# Using ML To Improve Insurance Policy Coverage

# Identifying Key Industries With Ethical Issues Regarding Machine Learning

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Rushil Korpol

November 1, 2021

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISORS

Kathryn A. Neeley, Associate Professor of STS, Department of Engineering and Society

Daniel Graham, Assistant Professor, Department of Computer Science

Rosanne Vrugtman Associate Professor, Department of Computer Science

**Title: The Tightrope Between Exploitation and Innovation: Machine Learning**

Machine learning is a branch of artificial intelligence and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy(IBM Cloud Institution, 2020). Massive amounts of data are fed into algorithms to produce models that serve a wide array of purposes, from predicting the weather to suggesting a new item to buy on Amazon. The use cases of machine learning are endless, but they require significant amounts of data. Oftentimes, this data is not ethically collected and comes at the expense and ignorance of the average individual.

A lot of ethical concerns arise with regards to the deployment of machine learning models, particularly in the spaces where personal information is heavily used such as in medicine or insurance. Oftentimes, people are not made aware of what data they are giving up to companies who can later sell that information or use it for themselves. The usage of private user data must be thoroughly regulated to preserve security and safety of people's information, and to prevent biases that may emerge from the use of such data. Simo(2001) expands on some of these concerns, stating that, "The ability to accumulate and manipulate data about customers and citizens on an unprecedented scale may give big companies with selfish agendas and intrusive/authoritarian governments powerful means to manipulate segments of the population through targeted marketing efforts, perform social control, and hence possibly negatively impact the course of our democracies, or do all sorts of harm" (pg 17). This is already seen with China's implementation of their social credit system, where citizens are constantly monitored and penalized for engaging in actions disagreeable with that of the state.

As it is now, the protection and ethical usage of data is often overlooked. Consumer data is amassed and sold in large scales by companies around the world, with little to no repercussions.The research paper “Ethical Challenges Posed By Big Data” (2020) highlights this mindset: “In general, it is believed that there is less of a need to protect publicly available information. This has resulted in participants being left unaware of the use, or purpose of use, of their information”(pg 1). Both a shift in mindset and improved legislation are needed to prevent such exploitation of information. The technical portion of this prospectus will reflect on the challenges faced when building an ethically considerate machine learning model for a business, and discuss the areas in which the current academic curriculum both prepared and not prepared the team for the task. The STS portion of the prospectus will expand upon the various applications of machine learning, and which face the greatest ethical challenges. This will be important for pinpointing where regulatory legislation should be prioritized.

**Technical Topic: Industry Machine Learning Experience vs Curriculum Based Experience**

During my internship at Markel, an insurance company, this past summer, I completed a research project with a team of fellow interns. We created a machine learning pipeline that would recommend coverage forms to attach to insurance policies. The purpose of this project was to provide support for underwriters, the individuals who craft insurance policies, so that they can do their job more efficiently, ultimately leading to greater profit for the company. While I had taken classes relating to machine learning, there were aspects that I felt sorely unprepared for.

Although the classes I had previously taken contained vital information regarding the general flow of machine learning from data to a model, they were lacking depth in several key areas, particularly in the enormous importance of the data gathering, transforming, and labeling step. As can be seen from figure 1 below, the steps involving data augmentation take up the vast majority of work. Data augmentation involves cleaning up the data and adding new parameters so that it is in a standardized and usable format. This was often gleaned over during class despite its clear practical importance.

