



Master’s in Applied Data Science portfolio

**Name: Sourabh Ghosh**

**SUID: 730661878**

**Contact Number : (315)-880-2597**

**Email :** [**sghosh08@syr.edu (current)**](mailto:sghosh08@syr.edu%20(current))**,** [**sg988079@gmail.com(primary)**](mailto:sg988079@gmail.com(primary))

**Contents**

[**1. Resume** ………..3](#_Toc66740666)

[**2. Introduction** 4](#_Toc66740667)

[**3. IST 718 Big Data Analytics** 5](#_Toc66740668)

[**a. Course and Project Description** 5](#_Toc66740669)

[**b. Reflection and Learning Goals** 5](#_Toc66740670)

[**c. GitHub Link** 6](#_Toc66740671)

[**4. IST 731 Artificial Neural Networks** 6](#_Toc66740672)

[**a. Course and Project Description** 6](#_Toc66740673)

[**b. Reflection and Learning Goals** 6](#_Toc66740674)

[**c. GitHub Link** 8](#_Toc66740675)

[**5. IST 687 Introduction to Data Science** 8](#_Toc66740676)

[**a. Course and Project Description** 8](#_Toc66740677)

[**b. Reflection and Learning Outcomes** 8](#_Toc66740678)

[**c. GitHub link** 10](#_Toc66740679)

[**6. References** 11](#_Toc66740680)

# 

# **Resume**

**SOURABH GHOSH**

[sg988079@gmail.com](file:///E:\\Disk%20partition\\Syracuse%20ADS\\Syracuse%20ADS\\FULL%20TIME%20Applications\\v6\\sg988079@gmail.com) | 315-880-2597 | <https://www.linkedin.com/in/sourabhghosh29/> | <https://github.com/sourabhghosh29>

**EDUCATION**

**Master of Science – Applied Data Science Aug 2019-May 2021**

**Syracuse University, School of Information Studies, Syracuse, NY**

Coursework : Big Data Analytics, Financial Analytics, Business Analytics, Data Analysis and Decision Making, Machine Learning, Data Warehouse & DB Management, Artificial Neural Networks, Cloud Computing, Deep Learning/NLP in Computational Social Science

**Bachelor of Engineering - Computer Science and Engineering Aug 2011-June 2015**

**Bhilai Institute of Technology, Durg, India**

**SKILLS  
Programming Languages:** Python, R, SQL, C++, Unix Shell Scripting, Javascript

**Datawarehouse & BI tools:** Teradata, Hive, Oracle, MS SQL Server, Amazon Redshift, PostgreSQL, MS Access

**Big Data & Analytics tools :** Spark, Hadoop, Airflow, Google Analytics, Elastic Search, MS Excel(VBA, Macros)

**Data visualization tools & Cloud tech :** Tableau, MS-Power BI, AWS, GCP, IBM Cloud, Docker, Kubernetes

**ML/Quant libraries:** Pandas, NumPy, Scikit-Learn, TensorFlow, Keras, Matplotlib, Caret, Dplyr, Quantmod, Spark-Mllib/SQL, Gensim, NLTK

**PROFESSIONAL EXPERIENCE**

**Artificial Intelligence Strategy Intern Sprint – T-Mobile June 2020 – Sep 2020**

* Built computationally cost-efficient model reducing the complexity for automatic memo generation based on user-agent conversation for quick issue resolution. Performed Cohen’s Kappa test for the training data reducing bias in AI algorithm, improving the prediction accuracy
* Improved customer’s AI journey completion experience using profiling and predictive modeling, boosting the AI Net Promoter Score for making a payment arrangement by a user using chatbot
* Conducted RCA, to analyze the channels and factors declining the KPI’s used for assessment of Sprint’s Virtual Assistant
* Improved F1-score(Macro) by 4% for widely used use cases involving customer intents using Conversational AI, IBM Watson
* Designed and automated AI - Quarterly Gross Progress report for “Bot Analytics” reducing manual labor of the analytics team by 95%

**Data Analyst, System Engineer Tata Consultancy Services Nov 2015 – July 2019**

* Developed predictive mining techniques for identifying threshold breach in KPI’s, facilitating automatic ticket creation in case of fault/issue
* Designed dashboard using AngularJS to track user engagement within different modules of web page, navigation, likes and comments
* Enhanced data uploading performance by 80% for a list of 1 million records for client critical application used to store, manage, create, and guide the flow of purchase orders using Java Microservices architecture
* Developed web app UI/UX which used speech recognized commands based on Web Speech Tool Kit, Synthesis, Google Dialog Flow
* Led proof of concept for automation and migration of the patch delivery application from Sun Solaris to Linux with 98% defect-free delivery

