Master of Science in Applied Data Science
Portfolio Milestone Report
Syracuse University

Srihari Busam | sbusam@syr.edu | SUID # 742162654

Table of Contents

Why I wanted to pursue education in Data Science

I have been developing software for almost 18+ years now. During the later part of my career, every organization I was part of got access to data scientist resources. In one project, our team planned to use sentiment analysis and classification experiments to improve our systems from engineering. I started managing the data scientist assigned to our organization.

Overall, I felt the results were not positive as I thought the data scientist provided a generic solution that does not apply to my team's business domain. Our team had data at scale, and data scientists struggled with cleaning and processing the data. I got an idea of the data potential with my interactions with the data scientist, and I have good familiarity with dealing with data at scale. I felt getting a traditional education in applied machine learning to bring out data insights would further my career. That is why I wanted to pursue an M.S. program in applied data science.

How the Applied Data Science program helped me

MBC 638 class is my first exposure to statistics. I did not do any formal training or course on statistics before. This class, in my view, is the foundation for the rest of the system. IST 772(Quantitative Reasoning for Data Science) reinforced the learning of MBC 638 and provided bayesian thinking. IST 772 also provided a playground to explore the concepts through R.

I explored big data processing using Python and Apache Spark and worked on my Project-related deep learning leveraging Tensorflow. IST 707 (Applied Machine Learning) helped me understand the techniques related to various regression and classification methods using the R tool. IST 718(Big Data Analytics) is where the learning took top gear with real-world problems and vast amounts of data.

IST 719 (Information Visualization) provided me with a great understanding of how important it understand the audience and present the data accordingly.

In today's world, data science deals with the data's 3 V's (volume, velocity, and variety). RDBMS may not scale enough to solve all data problems as information evolves. IST 659

provided(Data Administration Concepts and Database Management) traditional RDBMS management aspects. These base concepts are furthered by IST 769 (Advanced Big Data Management), which provided an overview of other No-SQL systems and introduced the idea of polyglot database systems that will help solve real-world problems.

How am I planning to further my learning?

The Applied Data Science program gave relevant theoretical aspects underpinning descriptive statistics and Machine Learning. Our engineering teams produce Petabytes of time series operational data at my current organization. There is a huge opportunity to mine the operational data, which can help improve processes and reduce costs for the organization through the forecasting and classification techniques I learned in the program.

Projects/experiences for the portfolio

Course	Project/Course overview	Learnings
MBC 638 **Data Analysis and Decision Making**	Process improvement Project: **Reducing minutes late to office**	Exposure to statistics, hypothesis testing, confidence intervals, different types of errors.
IST 718 **Big Data Analytics**	Image classification using deep learning: **Human Protein Atlas Single Cell Classifier**	Python, pandas, Jupyter notebook, Apache spark using python, Google Tensorflow 2.0, Understanding of Convolution neural networks and Transfer Learning concepts. Learned time series analysis using the prophet package.
IST 719 **Information Visualization**	Poster project **Sony Rules Gaming**	Principles of visual design. Layout basics. Understanding and targeting the audience. Adobe Illustrator. R Studio and grammar of graphics though ggplot package.
IST 769 **Advanced Big Data Management**	Hands-on experience on modern RDBMS and NoSQL Databases	Hands-on experience in advanced SQL Server, Hadoop, MongoDB, Cassandra, and Kafka.

MBC 638: Project about "Reducing minutes late to the office."

About the class and the project:

I have never taken a statistics class before. MBC 638 is the first formal class that introduced basic concepts about statistics. I attended this course from Whittman school of business and initially felt that the statistics aspects were more geared towards business processes(like any commercial business entity) than scientific processes. However, by the end of the course, I am very clear on applying the principles where ever data is available and the research question applicable.

Since I started my new job at Qualtrics, I am always late for the office. This course had a deliverable of "Process Improvement Project With Storyboard," which uses the DMAIC(Define, Measure, Analyze, Improve, and Control) framework to improve the existing process. The instructor suggested utilizing a process that the student wants to improve personally. I tried to use the tools learned in this class to measure existing processes, identify improvements, and analyze post-process improvements.

