IST 687: Introduction to Data Science

I took this course in my first semester (Fall 2019) and it introduced me to the world of Data Science and the fundamentals about data and the standards, technologies, and methods for organizing, managing, curating, preserving, and using data. It provided applied examples of data collection, processing, transformation, management, and analysis as well as a hands-on introduction to the emerging field of data science while discussing broader issues relating to data management, quality control and publication of data.

The project for the course was about analyzing the customer data of Southeast Airlines and help the company lower their customer churn and we were divided into groups. One of the key factors to measure customer churn is the Net Promoter Score (NPS), where a score less than 7 is classified as Detractors, 7 or 8 as Passive, and greater than 8 as Promoters. The data contained 129,889 observations and provided information about the characteristics of the flight (day of month, date, airline, origin and destination city, if the flight was delayed, etc.), the customer (age, gender, price sensitivity, the person's frequent flyer status, etc.), and a simple survey-based rating of each customer's likelihood to recommend the airline that they just flew as well as a field for open-ended text comments. The aim of the project was to help Southeast Airlines increase their NPS by determining which group of customers are likely to be Detractors.

We first cleaned the data and performed some exploratory analysis by plotting the different variables against each other (using grouping if necessary) to understand the relationship between them. We then performed association rules mining to find predictors with a strong association with the target variable (binned NPS score). The predictors identified in this process were then used to train classification models after appropriate feature engineering using Logistic Regression and SVM for different subsets of the data. The high accuracy of the models validated the assumptions about the correlation between the predictor variables and the target variable, and the information was used to suggest recommendations to the airline. All the analysis was done using R. The results were presented in a manner assuming the target audience to be a business executive and a report containing the method details along with the results was submitted.

The key learning goals achieved in the project and the course were –

- Collecting and organizing data.
- Identifying patterns in data via visualization, statistical analysis, and data mining.
- Developing alternate strategies based on the data.
- Developing a plan of action to implement the business decisions derived from the analyses.
- Demonstrating communication skills regarding data and its analyses for relevant professionals in the organization.

Github link: https://github.com/aatishsuman/IST687-southeast-airlines-customer-churn

IST 736: Text Mining

I took this course in my second semester (Spring 2020), and it taught me the techniques used for analyzing text data. The primary purpose of the course was to provide the fundamentals of text mining with emphasis on the traditional machine learning algorithms like Naïve Bayes, linear SVM, k-means clustering, LDA topic modeling, along with the techniques for preprocessing and vectorization of text data, using the open source tools Weka and Mallet, and the Sklearn package in Python. Methods for the collection of social media data using web-scraping tools and the ethical concerns related to the use of such data were also discussed. Most of the assignments were open-ended for the purpose of allowing the students to critically analyze and interpret the results. To further sharpen critical analysis skills, the course also included an advanced topic presentation, where students were asked to analyze the methods and the results of a research paper of choice and present their thoughts to the class.

We were divided into groups of 2-4 for the final project. For my project, we performed a comparative study of the different traditional algorithms used for sentiment analysis. The sentiment140 dataset containing 1.6 million tweets was analyzed using five traditional machine learning algorithms (Bernoulli Naïve Bayes, Multinomial Naïve Bayes, Linear SVM, Random Forest and Logistic Regression) with different vectorization (Boolean, TF, TF-IDF) and feature selection (unigram, bigram) methods, and their performances were evaluated to determine their suitability for the task.

The data was cleaned following the steps mentioned in the paper by the creators of the dataset (Go, A. et al 2009), and different versions of the vectorized data was prepared based on the input requirements of the algorithms. The data was split into training and test sets and the training set was used to train the models with 3-fold cross-validation while the test set was used to measure the performance of the models. The data was perfectly balanced between the two classes of positive and negative, and therefore, test set accuracy was used to measure the performance of the models. The results were analyzed to compare the performance of the models and were presented to the class, and a report prepared in research paper format was submitted.

The key learning goals achieved in the project and the course were –

- Collecting and organizing data.
- Identifying patterns in data via visualization, statistical analysis, and data mining.
- Demonstrating communication skills regarding data and its analyses for relevant professionals in the organization.
- Synthesize the ethical dimensions of data science practice.

Github link: https://github.com/aatishsuman/IST736-sentiment-analysis-twitter

IST 700: Deep learning, NLP, and Computational Social Science

I took this course in my third semester (Fall 2020), and it allowed me to dive deeper into the field of machine learning and its practical applications. Natural Language Processing (NLP) has become an important research approach for computational social science and the deep learning methods have revolutionized the field in recent years. The new end-to-end models have achieved high performance and can be trained without task-specific feature engineering. The course was designed to train research-oriented students to use deep learning-based NLP techniques as a research method for computational social science. I am planning to go for a PhD in a few years and this course perfectly suited my interests. As part of our assignments, we analyzed several research papers and learned to critique the methods, the results and the inferences drawn from the results. We learned to pay attention to research design issues such as controlling confounding variables and evaluating different kinds of algorithmic bias, and ethical concerns that may arise because of bias in our methods of data collection or analysis. We also learned some of the more advanced machine learning concepts like word-embeddings, attention-based models (transformers, BERT, etc.) and transfer learning.