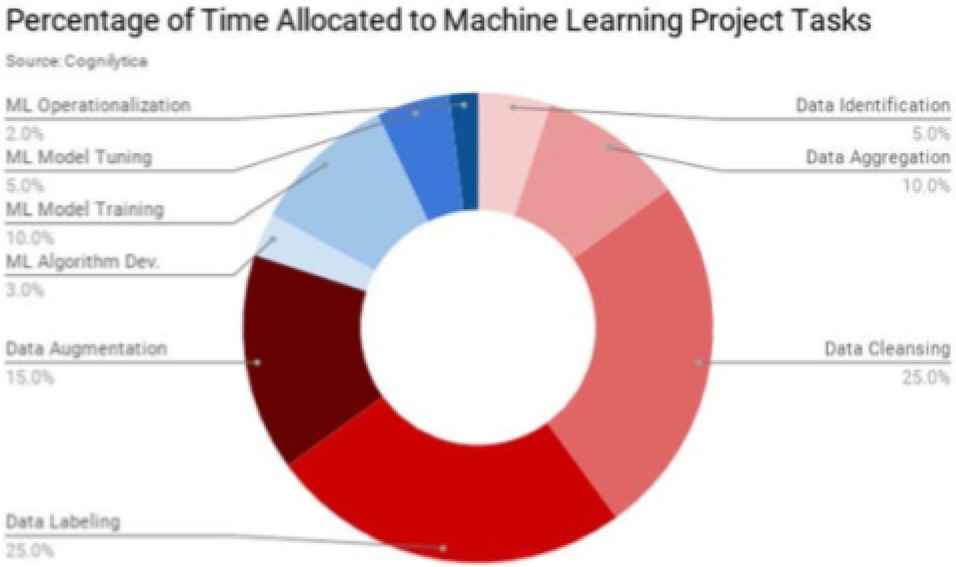

**Figure 1.** Percentage of Time Allocated to Machine Learning Project Tasks. A large majority of time is spent working on data augmentation and labeling (Cognilytica, 2020).

Another aspect was the lack of teamwork in the curriculum. In fact, according to Berry and Lingard(2001), “Employers often report that new hires typically do not know how to communicate and that they have insufficient experience and preparation for working as part of a team”. They go on to state that despite universities recognizing the need to include teamwork in their curriculum, it is often just thrown in and, “students seldom receive any specific training on how to function collaboratively before such assignments are given, and little attention is given to how teams are formed”. They further explain that this often leads to ineffective teams, and students gain little from the experience, with some even walking away with a more negative view on teamwork (pg 34).

Personally, much of this rings true; although there is often a major group project or two throughout the semester in several classes, they fail to simulate actual conditions of working in a team, where meetings are often needed, and clear schedules are created. This often led to a poor division of work, with some people not carrying their weight. The curriculum also addresses ethical concerns in only one unit, rather than being woven throughout its entirety. As Pancake(2020) states, "Teaching ethics at just one point in the curriculum doesn't really convey the extent to which we, as technologists, need to be alert to the ethical aspects of everything we do" (pg 5). The cost of these lacking aspects of the curriculum were felt two-fold. First, there was a slow ramp up as there was a lot of initial confusion on how to begin development as a team and divide up tasks. This ultimately led to a more rushed development of our machine learning pipeline, and as a result it was not as fully fleshed out as it could have been. Second, the lack of understanding of ethical considerations meant that additional time had to be spent for self-education, which further slowed down development.

Curriculum regarding machine learning should emphasize the data augmentation step more heavily as well as more team based work. It should also look at the ethical implications of machine learning, particularly in the spaces where it is significantly more important like medicine. For our project, we managed to ethically use the data we were provided by obfuscating it. This was a necessary step because our data consisted of private information about corporations and individuals, and thus obfuscating the data masked this important information, but still allowed for a model to be created. As a result, we were able to demonstrate that a practical, yet ethical, machine learning system could be created.

## STS Topic: Identifying Key Industries With Ethical Issues Regarding Machine Learning

Machine learning has slowly begun to trickle into virtually every industry, and with its wide scale adoption comes a greatly increased desire for data. Brynjolffson and McElheran(2017) explain, “New digital technologies have vastly increased the scale and scope of data available to managers. We find that between 2005 and 2010, the share of manufacturing plants that adopted data-driven decision-making nearly tripled to 30 percent”(pg 1). The adoption of such decision making requires vast amounts of data and results in the creation of a new network with data at the center. As a result, it is essential to pay attention to where this data is coming from and how it gets utilized. It is particularly important in fields that contain private personal information, such as in medicine and insurance.