**Graduate Research Data Scientist SOS & Computational Discovery Lab, Syracuse University Sep 2019 – Present**

* Optimized “Publications and Grants” search and recommender system leveraging topic modeling using TF-IDF for more than 32 million documents. Used Apache Airflow and Kibana Elastic Search improving data loading performance by 40% (<https://eileen.io/>)

**PROJECTS**

**AWS Cloud Infrastructure Based Online Betting System :** Flask | AWS (EC2, EBS, Sagemaker) | Python **Aug 2020 - Dec 2020**

* Created AWS cloud infrastructure involving Sagemaker for predictive analytics leveraging classifiers using UFC’s historical fights data
* Determined factors having the most telling impact on the outcome of a match, by engineering features across 882 attributes hosting analytics dashboard which could be utilized for Team Managers, Investors and Hires for selecting and betting on fighters

**Reinforcement Learning in Dialog Systems(Task Completion Chat Bot) :** Python | Q-Learning | Keras **Aug 2020 - Dec 2020**

* Train the chatbot with seq2seq model using LSTM based architecture to process dialogue input and generate textual response
* Leveraged reinforcement learning using Deep Q-Network for policy learning in chatbot that adapts itself based on user input
* Implemented reinforcement algorithms like Deep Q-network in a computationally efficient and statistically effective manner

**Tv News Commercial Detection :** PySpark |Logistic Regression | Random Forest | GBT **Aug 2019 - Dec 2019**

* Implemented a classification model to predict TV commercial from audio and visual features by building pyspark pipelines
* Used PySpark ML to train Logistic regression, Random Forest, and Gradient boosting trees models, performed hyper-parameter tuning and assessed generalization performance with best model having 95% testing AUC score

**Improving Airline Customer Satisfaction - Churn Prediction :** R | Classification | Clustering | Recommendations **Aug 2019 - Dec 2019**

* Analyzed airline customer satisfaction data to identify any potential gaps and generated actionable insights
* Applied several data mining techniques like Logistic Regression, XGBoost classifier, K-Prototype Clustering techniques and association rules mining in R to identify most critical features affecting satisfaction recommending improvement in marketing strategies

**LEADERSHIP**

* Spearheaded 3 teams for 3 applications, liaising with the client in development and maintenance projects in TCS Ericsson Telecom domain
* Managing class of 15 in understanding concepts of Global Financial System Architecture at Syracuse university

# **2. Introduction**

The depth of intelligence provided by the geometric growth of data and the way it could coerce an impact in the world has driven me to my ongoing quest to pursue a Master of Science in Applied Data Science from Syracuse University, ISchool. The learning outcomes from the coursework have sculpted me in a way that enables me to understand, strategize and build data science solutions around the problems that challenge modern-day organizations. Being from a strong Computer Science background and working in tech-industry for 3.5+ years, I built an interest in gathering insights and using data science to help an enterprise in taking strategic business decisions in an informed way. Studying at ISchool, along with all the courses were extremely beneficial for me to develop an understanding of the life cycle while building data-driven solutions and impacting business strategies through decisions.

The curriculum of the core and elective courses laid the foundation of the statistical concepts and technical frameworks that enabled me to understand the tools and approach while undertaking business problems. During my first semester coursework like IST 687, MBC 638 helped me to understand the basic statistics, reporting, visualization, and machine learning which is important for professional seeking learning in the field of data, whereas IST 718 helped me gain a broader understanding of algorithms and measures to judge the developed solutions. In the following semesters courses like IST615 emphasized the importance of Cloud Infrastructure Solutions, I also took Deep Learning in Computational Science using which I understood the impact of data science solutions in society and how to overcome it. Some of the elective courses that I took were Artificial Neural Networks that built my understanding of complex solutions and underlying advanced statistics that Neural Network black box are built on. The coursework also focused on subjects at the application level in collaboration with Whitman School of Business such as Business Analytics elaborating different business use cases and tools that could be leveraged to help make data-driven decisions using insights and recommendations. Financial analytics helped me understand why and how to deal with financial data underlying few techniques such as portfolio optimization.

