OXFORD GENETICS

"The Evolution and Future of *Data Science* and *MLOps* in the Corporate World"

Harshini Kavuru[1], Lakshmi Manasa Medepalli[2] and Akansha Hastimal Jain[3]

[1]Kavuru.7, kavuru.7@osu.edu
[2]medepalli.5, medepalli.5@osu.edu
[3]jain.787, jain.787@osu.edu

Abstract

This data visualization project aims to explore and analyze data science and machine learning operations (MLOps) trends and patterns in the business world. This project presents a comprehensive overview of the evolution of data science and MLOps, including their development, adoption, and integration into various industries using the dataset available on Kaggle. By leveraging various data visualization techniques, this project offers insights into the current state of data science and MLOps in the corporate world and their potential future growth and impact. The dataset used is a survey conducted among individuals interested in machine learning and data science. Ultimately, this project offers a valuable resource for business leaders, data scientists, and other professionals seeking to stay up-to-date with the latest trends and developments in data science and MLOps and make data-informed decisions for their organizations.

Keywords: Data visualization, Data analysis, Machine learning, Data Science

Motivation

Machine learning and data-science have been around for quite sometime and their use-cases have only grown significantly in various industries and sectors of the internet era. Today, machine learning operates and automates various workforce parameters across industries. These industries span from healthcare, finance, retail, manufacturing, transportation, and agriculture. This project aims to understand and analyze the data collected as a part of a survey conducted to understand how today's crowd infers and uses machine learning techniques. This data is then used to understand how machine learning gets introduced into small and large-scale industries, what technical stacks are used by companies and how deep they dive into it.

Materials

This project aims to understand, explore, analyze and present interactive visualizations to cater to deeper insights into correlations between ML and real-world.

Data set

The dataset that we use is a survey conducted by Kaggle to understand how people around the world wrap their heads around ML and Data-science. We have gained a topical understanding of the data and what it represents. The data roughly contains 24000 tuples and around 296 columns. Each survey question's choices are enrolled as a column for better readability of data.

Variables and Information

A brief description of the features of the dataset is as below:

- The age, gender, and demographics group of survey participants
- The online education platforms they used to dig into data-science and ML (Udacity, edX, Coursera, Kaggle, LinkedIn learning etc.,)
- Their highest degree of education
- Their academic publications (papers, conferences etc.,) and if they used ML, Data-science in their work.
- The programming languages they are familiar with (C, C++, python, MATLAB etc.,) and the IDE (development environments) they use.
- The visualization libraries that they use to represent their work (Plotly, seaborn, matplotlib, D3 etc.,).
- The frameworks they are familiar with (scikit-learn, keras, PyTorch etc.,) and their years of familiarity with these libraries and frameworks.
- The ML algorithms they have utilized and understand (Regression, Decision Trees, Bayesian methods, CNN etc.,) These are just some of the features (questions) in the survey. We will explore the dataset even further as we dive deeper.

Tools and Technology

Visualization Tools

We will code and generate visualizations using python and its inbuilt libraries like Plotly, Seaborn and NumPy. We will then generate a python notebook (using either google Colab or VS-code as platform editors) to accumulate our visualizations with needed explanations for audience to get a good grasp of it.

Initial Questionnaire

- How widely has Machine Learning been adopted in the corporate sector today?

- What percentage of companies are currently implementing data science and machine learning in their production processes?
- Is there a correlation between a company's size or industry and their adoption of AI? Are larger companies more inclined to deploy AI compared to smaller ones?
- Is it essential for businesses to adopt cloud-native solutions today?
- Which tools are currently the most popular for Data Storage, Data Management, AutoML, Business Intelligence, and other related tasks?
- Which frameworks and libraries are prevalent in the market for Machine Learning and Data Science?
- Can transfer learning methods be considered mature enough for implementation in the business environment?
- Is the usage of specialized hardware for ML model training necessary due to the prevalence of big data and deep learning methods in the industry?

Roles and Responsibilities

- Dataset and technique selection – The entire team cumulatively searched and agreed on this dataset and understood the crux and functionality of the data accumulated. We also agreed to use python as a programming language to build clean data and viz. models.
- Paper Literature review – Each member of the team has religiously searched for papers relating to how well we can build the plots and annotated them to their maximum understanding.
- Task List - The team sat through to ideate what questions can be potentially asked from the data and how these can be used for further data exploration.
- Data pre-processing- The team has decided to use python to run scripts and libraries and filter unwanted data.
- EDA
- Creating Viz. – Each member has decided to create 2-3 visualizations to answer the questions that we initially developed. We will refine/ re-define the questions as we pace through the project.
- Documentation – The proposal document was a team effort as we bundled different parts of the proposal done by different members together in the document.

Individual work done

Deliverables

Our delivery will include a Python notebook containing comprehensive exploratory data analysis, data visualizations, and insightful interpretations derived from the visualizations.

Alpha

1) Data Analysis: To fully comprehend the dataset's variables, an in-depth examination of the dataset is essential. In order to do this, each dataset column and its inputs must be examined, and accurate statistics must be drawn from the data. We will be able to fully understand the variables in the dataset using such a method.

