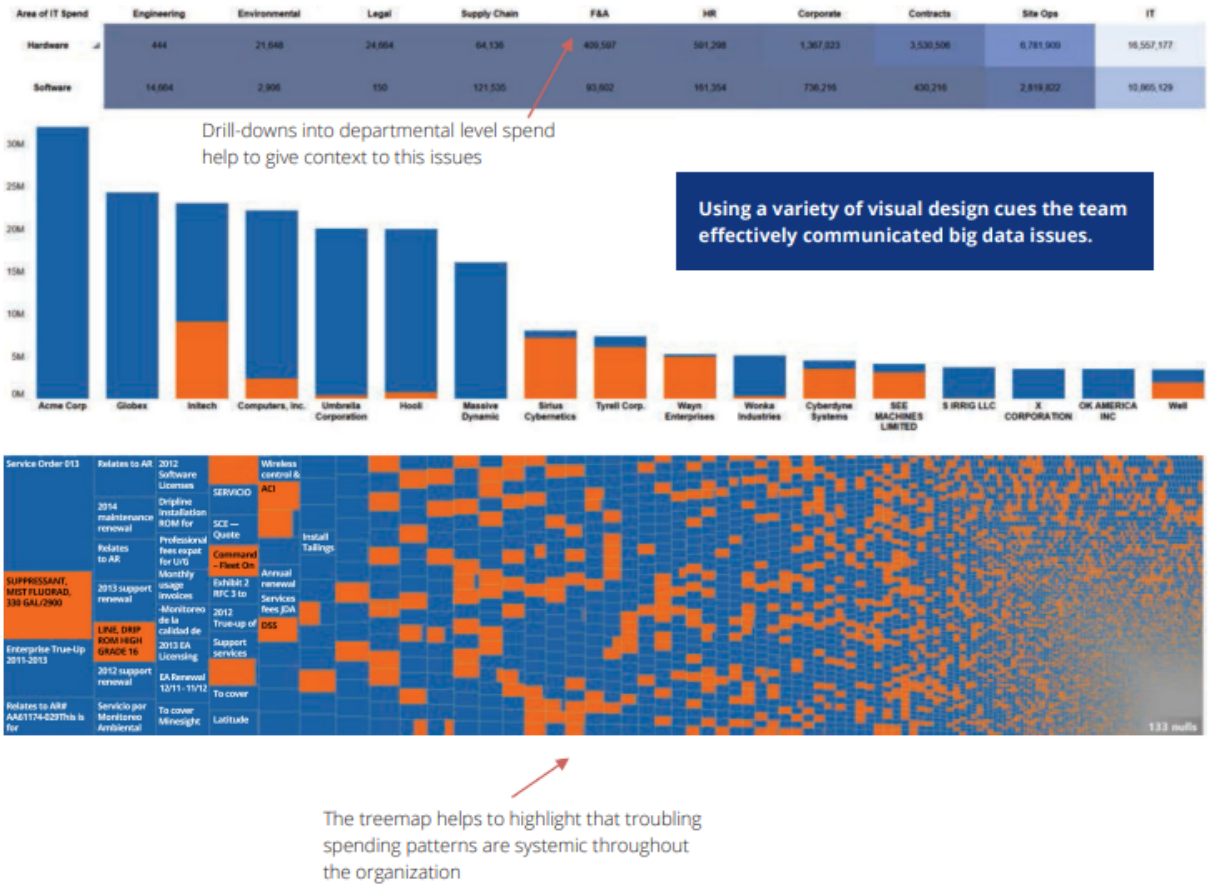
Jacob Wood (https://canvas.okstate.edu/courses/118118/users/214790)

Apr 26, 2022

Hello all,

As a CPA with a background in internal audit (IA), I've utilized descriptive analytics extensively to identify business process issues. I also recognize the value it can bring to organizations. If anyone is interested in data/business analyst type work, there is a lot of carryover from the IA world. I found a whitepaper from Deloitte which outlines best practices for embedding analytics into an audit group below. The whitepaper also includes high-level overview of a project to analyze IT overspend within an organization. This was a powerful example, that builds on the descriptive analytics we learned in this course. The IA team was able to visually show the departments which were going around established procurement channels with a tree map, and accurately project the added cost from not leveraging IT's relationships with large vendors. This is a great example of descriptive analytics delivering impactful visuals to communicate further opportunities for cost containment.