Ranging from extracting, cleaning, manipulating, visualizing, and machine learning modeling to building solutions to present and communicate useful business insights to the leadership the courses designed in the ISchool helps to gain key-skills required for all the tasks. Working on the final projects for the courses helped me put my theoretical learnings to hands-on experience and test my potential developing project deliverables and learnings through reports, code, application, and presentation. ISchool also provided me the opportunity to work with Prof. Daniel Acuna at the Science of Science and Computational Discovery lab and Prof. Jason Dedrick as a Teaching Assistant for the Global Financial System Architecture course. I also got the opportunity to work at T-Mobile during my summer intern 2020 as an Artificial Intelligence Strategy Intern wherein I used my skills acquired from my learnings in data science at ISchool. The blend of collaboration among peers in different projects, from different backgrounds and faculties with extensive experience, gave me broad proficiency and learnings, which I will take with me in my future ventures and expeditions. I am proud to be a SU student and satisfied with the learning outcomes that I gathered.

**Projects**

# **3.** **IST 718 Big Data Analytics**

## **a. Course and Project Description**

This course elaborates and explains the underlying principles of advanced machine learning algorithms and the statistical principles on which these models are based on. It also covers concepts and architecture for Big data that industries work on and working of distributed processing framework such as Hadoop and Spark with an in-depth understanding of concepts such as MapReduce on which this framework works. It is one of the best and key courses that covers techniques and best practices one should undertake while handling big data.

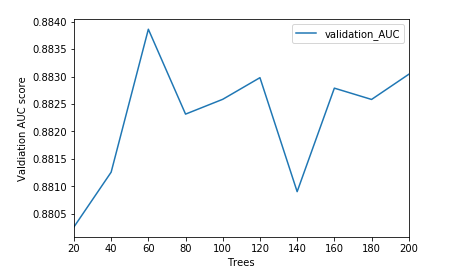
Project: TV News Commercial Detection

In this course, we were open to select a data problem of our choice working as a Data Scientist in a Big Data scenario. The goal in our final project was automatic identiﬁcation and extraction of commercial blocks in telecast news videos. Extraction of commercials is an important preprocessing step in broadcast media analysis, and evaluation of video analysis is crucial for competitive marketing analysis and advertising planning. News channels do not follow any particular news presentation format, have large variability and dynamic nature. Hence, automatic identification of commercial blocks using machine learning approaches provides a viable alternative.

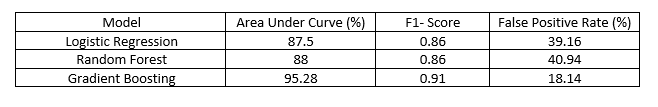
## **b. Reflection and Learning Goals**

I worked with a team including Mihir Deshpande, Yash Shah, and Mansee Agrawal to solve the data problem consisting of more than 1.2 million records and 205 features. The goal was to identify a block of video as commercial or not by looking at the audio and video features of the data. We performed advanced data pre-processing techniques such as mean imputation and stratified sampling to deal with the initial life cycle of data. Further, we implemented classifiers to measure their performance based on generalization such as Area under the curve, F1- Score, and False Positive Rate. The machine learning models implemented were further applied using an interactive widget which allowed the user to select model and inputs to determine if it was a new commercial video or not.

Experiment Outcome : Hyper parameter tuning for random forest resulting in tuned and optimized model



Model Comparison : Model performance comparison to determine the best model for this problem statement.



This course helped me test my skills coding in pyspark and implementing advanced machine learning concepts like regularization and hyperparameter tuning in classifiers to select the best model performance. It also helped me to understand scenarios where data has a high dimension of features/attributes and ways to reduce them to visualize and strategize which particular features in the problem statement contribute to decision making in a supervised machine learning scenario.

## **c. GitHub Link**

GitHub Link : <https://github.com/sourabhghosh29/news_commercial_bda>

# **4. IST 731 Artificial Neural Networks**

## **a. Course and Project Description**

Artificial Neural Networks was among one of the electives that I enrolled in and was important in understanding neural networks in depth. The course covered theoretical, use cases, statistical and practical concepts in solving complex problem statements and underlying neural network black box architectures used in them. It helped to gain practice in neural network frameworks such as TensorFlow and libraries such as Keras while building neural networks and covered a wide range of problems from Computer Vision, Forecasting, Classification in its assignments. The course also focuses on the evaluation of computational cost and determination of neural network parameters(number of neurons, hidden layers, etc.) in-depth that will lead to an efficient solution

