MSDS 6120 and MSDS 6130

Capstone Projects

Capstone projects are intended to provide an opportunity for you to apply the skills and knowledge that you have obtained from your studies in the Master of Science in Data Science (MSDS) program to solve an open ended problem while showing that your solution is sufficiently optimal. As a data scientist, you will be tasked to solve open ended problems with a goal of not just finding a solution, but finding an optimum (or at least optimal) solution. Many of these problems may be in application domains where you are not an expert. Knowing how to learn and quickly understand the important factors in a new domain are valuable skills for every data scientist, and your capstone project will challenge you to go beyond just simply applying some analysis techniques on a given data set.

In addition to solving a problem and showing the optimality (or at least goodness) of your solution, as part of your capstone work you are required to think about and address, at least in theory and argued in your final paper, the societal and ethical implications of your problem and your solution. As data scientists, ethics are an important part of our everyday lives, yet we rarely discuss the ethical implications of our actions and our inactions. Nearly everything that we do with data, from the collection to the cleaning to the analysis to the use of the information therein obtained, can be used to impact our lives, those around us, and even society in general. Consequently, we must consistently act and perform our roles as data scientists in an ethical manner.

In your capstone project, you will perform research in an area focused within the broad domains of data science. As part of your research, you will address the intersection between your specific problem, data, ethics, and society. These themes may be an integral part of your capstone work or they may be a singular addition to your main capstone work. Your final paper will have at least one section where you present at least your ethics analysis related to your work.

Exemplar broad research topic domains are given from which you will identify, in collaboration with the course professors, advisors, and any project sponsor(s), a specific research *problem* that you will address for your capstone project. The general themes involved in each of these exemplar topics include ethics, data, and society and the interactions between these three. However, an examination of the topic domains will clearly reveal that the possible topics for your capstone project are limited only by your imagination.

Your research on your specific problem will address all three of the basic themes (data, ethics, and society) in addition to the primary problem addressed in your research. *Specifically, as part of this research, you will perform an ethical analysis of the problem, the solution, the impacts of both or any combination of these and include this analysis as one section in your final research paper*. This ethical analysis will be performed regardless of the problem being solved for the capstone project. Similarly, you will perform an analysis of the impact on society of the problem, the solution, etc. and include that analysis within your final research paper. The ethical

and societal impact section(s) should be a substantive section within your final paper.

While your research will be performed largely by you and your research group, you will receive feedback on your work over the course the capstone project. This feedback will come from the professors, your advisor(s), any project sponsors, and from your peers in the course. It is *required that you obtain one or more capstone advisors* to act as guides for your research, sounding boards for your ideas, or other sources of support. Advisors may be professors, industry researchers, or other qualified persons. Your advisors should be in addition to any project sponsor(s) that you may be working with. Your advisor(s) will be listed as coauthor(s) on your published research paper.

The documentation of your research and findings will consist of a research paper to be published in the *SMU Data Science Review* journal, a poster presentation of your work to be presented at the Symposium on Data Science conference, and a presentation of your work at the conference. The lasting documentation of your research will be the research paper. The *SMU Data Science Review* journal is an open access journal published by SMU to act as a searchable and accessible record or your capstone research.

Example Capstone Project Topics

The following list of capstone project topics is an example list from which you may develop a specific target problem for your capstone research. This is not an exhaustive list. If you have a topic or problem idea (especially one that involves the intersection of ethics, data, and society), then that topic or problem is relevant and a possibility for your capstone research. Please discuss your problem ideas with the course professor before pursuing them too deeply.

Corporate Activities – Corporations perform a wide variety of data acquisition and analysis of their customers. Some of this acquisition and analysis is widely disclosed and intended to provide better services to customers. However, sometimes this activity is not disclosed or is compelled under secrecy rules. What is the impact on society of corporate activities such as scanning emails and monitoring their users both with and without their knowledge?

Unintended Consequences of Location Information – location information is commonly used by advertisers and social media companies to provide contextually aware advertisements and recommendations. However, these recommendations have the potential to violate privacy, for example, by recommending connecting with other people who frequent the same locations. What are the privacy, societal and ethical implications of using location information for advertisement and social media applications?

Ethical Implications of Perfect Recall – Much of our lives is being captured on cameras. Each still image or video shows an incomplete view of a portion of our activities. Other portions of our lives are being captured, however imperfectly, by ourselves and others in social media posts. How do digital memories of imperfect data, both implanted and recorded, impact law and society?

