Kaggle: Exploring the Data Science user-community

ABSTRACT

Despite the abundance of research going on in Data Science and Machine Learning, relatively little exploration and research has been done towards understanding the data science user-community itself and the reasons for the dramatic growth in the use of data-driven methods. In this paper, we take on the *Kaggle Machine Learning and Data Science Survey Challenge* and conduct an exploratory study to present a comprehensive view of the data science and machine learning user-community. We look at the responses from over 23,000 people who participated in a comprehensive survey answering several questions in the categories of educational and professional qualifications, data science practices in the work place, their level of involvement with data science and machine learning techniques and their views and understanding of salient issues in the data science sphere.

KEYWORDS

Kaggle, survey challenge, data science, machine learning, user-community

ACM Reference Format:
. 2022. Kaggle: Exploring the Data Science user-community. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, Article 4, 9 pages. https://doi.org/xx.xxx/xxx_x

1 INTRODUCTION

The democratization of data and our increasing ability to discern data is transforming every facet of academia and the corporate world [11]. Over the last decade, there has been a dramatic rise in job titles that require people to be able to analyze and deduce meaningful insights from data [19]. Even industries like journalism and advertising that have not traditionally employed data-driven methods are seeing new job titles emerge such as computational journalism [6] and computational advertising [30]. Data-driven science is changing how we think about technology and its purpose. Data is no longer considered to be just a consequence of the applications of technology but a compelling force that drives the understanding and development of the next generation of scientific applications and commercial products [23]. An example that truly captures the powerful impact of data-driven methods comes from computational biology where a deep learning model won the 2012 MERCK Molecular Activity Competition. This protein structure prediction model was developed by a team that had purely machine learning expertise and no domain knowledge in biology or protein structures [26]. Biological research has seen several advancements and cutting-edge breakthroughs because of data-driven methods and recently a paper published in PLoS Biology argued the following [18],

> *"Computational thinking and techniques are so central to the quest of understanding life that today all biology is computational biology [17]"*

1.1 Data Science Ecosystem

The data science ecosystem incorporates three essential components, namely, freely accessible datasets, easy-to-use programming languages/tools and an active user community. This thriving ecosystem is a consequence of the progress and development made in all three of these components.

(1) Freely accessible datasets - The ecosystem has greatly benefited from the increasing availability of free datasets and even the corporate world is starting to open up and share their data. Research shows that firms can profit immensely from being open and proactively be involved in the innovation processes instead of passively playing catch-up to stay on pace with the ever changing technology [3].

(2) Easy-to-use programming languages and tools - Easy and free access to tools used for manipulating, cleaning, analyzing and visualizing data has abetted the informed use of data-driven techniques. Python is the fastest growth programming language that is used for web development, application development and now data science as well [27]. Developing data science libraries in Python such as Pandas has really helped propel the data science user community because people working in different domains of software engineering were able to transition into the data science ecosystem [27]. Figure 1 depicts the growth in use of Python packages by accessing the number of Stack Overflow question views per month. Data science packages such as *Pandas*, *Matplotlib* and *Numpy* are seeing a dramatic rise in number of question views since 2012 while packages designed for other purposes continue to get views at roughly the same rate as they did in 2012. Another popular language used within the data science community is R, a statistical computing language that is freely available and has an active and robust user-community [31].

(3) Active user-community - The next good indicator of a thriving ecosystem is the people, that is, the user-community itself. To assess the activeness and level of expertise of different user-communities on Stack Overflow, researchers conducted a survival analysis study that looked at the *"Life-time"* of questions posted for several programming languages. They considered several factors such as the time till the first answer, time till the first accepted answer, number of views and

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/xx.xxx/xxx_x

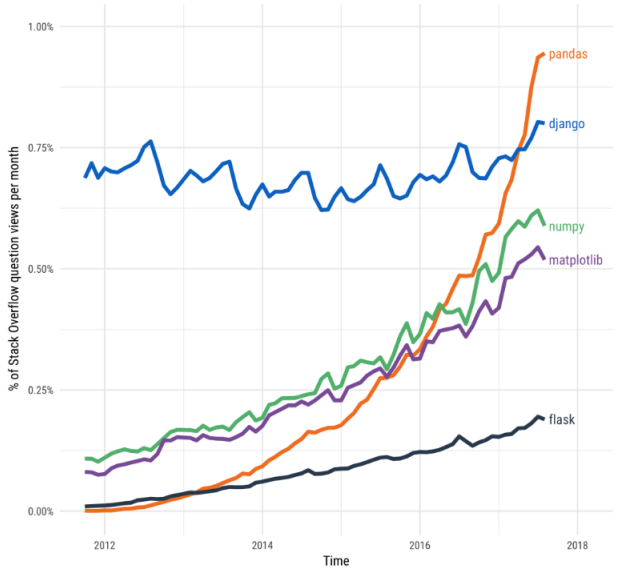

Figure 1: Growth in the use of Python packages [27]

number of responses. They found that *Python* demonstrated the best overall answer rate and *R* demonstrated the best-accepted answer rate [14]. Another study focused on the R community and showed the existence of prolific contributors who act as initial knowledge creators and bridge the knowledge gaps. Their active involvement has lead not just to knowledge creation but knowledge curation that helps future users [31].

1.2 Kaggle - The Data Platform

Addressing the needs of the data science ecosystem and its three essential components, Kaggle has become the ideal platform that allows all three components to coexist and thrive together. Kaggle is an online community designed for data scientists and machine learning engineers [29]. It allows the data science and machine learning community to easily share their datasets and models with each other. It allows them to share their Python notebooks, run their code on the platform, work with other collaborators and seek advice from the user-community. Realizing the enormous potential that was being generated, Google acquired Kaggle in 2017 [15]. Kaggle continues to be the most influential platform that brings together data scientists and machine learning engineers from very diverse domains.

