|  |
| --- |
| Syracuse University ISchool |
| MS ADS Final Portfolio |
| |  | | --- | | Joseph R Maugeri IV  6-9-2023  <https://github.com/jrmaugeri/MS-ADS_Portfolio/tree/main> |   IST 782 Applied Data Science Portfolio |

Contents

[Introduction 2](#_Toc134987612)

[1. Collect, store, and access data by identifying and leveraging applicable technologies. 3](#_Toc134987613)

[2. Create actionable insight across a range of contexts (e.g. societal, business, political), using data and the full data science life cycle 5](#_Toc134987614)

[3. Apply visualization and predictive models to help generate actionable insight 7](#_Toc134987615)

[4. Use programming languages such as R and Python to support the generation of actionable insight 8](#_Toc134987616)

[5. Communicate insights gained via visualization and analytics to abroad range of audiences (including project sponsors and technical team leads 9](#_Toc134987617)

[6. Apply ethics in the development, use and evaluation of data and predictive models (e.g., fairness, bias, transparency, privacy) 10](#_Toc134987618)

[References 11](#_Toc134987619)

# Introduction

I always tried to bring my curiosity to the interaction with the curriculum and my fellow students. Through many unique models and repositories, I’ve mostly focused on scientific applications. I hope to apply my work to enhance the techniques of others and provide clear information accessible to all. I mostly pursued work that sought to improve or illuminate important real-world relationships that human behavior could play a big part in (or be affected by). I feel that I was able to enhance the development of my professional skills in this kind of challenge. When I began the program about 2 years ago, I was primarily exposed to many elements of data-science through my undergraduate biology-psychology dual degree. The work studying of life sciences and human nature, are followed naturally with data science as I had many questions about the differences in the two fields of underlying research that I studied. Not only that, but I desired to learn and be able to apply a variety of statistical methods to any subject matter of my choosing to generate more knowledge. After graduating and entering a job in the health-sciences sector, I was able to spend a year designing data entry systems in excel for the non-profit sector. I was incredibly curious and eager to expand my data-skill set to a wider variety of methods and contexts and felt that my skills with technology paired well with a career centered more around computer science research. To the present day, I have been pursuing much more than just research and modelling capabilities. Many questions that any business seeks to answer come far before data science or “Research” can truly begin. I hope to operationalize the answering of those questions in continuing my career in the technology sector. It’s evident everywhere that businesses and researchers across the globe have seen an exponential number of questions that must be answered, and there shouldn’t be distance between those with subject matter expertise, and competent technological and analytical services. At the time of writing this paper and near completing the program, many worry the advent of AI will come to replace their jobs. While the future is not set in stone, there will always be a need to provide novel unique input to the data-science process. I hope that I can share in the many benefits from this process with many people, and you will see that reflected in the subject matter of the projects I selected. I hope that someday through my work, I will be able to learn what I truly would like to research in the long term. I see all too often how exponentially more important that aspects beyond performing analysis (such as dissemination, accessibility, and reception) are key components for effective knowledge sharing. What I've ultimately gathered from a course is that a good data scientist will need to just not forget to ask questions. It’s also critical to effectively communicate insights and recommendations to their intended audience. Creativity enhances engagement and often the effectiveness in communicating, as well as memory. Many if not all pieces that have been included had some level of creative risk. This is a regular process that happens every day because analytics are a collaborative effort where novel ideas are welcome. What you start with may not always be what you ultimately produce, but what won’t be lost are the relationships built that enable true creativity. Hopefully, this selection of projects will highlight not only my capabilities with applied data sciences, but my love of sharing and expanding the sciences.