Right to be Forgotten – The European Union allows private citizens to request that certain information about them be removed from Internet search engine results. The effect of this removal is to make it very difficult to find the information about an individual, effectively erasing a portion of an individual's history. Public individuals do not have this protection; however, any information "forgotten" while an individual is a private citizen is not automatically "remembered" when they become public individuals. What does the right to be forgotten do to true memory, the collective memory of society and to history?

Privacy Issues in a Cashless Society – A cashless society relies solely upon digital transactions for the purchase of goods and the exchange of wealth. All of these financial transactions, in accordance with standard financial practices and auditing laws and rules, are traceable. When all financial transactions are digital, does the resulting cashless society reduce our privacy?

The Return of HAL – We are currently witnessing the rise of intelligent machines through the use of artificial intelligence, machine learning and advanced analytics techniques. What are the ethical and societal implications of human intelligence in machines?

Stingray and Big Brother – The Stingray cell tower has been widely used by law enforcement to monitor the cell phone activities of large numbers of citizens in the pursuit of criminals utilizing their cell phones for illegal activities. What are the consequences, actual, likely and/or possible, on society and the technologies that we use due to this widespread indiscriminate surveillance?

Wheat from the Chaff – Anonymized data sets are commonly made available for research purposes, particularly for drug trials and other medical and financial research purposes. However, recent research has shown that when combined with other public data sets, individuals in the anonymized data sets can often be uniquely identified. One approach to counter this type of analysis is to modify the anonymized data. How much can the anonymized data be modified in order to mitigate the identification of individuals in the data set while allowing for the same analysis results of the data to be achieved?

The Rights of People to Their Data – Every person generates huge amounts of data through their everyday lives and activities. All digital activities are monitored and recorded to greater and lesser degrees, leaving at least bread crumbs to indicate the passage of a particular person through the ether. Even travel through the physical world leaves electronic bread crumbs through the numerous cameras and electronic devices that travel with and around us. Most of these digital bread crumbs (some of which are large loaves of bread) collected about an individual are not under that individual's control, are not collected by that individual, and are not accessible by that individual. Given the ever increasing digital life led by everyone, what are an individual's rights to the data collected about him/her? What laws are in place to protect an individual? What laws protect the collector of the data? Should these laws by changed?

How much Privacy do we have Today? – Much of our lives are available online for anyone that cares to look at us. But, how much can someone find out about an individual just from publicly available information? In this project, select a public or semi-public individual, such as a politician, a judge, or a celebrity, and develop a detailed record of their life.

Air Pollution and Increased Death Rates – A number of studies have found a positive correlation between high air pollution and increased mortality. Using recent data, evaluate the impact of air pollution on the daily health (and death) of a major city.

Fake News and the Bots that Make It – Bots have been deployed widely to generate news articles, tweets, and other sources of information. They have also been deployed to follow users and other bots on social networks, blogs, and news sources. In the process, bots have been used to generate both accurate news, based upon facts, as well as fake news, sometimes based on facts, sometimes not. Using an array of investigative techniques and data, evaluate the extent of bot-generated articles and of "fake news" proliferated by these bots.

The Rise and Fall of Popular Baby Names – Names, specifically first names, have a popularity that ebbs and flows throughout history, and some names may even cross over in popularity from one gender to another across generations. In this project, the investigators will evaluate the events and societal changes that impact the rise and fall of baby names over time. One source of baby name information is Babynamewizard.com . (Possible advisor: Dr. Monnie McGee.)

How Much Data is There in the World? – In 2012, Martin Hilbert published a set of paper that purports to answer the question "How much information is there in the "Information Society"?" The world has changed since this paper was published half a decade ago with even more storage storing data being generated at an exponential rate compared to 2012. Picking up where Martin Hilbert left off, the question today is, How much data is there in the world? Using significant research and data analysis techniques, quantify the amount of data generated and currently stored in the world today. (possible advisor: Dr. Daniel Engels.)

The (Loss) of Fortunes Over Generations – What is the effect on the fortunes of second-born sons (and third, fourth, etc.) due to the primogenitor laws? How long does it take (in terms of generations) for a once wealthy family to become poor when these rules are enforced? One could imagine a simulation with various factors to answer this question. It may be hard to find actual data. (Possible advisor: Dr. Monnie McGee.)