Project details/execution

I collected live data for the project. I collected about four weeks of data for my project. The first two weeks of the data are about collecting data about the existing process, and the last week's information is collected after implementing the process improvement. For each day, about ten observations are collected. All the observed variables are continuous. All measurements made are in the same unit (minutes). My overall goal was to reduce the "minutes being late" to the office by 50%, and by the end of process improvements, I got a 69%

reduction in "minutes being late " to the office." I want to outline the project synopsis in the DMAIC format below:

Define:

This is the project's first phase where I have defined the problem statement as "reduction of minutes being late to the office by 50%"—created a process map to give a clear understanding of the current process.

Measure:

Measure phase helped how to optimize my current process. In this phase, I have collected the data before process improvements. I calculated the mean differences before(37.46min) and after the process improvement(12 min). I have used tools like the Pareto chart to understand 80/20 aspects of the data about where I am spending more time in my process. This phase also helped me identify what tasks related to phone browsing checking emails/news steps are redundant, as I will do that once I reach the office anyway.

Analyze

I did a hypothesis testing and rejected the null hypothesis(Ho= there is no change in minutes being late to office). In this phase, I collected the data before and after process improvements. The p-value from the analysis is about 2.88E-08, which is lower than the alpha(0.05) defined. I am also able to build a multiple regression model. With the model, I got Multiple R squared as 0.82, which provides the variance in minutes being late to the office is explained by the predictors.

Improve

In this phase, I have identified which tasks need to be optimized. I felt that checking mobile on the bed and checking office mails is a redundant process. Once I removed those processes, my process map became more apparent. I calculated my process's "mean" and SQL(Sigma Quality Level) for the improvement validation. SQL improved from 2.91 to 3.4. mean minutes late for office was reduced from 37.46 to 12 min

Control

I used the X-bar chart and Moving Range chat to see whether my process is in the Upper Control Limit(UCL) and Lower Control Limit(LCL) to ensure no anomalies.

What is my overall learning?

This class introduced me to so many new concepts. I learned many processes like DMAIC and tools like process maps and visualization tools like Pareto charts. I learned the concepts of hypothesis testing and how to build a null and alternative hypothesis. I understood "statistical significance" and how to prove it using the calculations. I also understood the importance of defining a clear problem statement. Understood the concept of variation and how it is everywhere, including the measurement process. The control chart tools like XmR charts provided me with how to assert anomalies in the process. Overall, I am confident now to represent and deliver a strategy to improve the process and analyze whether it helped the goal. This course gave me all the essential tools to evaluate a process.

Overall, I felt that one of the best things was using my data and seeing how the tools helped improve the overall process.

IST 718: Human Protein Atlas Single Cell Classifier

About the class and the project

IST 718(Big Data Analytics) is the most demanding class I took M.S. Applied data Science stream in my view. The instructor has academic background and works in a similar domain and the real-world experiences shared were very useful and relatable in my work field. The instructor encouraged us to select real-world, large-scale problems in this class and provided pointers on Kaggle. I partnered with my classmate Sharat Sripada on this Project. Sripada and I were interested in experimenting with image classification using deep learning. We both finally chose the active Kaggle competition to do a single cell classification of human protein outlined here: https://www.kaggle.com/c/hpa-single-cell-image-classification

Project details/execution

The images in the competition had about 19 types of cells. When we first looked at the competition, the leaderboard had about 44% success rate, incorrectly detecting the cells. Our project goal was to produce at least 44% accuracy in classifying cell images.

The training data set has about 21k images with 2048 X 2048 resolution. Each image has more than one cell image. The training data set has what type of cells are present in the picture. The training dataset does not have any information on what part of the image belongs to which cell.

Sharath took most of the data's EDA(Exploratory Data Analysis) and built the report. I was responsible for building the suitable dataset for the training and building the deep learning model as I have an NVIDIA RTX 3090 graphics card and used the Google Tensorflow framework to build the models.

Exploratory analysis of the data:

From EDA analysis, we found that Cytosol, Nucleoplasm, Plasma membrane are the most common type of cells present using the histograms. Association data analysis on the training set was also done, which provided a strong association between "Microtubules -> Mitotic spindle" and "Nucleoplasm -> Cytosol -> Mitotic spindle" cell combinations in the given training images.