# 1. Collect, store, and access data by identifying and leveraging applicable technologies.

The management of data is an essential art required for providing business value to an enterprise. We are unable to truly deploy effective machine learning, optimization, or other complex methods without proper control and understanding of the data sources that feed their models. Despite many enterprises possessing different needs at different stages of developing a data informed culture, effective management and storage strategies are a key part of any data-scientists toolbox. It may be that your company has a well-defined business warehouse but is looking to store additional data from outside sources to provide more insight. That may require an analyst to design a system that can hold the information needed to answer the questions of the stakeholders. On the other hand, it may be that a company is still working through a mass of excel data sheets, and in need of a simple database for end users to collaborate effectively in. The manifestation of collection, storage, and access are all dependent on the use case at hand, which may call for different approaches from one analytical team to the next. However, one of the most important common elements that any business needs to consider when making data-based decisions are the governing policies that a company uses to standardize their interactions and responsibilities related to data.

1.a. IST 615 Cloud Management Final Project – S/4 HANA Cloud Migration Solution

<https://github.com/jrmaugeri/MS-ADS_Portfolio/tree/main/1%20-%20Collect_Store_Access/IST615_CloudChasers_FinalProject>

This class involved experience and exposure to modern cloud solution providers like Microsoft Azure, and Amazon S3, however there was an emphasis on applying the usage of these technologies to a specific business scenario, or multiple scenarios. Current events were frequently integrated to the curriculum as cloud management concepts and technologies are continuously evolving. Systems architectures must be able to evolve to match a rapidly changing landscape of requirements and threats. Diligent work is required to choose and apply a solution that meets the requirements of the current state and prepares to meet the foreseeable changes that will inevitably come.

Through the nature of the final project, I realized the fortune of having real-world overlap with my career and with the topic. My employer at this time of writing is in the early stages of developing a cloud migration strategy. This entity was a high-tech manufacturing company and was used as a referential basis (with altered names and figures) for the modeling of a cloud migration. Given that from 2020 to 2022, SAP saw a major shift in revenue to cloud technologies over their software, and that SAP is the system I work out of every day, it felt incredibly valuable to be able to dig into a real solution with my team as best as we could. The final project’s recommendations leveraged informed hardware estimates for an international manufacturing company, cloud migration tool demos, and TCO estimators to model a solution for migrating a on-premises hosted SAP ERP business warehouse to the SAP S/4 HANA cloud platform. There would be many other things to consider beyond the scope or accessibility of someone in my role, but we were able to capture several key issues that lend credit to a migration. The company's rapid growth and reliance on legacy systems have resulted in inefficiencies in their operational management. Issues include inconsistent data availability, limitations on the SAP development team, and increasing costs of hosting and management. In the context of the class to provide cloud technology as a solution to an individual business use-case, we recommended migrating the entire business warehouse to the SAP S/4 HANA Cloud. The migration would provide enhanced capabilities for resolving current and future issues while reducing ownership costs. The are more complexities that would truly have had to been resolved in the underlying data and governance policies prior to expecting a complete resolution from a new software. Additionally, not every component of the business resided within the ERP warehouse, and there were other policies that needed to be implemented around the management of data prior to this type of undertaking.

The potential resolutions, including improved performance for analysts, developers, and auditors through data consolidation and the utilization of AI, machine learning, and robotic process automation. This would help to close the gap between varying stakeholders and allow solutions to occur on a more real-time basis. This would be because of the improved communications around systems, and the improved experience of the data managers and consumers that elevates their communication from figuring out what they agree on to what actions they will take. Furthermore, the migration would decrease overall management costs, improve access to visualizations, and provide increased capabilities for SAP developers. There were already several advantages of choosing SAP S/4 HANA Cloud over other cloud providers, given the existing strong relationship with SAP and their commitment to optimizing the product for the cloud environment. The presentation was to go as far as a financial analysis of savings for the company but was limited in the information available to it. The proposed cloud-based solution design incorporates a hybrid cloud model, allowing for future expansion and integration of additional resources. The services we listed as compatible were SAP Cloud Integration for Data Services, SAP Cloud Portal Service, SAP Cloud Transportation Management, and SAP S/4 HANA Cloud for data storage, processing, and federation.