Sampling-Based Estimators of the Number of Distinct Values of an Attribute – Query optimization methods in databases, particularly relational and object-relational database systems, require a means of assessing the number of distinct values of an attribute in a relation or within the database. Accurate assessment can lead to efficient query planning, while inaccurate assessments can lead to significantly degraded query performance. Using existing and novel sampling-based estimators, determine their effectiveness in a range of NoSQL databases and their usefulness in developing query optimization methods for these databases.

A Data Driven Visualization of Something – We are awash in data and approaches to analyze that data. Yet, humans can glean information from large quantities of data when it is presented in a visually informative manner. Using large quantities of data, generate a series of informative visualizations that tell a story. (possible advisor: Prof. Ira Greenberg or Dr. Daniel Engels.)

Sponsored Capstone Project Topics

Capstone, Spring2018/Summer 2018: Models and Data Structure Comparison

Principle Contact: Bivin Sadler

Description
A common question that keeps popping up in my head is, "When does a Random Forrest perform better than other, more traditional, competing methods?" Specifically, it would be interesting to investigate when random forest classification performs better than logistic regression? In order to research this question, one could simulate data of various complexity (number of features) and various correlation structures. It stands to reason that if you simulate data from a linear model with known parameter estimates and independent features, that logistic regression would work best since it is actually estimating the true structure the data was pulled from. A) Is this true? I often assume that something will be true and/or easy just to find something amazing after a little research. B) What if we add a non-independent correlation structure to the linear model? C) What if generate the data from a a non-linear model … maybe research a known quantity that has a non-linear (and non curvi-linear) relationship with the logit and compare and contrast logistic regression and a random forest model in classifying these data? It would be nice to be able to give practitioners and clients from different fields some guidance as to which method they should use if they know something about the structure of their data.

Note: This project could be spun off in quite a few different directions. Another group could look at differing sample sizes, while other groups could simply pick off a different correlation structure or model complexity that that is commonly found in a different field. In addition, under the same umbrella but in a slightly different direction, a group cold look at continuous responses and contrast linear models (MLR) with random forest regression. Of course, in the end, there are many more models to compare than simply these 4 … groups could compare and contrast nearly any modeling/predictive method with respect to the above framework.

Personnel
An ideal team of students will have familiarity with and interest in classification models including regression models and random forest models.

Deliverables
We will work with the team to develop a set of concrete deliverables to answer the questions above.

MSDS Capstone, Spring2018/Summer 2018: Education Attainment

Principal Contact: Jake Drew

Description
What drives educational attainment? Can we build machine learning models to predict dropout rates, SAT scores, or how many high school students will enroll in college within 16 months of graduation? Is segregation in public schools a problem in 2017? Do minority birthrates and immigration explain the growing number of majority-minority schools over the past 5 years?

The State of North Carolina maintains one of the most comprehensive collections of data on public schools in the nation. The North Carolina School Report Card and Statistical Profiles databases comprises over 300 school level features which are available for mining. Student level data may also become available as the project progresses. Join Dr. Drew in a deep dive into the world of public education. This project has the potential to initiate meaningful changes to the public education system in North Carolina. Be a world changer and earn your capstone credits at the same time!

Datasets available at: https://github.com/jakemdrew/EducationDataNC

Personnel
An ideal team of students might consist of: familiarity working with text data, traditional machine learning methods, and convolutional/recurrent neural networks (or desire to learn them within about three weeks of the project start date). If this intimidates you, please do not select this project. Teams should consist of members that all received a B+ or better in the MSDS Data Mining course.

Deliverables
We will work with the team to develop a set of concrete deliverables to answer the questions above.

MSDS Capstone, Spring2018/Summer 2018: Classifying Compatibility

Principal Contact: Brent Allen (brenta@smu.edu)

Description
Many people want to find love and, hopefully, some of you have. The difficulty is finding not just someone you love, but someone with whom you are compatible. Can we measure how compatible a person is with another for a relationship? Many methods have been developed and tried and are used in commercially successful businesses (e.g., Match.com). Can we as Data Scientists use existing methods to create new insight or create a new method to help a person find compatibility in a relationship? The goal of this project is to use two personality and relationship algorithms (Meyers Briggs and The 5 Love Languages) to classify the type of compatibility for a person. You will have to research each of these algorithms, determine how you will classify each person within each algorithm, create a method to collect and store the data, create compatibility classifications from these two algorithms and visualize the results.

Personnel
An ideal team of students might consist of: familiarity working with text data, familiarity with social and personality categorizations, and familiarity with neural networks. Teams should consist of members that all received a B+ or better in the MSDS Data Mining course.

Deliverables
We will work with the team to develop a set of concrete deliverables to answer the questions above.