2) Simple visualizations: We will create several straightforward visualizations to help users understand the crucial connections between the dataset's elements. They comprise histograms, bar graphs, donut charts, and Spyder graphs. We hope to gain insightful information from these straightforward visualizations that will serve as the basis for creating more complex representations.

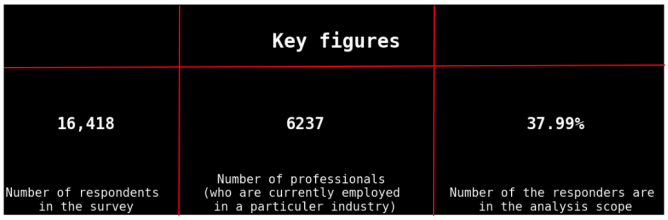

Figure 1 Output of Initial Screening of dataset.

3) More deep and complex visualizations: We realize that thoroughly comprehending the dataset is essential as we start our data analysis journey. To do this, we intend to create more complicated and complex visualizations that let us go deeper into the data. With the aid of our visualizations, we will be able to understand better the underlying data as well as spot trends and patterns within the dataset.

To do this, we will use various data visualization techniques, including bubble charts, treemaps, sankey graphs, and more sophisticated visualizations. These visualizations are designed to draw attention to the connections between the dataset's various columns, giving us important new information about the causes influencing the patterns in the data. Also, by looking at the data differently, we expect to spot any biases or poor data quality concerns that could influence our research, assuring the accuracy and dependability of our results.

We can spot any outliers or anomalies in the data using our sophisticated data visualizations. These discoveries might change how we see the dataset by exposing hidden patterns or trends that were previously unnoticed. By examining the data from many perspectives, we expect to find insightful information that will guide our decision-making.

We want to gain a clear and thorough grasp of the dataset's structure through our attempts at data visualization. Furthermore, we aim to produce aesthetically interesting and instructive visualizations that enable us to get valuable insights from the data. In the end, our efforts in data analysis and visualization will assist us in reaching conclusions that are well-informed and based on a thorough comprehension of the underlying facts.

Beta

We want to spend much time and energy perfecting our graphics as we get closer to the beta release. We want to produce a final product that is visually beautiful and instructive and successfully communicates the key features of the dataset. To do this, we will thoroughly examine our current visualizations to improve their readability, accuracy, and clarity.

We will produce a final visualization that encapsulates the core of the dataset when we have improved our first representations. This visualization will be a Heatmap and Spyder chart that uses all of our analytical tools and offers a thorough investigation of the subtleties of the dataset. Using this graph, we want to demonstrate the intricate correlations between the dataset's many variables and convey our findings engagingly and compellingly.

We will talk about the conclusions we came to from our investigation in addition to the visuals. We will examine the patterns

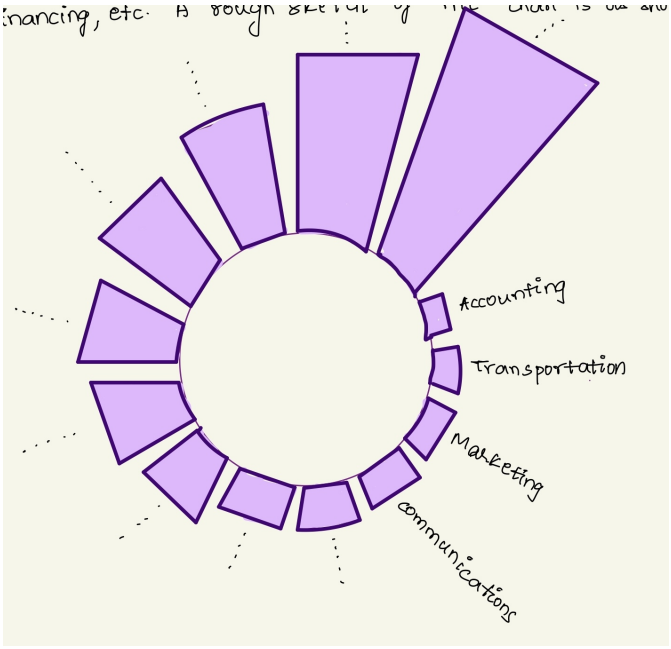

Figure 2 Output of Initial Screening of dataset.

and trends in the data and pinpoint the most important lessons to help us make wise decisions. These insights will guide our decision on the best visualizations and outputs, ensuring that we produce a final product that is both educational and aesthetically pleasing.

Final

In the final stage we will go through all our visualizations and discuss among ourselves regarding the insights that we gained from them. We will brain strom all the possible useful information from the graphs and produce them as final results.

Evaluating the usefulness:

An evaluation of a data visualization project's usefulness includes determining if the project's goals have been met and whether the visualizations have improved the data analysis procedure. Here are some key considerations when evaluating the effectiveness of a data visualization project:

1) Alignment with objectives:

Assessing whether a data visualization project is aligned with the overall goals of the data analysis endeavor is the first step in evaluating the project. For instance, if the goal is to spot trends in sales data, the visualizations should emphasize these trends.

2) Quality of visualizations:

Another crucial aspect to consider is the visuals' caliber. The visuals have to be understandable, precise, and straightforward. In addition, the desired audience should be reached by them successfully.