In conclusion, an urgent lens was put on the project as *CloudChase* Technology continuously innovates new products and technologies that require unique quality assurance and data storage methods. By pursuing a cloud migration solution to optimize its operational management, it would be prophetically fulfilling for a company bearing the name *CloudChase* Technology. This project was one of those that drew nearest to my current career’s subject matter and involved almost no handling of real data or application of statistical methods. That highlights the importance of data scientists’ knowledge of the competitive technologies and services out there and their capabilities pertaining to data. This project stretched beyond database organization and table structure but is focused on a fundamental component of a data scientists’ role. That is being able to identify your system’s capabilities and provide value added solutions and recommendations to their weaknesses.

# 2. Create actionable insight across a range of contexts (e.g., societal, business, political), using data and the full data science life cycle

2.a. IST 707 Final Project – Mushroom Classification using ML Algorithms

<https://github.com/jrmaugeri/MS-ADS_Portfolio/tree/main/2%20-%20Create_Actionable_Insight/IST707_Final_Project>

My love of biology was able to take center stage in this group project. The application of data science to novel scenarios may be more of a hobby than a career for me at this time. However, with adequate data on a particular subject of interest, there is a chance that you will be surprised by the results. In my personal life I enjoy walking in the woods and fields of Upstate NY. In the community I have lived in my whole life, there are still undiscovered truths. One of those being the finite classification of mushrooms into edible and poisonous categories. A question for me was that; Using algorithms, and if so, which ones are best suited for edibility classification. While this is of high interest to me, it may also be culturally significant. In the introduction, a history of mushrooms and their significance in human culture is given, highlighting the diversity and complexity of mushroom features, and their documented uses since prehistoric times. The intended effect was an overtone in the importance of the natural world as part of humanity, and how humans and ecosystems shape each other over time. Also, to highlight the importance of preserving human culture and the insight it’s held for long durations of time. It highlights the long-standing reverence for mushrooms in human culture, with evidence of their use dating back thousands of years. It is of high interest to me to apply my interest of humanity’s knowledge to my data science skills, to better contextualize questions that may lead to valuable answers for us at a societal level. Through this I see how a love of literature and research is a great strength to a data-scientist, even if it may not directly pertain to the computer. We must always be looking for creative ways to join information together to provide insight, and much of what I see in the world are datasets that are just waiting to be joined to reveal insight. The ability to dig into subject matter bravely and bring a new lens of analysis to topics I always had great interest in. The final project comprised of a presentation given and a formal write-up paper.

Features of mushrooms are complex and overlapping across species, making classification challenging, even for experts. A general introduction to features of mushrooms is provided along with Agaricaceae family images as examples from the dataset to illustrate the morphological variations. The “mushroom dataset” used for analysis contains information about 23 species of mushrooms and their edibility. Specimens came from the Agaricus and Lepiota families residing in the U.S., obtained from the National Audubon Society Field Guide to North American Mushrooms. The data was pre-processed in R and drew lines of analysis along various classification and data mining techniques. These included use of K-means algorithm clustering, Hierarchical clustering, Association Rule Mining, and Decision Tree Algorithm partitioning. Confusion matrices are used to evaluate the accuracy of the models.

Firstly, it was found that algorithms such as K-means and hierarchical clustering were not appropriate for the dataset. We did not have elements from the 21 features of available data that aligned with perfect categories one way or another. This makes sense, as the literature provided suggests that many species of mushrooms often appear like others in different stages of development. That is specifically why this dataset dealt with known mushroom species only of very closely related families. It is possible that in larger sampling and analytical efforts, mistaken identification occurs along the feature capturing process as the features of individual species of mushrooms are highly specific.

Secondly, I delve into the details of hierarchical clustering, association rule mining, and decision tree algorithms. They explain the steps involved in each technique and present the results obtained from the analysis. I then provide visual representations of the clustering results and association rules generated.

Thirdly, decision trees and random forest algorithms. They explain the process of training and testing the models. Decision tree and random forest algorithms are also applied to the dataset. The decision tree model is trained and tested using a split of the data, and the results are evaluated using a confusion matrix. The paper includes a visual representation of the decision tree. The random forest algorithm is used with the same training and test data, and the number of decision trees is adjusted to achieve high prediction accuracy.