2 RELATED WORK

Kaggle hosts several surveys and competitions each year to generate their own original data sets, but also, to help the user-community tackle hard machine learning problems through crowd-sourcing the solutions. The platform offers the winners prize money and participating in these competitions allows data scientists to compete with users around the world and hone their own skills. Winning these competitions is a prestigious accomplishment and is highly regarded within the user-community. The research coming out of these competitions is often published in top conferences. Most competitions hosted by Kaggle are in specific domains and aim to solve a specific problem. [22] is a study that was published in *International Joint Conference on Neural Networks* that won Kaggle's Social Networks Challenge intended to promote research into real-world link prediction. [28] is a study that won Kaggle's Load Forecasting competition and was published in the *International Journal of Forecasting*. Kaggle's Large Scale Hierarchical Text Classification competition was another challenge targeting a niche of researchers working in Natural Language Processing. [25] was the study that won this competition. [12] was published in response to Kaggle's Satellite Imagery Feature Detection Competition.

Kaggle competitions continue to target specific niches in the data science and machine learning user-community and propel new research in these domains. The *2018 Machine Learning & Data Science Survey Challenge* is Kaggle's attempt towards achieving a comprehensive perspective of its entire user-community. This is the second annual survey released by Kaggle, with respondents doubling over the survey from 2017. The 2018 survey is also quite extensive, including several categories of questions that were not covered in 2017. Our data is sourced from the 2018 survey, and was the impetus for the Kaggle challenge. Several studies in the past have sought to understand user-communities of programming languages and what propels the activeness in these specific communities. [4], [16], [20] and [9] are study that focus on user-communities and different aspects of engagement such as gamification, achievements, reputation, knowledge creation and curation. However, these studies look at specific user-communities with well-defined job descriptions such web development, database management and embedded programming. This is intricately different from the data science user-communities where the *"data scientist"* is still an ambiguous term with job descriptions differing significantly in different sectors of the corporate world.

Data Science is an intersectional approach to data-driven decision making [24]. The exact boundaries of what qualifies as data science, however, are debated within academia. Within industry, data science is a widely used buzz word, often thrown around to describe a variety of practices from standard statistical analyses to big data aggregation. There are data scientists employed in the wealth management industry that are practising a different set of skills than data scientists employed in journalism as computational journalists or data scientists employed in engineering firms and working in computer vision or natural language processing. Data science teams are becoming increasingly common in industry. Companies that implement data-driven decisions have 5-6% higher output than other companies, even when accounting for other advantages [5]. This escalates the pressure on companies that have not yet invested in data-driven practices to hire data scientists even if the need and/or roles for these individuals within the company are not well defined [19]. [21] recently defined the *"data scientist"* as someone who learns from data and solves problems with data; an incredibly vague definition.

The methods of practising data science are largely well known and understood. Despite the obfuscating layers of machine learning algorithms, the concept of machine learning is understood from a more general stance. So what make *"the data scientist"* such an

enigmatic job title? In this paper, we seek to reconcile the idea of a *"data scientist"* and examine if this job title has evolved to identify with a certain skill set. Secondly, explainability of black-box models and algorithmic bias concerns are known to exist in every industry that is implementing machine learning algorithms [10]. It is unclear how data scientists and machine learning engineers are addressing problems with black-boxes, biased models and data. With these two matters in mind, we have developed two main research questions:

(1) **RQ1:** Who identifies themselves as a data scientist? Where do other job titles fall with respect to *"the data scientist"*?
(2) **RQ2:** How does the data science community perceive black-box models? Who cares about model explainability and algorithmic biases?

3 DATA AND METHODS

Data was sourced from the results of the *2018 Kaggle Machine Learning & Data Science Survey*. This data set is the result of a survey administered by Kaggle from their platform as a community of data scientists and machine learning engineers.

The data set is organized in three components:

(1) Multiple choice responses: The bulk of the survey, this includes the standard responses to each question as well as the time a respondent took to complete the survey.
(2) Free form responses: Custom responses respondents typed when answering "Other" on a multiple choice question. There is no way to connect these responses to a set of multiple choice responses, and responses were randomized such that a given row would not necessarily be from the same respondent.
(3) Survey schema: This item shows how many participants were asked a particular question, as well as the factors that would disqualify a participant from seeing a question.

Our examination will be primarily based around the responses found in the multiple choice response section. The free response answers are difficult to meaningfully incorporate due to their extensive anonymized nature, and the sparsity of responses makes specific analysis of this section a lower priority given the limited resources of our project. The survey schema acts primarily as a reference guide as to the users who actually received responded to a question. Though it does contain important meta details about the data set, it affords less insights into out research goals.

The data set underwent very little cleaning on our end. The overall table was reorganized to work more cohesively and clearly within a pandas dataframe. We were able to avoid an intensive data cleaning process because Kaggle presented the data having already underwent significant cleaning. Spam responses were identified and removed, though the specifics of detecting spam were not disclosed. The free form and multiple choice responses were already separated, with the free responses purposely randomized for privacy purposes.

The data set contains 23,859 unique responses. Of these, 81.4% were male while only 16.8% were female. This overwhelming majority of men, while perhaps reflective of the male dominated state the technology industry, means that most of the analyses will be biased towards the male perspective. The country demographics, on the other hand, were not so one-sided. The United States of America was the most represented country, accounting for 20% of responses, with India being a close second at 19%. This is close enough that the perspective from within the United States will not so wildly dominate the overall trends in the survey. Finally, the most represented age demographics were 25-29 and 22-24 comprising 26% and 22% of the respondents respectively. Nearly half of all respondents' ages lie within this 8 year range, which will bias our data towards the point of view of those in their twenties, the late millennial. Taken together, the "average" respondent is a young adult male from the United States or India. It is important to keep this bias in mind when interpreting outcomes that do not control for these variables.