3) Insights gained:

A data visualization project aims to find patterns and insights in the data. It is crucial to evaluate whether the visualizations have offered insightful information that was impossible to learn from other sources.

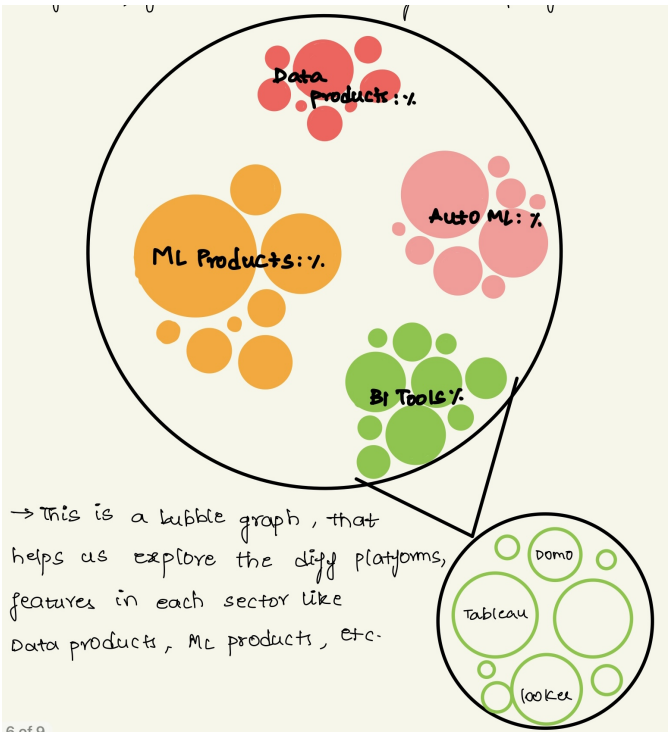

Figure 3 Rough Sketch of Bubble chart.

4) Actionable recommendations:

A practical data visualization project should offer takeable advice that can direct decision-making. Stakeholders should be able to make wise decisions using the insights provided by the visualizations.

5) Impact on decision-making:

Finally, a data visualization project's influence on decision-making processes may be used to gauge its efficacy. The project might be successful if the visualizations have produced well-informed decisions that have improved business outcomes.

Annotated Bibiliography

PAPER 1

This research proposes an artificial intelligence method for diagnosing cardiac disease based on machine learning algorithms. The authors evaluated the proposed model using several machine learning (ML) techniques, including Random Forest, K-Nearest Neighbors, Support Vector Machines, and Naive Bayes, on a dataset from the UCI Machine Learning Repository. The study found that the proposed model outperformed the individual ML algorithms at detecting heart disease. The authors emphasize the significance of early discovery because cardiac disease is one of the leading sources of mortality in the world. Their proposed strategy could aid in accurate diagnosis and quick treatment, potentially improving patient outcomes. The methodical strategy taken by the writers and the use of numerous machine learning techniques provide a trustworthy evaluation of the proposed model.

PAPER 4

The study's authors used a variety of visualizations to gather information and create better precision. Age, chest pain, blood

pressure, and cholesterol are some of the variables that have a significant connection with heart disease, and a heatmap shows the link between the variables in the dataset. The accuracy of different machine learning methods used on the dataset is contrasted in a bar graph. The authors state in their conclusion that the model's precision could be improved and that it could be applied to actual circumstances. This paper is relevant to the disciplines of healthcare and machine learning because it presents a positive solution to the problem of cardiac disease detection using AI models.

PAPER 2

A new instrument to assist basketball coaches and researchers in decision-making processes is introduced in the article "BKViz: A Basketball Visual Analysis Tool" by Antonio et al. The application offers interactive visualizations, such as scatterplots, heatmaps, and shot charts, to examine and evaluate various facets of basketball contests using data visualization, machine learning, and analytics methods.

Users can compare the data of various individuals or teams over the course of several games thanks to person and team performance displays. Users can compare shooting percentages, rebounds, and other performance measures using the tool's live graphics, which include bar charts and heatmaps. Users can follow the positions of specific players on the arena using trajectory visualizations, which shed light on player motions. In order to evaluate player motions during particular game scenarios, such as quick breakouts or pick-and-rolls, the application provides 2D and 3D trajectory images.

By performing a user study with coaches and analysts, the authors assessed the tool's efficacy and discovered that it was useful for evaluating player performance and finding tactics to boost team performance. The study, which shows the potential for cutting-edge methods to enhance performance analysis in sports, makes a major addition to the fields of sports analytics and visual analytics.

PAPER 3

Mahipal Jadeja's article "Tree-map: A Visualization Tool for Large Data" introduces a cutting-edge visualization tool made to assist users in visualizing and analyzing large datasets. The tool offers a tree-map structure with a hierarchical view of the data, where each node reflects a part of the data and its height is inversely proportionate to its weight or significance. The Enron email dataset and the KDD Cup 1999 network attack dataset were two of the big datasets used to evaluate the tool that the author created using Java and JavaFX. Users can examine the data and spot patterns and trends using the tool's dynamic features, which include zooming and moving. The article emphasizes the value of visualization in managing big datasets and the potential of such tools to increase performance and decision-making. The author also talks about the tool's drawbacks, like the need for more study on how well the tree-map structure works for displaying big databases. In conclusion, the article offers a thorough graphical tool that can help users explore and analyze sizable datasets. The tree-map arrangement employed by the author offers a distinctive and useful method for displaying hierarchical data. The article has significance for data visualization because it emphasizes how visualization tools can improve data research and decision-making.