Through iteratively testing and re-testing different algorithms in different scenarios, our team was able to demonstrate the effectiveness of one algorithm over another in application to the unique classification problem. Using mushrooms as an example provides an interesting metaphor to the work that one may find building a classification model for a business or organization. Aside from data formats and a binary result, there is no roadmap that dictates one model will be better than another, and this assignment demonstrates that some are in fact not. Those were however included in the final presentation to showcase the nature in which effectiveness of machine-learning algorithms is determined and communicated to an audience. Since the selected algorithms for presentation required different formats, a heavy amount of separation and preprocessing in R was required, which is where performing work in groups became an asset.

The paper not only provides a deep analysis of the machine-learning effectiveness, but also a contextual overview of the morphology of mushrooms and how that adds complexity to the issue. It may also explain why our results were very limited to find abundant support. It highlights the challenges in mushroom classification and the potential of data mining methods to gain insights into mushroom characteristics and edibility.

# 3. Apply visualization and predictive models to help generate actionable insight

IST 707 Homework 5 – Time Traveler Seeks Madison

<https://github.com/jrmaugeri/MS-ADS_Portfolio/tree/main/3%20-%20Apply_Vis_and_Pred_Models/IST707_Homework_5>

The paper discusses a fictional scenario where I take on the role of a Time Traveler from an alternate timeline in which the Federalist papers authorship became some mysterious critical component of modern society. Critical enough that it warranted the usage of time travel back to a time one week after the death of Alexander Hamilton when a majority of what are considered the “Federalist Papers” have been written. The report takes the form of a Newspaper Article, in which I was able to publicize the research and beckon Jefferson and Madison to come forth and verify the findings. Ignoring the potential time rift this causes, I seek to ensure that Hamilton's work is properly immortalized, and the authorship of the disputed papers is resolved, as it is crucial for preserving the United States.

While it may also be a far stretch to discuss Machine learning with 19th Century Americans, I hoped that the significance of a U.S. collapse would overshadow and also rally the locals to aid me in my cause. Through using natural language processing, a corpus was generated for each of the federalist papers, through which the speech and word usage were conformed into a format ideal for data processing. This involved removing stop words, truncating to the base form of a word, and omitting any other characters or text that does not indicate an author’s style. In running decision trees on training and test sets of known authored data, the model was then applied to the set of disputed federalist papers to predict authorship. However, the results are inconclusive which ultimately set the theme for a story in my mind.

Influences such as the dominance of certain authors in the dataset, and the lack of variation in the models, and the over reliance on specific elements in the texts are what drew these conclusions for me. The complexity of the dataset used in the papers led to far too simple an outcome. Their decision tree method always resulted in pure leaf nodes, indicating that our model was able to classify papers solely based on whether they were authored by one author or another, using only the presence of the word “Alexander”. In both cases, the model suggests that papers with a frequency of "alexander" above 50% are authored by Alexander Hamilton. This high frequency raises the possibility that Hamilton either referenced himself frequently or wrote his name extensively in his papers.

Examining the confusion matrices, we observe that our constructed model made no incorrect predictions. All disputed papers were identified as being written by "Hami," referring to Hamilton, or “Madi”. However, it is important to be cautious about trusting this high accuracy without further testing and verification. While the employed methods are relatively reliable, they may not be perfect for determining authorship. The fact that the word "alexander" is the sole determining factor in both instances raises concerns. If this holds true for the disputed papers as well, it becomes questionable why there would be any uncertainty about ownership when Hamilton seemingly included his name throughout his work.

Another feature of the data that supports this hypothesis the data is the imbalance in the number of authors within the papers. Hamilton's contribution accounted for approximately 60% of the entire body of federalist papers, which means his writing is more likely to match any individual paper simply because he has a larger pool of words to draw from compared to Madison. Additionally, there were papers co-authored or written by infrequent authors which had to be removed at the start of the analysis. While that may normalize effects, doing so compromises the complexity of the dataset which may have biased results to favor Hamilton. This instance appears to have experienced overfitting since there were no failures, and a single node served as the decision maker.