The majority of our methods will be relatively simplistic analyses, which will be described in more detail in the next section. We did not create a predictive model nor a classification model. This is in part due to time constraints, but also due to the broad insights available from "just" analyses. Only after garnering a thorough understanding of the relations and trends within the data can we design a useful and accurate model of some sort. Blindly trying to create a model would likely lead us into significant ethical pitfalls.

4 DATA ANALYSIS

4.1 Who is a data scientist?

Firstly, we wanted to gain a general understanding of the Data Science user-community. Respondents from many different countries and diverse backgrounds participated in the survey. Respondents included students, researchers in academia, engineers, business professionals in accounting, finance, marketing and among many other categories.

First, we focused on the people with the job title of a data scientist and similar job titles that are believed to be generally close to a data scientist. Figure 4 is a *Linear Discriminant Analysis model* depiction of the range of respondents to this Kaggle survey. Across the board, there are Data Scientists, Data Analysts, Researchers, Statisticians, and other occupations. This visual is an attempt at displaying the cross of engineering and math with business and code and where the skill-set of data scientists lie on this cross section. If an individual listed several programming languages such as C, C, Java, Python; they would fall towards the *More Code* side of the graph. If they listed tools used in the business sector such as PowerBI and Tableau; they would fall on the *More Business* side of the graph. Mathematical skills such as regression analysis and time series analysis will push a person to the *More Math* side of the graph and engineering skills/tools such as Visual Studio and MATLAB push a person to the *More engineering* side of the graph. The more code and engineering is involved in a person's day to day job, the closer they are to being a Data Engineer while the more math and business involved in a person's job, the more likely they would be a Statistician. This visualization provides a good representation of what a Data Scientist's job is like: it requires good knowledge of mathematical concepts but also knowledge of programming languages. Furthermore, this visualization provides an attractive representation of the mix of a Statistician, Software Engineer, and Research Scientist skills that are required for a Data Scientist. A Data Scientist fits almost nicely between a Statistician

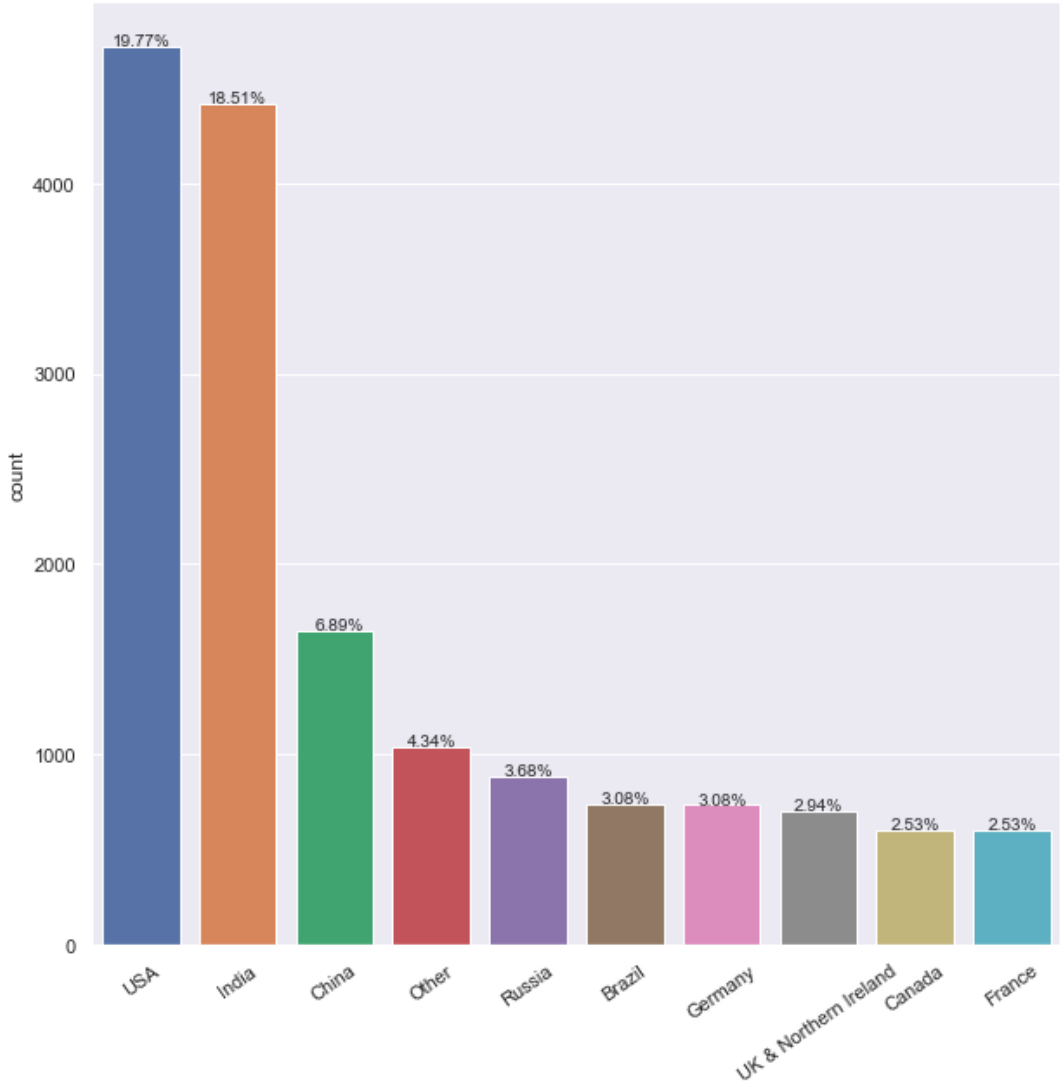

Figure 2: Countries with the most respondents with included percentage of the amount of people who answered

and a Software Engineer, yet is more towards the Software Engineer side as more coding knowledge is required of a Data Scientist.