Deep learning modeling on images:

Our first challenge was to find out which image refers to which cell. The Training image has multiple images and information given on what cells it contains, but it was not clear which part of the image belongs to which cell. I used the Cell Segmentor utility referenced in the Kaggle competition. The cell segmentor will detect each cell boundary. To build single cell to cell type mapping, I choose the training images with only one type of cell. I used python programming and cell segmentor utility to split the images from the source image. I saved each split image as the specific cell type. As I chose the images with only one type of cell present, I tagged all the images to the specific cell type. Out of the 21k images I had, about 10,412 images had a single cell type. I got about 151,707 single-cell images with correct cell type mapping with the approach above.

After the image library building, I removed the images under 4kb as those images do not have much detail. I also removed images larger than 1 M.B. as compression of those images will create artifacts that could mislead the model building. I also restricted our scope to work on only 12 classes of cells instead of 19 classes as we did not have a good amount of samples for other classes.

Out of 151,797 samples, we split them into 113,780 training samples, 28,445 as validation samples, and 9,482 as test samples. We used stratified sampling over all the 12 classes.

For building the model, we used the Transfer learning approach to use the base model as **VGG16**, **ResNet101,** and **EfffficientNetB6** models as our base models. The system which ran the modeling has AMD 5800x 8C 16 thread CPU, 128 GB RAM, and NVIDIA RTX 3090 with 24 G.B. dedicated video memory.

Of all the models built, EfffficientNetB6 was the top-performing model. This model detected the Nucleoplasm with 82% accuracy, Nuclear speckles with 79% accuracy, and Mitochondria with 52% accuracy. The lowest accuracy was found for the Endoplasmic reticulum at 52%, which also beat the original goal we had set. Each model build took about ~3days continuous run.

My learnings

In this course, the instructor challenged us with real-world problems. In this course, I experimented with data science using python language. I learned the how-to-use spark and did spark batch processing. I learned deep learning principles and real-world frameworks like

Tensorflow. I also learned how to use Jupiter notebooks in data science projects, which are phenomenal. While other classes either took the approach of a small sample or academic dataset that fits within a spreadsheet to do analysis, this course dealt with enormous quantities of data that may not fit in a single computer. The examples are practical and real-world problems. Each experiment took time to do, so planning was also important. After this class, I felt I could handle real-world big data data science problems.

IST 719: "Sony Rules Gaming" poster project

About the class and the project

This class is about understanding the audience and providing appropriate visualizations to provide insights. The instructor requested that we find a large dataset and build a data visualization poster using Adobe illustrator. I choose Video Games Sales Dataset from Kaggle. My poster provides visualizations about why SONY corporation is the king of the video game industry. The dataset has about 16,720 observations. Each observation has about 13 features.

Project details/execution

Data for the Project was collected from Kaggle. The data is about video games on three major platforms(Sony, XBOX, and Nintendo). There are about 16,720 games present in the dataset. The project was done in 2 phases. The first phase provides the proposed visualization's Work In Progress(WIP). The final project is about posted created using illustrator and exported into PDF.

I planned a layout for the Work In Progress and provided the poster proposal by taking a picture of the whiteboard diagram I drew using markers.

For the plots in the project, I used R with the ggplot package. I used additional packages like word cloud for a word cloud plot. I used dplyr, tidyr for data cleanup, and data transformations.

Data cleanup: I filtered the data to have consoles and handhelds only from Sony, XBOX, and Nintendo as I scope my focus to look into these three corporations. I found the large title stings for word cloud generation and converted them to short acronyms to fit correctly in the word cloud. I filtered the data to use data only up to 2016 and removed games where the year was not present.

The final poster was developed using a 2 column layout. The central theme of the poster was an image of sales aspects of 3 key console brands that drew farther audiences towards the poster. I leveraged Pareto charts and word clouds to provide more context on which consoles were the best and which gaming company is popular for a nearer audience. Overall, I offered data visualization as to why sony was the king of video games.