In conclusion, going back to 1804 was a rectifying solution for the problems the data faced, as we could collect more samples of each author’s work. While the work may not closely match the prose in the Federalist papers, it was one of the only solutions available to better understand the authors' styles and determine if word usage alone can establish authorship. The paper finishes with a final plea of delivering the information to James Madison and the Hamilton family, emphasizing the importance of their role in rebuilding the nation. This may draw parallels to some actions that can be taken in a similar scenario in modern day. If we are unable to trust the results of one model due to a limitation of data or special circumstances, then it may be time to go out and fetch more data to build and apply the model on.

# 4. Use programming languages such as R and Python to support the generation of actionable insight

IST 652 Final Project – Associating Elon Musk’s Tweets with Crypto Prices

<https://github.com/jrmaugeri/MS-ADS_Portfolio/tree/main/4%20-%20Use_R_and_Python/IST652_Final_Project>

This project focused on joining natural language processing performed in python with crypto currency and fiat currency price index data. The project obtained data for NLP through a twitter scraping API called snscrape, which is one of many APIs that can perform web-page based data mining. Snscrape is one of the most popular and best APIs because of its lightweight nature and inherent flexibility to query data. Snscrape stands for Social Networking Services, and we pointed our calling of the API to a specific user and dataset present in the twitter networking environment. We all know who Elon Musk is, for better or worse. Humanity has not seen many figures if any that can match the overall wealth, and broad range of topics that Musk has managed to insert into his life. There may never be another polymath quite like Musk, however it never is good when one person’s own decisions and preferences have massive and often unchecked influence on the rest of humanity at large. What I mean by this is that Musk, despite being a great physicist, is not always the greatest philanthropist. You will recall that back in 2020 during COVID and when cryptocurrencies were receiving increased attention, Musk felt so inclined to attempt at fingering the scale. The popular currency “Bitcoin” was skyrocketing in USD value; however, Musk took an oppositional stance to the currency based on the façade of a lacking in eco-friendly practices. We found out later that Musk had secretly been coordinating with the developers of DOGE a.k.a. DOGE coin and was attempting to use his own influence to increase its popularity, or at least detract from that of Bitcoins. Musk often uses his name and status on the twitter platform to push his own agenda. This project investigated possible associations between Musk’s twitter activities and specific word choices he used, compared against fiat and crypto price indexes for that day. It was hypothesized that during the period for which he showed the most interest in cryptocurrencies via twitter, that there would likely be increased volatility or even corollary movements between the sentiment of musk’s tweets towards crypto, and its ultimate price. This was ultimately not found to be the case for several reasons, and the project resolved that a more complicated model would need to be developed to investigate this further. This is an ever-important part of the analytical process because we are not destined to get everything right the first time. It’s important to recover from these moments and use them for what they are, more knowledge for the second attempt at the process. The general methods of the endeavor were executed as follows.

Twitter data for user @ElonMusk was accessed via TwitterScraper from snscrape and processed through use of the nlp package for python. All tweets from 2020 and 2021 were extracted and tokenized to investigate the prevalence of references by Musk to various cryptocurrencies such as ETH, BTC, and DOGE. In addition to that, yahoo finance was used to obtain market indexes of the same coins for the same period. It was conceived as valuable to obtain the stock index price of TWTR for the same period, but this was ultimately not used in the current analysis. A review of Musk’s most used tokens after reducing them to roots was that Musk not only discussed crypto, but also several other topics which muddied the waters for analysis. He spoke often of tesla, SpaceX and memes, but ordering the tokens by frequency found that DOGE was only used 50 times during this period, and Bitcoin only used 27 times. No formal model was developed on top of this data aside from investigating the nature of musk’s tweets and the market data which did not have any apparent relationships or even association to each other. In fact, there was only the inkling that popularity of Musk’s tweets was entirely conditional on luck, and Market data was entirely conditional on the market, which were really two separate things at least at up to this point in time. There were also not enough references by Musk to our variables of interest (such as ETH, BTC, or DOGE) that laid grounds for providing a significant study. However, it was interesting to find that Musk’s twitter activities are just like that of another twitter user. Not professional, and not terribly concerned with always communicating a clear message about one thing or another. Most of Musk’s original tweets were either out of touch memes or hype for his own pursuits with tesla/SpaceX.