Next, we present a more "functional" *Linear Discriminant Analysis model* visualization in Figure 5. In this visualization, the size of the bubble represents the size of the corresponding group. The placement of each bubble corresponds to the skill set of each group moving horizontally on the scale between *More Business* and *More Code* and vertically on the scale between *More Engineering* and *More Math*. This visualization develops a better realization of what skills are required of each job role. Interestingly, Research Scientist is engulfed within the Data Scientist bubble. That being said, the academic skills required of a Research Scientist corresponds to that of a Data Scientist as many of the technical requirements are shared between both categories. Research Scientists spend a good proportion of their time developing qualitative and quantitative skills required for analyzing data in their specific domains. Additionally, both positions ask the person to have good coding and math skill sets that are needed to discover trends and patterns in the data.

Outside of academia, statisticians are the only professionals that are vertically at the same height as the data scientists. This implies that statisticians have the same mathematical skill-set as data scientists. However, horizontally, statisticians are all the way to the left whereas data scientists have transitioned towards the center by acquiring programming skills such as R, Python Pandas, NumPy, SciPy, StatsModels etc. This further implies that statisticians are the most suited professionals to transition to data science roles because they have the foundational knowledge that takes years to acquire. Acquiring some coding skills and learning Python libraries takes

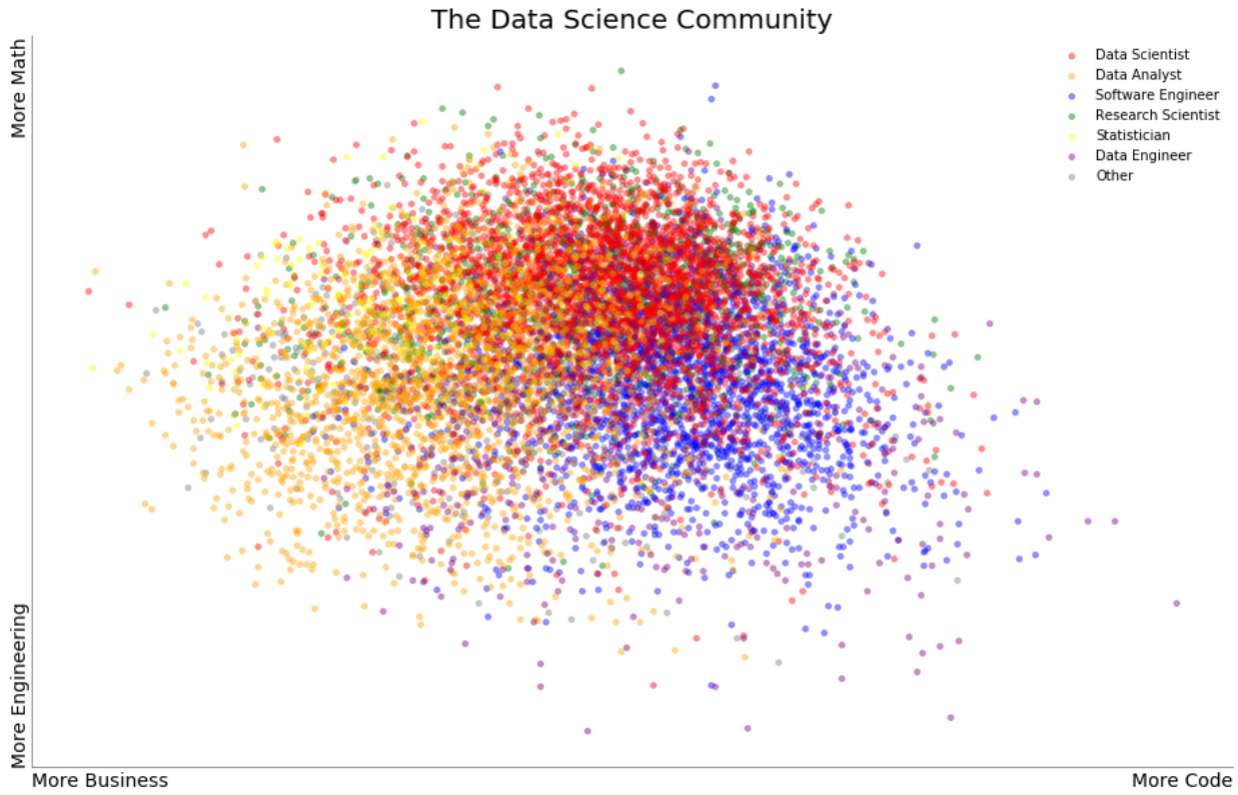

Figure 3: Linear discriminant model of respondents of survey involving their occupation's skill set

a lot less effort. This is a limitation of our visualization that does not account for the fact that moving vertically up on the visualization is much harder than moving horizontally to the right. For example, Marketing and Data analysts are vertically half-way up as compared to statisticians. It is easier for them to add coding skills to their resume and move further to the right. However, moving vertically up to the level of a statistician would probably require them to go back to school for a graduate degree.