Figure 6: Delivering the message through visual cues



https://www2.deloitte.com/content/dam/Deloitte/us/Documents/risk/us-risk-internal-audit-analytics-pov.pdf. (https://www2.deloitte.com/content/dam/Deloitte/us/Documents/risk/us-risk-internal-audit-analytics-pov.pdf)

(http

Paul Davis (https://canvas.okstate.edu/courses/118118/users/194957)

May 1, 2022

Jacob - this is a great example. I love the tree map and how visually compelling it is.

(https:/

Nithya Satheneni (https://canvas.okstate.edu/courses/118118/users/187996)

Apr 27, 2022

Hello Class,

I discovered a wonderful article that explains neural networks and how they are referred to as "deep learning," revitalizing a concept that was first proposed 70 years ago. If you're interested in learning more about the history of this fantastic method, please take a look at here.

https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414 (https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414)

Thanks,

Nithya Satheneni

(http **Rudrakumar Ankaiyan (https://canvas.okstate.edu/courses/118118/users/190716)**

Apr 29, 2022

Hi Nithya,

Thanks for the article on neural networks. I read the article and I got to know the history of how the neural network idea was formed. A neural net, which is roughly designed after the human brain, is made up of hundreds or even millions of basic processing nodes that are tightly linked, as indicated in the article. The majority of today's neural networks are arranged into layers of nodes and are "feed-forward," meaning that input only flows in one direction through them. A single node may be linked to numerous nodes in the layer below it from which it receives data, as well as several nodes in the layer above it from which it transmits data.

Thank you,

Rudrakumar Ankaiyan

(https:/ **Paul Dreyer (https://canvas.okstate.edu/courses/118118/users/183234)**

Apr 27, 2022

Hey everyone,

I saw this and thought it might be helpful for anyone that plans on continuing to use KNIME

https://registration.events.knime.com/events/knime-spring-data-talks-2022/registration?utm_source=website&utm_medium=text&utm_term=listing&utm_content=event&utm_campaign=Data-Talks

(http

Ben Lewis (https://canvas.okstate.edu/courses/118118/users/211833)

Apr 27, 2022

Thanks for sharing Paul! Looks cool. I went to the KNIME meetup earlier in the semester and had a positive experience. It looks like some corporations are going to be there, I'm curious to hear how KNIME would be used in a work setting with a data science team.

(http

Chitra Boorla Boorla (https://canvas.okstate.edu/courses/118118/users/193742)

Apr 28, 2022

Thanks for sharing it here Paul. At first i dint find knime that interesting but after working with it for couple of assignments and for our term project, i believe it does a very good and user friendly job at predicting data. And honestly it quite fun to play around with different model's.

(http

Ashish Kumar Pampana (https://canvas.okstate.edu/courses/118118/users/24369)

Apr 30, 2022

Thank you Paul for sharing the link. Those KNIME events are very useful if you want to keep up with the data science world. I hope to find some of my friends from this class to be there in the event!

(https:/

Rithik Ponugoti (https://canvas.okstate.edu/courses/118118/users/189064)

Apr 27, 2022

Hello Class,

As we use Decision Trees heavily, we often think they are ideal models to use. However, here are a few disadvantages and possible solutions to address

Decision-tree learners can create over-complex trees that do not generalize the data well. This is called overfitting.

Mechanisms such as pruning, setting the minimum number of samples required at a leaf node, or setting the maximum depth of the tree are necessary to avoid this problem.

Decision trees can be unstable because small variations in the data might result in a completely different tree being generated.

This problem is mitigated by using decision trees within an ensemble.

Predictions of decision trees are neither smooth nor continuous, but piecewise constant approximations. Therefore, they are not good at extrapolation.

If you want to dive deep into decision trees, feel free to dive into the following link

https://scikit-learn.org/stable/modules/tree.html (https://scikit-learn.org/stable/modules/tree.html)

Best Regards,

Rithik Sai Ponugoti

Neeraj Kankani (https://canvas.okstate.edu/courses/118118/users/190006)

Apr 27, 2022

Thank you for sharing information on a tree-based algorithm, Rithik!

This is very timely since tree-based algorithms are extensively used for classification tasks like that of our final project. Also, they work reasonably well on categorical data making the model very easy to interpret because of the creation of split rules. Making them a very handy tool in your classification toolbox!

Neeraj Kankani (https://canvas.okstate.edu/courses/118118/users/190006)

Apr 27, 2022

Good evening!

We all have utilized feature selection techniques in some way or form in our **final project** to reduce the dimensionality of the dataset. Although feature selection is a very important sub-section of the data science process to tackle any business problem, it all usually starts with a knowledge-rich dataset. A dataset that is full of features that are related to our business problem makes our job as data scientists a lot easier. But then again like Professor mentioned in class if it is easy then essentially there is no job security!