# 5. Communicate insights gained via visualization and analytics to a broad range of audiences (including project sponsors and technical team leads)

IST 719 Information Visualization Final Project – Top 20 Hazardous Spills in New York

<https://github.com/jrmaugeri/MS-ADS_Portfolio/tree/main/5%20-%20Communicate_Insight/IST719_Final_Project>

The focus of this project was to generate a focus on the state of New York and the data that was available to it in terms of Hazardous Spills. The dataset obtained from New York’s Open data platform was a incident level history of hazardous material spills from 1905 to present day. Through using R code and packages like ggplot, lattice, dplyr, and more, I perform various transformations and data manipulations to prepare and present the data for a poster presentation to be viewed and interpreted graphically. One of the often-forgotten aspects of analytics is making the results accessible and easy to understand to non-experts. Through clever techniques and following artistically based rules such as the Golden Rule, the poster generated provides an interesting look at what varieties of material have been spilt and where. The data was however limited to the past 10 years for feasibility, but that likely makes the results and analysis more relevant to a likely audience member. When we are handed large datasets like this and given a question, oftentimes we do not need all the data, and the answers to the question lay in a small portion of it, for which our stakeholder has a question on. Discovering that source of information and presenting it in a salient way with respect to the rest of the data can answer many questions at a time, and bring the stakeholder right to the key pieces of information that they seek. Within the context of hazardous spills, we are often looking to separate the catastrophic events from the average everyday spills that occur. This is because they may affect visualizations in a stark way that can throw our viewers off the insights they need. This is especially important in the context of hazardous spills and attempting to triage resources to an area that needs more infrastructure or personnel support. A one-off in one area may offset critical funds that could be better used in an area that has more frequent spills. This was discovered in the forthcoming poster presentation.

The R code file that generated all graphics in the poster stores the data in a data frame called "spills”, transforming the date columns to the appropriate date format. Feature engineering was performed on the data to provide more information to audience members for questions they may ask, such as how long did it take to contain the spill? The data had information in the form of "Close.Date" and "Spill.Date" columns, which were cleaned and operated on to create the result column called "Diff.Days.O2C." A similar calculation for the "Received.Date" and "Spill.Date" columns, storing the result in the "Diff.Days.R2C" column was performed. Through visualizing the results of this data across other variables, there was not enough data present and no apparent way to visualize this information. Therefore, it was omitted from the final poster presentation. This practice happens often as it is incredibly important to know when and where to omit unnecessary information. While it may provide good context, you run the high risk of getting in the way of the key insights laying at the heart of the analysis.

Within the poster itself, the centerpiece is a large heatmap of NY state with quantity of material spilled, and highlights by numbered list that reference spill outliers that are orders of magnitude beyond entire county aggregates for the entire time. This cuts to the point of needing to present information about what it is. It would not be appropriate to showcase Onondaga county as higher on this heatmap than many others for the one spill that it received in 2005.

It creates separate data frames for spills with units of "Gallons," "Pounds," and those with missing units. In addition to this, supporting multi-dimensional viz were constructed and formatted such that they work well on a 30 X 36 poster. They were a scatter plot showing the five largest spills as outliers, scatter plot showing the top 20 spills since 2000, a box plot showing the distribution of spills by units, a line plot showing the quantity of spills over time by material family, and three bar charts showing the quantity of spills by source, separately for all units, gallons, and pounds.