4.2 Algorithmic Bias and Fairness

Data scientists are being employed across several industries ranging from wealth management to health care to engineering companies building artificial intelligence systems. However, what remains innate to datasets is the existence of bias and all data assimilation processes are imminently affected by biases [7]. Biases can originate in datasets from several sources and consequentially find their way into algorithms trained using these datasets [13]. Therefore, understanding the sources of biases and knowing how to tackle them should be in the job description of every data scientist. Algorithmic bias and fairness is a quickly increasing topic of interest within the data science user-community. In the Kaggle survey, respondents were asked *"Approximately what percent [of time] of your data projects involved exploring unfair bias in the dataset and/or algorithm?"* Surprisingly, the top answer was - *0-10% of respondents' time was spent searching for unfair bias.* As algorithms are growing in usage and being employed in the public sphere and interacting with people in their everyday life, with systems as detrimental as the *COMPAS risk and needs assessment system* being used by the judicial system [8], this answer was truly shocking.

As algorithmic fairness and bias concerns are just beginning to rise in popularity among the user-community, it would make sense that the people already in the work-force for a few years but graduated in the data science era, that is, the age group 25-29, would be the group looking most actively into algorithmic biases. Figure 6 depicts the percentage of different age groups of people that are looking into algorithmic biases. The main group of people who would be learning more in depth of bias and fairness would be college-aged people, ages 18-21, yet until the college-aged group is reached, there are still four large groups who still surpass those who have a higher chance of learning about algorithmic fairness and bias. Arguably the most damaging aspect of this situation is the idea that those who do not practice searching for bias within their dataset or algorithm have been working towards building popular models for quite some time. Digging deeper, analyzing the level of education of these groups of people was our next step. Interestingly, those who have attained their Doctorate degree had one of the lowest count of people who spent 0-10% of time searching for algorithmic bias (Figure 7). This is once again a sign of a very troubling scenario. Data scientists with doctorate degrees are lead and principal innovation engineers in their respective companies

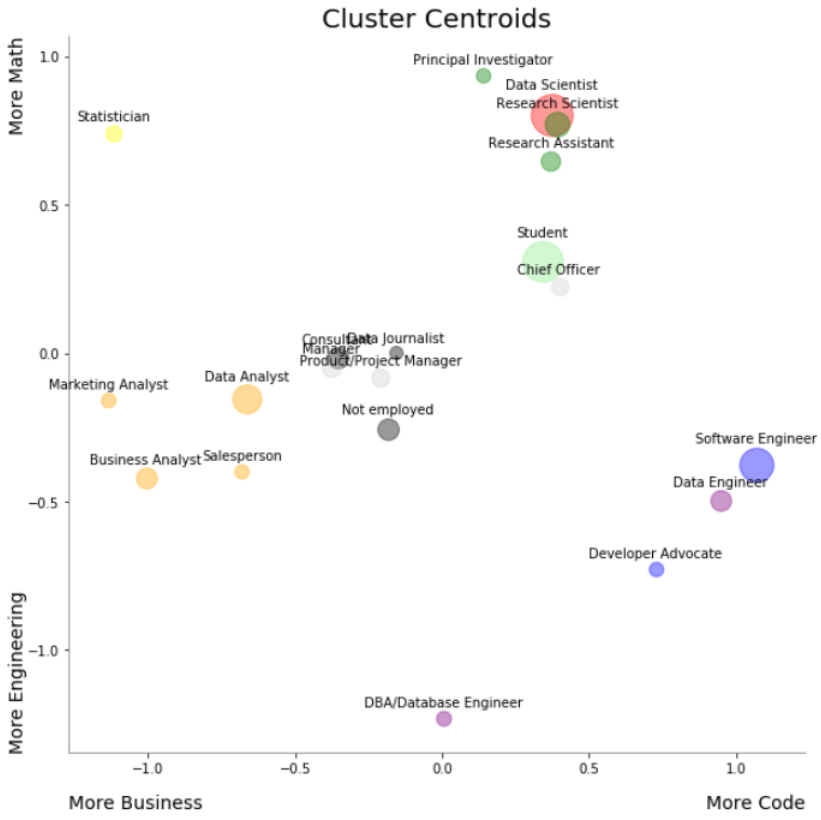

Figure 4: Linear discriminant model of respondents of survey involving their occupation's skill set

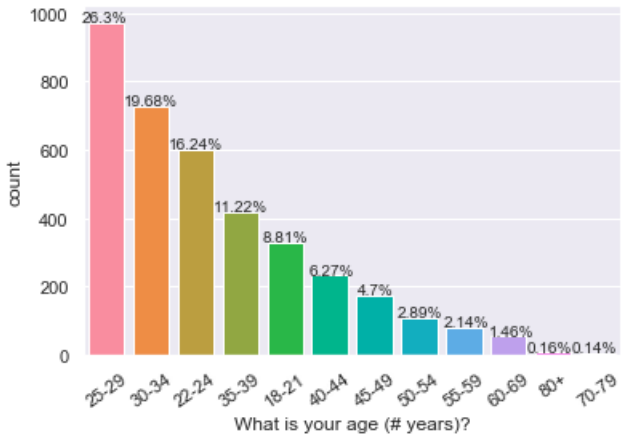

Figure 5: Count of the age groups of people who only spend 0-10% of project time looking for bias

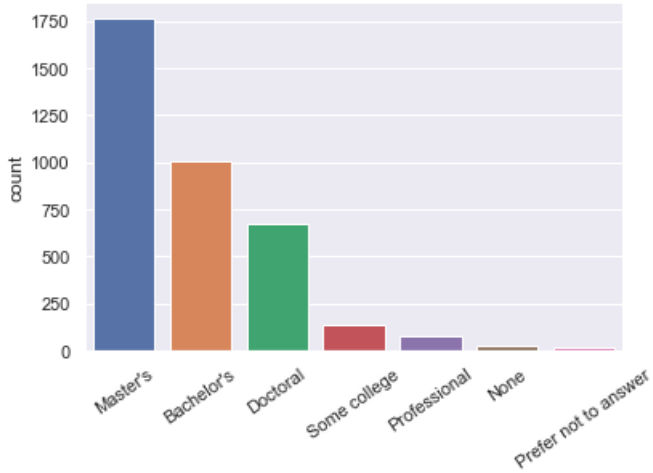

Figure 6: Education level of people looking into algorithmic biases

and the decision to devote time towards searching for and tackling biases must come from the leadership.