In industries such as medicine and insurance, personal information such as name, address, birthdate, and more is often used as data. Although this has allowed for great progress in these fields, such as with machine learning models being able to predict whether patients have certain diseases, it has come at a cost. In fact, according to Vayena(2016), a recent survey conducted in the United Kingdom revealed that 63% of the population was uncomfortable with their personal information being used for machine learning systems. Even more importantly, it is necessary to properly secure consent to use information, as only then could the highest quality of data be achieved. This is essential because if data is not held to a high standard, then you get, as Vayena puts it, “garbage in and garbage out”(pg 2). Bias emerges with poorly representative data, and this could only serve to harm certain demographics. One example of the current lack of regulation regarding data usage was with Cambridge Analytica during the 2020 elections. As Miller(2019) details, “Cambridge Analytica was a firm that used machine learning processes to try to influence elections in the US and elsewhere by targeting ‘vulnerable’ voters in marginal seats with political advertising. They had access to the personal information of millions of voters, and developed detailed, fine-grained voter profiles that enabled political actors to reach a whole new level of manipulative influence over voters"(pg 1). This is but one incident where the average individual gets exploited through the misuse of their own data.

In "Ethical Issues in Big Data Analytics: A Stakeholder Perspective"(2019), the authors assert that,"We view big data analytics as interactions among stakeholders (individuals, organizations, and society) (Zuboff, 2015). The various interactions between stakeholders may not equitably distribute big data analytics' costs and benefits"(pg 721). Future research must examine the relationship between stakeholders across various industries to identify effective legislative solutions. Priority should be given to preserving individual privacy and enforcing ethical usage of data. However, as Hands(2018) describes, a balance must be achieved between regulations and progress: "On the one hand, overlooking ethical issues may prompt negative impact and social rejection …On the other hand, overemphasizing the protection of individual rights in the wrong contexts may lead to regulations that are too rigid, and this in turn can cripple the chances to harness the social value of data science" (pg 1). An overemphasis of one or the other will only serve to prevent progress and lead to stagnation. Thus, further research must be done to clearly identify where such a balancing point may lie.

**Conclusion and Intended Outcomes:**

The anticipated deliverable of the technical project was a machine learning pipeline that could recommend forms to attach to insurance policies for underwriters. A proof of concept was created in Azure Machine Learning that achieved an accuracy of 83%. This meant that of the forms recommended, 83% were a reasonable form to recommend to be attached to a given policy.

The final product showcased that it is possible to ethically generate machine learning models without stepping over an individual's privacy as it made use of obfuscated data. Thus, it shows that ethical concerns are not necessarily a full stopgap on technological progress. It was also able to demonstrate future potential for a larger product that could ultimately be utilized to maximize revenue. The deliverable of the STS project will be a corpus relating to industries where data regulation is pivotal and potential approaches to promote both progress and security.

WC : 1832

## References

Berry, E., & Lingard, R. (2001). Teaching communication and teamwork in engineering and Computer Science. *2001 Annual Conference Proceedings*. https://doi.org/10.18260/1-2--9855

Brynjolfsson, E. and K. McElheran (2017) The Rapid Adoption of Data Driven Decision Making, *American Economic Review*, 106(5), 133-139

By: IBM Cloud Education. (n.d.). *What is machine learning?* IBM. Retrieved October 14, 2021, from https://www.ibm.com/cloud/learn/machine-learning.

Florea, D., & Florea, S. (2020). Big Data and the Ethical Implications of Data Privacy in Higher Education Research. *MDPI*.

Hand DJ (2018) Aspects of data ethics in a changing world: where are we now? *Big Data* 6:3, 176–190, DOI: 10.1089/big.2018.0083.

Howe Iii, E. G., & Elenberg, F. (2020). Ethical Challenges Posed by Big Data. *Innovations in clinical neuroscience*, *17*(10-12), 24–30.

Miller, S. (2019). Machine Learning, Ethics and Law. *Australasian Journal of Information Systems*, 23. https://doi.org/10.3127/ajis.v23i0.1893

Pancake, C. M. (2020). New ways to think about CS Education. *Communications of the ACM*, *63*(4), 5–5. https://doi.org/10.1145/3382126

Simo, H. (2021). *Big Data: Opportunities and Privacy Challenges*. arxiv.org. Retrieved 2021, from https://arxiv.org/pdf/1502.00823.pdf.

Someh, I., Davern, M., Breidbach, C. F., & Shanks, G. (2019). Ethical Issues in Big Data Analytics: A Stakeholder Perspective. *Communications of the Association for Information Systems*, 44, pp-pp. https://doi.org/10.17705/1CAIS.04434

Vayena, E., Gasser, U., Wood, A., O'Brien, D. R., & Altman, M. (2016). Elements of a New Ethical Framework for Big Data Research. *Washington and Lee Law Review Online*, *72*(3).