The graphics generated by the code provide visual representations of the spill data, allowing for a better understanding of the distribution, trends, and outliers in the dataset. These visuals can help identify patterns, correlations, and potential areas for further analysis or investigation. The information presented in the context of separating outliers from the rest of the data was key for informing the right stakeholders of our takeaway message regarding Hazardous spills, and proper resource allocation for prevention, recovery, and maintenance.

# 

# 6. Apply ethics in the development, use and evaluation of data and predictive models (e.g., fairness, bias, transparency, privacy)

It’s necessary across all lines of work to consider the approach taken to any kind of findings or truths. Whether outside or internal methods is under question, determines the extent to which you can be informed and influence the ethics and practices around data. It’s also important to consider the effectiveness of data practices across the globe and the organizations and activities that they support. We can often turn our attention to social media or other web platforms and examine their behaviors and where they may have overstepped to provide clear examples of where care must be taken.

6.a. IST 623 Information Security Final Project

<https://github.com/jrmaugeri/MS-ADS_Portfolio/tree/main/6%20-%20Apply_Ethics/IST623_Final_Project/ROBLOX_Security>

This Final project was a group assignment in the form of a presentation, however numerous hours of teamwork and individual research went into the project to acquire documentation and articles that supported the timeline we painted of Roblox’s ethical usage of data. For context, Roblox is an international company that stemmed from a video game platform released by two physicists, Dave Baszucki and Erik Cassel, in 2006. At the time, the case of the ethics of a platform like Roblox was incredibly limited as there was hardly anything one could do aside from build virtual models using bricks and other components. Imagination ran wild on the platform in both its users and its creators, for which ethics can be sort of subjective and dependent on various factors. The historical overview of Roblox from 2006 to present reveals that Roblox was not always plagued with the same problems they face today, such as cybercrime and child exploitation. Over that time however, Roblox went from a simple physics engine to a content creation platform, social media network, and eventually a virtual item trading environment, in which the in-game currency ROBUX was listed on the NYSE. It became evident that greed by the owners and uncontrolled growth in popularity on the platform were the only available conclusions as to why Roblox evolved as it did. It was an incredibly sensitive topic after digging into it, with our team ultimately finding large areas of oversight by both Roblox and Legal Authorities that ultimately ended up hurting children. These areas included the following:

1. Inconsistent execution of security policies and procedures across different parts of the user base through third-party security vendors, ultimately putting the end user’s information at risk.
2. Unregulated behavior/content and injection of malicious content by users and/or bots targeting/affecting users under the age of 13.
3. Exposing young children to the risks of gambling via encouraging purchase of limited items that fluctuate in value on a global market, without providing guiderails to prevent loss or theft of those items.
4. Profiteering off content creators by compensating labor via in-game currency and penalizing the trade to fiat currency, as well as encouraging microtransactions in an unregulated environment.
5. Providing an unregulated gateway to access off-site via shared links, content-creators websites.
6. Lack of action / delayed action to address child exploitation on the platform.

Our group walked through several key points of evidence that pointed to how Roblox had been mostly pursuing a for-profit business model and was caught off-guard by the rapid popularity increase that occurred over the 2016-2020 period. Prior to this time, Roblox had been consistently introducing Here are the key points we delivered on assessing ethics of Roblox:

Safety and Protection: Roblox has made efforts to provide safety features, parental controls, and moderation tools to protect users, particularly children, from inappropriate content and interactions. However, there have been incidents where users have been exposed to harmful content and exploitative situations, for which Roblox has tried to deny liability and point to its privacy policy and terms of use for coverage.

Privacy and Data Handling: The handling of user data and privacy is an important ethical consideration. Roblox collects user data and uses it for various purposes, as outlined in its privacy policy. It is crucial for Roblox to ensure the security and responsible use of user data. There have been occurrences of user data being hacked from Roblox’s own infrastructure, and little safeguards are in place to protect more vulnerable users from being exploited for their accounts and items.

User-generated Content: Roblox is largely based on user-generated content, allowing users to create their own games and experiences. While this fosters creativity and innovation, it also raises concerns about the potential for inappropriate or harmful content being created and shared on the platform. While it is difficult to constantly monitor images and creations placed on the platform, it is possible to avoid marketing a real gateway to danger on the web to users under 13 if it cannot be secured properly.