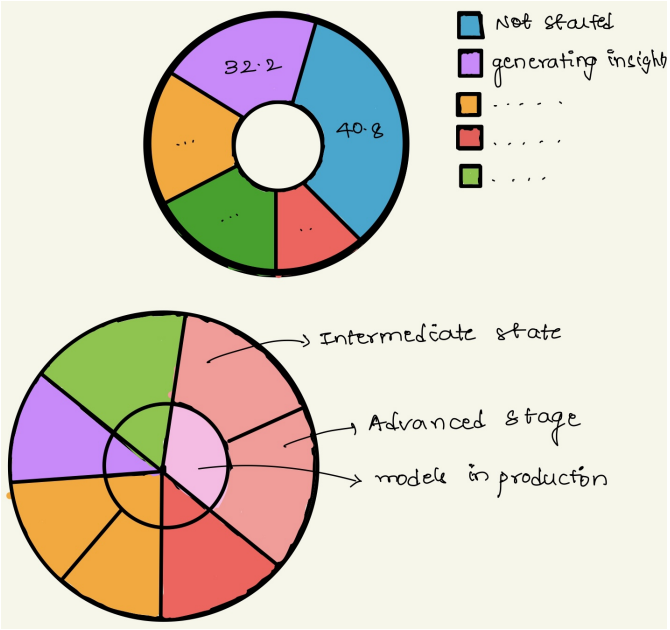

Figure 4 Rough Sketch of Donut-chart.

PAPER 4

An overview of modeling methods used to forecast the development of new technologies and the market penetration of electric cars is provided in this piece. (EVs). The writers go over the various aspects of market penetration for EVs, such as technical developments, governmental regulations, customer preferences, and infrastructure growth. They look at various modeling approaches, such as agent-based models, system dynamics models, and economic models, that are used to forecast the uptake of EVs.

The writers emphasize the difficulties still faced in creating efficient modeling tools while also providing instances of successful EV modeling projects. They make suggestions for future research directions, such as the creation of combined models that take into account numerous variables and the application of machine learning strategies to increase model precision. The research also examines how EV market penetration is impacted by charging facilities, battery expense, and range anxiety. The authors present the market share of EVs in various scenarios using spider charts, and they show the sensitivity analysis of various variables influencing the EV market using bubble charts. The report ends with suggestions for stakeholders and lawmakers on how to encourage the adoption of EVs in the future.

Using bubble charts to show the sensitivity analysis of various factors affecting the EV market and spider charts to visualize the market share of EVs in various scenarios, the authors analyze the impact of charging infrastructure, battery cost, and range anxiety on the market penetration of EVs. For researchers and decision-makers who are interested in the adoption of electric cars and the creation of sustainable transportation networks, this paper offers a useful resource.

Initial Screening

We want to make sure that the dataset for this project is as representative of the professional population in the target sector as feasible. In order to do this, we've chosen to filter the dataset

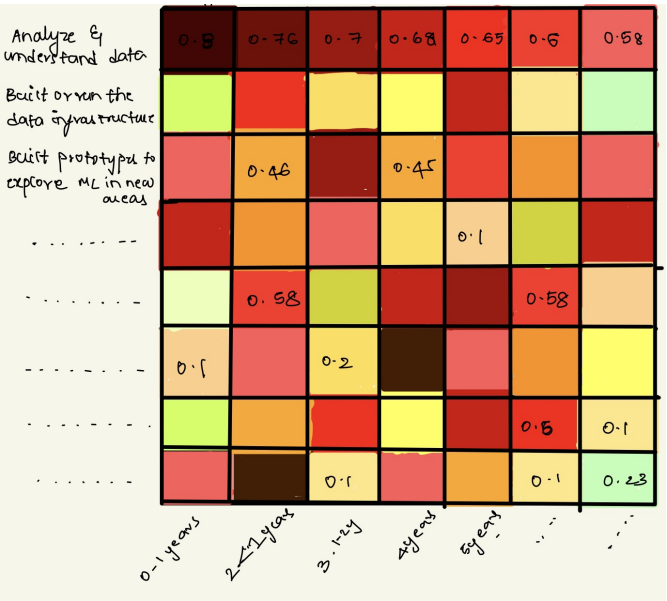

Figure 5 Rough Sketch of Heat-map.