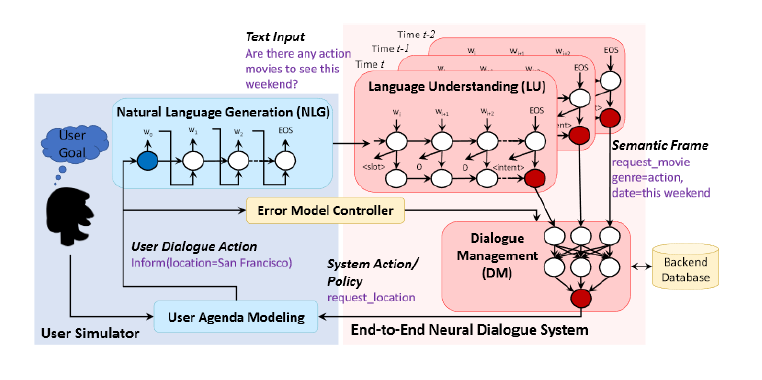
Project: Reinforcement Learning in Dialog Systems

I worked with Akhil Nair in the final project which was to build and optimize a task completion chat-bot proving and implementing reinforcement learning in a dialog system. Our strategy was to create a dialogue system(task completion chatbot) where the task is to book a movie ticket and further optimize the performance by simulating noise in the dialogue system components to achieve the desired goal of the user. Although this limitation can typically be managed in the context of dialog systems with adequate monitoring changes in behavioral flow, these errors may be noises in components such as Input from the User, Natural Language Processing components that deal with their own mistakes when understanding and manipulating user dialogues. To achieve the user objective, we also focus on the dialog framework that decides the appropriate policy given the dialogue state. These strategies are generally framed as a task of reinforcement learning in which the goal is to maximize the reward for those actions that result in a successful conversation.

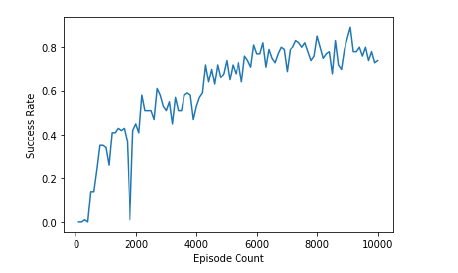
## **b. Reflection and Learning Goals**

The implementation was done using a user simulation module and a dialog manager consisting of Dialog State Tracker to incorporate agent state and user action(conversation), an Error model controller that introduces Natural language Understanding errors. Our experiment was around the DQN agent that inputs the state created in the state tracker which maintains and keeps a track of current conversation and selects the most optimum action for the dialogue. To approximate the value of the Q function, the agent uses a deep neural network (DQN) built using Keras and conducted several experiments with 1 and 3 hidden layers along with 5k, 10k episodes of dialog between simulated user and bot. Concluding 3 hidden layers Neural network trained on 10000 episode counts, the success rate crossed the threshold after 2000 episode count in contrast to that of 1 hidden layer trained network on 10k episode which crossed the success rate threshold in the first 15 episodes itself. The highest recorded success rate was 0.83.

High-level Functional Diagram of the Neural Dialog System representing interactions with different components implemented.



Experiment Outcome : Performing experiments with neurons and hidden layers to reach at statistically sound results for the systems performance.



This course helped me build strong fundamentals of Artificial Neural Network and helped me work in an unexplored territory of the Reinforcement Learning problem. It also enabled me to identify the situations where complex black box models like neural networks could play an efficient role.

## **c. GitHub Link**

GitHub Link : <https://github.com/sourabhghosh29/RL_Dialogue_Systems_GO-BOT>

# **5. IST 687 Introduction to Data Science**

## **a. Course and Project Description**

This course laid the foundation for understanding data science life cycle and phases such as data munging, manipulation, visualization, and modeling using R. I developed a base in understanding underlying statistical concepts and confidence in dealing with a scripting language like R through this course.