For the final project, we were divided into groups of 4-5. My project was designed to answer the research question – what is the nature and frequency of Social Presence, as defined by the Community of Inquiry framework, in voluntary online platforms. We analyzed the comments of the members of the Gravity Spy project on Zooniverse, an online platform for hosting citizen science projects, to understand the kind of learning happening on such platforms. In our case, learning was defined by the Community of Inquiry framework, and we focused on the kind of social presence exhibited by the members of the project in the discussion forums. The literature on the topic indicates that in citizen science projects having social interaction improves individual's participation (Curtis, 2015), their motivation to participate (Price and Lee, 2013), improvements in scientific literacy and eventually contributes to their knowledge and self-confidence (Jennett, et. al., 2015). It also improves the individual's command on the topic and willingness to participate in and learn from the communal interactions (Jackson, et. al., 2020). Considering the importance of social presence, the research was intended to contribute to literature by analyzing text-based interactions and identify the kinds and frequency of social presence between the volunteers of the project.

The data was collected using web-scraping tools from the discussion forums on Zooniverse, annotated and the intercoder agreement was measured to assess the quality of the data and the applicability of the annotation schema. For the annotation, we went through several rounds of training until the intercoder agreement was sufficiently high. The annotated data containing 400 comments was then used to train classifiers using BERT-based and rule-based methods.

The rule-based approaches were used in cases where the number of positive observations were too low to allow a machine learning model to learn from them. These classifiers were then used to make predictions on a larger collection containing 90,501 comments (~85% of all the comments for Gravity Spy) to predict social presence in the discussions and the results were analyzed using statistical tools. They were presented to the class, and a report prepared in research paper format was submitted.

The key learning goals achieved in the project and the course were –

- Collecting and organizing data.
- Identifying patterns in data via visualization, statistical analysis, and data mining.
- Developing alternate strategies based on the data.
- Demonstrating communication skills regarding data and its analyses for relevant professionals in the organization.
- Synthesize the ethical dimensions of data science practice.

Github link: https://github.com/aatishsuman/IST700-zooniverse-online-learning

IST 615: Cloud Management

I took this course in my third semester (Fall 2020), and it help me build a foundation in cloud computing and its use for data science projects. With the amount of data being generated and the computational complexities of the state-of-the-art models being used today, most of the companies have resorted to using cloud services for their data science projects. Cloud computing offers great potential benefit to users, such as lower cost, greater flexibility and scalability, and the ability to focus on core expertise and less on managing complex infrastructure. The course equipped me with skills in cloud enterprise service creation and management, and technical and business knowledge required for assessing opportunities and risks of cloud services as well as understanding data security and privacy issues when dealing with data stored in a cloud. It focused on developing skills to manage cloud instances through lab assignments in the simulated environment provided by top Cloud vendors in the market today (Amazon Web Services and Google Cloud Platform). Some of the technologies learned in this course are Docker, Kubernetes, Google AutoML, AWS SageMaker, AWS CloudWatch.

For the project for the course, we were divided into groups of 4-5. In my project, we created an end-to-end cloud hosted service for detecting face masks in live feed images. The idea of the project was to formulate a plan for a startup based on the service and create a working implementation of the service hosted on a cloud platform, including a cost and revenue model, a user interface, a machine learning model to detect the presence of face masks in images along with APIs for using the model using the interface. Common use cases of the service would

be any public location where the use of a face mask is mandatory like shopping malls, airports, hospitals, restaurants, offices, etc., and the entrance points would be controlled via the service.

We chose AWS as our cloud platform. To generate the data for training the models, we recorded short videos of ourselves with and without masks and generated images from them. The data was deployed to the platform along with the code for training a CNN model and the model was trained on the platform. We also developed a UI which would capture the feed from the cameras of our laptops and an API which would accept these images as its input and return the prediction of the trained model as its response which would be displayed on the UI. Both the UI and the API were also deployed to the platform. A live demo of the application along with a realistic cost and revenue model was presented to the class.

The key learning goals achieved in the project and the course were –

- Collecting and organizing data.
- Developing a plan of action to implement the business decisions derived from the analyses.
- Demonstrating communication skills regarding data and its analyses for relevant professionals in the organization.
- Synthesize the ethical dimensions of data science practice.

Github link: https://github.com/aatishsuman/IST615-FaceIt