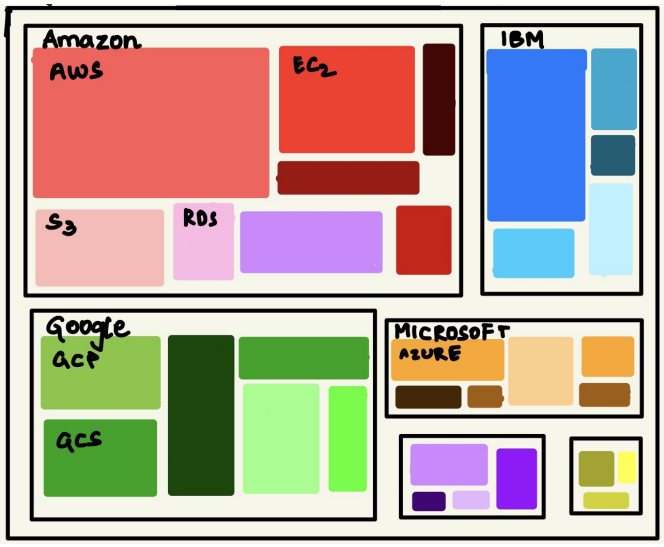

Figure 6 Rough Sketch of Tree-map.

to only include respondents who satisfy certain requirements pertaining to their job and educational status.

First off, we want to exclude out those responders who are presently enrolled in school since they can have different requirements, life experiences, and viewpoints than those who are employed in the field. As a result, we will only count respondents who said "No" when asked if they were presently enrolled in school. (Q5).

Secondly, we should only concentrate on professionals who are actively employed because our study is most likely to benefit from their experiences and perspectives. To do this, we will exclude any respondents who choose "Currently not employed" as their answer to the inquiry, "What is your employment status?" (Q23).

Last but not least, we want to make sure that our dataset only contains respondents who truthfully answered the inquiry, "What industry do you now work in?" (Q24). By doing this, we can make sure that the variety of sectors and job positions that make up our target profession are fully and accurately represented. Figure 1 represents the results of the initial screening of the data set. From the figure we can observe that the total number of people who filled the survey are 16,418. the total number of professionals are 9094. Therefore, we can say that 37.99 percent of people who filled the survey are professionals.

We will be able to improve our research and get more insightful knowledge about the viewpoints and experiences of experts in your chosen sector by applying these criteria to the data set.

Rough Sketches

Bar-graph

In the first visualization we want to analyze and explore the sate of Machine Learning and Data Science in various sectors or industries like education, Technology, Financing, etc.

A rough sketch of the bar graph is shown in 2

Donut-chart

In this visualization we want to take a look at the state of Machine Learning usage or adoption in different companies.

Some have models in production, while some didn't start yet , some are not known, and some companies are still exploring.

A rough sketch of the donut chart is shown in 4

Bubble Chart

We would like to create a graph that helps us understand the usage patterns of all the techniques, platforms, frameworks, used by professionals.

The rough sketch of how the bubble chart will look like is shown in 3

Each circle in the 3 represents different products like ML products, Data products, BI tools, etc, which again have divisions in them like Tableau, Looker, etc.

Tree-map

In this visualization we want to take into consideration only the giant Tech companies like Amazon, Google, etc, and understand their top products.

A rough sketch of the tree map is shown in 6

We can observe from the sketch that each branch belongs a tech company like Amazon, IBM, etc. Each branch has leaves that show the various products of these companies and the size of the squares represents the popularity of these products.

Heat-map

Being students we wanted to explore the correlation between the different Machine Learning responsibilities and the years of experience.

For this we though that the heat map would work the best.

Arough sketch of the heat map is shown in 5

Sankey graph

In this section we would like to know what tasks a person in some of the job roles perform.

Job roles like Data Analyst, Data Scientist, Machine Learning Engineer, Research Scientists.

Some of the tasks performed are "analyze and understand data", "Build and run data infrastructure" , "Create ML to explore new areas."

The rough sketch of the sankey graph is as shown in 7

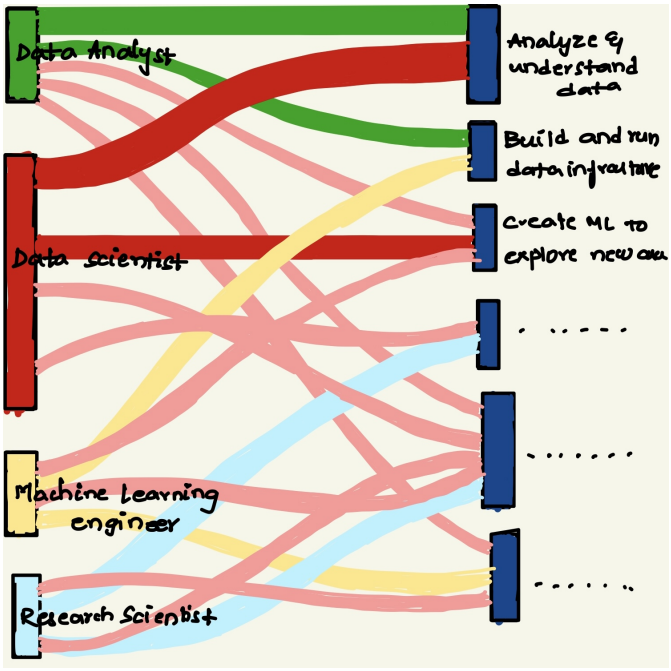

Figure 7 Rough Sketch of Heat-map.

Figure 8 Rough Sketch of Heat-map.