Data scientists with doctorate degrees should be the top category among people who are spending time searching for biases. As these people have been in the industry for more time than others, it is good that this group of people have understood the magnitude that algorithms will have on our daily life. On the other hand, as depicted in Figure 7, people with Master's degrees are spending more time looking into algorithmic biases than doctorates. This is a

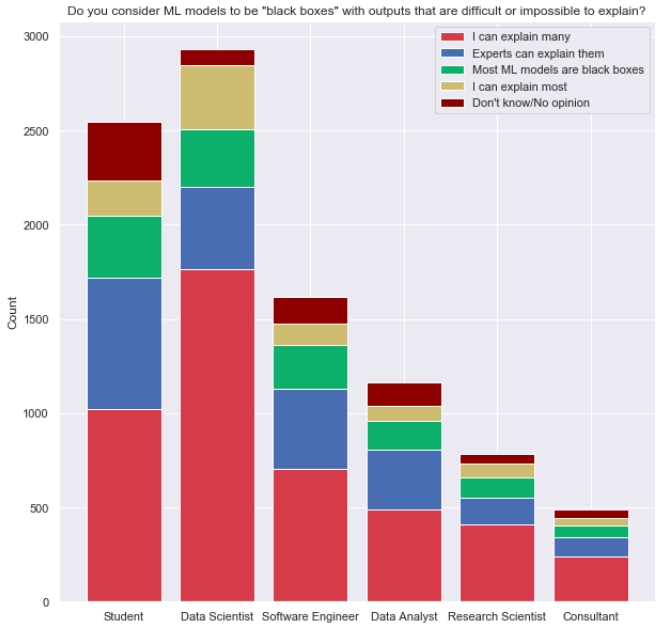

Figure 7: Stacked counts of people's knowledge of black boxes grouped by occupation.

positive sign but we need to remember that this is a small fraction of people that only spend about 0-10% of their time looking into plausible biases. As searching for bias within a dataset or algorithm is becoming a prominent problem, more data scientists should get involved and begin to explore ideas and raise concerns about how increasingly important of an issue algorithmic bias is becoming in the data science user-community.

4.3 Black-Box model explainability

Next, we decided to delve into the idea of how the data science user-community perceives black-box models. The corresponding Kaggle survey question was as follows - *"Do you consider ML (Machine Learning) models to be "black boxes" with outputs that are difficult or impossible to explain?"*. To display the community's perception of black boxes, we created a visualization depicted in Figure 8 that shows the knowledge of black boxes across different industries and each group's perception of black-box models.

We expected and were hoping that the majority of respondents to this question were Data Scientists, the supposed experts in the field. Machine Learning black-box models by definition of being black-boxes cannot be perfectly explained. By analyzing the outputs against inputs, we can make intelligent judgments and guesses about how a certain model is expected to work. Since they are black-box models, we do not know the weights assigned to variables and can only make intelligent guesses. As depicted in Figure 8, most data scientists believe that they can explain most machine learning models and only a small fraction of data scientists acknowledge machine learning models for what they really are, that is, black-boxes. We saw a similar trend in the *Students* group as well where about half the students believe that they can explain most machine learning models. This raises several educational as well ethical concerns. Are the data science instructors in colleges themselves oblivious to the obvious reality of machine learning models? Since data science is a new and emerging field, what is the educational and/or professional background of instructors at most universities?

The answer to the survey question concerning explainability of machine learning black-box models allows us to circle back to the previous question about who is looking into algorithmic biases and how much time are they spending towards it. If most data scientists believe that all machine learning models are perfectly explainable then that eliminates the need to spend time towards understanding, searching for and tackling algorithmic biases. This further raises concerns about the level of expertise of the supposed data scientists.

5 DISCUSSION

Our first research question embarked upon defining *"the data scientist"* and understanding what this job title entails. From Figures 3 and 5, we saw that there are several job titles that cluster around the data scientist but we are able to develop a vivid understanding of what it means to be a data scientist. Over the years, data science has been called a *"buzz word"* by leaders in many industries but we believe that the role has distinctly evolved into its own entity. We define a data scientist as someone who has the foundational mathematical knowledge equivalent to that of a statistician and coding knowledge that allows them to manipulate and analyze datasets. They should also have enough programming knowledge to be able to work with cloud platforms such as Amazon Web Services and Microsoft Azure. We also deduce that statisticians are best suited to transition into data science roles because learning statistical languages like R and Python libraries such as Pandas, NumPy, SciPy and StatsModels is much easier than trying to acquire fundamental mathematical skills in the same time frame. Active and robust Python and R user-communities also make the learning curve easier to attain because of the incredible knowledge curation that has occurred on platforms such as Kaggle and Stack Overflow [14]. As a new user, if you run into obstacles, it is very likely that other users have already asked the same questions and the right answer is already curated and accessible to you [31]. This is also a good recommendation for companies looking to hire and/or train data scientists. It will be relatively painless to hire statisticians and buy DataCamp [1] or Pluralsight [2] subscriptions so that the statisticians can train themselves and hone their coding skills in a few months than hire software engineers with feeble mathematical backgrounds.