My learnings

Before this course, I was never into visualizations. I took this course to understand how to tell good stories with data. The course helped me understand the visual design principles and helped me learn Adobe illustrator. The course also taught me the target audience and prepared the right visualization for the specific audience to get the proper attention. This class also taught me how to enhance standard plot output from R to provide striking visuals to convey information.

IST 769: Advanced Database Management

The Three V's of Big Data: Volume, Velocity, and Variety. This class, in my view, is a window to different practical databases (tools) available in the current market that can handle various aspects of the 3 V's of the data.

This class provides a detailed view of the CAP theorem. This class also introduced the concept of polyglot persistence and how to achieve scale and performance in processing petabytes of data. This class also taught NoSQL databases specialized for specific purposes and how they fit in polyglot persistence to help in data processing and analyzing.

This class provided advanced aspects of modern RDBMS systems like temporal tables to help with a point-in-time analysis. Provided the need for transactions and implemented them in MS SQL Server systems.

The next section of the class introduced a map-reduce framework to process large amounts of data using Hadoop. A great deal of time spent understanding MapReduce and map reduces programming through Pig scripting. Then the class introduced the other tools like Hive and Impala that plug into the Hadoop framework to help with analytical queries using familiar SQL language and leveraging high-performance query engines over Hadoop.

The last section of the class introduced various NoSQL databases like Mongo, Redis, Cassandra, and distributed event platform Kafka. The class provided how each database is unique and its possible role in polyglot persistence.

My learnings

I am exploring a career path in Data Engineering combined with Data Science. I am currently working as a software engineer. I have some familiarity with RDBMS and NoSQL systems. Having first-hand experience in current generation database technologies will help further my skills in the field. This class gave a practical overview of existing database systems and how polyglot persistence help solve data storage and processing challenges. This class helped me equip myself with the details to build an appropriate data platform for customer needs.

Conclusion:

My journey with Syracuse Masters in Applied Data science started first to understand the basics of statistics. The foundation of hypothesis tests, understanding of confidence intervals, Bayesian approach were crucial to start the Machine Learning journey. The initial parts of the Machine learnings started with Linear modeling and later expanded to Data Mining to explore Association rules, clustering, and text mining. Everything was hands-on using the R and R Studio ecosystem. Then the learning stepped up, discussing the big data. Processing large datasets reinforced the root concepts to clean the data and leverage pipelines like Apache spark to process millions of records. The projects in big data class taught me python programming and how to leverage Python libraries for Data Science. Data visualization class taught me how to present the analysis based on the audience. Advanced Big Data Management class provided details about the current DataBase systems and how polyglot architecture is helping to solve some of the large big data workloads.

Overall I learned the following during my tenure:

- Understanding the statistics needed for data science.
- Importance of data cleanup, transformation, and scaling.
- Ability to do appropriate sampling for the research task
- Conduct hypothesis tests and provide evidence
- Machine learning leveraging R and Python, Google Tensorflow, R Packages(ggplot .. etc.), pandas, Apache Spark. Databases using docker containers.
- Tools like R Studio, Jupyter notebook, Adobe Illustrator
- Developing Machine learning models. Evaluation of models and optimization.
- Interpret results and provide visualizations to the target audience

The applied part f the course is what made the real difference. The early classes have the right amount of theory related to Data Science fundamentals, and after that, each class is followed by intensive hands-on exercises and projects. I did some projects individually and did some projects collaboratively. During the collaboration, I learned about a new point of view from non-engineering background co-students which provided more insight to me to understand the audience better. It helped me build common ground through the learnings used from data visualization.

After the course, I am confident that I can outline a clear research problem. I can clean munge data appropriate to the research scenario. I can build models to help answer the research question and evaluate them. I am confident that I can provide results based on the target audience.

Student Information

- Name : Srihari Busam
- SUID : 742162654
- NetID : SBUSAM
- Email : sbusam@syr.edu
- Program: M.S in Applied Data Science
- Linkedin : https://www.linkedin.com/in/srihari-busam/
- GIT repo: https://github.com/srihari-busam/MS-ADS-Portfolio
    - Repository is public
    - **RESUME:** https://github.com/srihari-busam/MS-ADS-Portfolio/blob/main/SrihariBusam-Resume-2022.docx