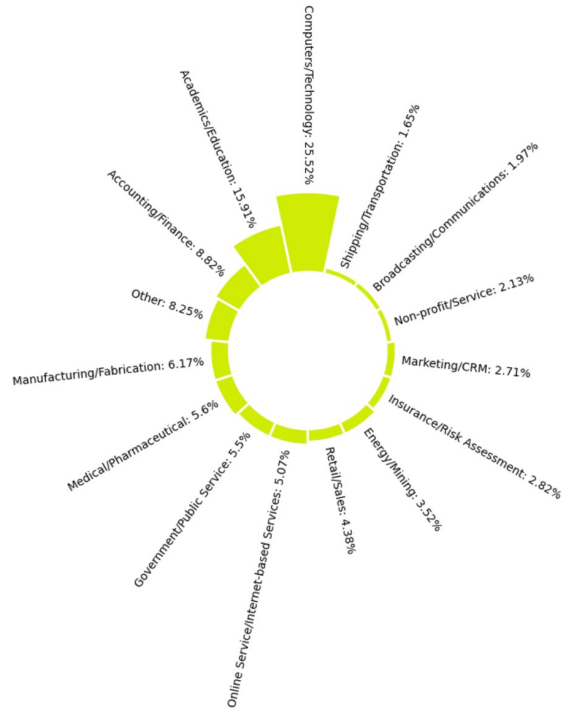

Figure 9 Circular Bar-graph.

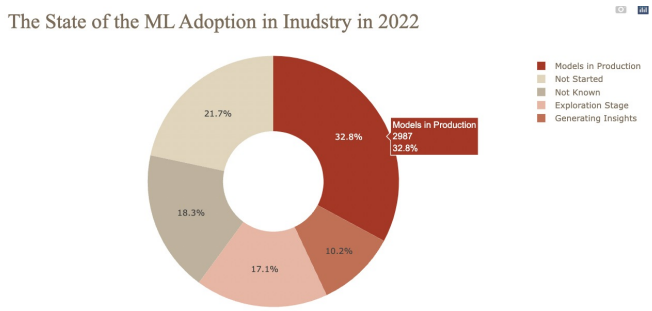

Figure 10 Donut-graph

The output circular bar-graph is shown in 10

Visualizations

Circular Bar-graph

The study findings indicate that ML usage is spreading across a variety of sectors. Tech businesses are leading the way in integrating ML into their business procedures, as seen by the fact that 25 percent of responders originated from them. Given that IT businesses have long been at the forefront of technical innovation, this is not unexpected.

It's also important to note that about 16 percent of responders were academics. This is most likely a result of the academic community's increased interest in data science and machine learning (ML) as well as the expanding significance of these subjects across a range of research topics.

The findings of the survey also show the variety of sectors using ML. This is significant since it shows that ML is likely to become more widely used and utilized. For instance, the financial sector has been adopting machine learning (ML) to spot fraud, while transportation industries are employing it to streamline their supply chains and boost consumer satisfaction.

Overall, the survey's findings show that ML is no longer restricted to a single business or branch of academia. Instead, it is evolving into a technology that is widely used by a variety of businesses.

Donut-graph

The data indicates that around one-third of the survey participants reported having Machine Learning models in production, either in an advanced or intermediate stage of development, and approximately 10.2 percent reported using ML techniques for generating insights. However, a significant portion of the respondents, about 21.7