Our second research question sought to explore the current state of algorithmic bias and fairness concerns within the data science user-community. The reality was a lot more grim than we could have imagined and that raises several concerns ranging from the general level of expertise within a user-community that is supposedly very good at knowledge creation and curation [31] to the impact of a user-community that is growing exponentially every day. Only a fraction of data scientist consider machine learning models to be black-boxes and consequently devote very little time towards algorithmic bias concerns. The data science user-community is very

active and are producing newer libraries used for data science everyday making it easier to use other peoples' packaged code instead of writing it yourself. This implies that most new users that are entering the data science discipline are just "pushing buttons" on freely available datasets. This raises several concerns, for example, *"Does a larger user-community actually benefit the discipline?"*, *"Who are the supposed gatekeepers that ensure misinformation is not entering a discipline?"*, *"How do you ensure that a user-community is well informed and employing good practices?"*. We intend to formulate and address these questions in a future study.

6 SPECIAL CONSIDERATIONS: DEON CHECKLIST

6.1 Data Collection

Survey respondents freely chose to respond to the survey, and self-selected into the data pool. Users would have knowledge that the data would be made publicly available and analyzed by the community. Therefore participants had ample informed consent.

Due to the self reporting nature of the survey questions, responses will be biased by subjective perceptions of human participants. This type of bias is inherent to this type of self-reported performance related questions. The survey was spread through Kaggle related channels of their main site, forums, and social media. Because of this, people who found the survey were only those who spend a lot of time around Kaggle, which limits who would respond. This would help to keep the survey focused on those who are familiar with the subject matter, but also prevent data scientists who do not follow the website from participating. Not all questions were shown to every respondent, however, in an effort to keep responses limited to those who at least purported to have relevant knowledge or experience. For example, those who reported they were "not employed" would not receive questions about their work or income.

The protection of personally identifying information was strongly valued and thoroughly designed before the release of the data. No names or addresses were included. Respondents from countries with fewer than 50 participants were added into the "Other" category in an effort to maintain anonymity. Further, all free responses were randomized such that entries found in a given row were not all necessarily from the same respondent. With these protections and choices put in place, there is no realistic way to identify participants based on responses.

6.2 Data Storage

Due to the public nature of the data set, there is very little security. Anyone can access the anonymized version of the data, with no controls to keep it from spreading. Unfortunately, we do not know if another version of the data set with identifying information is being held, nor would we know the details of Kaggle's security for it.

There does not appear to be a clear means for someone to have their data removed from the data set, especially in copies since it was intentionally widely and freely distributed.

Again, the public nature of the data set and its position as the cornerstone of a very public Kaggle competition, there are no plans to delete it in the future. Given how it has been cloned to, at minimum, every contest participant's machine, deleting the data is an intractable problem.

6.3 Analysis

The data set unfortunately has a number of significant underlying biases. The majority of respondents are from the United States, India, and China. While this might reflect the general distribution of the data science community, it privileges the views of people from within the sub-community of data scientists in these countries. Furthermore, due to the anonymization of data from countries with few respondents, the view of the community from these countries is obscured to the point of effectively being missing. Due to the time and funding constraints of this project, as well as the limited scope, no effort was made to contact and collect missing viewpoints.

Our analyses attempt to include all facets of the community to the best of our abilities. Furthermore, our code is entirely available on a Github repository and can be easily reproduced and audited. This should allow our work to transparently and easily checked for unfairness and, hopefully, corrected.

Personally identifying information is completely absent from the analysis. This is largely due to the absence of such information from the data to begin with. But even if said information was in the data, the analyses we chose to perform would not have utilized those features.

6.4 Modeling And Deployment

Strictly speaking, no model, classification or prediction, was made by our group. We limited the scope of our work to interpretations of more descriptive analyses. Therefore, there is little possibility for any further ethical concerns as described in the Deon checklist.

That said, it will be hard if not impossible to truly account for unintended use or concept drift. With the public nature of our code and paper, anyone could use our analyses and interpretations for their own purpose. Though this may seem to be a minor issue, there is a possibility of a serious ethical concern. For example, a company may highly value spending time to identify biases in their model. This, taken with a finding in our paper that a certain a given age group is less likely to check for biases, could lead to that company being less willing to hire from that age group. This is a series of very realistic cause and effects that, despite our intentions, leads to age discrimination being justified by our analyses. In theory, similar employment repercussions could accidentally result from effectively any of our group-wise analyses.

7 CONCLUSION

Examining the 2018 Kaggle Machine Learning And Data Science Survey, we were able to gain some valuable insights into the data science community and their practices. We developed a good definition of a data scientist as someone who possessed the foundational knowledge of a statistician but also a certain set of programming

skills. Beyond the general descriptive statistics, we found a disproportionate tendency amongst older respondents to spend only 0-10% of their time considering data and algorithmic biases in their models. We also found that the programming languages used by data scientists were relatively mathematical, being focused on performing calculations rather than building a software product. Data scientists wrote most of their code manually using R and Python but data analysts in the business world mostly used pre-built user interfaces such as PowerBI and Tableau.

Furthermore, we raised several concerns about the data science user-community and their understanding of black-box models and algorithmic biases and fairness. We have tried to act ethically in the course of this project. Data was collected voluntarily with an emphasis on respondent's privacy. We tried to be fair and accurate in our analyses. We attempted to avoid relying on features that could be proxies for protected classes. We have no models which can negatively impact a community. Maintaining a thoroughly ethical paradigm was of utmost importance to us.