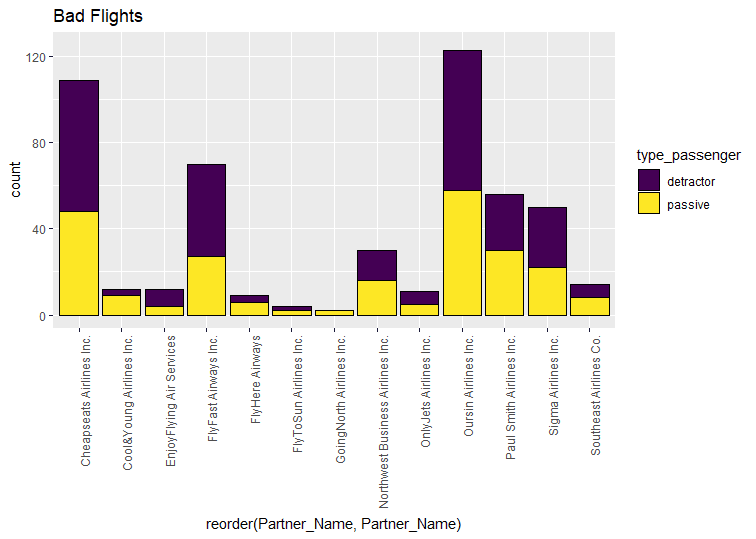
Project: Improving Customer Airline Churn

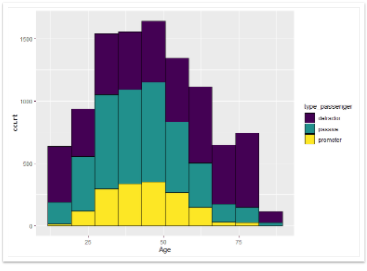
The final project tasked our team to act as Data Science Consultant working for South East Arline data which required recommendations and solutions to improve their customer satisfaction index(likelihood to recommend) decreasing churn rates. I worked with members from different backgrounds in ISchool Jimit Mistry, Vasundhara Patil, Sanjana Bhot, and Sanyukta Shandilya. We were provided with data consisting of 11000 records and attributes capturing 26 characteristics of the flight (ex. day of the month, date, airline, origin and destination city, if the flight was delayed), the customer (ex. age, gender, price sensitivity, the person’s frequent flyer status). The expected goal was to identify patterns and attributes that resulted in decreased customer satisfaction and recommend solutions to overcome/improve the Net promoter Score for South East Airlines.

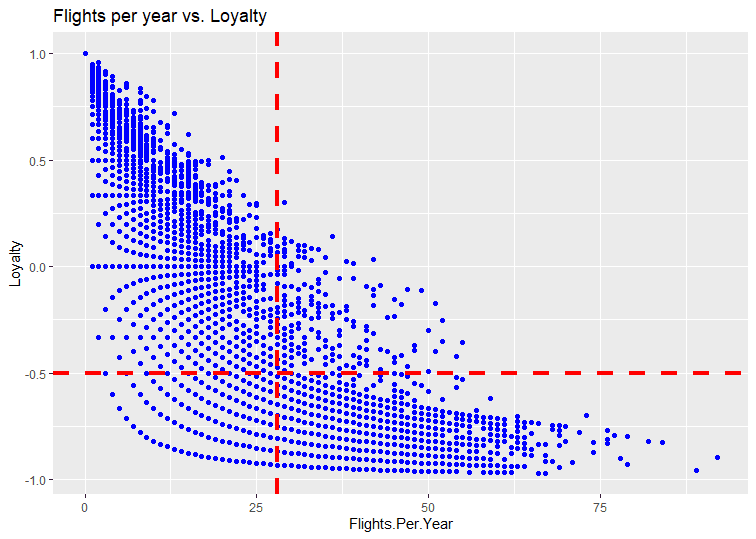
## **b. Reflection and Learning Outcomes**

We applied data pre-processing techniques such as outlier removal, data imputations, discretization, and visualization for understanding granularity and telling a clear story using the data in our presentation. We also implemented several data mining and classification techniques such as Logistic Regression, Random Forest, XGBoost, and Apriori Algorithm to understand pattern demoting likelihood to recommend and increasing churn. We came out with several recommendations based on customer flying patterns and demographic attributes affecting satisfaction, exploiting potential gaps, and developing better marketing strategies.

Recommendations visual representation based on deducted data mining results :







This course helped me understand the life cycle that data undergoes while building solutions in an organization. I understood the process of converting data to information and information to knowledge. Also, this course helped me gain hands-on expertise in the R programming language and added key skills to my experience.

## **c. GitHub link**

GitHub Link : <https://github.com/sourabhghosh29/airline_customer_churner>

# **6. References**

* <https://github.com/sourabhghosh29/news_commercial_bda>
* <https://github.com/sourabhghosh29/RL_Dialogue_Systems_GO-BOT>
* <https://github.com/sourabhghosh29/airline_customer_churner>