Thesis Project Portfolio

Using ML To Improve Insurance Policy Coverage

(Technical Report)

Identifying Key Industries With Ethical Issues Regarding Machine Learning

(STS Research Paper)

An Undergraduate Thesis

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Rushil Korpol

Fall, 2022

Department of Computer Science

Table of Contents

Sociotechnical Synthesis

Ethical Data Usage and its Challenges

Data has become an increasingly important commodity in recent years. In particular, there has been a rapid increase in amount of data collected and analyzed due to machine learning. However, most this data is not ethically gathered and leads to an invasion of privacy in individuals as their personal data becomes commoditized without their knowledge. My STS research looks into identifying the various industries where machine learning and data have seen a growth in usage and the ethical implications that arise as a result. My technical topic is about my experience in the insurance industry creating an ethical machine learning pipeline to recommend forms to attach to insurance policies. The technical topic highlights the lack of ethical considerations in current machine learning curriculum, which is directly relevant in the STS topic, as when individuals go on to the workforce, this lack of ethical education enables them to act unethically without any second thoughts. STS is vital to these topics regarding engineering practice because it educates individuals on important ethical notions such as personal privacy and consent.

The technical portion of my thesis produced a proof of concept machine learning pipeline that would recommend forms to attach to insurance policies. This would reduce the time underwriters would spend parsing through thousands of optional forms available for each insurance policy, thereby promoting greater efficiency. While the design of the pipeline itself was straightforward, one important point was that the data was obfuscated prior to being provided as an input. This was a challenge during development due to a lack of prior knowledge regarding data privacy and its ethical importance. However, through this I was able to show that machine learning can used both effectively and ethically.

My STS research analyzed the positive and negative impacts of the widespread use of data by companies, and identified the various ethical challenges that come with using data. In particular, it identifies a rapid rise in data usage across all industries, which has also led to an increase in unethical actions by companies. It looks at 2 scenarios: Amazon's use of data in the commerce industry, and Cambridge Analytica in the social media industry. It compares these two situations to show how a company is able to ethically use data to make more appealing recommendations to its customers through the Amazon case, and demonstrates how a company can exploit data to influence individuals to vote in a specific way in elections through the Cambridge Analytica case. In addition, it highlights an emergent law around data privacy in Europe known as the GDPR, and the current lack of regulations in the United States.

The development of my technical project brought up many important ethical notions that I hadn't previously considered. This made me want to dive deeper, and as a result, my STS research began to focus around answering such questions for myself. Through this process, I was able to gain valuable insight into the numerous ways companies unethically harvest peoples' data and commodify it so that it can be used to further improve their influence and profits. The STS research also helped me to fully understand just why techniques such as obfuscation of data are vital in order to prevent unethical practices within a company. By doing both the technical and STS portions of my thesis, I was able to create a solid framework of understanding about the ethics involving data and the challenges that have yet to be solved in this area.