Ultimately, we have performed a data driven examination of the data science user-community with a special focus on ethics.

REFERENCES

[1] 2018. Learn Data Science online. Retrieved December 14, 2018 from https://www.datacamp.com/

[2] 2018. The technology learning platform. Retrieved December 14, 2018 from https://www.pluralsight.com/

[3] Oliver Alexy. 2009. *Free revealing: How firms can profit from being open.* Springer Science & Business Media.

[4] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2012. Discovering value from community activity on focused question answering sites: a case study of stack overflow. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 850–858.

[5] Erik Brynjolfsson, Lorin M. Hitt, and Heekyung Hellen Kim. 2011. Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance? *SSRN Electronic Journal* (2011). https://doi.org/10.2139/ssrn.1819486

[6] Sarah Cohen, James T Hamilton, and Fred Turner. 2011. Computational journalism. *Commun. ACM* 54, 10 (2011), 66–71.

[7] Dick P Dee. 2005. Bias and data assimilation. *Quarterly Journal of the Royal Meteorological Society* 131, 613 (2005), 3323–3343.

[8] Anthony W Flores, Kristin Bechtel, and Christopher T Lowenkamp. 2016. False Positives, False Negatives, and False Analyses: A Rejoinder to Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks. *Fed. Probation* 80 (2016), 38.

[9] Scott Grant and Buddy Betts. 2013. Encouraging user behaviour with achievements: an empirical study. In *Proceedings of the 10th Working Conference on Mining Software Repositories*. IEEE Press, 65–68.

[10] Sara Hajian, Francesco Bonchi, and Carlos Castillo. 2016. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2125–2126.

[11] Tony Hey, Stewart Tansley, Kristin M Tolle, et al. 2009. *The fourth paradigm: data-intensive scientific discovery*. Vol. 1. Microsoft research Redmond, WA.

[12] Vladimir Iglovikov, Sergey Mushinskiy, and Vladimir Osin. 2017. Satellite imagery feature detection using deep convolutional neural network: A kaggle competition. *arXiv preprint arXiv:1706.06169* (2017).

[13] Keith Kirkpatrick. 2016. Battling algorithmic bias: How do we ensure algorithms treat us fairly? *Commun. ACM* 59, 10 (2016), 16–17.

[14] Laurel Lord, John Sell, Feyzi Bagirov, and Mark Newman. 2018. Survival Analysis within Stack Overflow: Python and R. In *2018 4th International Conference on Big Data Innovations and Applications (Innovate-Data)*. IEEE, 51–59.

[15] Matthew Lynley. 2017. Google confirms its acquisition of data science community Kaggle. Retrieved December 14, 2018 from https://techcrunch.com/2017/03/08/google-confirms-its-acquisition-of-data-science-community-kaggle/

[16] Lena Mamykina, Bella Manoim, Manas Mittal, George Hripcsak, and Björn Hartmann. 2011. Design lessons from the fastest q&a site in the west. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2857–2866.

[17] Florian Markowetz. 2017. All biology is computational biology. *PLoS biology* 15, 3 (2017), e2002050.

[18] Wes McKinney. 2012. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython.* " O'Reilly Media, Inc.".

[19] Steven Miller and Debbie Hughes. 2017. The Quant Crunch: How the demand for data science skills is disrupting the job market. *Burning Glass Technologies* (2017).

[20] Dana Movshovitz-Attias, Yair Movshovitz-Attias, Peter Steenkiste, and Christos Faloutsos. 2013. Analysis of the reputation system and user contributions on a question answering website: Stackoverflow. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM, 886–893.

[21] Ben Murphy. 2018. Demystifying Buzzwords: Using Data Science and Machine Learning on Unsupervised Big Data. *SAS: Analytics Software and Solutions* (2018).

[22] Arvind Narayanan, Elaine Shi, and Benjamin IP Rubinstein. 2011. Link prediction by de-anonymization: How we won the kaggle social network challenge. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*. IEEE, 1825–1834.

[23] Cathy O'Neil and Rachel Schutt. 2013. *Doing data science: Straight talk from the frontline.* " O'Reilly Media, Inc.".

[24] Foster Provost and Tom Fawcett. 2013. Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data* 1, 1 (March 2013), 51–59. https://doi.org/10.1089/big.2013.1508

[25] Antti Puurula, Jesse Read, and Albert Bifet. 2014. Kaggle LSHTC4 winning solution. *arXiv preprint arXiv:1405.0546* (2014).

[26] Ladislav Rampasek and Anna Goldenberg. 2016. Tensorflow: BiologyâĂŹs gateway to deep learning? *Cell systems* 2, 1 (2016), 12–14.

[27] David Robinson. 2017. The incredible growth of Python.

[28] Souhaib Ben Taieb and Rob J Hyndman. 2014. A gradient boosting approach to the Kaggle load forecasting competition. *International journal of forecasting* 30, 2 (2014), 382–394.

[29] Wikipedia. 2017. Kaggle. Retrieved December 14, 2018 from https://en.wikipedia.org/wiki/Kaggle

[30] Yanwu Yang, Yinghui Catherine Yang, Bernard J Jansen, and Mounia Lalmas. 2017. Computational Advertising: A Paradigm Shift for Advertising and Marketing? *IEEE Intelligent Systems* 32, 3 (2017), 3–6.

[31] Alexey Zagalsky, Daniel M German, Margaret-Anne Storey, Carlos Gómez Teshima, and Germán Poo-Caamaño. 2018. How the R community creates and curates knowledge: an extended study of stack overflow and mailing lists. *Empirical Software Engineering* 23, 2 (2018), 953–986.