So, waiting for someone to make a feature-rich dataset to solve your business problem is like asking for someone else to do the work for you. To navigate that, data scientists use feature generation techniques to make sense of the abundant unstructured data, to solve their business problems.

I've been fascinated by feature generation since the beginning of my data science journey, and in fact, I've used it extensively in generating features from marketing log data (about 40 million rows) from one of the organizations I was fortunate to work for. I've used a combination of SQL grouping and aggregations along with a sense of business understanding to create columns like the most recent visit (to account for recency), the number of visits on the platform per week (to account for the frequency of visits), retention time (as a measure of time spent/monetary gain),

and many more to understand the likelihood of a customer to request a demo on the platform. This real-time analysis can help in identifying better advertising placement strategies, effective marketing campaigns, and a better customer conversion rate.

I've found a couple of resources that helped me get a sense of feature generation via conventional but unstructured data. I will link those at the end of this thread.

Please feel free to reach out if you would like to converse more about feature generation!

A couple of resources on feature generation to get started -

https://www.explorium.ai/resource/feature-generation-the-next-frontier-of-data-science/ (https://www.explorium.ai/resource/feature-generation-the-next-frontier-of-data-science/)

https://turintech.ai/insights/feature-generation-what-it-is-and-how-to-do-it/ (https://turintech.ai/insights/feature-generation-what-it-is-and-how-to-do-it/)

Thanks for reading!

Neeraj Kankani

Seonwoo Ko (https://canvas.okstate.edu/courses/118118/users/194258)

May 1, 2022

Hello Neeraj,

Thank you for your great resources! :) I agree that if is easy, everyone can utilize dataset easily and we won't need a data scientist. Your current work looks really interesting! I have never used SQL grouping but I think it is beneficial to manage databases and perform diverse businesses! I have never heard about feature generation before, and your articles looked really interesting that it can improve accuracy and perform new features from existing features. Thank you again for sharing interesting experience and knowledge!

Thanks,

Seonwoo

Pranjali Pingale (https://canvas.okstate.edu/courses/118118/users/190864)

Apr 27, 2022

Hi everyone!

For the past couple of weeks, I've had the pleasure to learn a lot via this discussion board. Therefore, I just wanted to pop in and share an interesting read with you all!

Although neural networks have a deep history, coming from perceptron in the 70s. Yet they are the most widely used algorithm to map non-linearity in the dataset.

It's very fascination that a technology that is has been working so efficiently in our body since the beginning of human evolution (aka brain) can be the inspiration of one of the most robust machine learning algorithm.

Nature really is beautiful and a source of inspiration for many!

I came across this paper on animal migration optimization based on the movement of swarm of bees. Feel free to give it a read when you can!

https://www.researchgate.net/publication/273187372_An_Improved_Animal_Migration_Optimization_Algorithm_for_Clustering_Analysis (https://www.researchgate.net/publication/273187372_An_Improved_Animal_Migration_Optimization_Algorithm_for_Clustering_Analysis)

Thanks!

Pranjali Pingale

(http

Jacob Wood (https://canvas.okstate.edu/courses/118118/users/214790)

Apr 28, 2022

Hi all,

I found an article which is viewing Machine Learning through a slightly different lens. Throughout the semester I've spent a fair bit of time thinking about machine learning, it's implications, and potential downside risks in some scenarios. Harvard Business Review wrote a great article covering some risks associated with Machine Learning from the perspective of senior leadership. It's really interesting to think through what might happen in an organization when an algorithm performs poorly or unethically, in particular with unsupervised learning. This article explores that a bit as well as other considerations, and gives some strategies for managing ML.

https://hbr.org/2021/01/when-machine-learning-goes-off-the-rails (https://hbr.org/2021/01/when-machine-learning-goes-off-the-rails)

Enjoy!

(http

Seonwoo Ko (https://canvas.okstate.edu/courses/118118/users/194258)

May 1, 2022

Hello Jacob,

Thank you for sharing the article! This article was interesting that most people think and state machine learning is beneficial in our lives. However, this article shows some risky parts that we have to consider. I think that we really have to think about the ethical issues and challenges. I was surprised that facial-recognition algorithms still have some issues to identify and differentiate skin color, deeming unfair to a certain group. I also agree with the author that businesses should come up with some plans for certifying ML and make a regulations before launching the market. We cannot say which way is right or wrong, but we have to keep in mind and continuously follow how we can handle some negative possible chances. Thank you again for sharing the interesting article!

Thanks,

Seonwoo