Monetization and Economy: Roblox provides opportunities for developers to monetize their creations through in-game purchases and virtual currency (Robux). This can be seen as an ethical concern if it leads to exploitative practices or encourages excessive spending, particularly among young users. In addition, it is not an economy geared towards the success of end-users but rather the success of Roblox through use of end-user labor and content.

Community and Behavior: The behavior of users within the Roblox community is an important aspect to consider. Instances of cyberbullying, harassment, and inappropriate behavior can raise ethical concerns and emphasize the need for effective moderation and community management.

Cybercrime & Malware: The unique environment that Roblox is built on is a scripting engine that is responsive to code. Any user is given free access to the engine and the ability to generate scripts that will produce actions within a Roblox server environment, often with success. This has spawned an active community on both the surface and deep web to provide exploits, cheats, and malicious content, along with several imitators looking to steal the information of novice end-users.

Despite the many security warning signs, the call was not heeded by Roblox. A story was published by A. Belanger, 2022, that a young girl’s life had been thrown off by contact with adults sustained through the Roblox period over a two-years span of time starting from when the user was only 10 years old in 2020. Through her solicitation by an adult user, she eventually was hired for the Robux currency to perform work developing a Roblox game. The relationship she formed between the Adult grew and expanded and was taken off platform. The girl was manipulated and blackmailed into sharing explicit pictures of herself at this young age, which eventually took a toll on her mental health. The girl eventually attempted her own life, without success, but has had to withdraw from school and her dream aspirations of a government contracted job. Roblox has remained relatively silent on the issue since.

Ultimately, the ethics of Roblox depend on how effectively the platform addresses these concerns and protects its users. It is important for Roblox to prioritize user safety, privacy, and responsible content management to maintain an ethical environment. It is just as necessary for data scientists to consider the security and ethics of the systems as it is with cyber security professionals. The knowledge base and application of the two fields draw closer together as security and ethics continue to be an essential function of business. They must be at the heart of the considerations of a data scientist whose duty is to lead decision makers to the insights needed to garner best practices in the world.

# Conclusion

Based on the summaries provided, it is evident that I, as a data scientist, have undertaken various projects encompassing highly varied domains and subjects. All projects involved extensive research, data analysis, and on top of that the application of analytical techniques to derive insights and make informed decisions. It often takes much more than that and more than once for the insights provided to generate meaningful actions. The range of topics covered, including natural language processing, machine learning algorithms, social media platforms, modern technology solutions, and ethical considerations, highlight the breadth of my expertise as a data scientist.

Throughout these projects, I have demonstrated the ability to collect and process data from various sources, employ appropriate analytical tools and methodologies, and draw meaningful conclusions from the analysis. The findings and recommendations presented in my portfolio work represent a systematic and meticulous approach to understanding complex problems and generating valuable insights.

As a data scientist, I have showcased proficiency in utilizing programming languages such as Python, employing data mining techniques, conducting statistical analyses, and employing machine learning algorithms. The projects also emphasize the importance of critical thinking, problem-solving, and adapting research methodologies to suit the specific objectives and challenges at hand.

In conclusion, I fully believe the portfolio highlights my skills as a data scientist, and my skills as a researcher, neither of which are done alone or in a vacuum. The projects presented a range of mostly tangible challenges, requiring a combination of technical expertise, domain knowledge, and ethical considerations. As a data scientist, I am committed to leveraging data to drive informed decision-making, uncover patterns, and contribute to the advancement of knowledge and best practices for many fields and for many people who deserve to have their own hard work and passion, hold up to the rigorous standards that these technologies and methods call for us to uphold.

# References

All references are included in each of the subsequent projects references page, no additional references were used to compile this portfolio.

You can find all the resources referenced here:

<https://github.com/jrmaugeri/MS-ADS_Portfolio/tree/main>

.