The output Donut graph is shown in 10 and 11

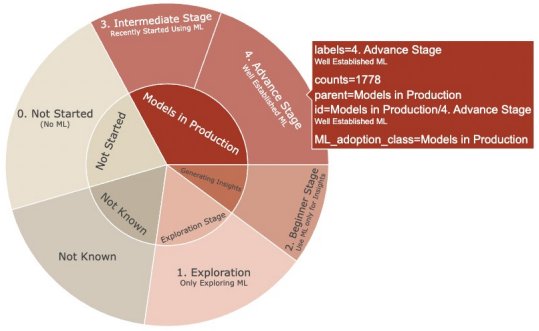

Figure 11 Donut-graph

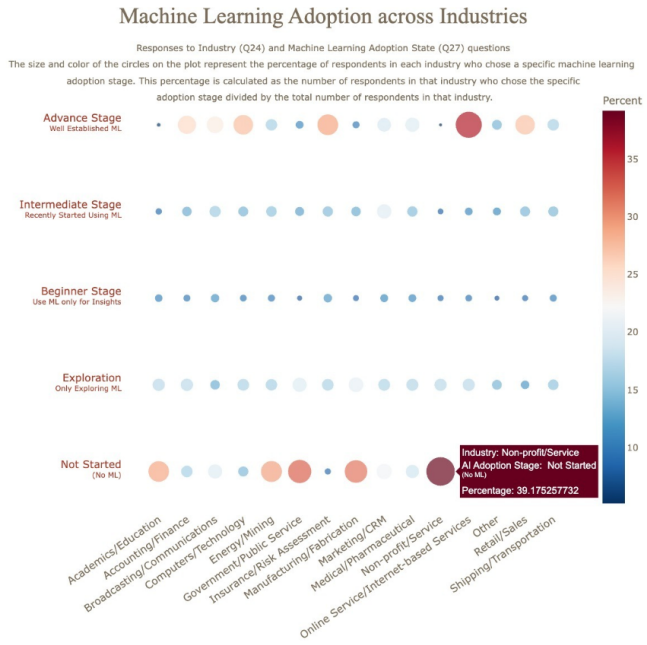

Figure 12 Scatter-plot

Scatter plot

Based on the chart, it's evident that internet-based service providers have the highest adoption rates of Machine Learning and Data Science, followed by insurance companies. However, non-profit organizations and government sectors have significantly lower rates of adoption due to bureaucracy and established processes that take longer to change. Employees in these sectors may also be less inclined to take risks and innovate. In contrast, private sector employers tend to focus on experimentation, innovation, and growth, such as gathering data from user online activities and applying analytics to improve their services. The insurance sector uses AI for various purposes, including insurance advice, underwriting claims processing, fraud prevention, risk management, and direct marketing. Retail is also using machine learning methods, with 27 percent of professionals in the industry reporting that their companies have implemented these methods. AI technologies can create value, reduce costs, increase efficiency, and enhance customer satisfaction and trust.

The output scatter plot is shown in 12

Figure 13 Bubble-chart.

Bubble plot

These graphs provide an overview of how professionals utilize various tools, techniques, databases, platforms, and frameworks.

The output bubble plot is shown in 13 and 14

Tree-map

Companies have their own unique technology stack with preferred software to use with their proprietary data, including different platforms for Visualization Analytics, Computation, Storage Distribution Data Warehouses. While there are numerous platforms available, this illustration focuses on popular cloud computing services and products offered by the top four giant tech companies: Amazon, Google, Microsoft, and IBM.

In terms of Amazon, the most widely used product is the Amazon Web Services (AWS) cloud computing platform, followed by Amazon Simple Storage Service (S3). For Google, Google Cloud Platform (GCP) is the most popular, with Google Cloud Compute Engine and Google Cloud Storage following close behind. As for Microsoft, the top products in the market according to survey respondents are Microsoft Power BI and Microsoft Azure, which rank third among the top cloud computing platforms. From IBM products, the most popular ones identified in the survey are IBM Watson Studio and IBM Cloud/Red Hat.

The output Tree-map is shown in 8

Sankey-graph

The above analysis shows that data analysis and machine learning are the most dominant activities in the roles of Data Scientists, Data Analysts, Machine Learning Engineers, and Research Scientists. It is also interesting to note that experimentation and iteration to improve existing ML models and building ML services are also important tasks, mainly performed by Machine Learning Engineers. Additionally, research to advance the state of the art of machine learning is undertaken mostly by Research Scientists. Building and running data infrastructure is an activity that all four roles contribute to almost equally, but it is a

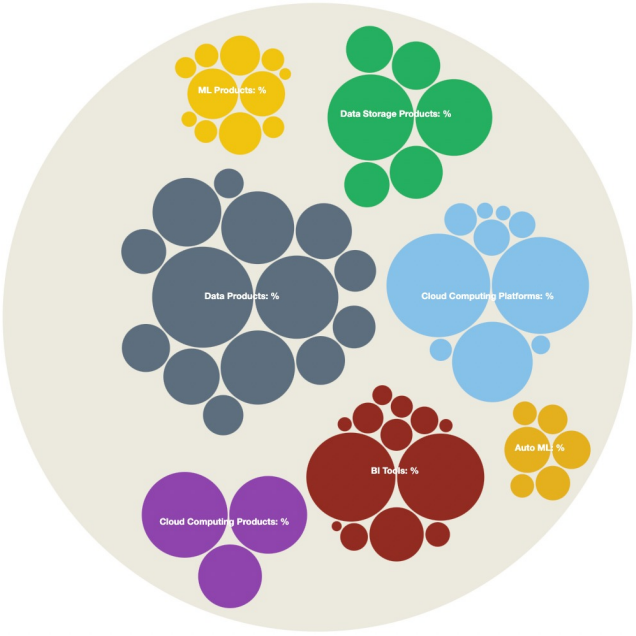

Figure 14 Bubble-chart.

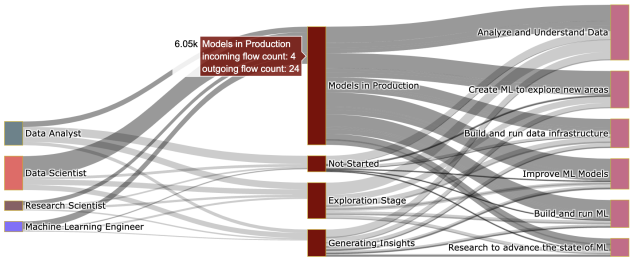

Figure 15 Sankey graph.

less common task overall. These insights can help companies understand the most important tasks for each role and allocate their resources accordingly.

The output Sankey-graph is shown in 15

Heat-map

The chart presented depicts the percentage of respondents with specific levels of Machine Learning experience for different responsibilities. By analyzing this chart, we can gain insights into the required level of ML expertise for performing various tasks.

One of the key takeaways from the chart is that data analysis activities require a higher percentage of individuals with 2-3 years or more of experience in Machine Learning. This suggests that having a strong understanding of ML techniques is critical for performing data analysis tasks effectively.

Another important finding from the chart is that tasks related to applying ML methods to new areas and improving existing ML models require a greater percentage of respondents with higher levels of experience. This highlights the need for individuals with advanced ML skills to tackle complex ML challenges and develop innovative solutions.

Overall, the chart demonstrates the importance of having a solid understanding of Machine Learning concepts and techniques to perform various responsibilities effectively. As ML continues to evolve and become more prevalent across industries, having a strong background in this field can open up new career opportunities and help individuals stay ahead of the curve.

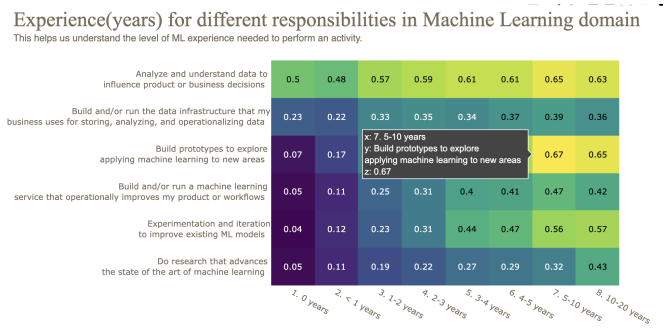

Figure 16 Heatmap.

The output Heatmap is shown in 16

Spyder graph

In the private sector, companies often prioritize experimentation, innovation, and growth in order to remain competitive. This is particularly true for internet-based services, which can collect vast amounts of user data that can be analyzed using advanced techniques to improve their services. For example, analytics and other innovative ideas can be applied to the data to uncover insights that can be used to optimize their products or services.

The insurance industry is another sector that is leveraging AI technologies for various purposes such as insurance advice, underwriting claims processing, fraud prevention, risk management, and direct marketing. With the increasing availability of customer data and advances in technology, AI is being increasingly used in the insurance market to create value, reduce costs, increase efficiency, and improve customer satisfaction and trust.

Retail is also embracing AI technologies, with a significant number of professionals in the sector reporting that their companies have well-established machine learning methods in production. These technologies can be used to analyze customer behavior, optimize supply chain management, and personalize the shopping experience for customers. The use of AI in retail can help companies stay competitive by providing them with the ability to quickly adapt to changes in customer preferences and market conditions.

In summary, the private sector is increasingly leveraging AI technologies to drive innovation, improve efficiency, and enhance customer satisfaction. As AI technologies continue to evolve, we can expect to see even more widespread adoption of these technologies across various industries.

The results of the survey indicate that larger companies, with 1000-9,999 employees or more than 10,000, are the leading adopters of AI. There are several reasons why larger companies tend to outpace smaller ones in AI adoption.

One significant factor is that large firms typically serve large markets, which allows them to better amortize the high fixed costs associated with employing AI production technologies over a higher volume of sales. As a result, larger companies can invest more heavily in AI and reap the benefits of increased efficiency and productivity.

Another factor is that larger companies are able to attract and retain top AI talent by offering higher wages and more benefits. This gives them a competitive advantage in accessing

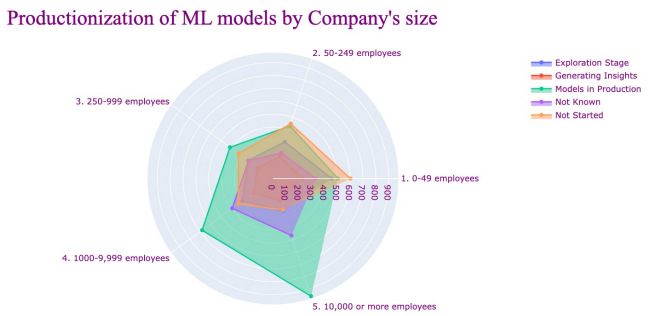

Figure 17 Spyder.

the expertise needed to effectively develop and implement AI systems.

Furthermore, vendors of AI systems often prioritize supplying companies with the largest consumer base, which tends to be larger firms. This can result in vendors focusing on building relationships and contracts with larger companies, which in turn exposes these firms to the value that AI systems can bring to their businesses.

Overall, the survey results suggest that larger companies are leading the way in AI adoption due to their ability to better amortize fixed costs, access top AI talent, and benefit from the attention of AI vendors. However, as AI technologies become more accessible and affordable, smaller companies may increasingly be able to leverage these technologies to their advantage.

The output Spyder Graph is shown in 17

Literature cited

[1] Chang, V., Bhavani, V. R., Xu, A. Q., Hossain, M. (2020). An artificial intelligence model for heart disease detection using machine learning algorithms. Future Generation Computer Systems, 105, 713-722.

[2] Antonio, G. L., Therón, R., Santos, A. B. (2019). BKViz: A basketball visual analysis tool. Journal of Visual Languages Computing, 53, 65-72.

[3] Jadeja, M. (2015). Tree-map: A visualization tool for large data. International Journal of Computer Applications, 129(7), 12-15.

[4] Ahmad, M. S. B., Pesyridis, A., Sphicas, P., Andwari, A. M., Gharehghani, A., Vaglieco, B. M. (2021). Electric Vehicle Modelling for Future Technology and Market Penetration Analysis. Frontiers in Energy Research, 8, 605229.