Data Science Portfolio

Tamilselvan Tamilmani

[1 Introduction 2](#_Toc97063352)

[2 introduction to data science | Prof. Gary Krudys 2](#_Toc97063353)

[3 Business Analytics | Prof. Kimberly Pendelberry 2](#_Toc97063354)

[4 Data Analytics | Prof. Jeremy Bolton 3](#_Toc97063355)

[5 Text Mining | Prof. Amy Gates 4](#_Toc97063356)

[6 Data Analysis & Decision Making | Prof. Kimberly Pendelberry 5](#_Toc97063357)

[7 Marketing Analytics | Prof. Raghavshyam Ramamurthy 6](#_Toc97063358)

[8 Big Data Analytics | Prof. Jonathan Fox 7](#_Toc97063359)

[9 Information Visualization | Prof. Raghavshyam Ramamurthy 8](#_Toc97063360)

[10 Natural Language Processing | Prof. Michael Larche 8](#_Toc97063361)

[11 Quantitative Reasoning | Prof. Gregory Block 9](#_Toc97063362)

[12 Conclusion 10](#_Toc97063363)

# Introduction

The Masters in Applied Data Science Program at Syracuse University offered me an opportunity to learn the end-to-end aspects of the data science practice. It helped me identify the business problem, develop strategies for the problems, collect and organize the data, analyze the data systematically by doing EDA, statical analysis, and data mining.

It also helped me effectively communicate the findings, both good and bad. Introduced the tools necessary to implement the business decisions. We will discuss in detail the learnings I had the opportunity to complete.

# introduction to data science | Prof. Gary Krudys

The main objective of this course is to perform data processing and modeling using R. And determine appropriate techniques for analyzing data.

Professor Krudy’s hands-on approach to R made language learning easy. He showed us the power of the ggplot package, which became my favorite plotting library. Below are the topics covered during the course

* Data Collection & Sampling
* Data Exploration – EDA - ggplot
* Geospatial Data Visualization - ggplot
* Linear Regression
* Data Mining
  + Association Rule Mining
  + Support Vector Machines
* Text Mining
  + EDA- Word Clouds
  + Sentiment Analysis
  + Topic Modeling

We utilized the data processing steps learned in the course to clean up the survey data and applied different machine learning techniques like SVM and linear regression on our final project. We uncovered more insights using association rules for this analysis, which gave us an idea of what real-world problems will look like. Sophisticated algorithms might not matter, just a systematic analysis to uncover the insights.

# Business Analytics | Prof. Kimberly Pendelberry

The course's main objective is to do data analysis and mine data, develop a strategy based on the data, and make decisions. This covered all three aspects of analytics: descriptive, predictive, and prescriptive.

Professor Pendelberry patiently answered all my questions. She introduced the basics of descriptive analytics, predictive model importance like R square, pitfalls like multicollinearity, serial correlation, and the prescriptive side of modeling using techniques like sensitivity analysis. Showed us the difference between linear, logistic, and neural networks. The in-class discussion about the MIT Sloan Management Review articles showed us a glimpse of how organizations are using analytics. Below are the topics covered during the course

* Google Analytics
* Tableau Dashboards
* Optimization
* Univariate & Multivariate Regression
* One-Way & Two-way Sensitivity Analysis
* Statistical analysis
  + Visualization – Histogram, Box, Scatter, Mean, XY Plots
  + Statistical Summaries, Correlation
  + ANNOVA
  + Regression
    - Dummy Variables
    - Moderating Effects
* Regression Diagnostics
  + Linearity Test
  + Collinearity Test
  + Heteroscedasticity Test
  + Serial Correlation Test
  + Outlier Test
* Bedford’s Law
* Decision Trees
* Choice Models
  + Logit & Probit Models
* Perceptron & Neural Networks

We get to practice descriptive analytics using google analytics with actual Whiteman site data. Predicted book prices and recommended actions for bank loan data using prescriptive analytics.

# Data Analytics | Prof. Jeremy Bolton

The objective of the course is twofold: one is applying data mining techniques to the real-world data set, the other is to effectively tell a story about the analysis in a valuable way.

Professor Bolton set up a high bar on the storytelling part, specifically making sure we submitted all our reports in a professional format. Below are the topics covered during the course

* Data Preparation - Summary Statistics, Aggregation, Transformation
* Classification
  + Decision Tree Analysis
    - C4.5
    - Information gain, overfitting & pruning
  + Naïve Bayes
    - Smoothing – Laplacian smoothing(add one)
  + SVM
    - Soft & Max Margin
    - Kernels – Polynomial
  + Random Forest
  + Ensemble Learning
* Clustering Analysis
  + K – Means, Tuning & Centroid Interpretation
  + HAC Algorithm - Hierarchical Agglomerative Clustering
  + kNN
* Association Rule Mining
* Text Mining
  + Tokenization
  + Vectorization

In the Federalists Paper homework, we applied all forms of clustering like k-means, hierarchical, and EM(Expectation-Maximization) and different distance measures like Manhattan, Euclidian, and Minkowski.

In the Handwritten recognition homework, we had changed to explore classification methods like a decision tree and naïve Bayes. We got introduced to feature selection/reduction methods like PCA.

We explored a time series data set on our final project and applied all the data mining techniques we learned. We started from association rule mining, k-means, KNN, SVM, decision tree, and random forest. We have tried new feature engineering techniques like lagged features.

# Text Mining | Prof. Amy Gates

This course aims to learn the basics of text mining like document representation, information extraction, text classification and clustering, and topic modeling.

Professor Gates is efficient in her approach, and she encouraged us to go beyond the curriculum and explore the state-of-the-art NLP models. And she set the highest bar in terms of report submission. Below are the topics covered during the course

* Document Vectorization
  + Tokenization
  + Vectorization
* Corpus Analytics
  + Regex
  + Lexicons
  + Comparative & Trend Analytics
* Text Classification
  + Multinomial Naïve Bayes – feature ranking
  + Decision Tress
  + SVM
  + Evaluation Metrics
* Scikit-learn
* Document Clustering & Topic Modeling
  + K-Means
  + LDA
* Bias in AI

We did sentiment analysis and topic modeling for the data from Twitter, Reddit, and news sites.

# Data Analysis & Decision Making | Prof. Kimberly Pendelberry

The objective of this course is to select the appropriate statistical technique for a given set of conditions to answer a particular question.

We explored hypothesis testing for continuous variables chi-square test for discrete variables. Get to try out the kappa agreement score using Amazon Mechanical Truck for sentiment data. We got an introduction to time series data and process controls from Professor Pendelberry. Below are the topics covered during the course

* DMAIC (Define, Measure, Analyze, Improve, and Control)
* Sigma Quality Levels
* Kappa Technique
* Probability Distributions & Hypothesis Testing
  + Normal & Binomial distribution
  + H0 & Ha Scenarios
  + Alpha & Beta risks
* Categorical Data Analysis
  + Chi-Square Test
* Confidence Intervals
* Simple & Multiple Linear Regression
* Process Control Charts
  + X-bar/R, ImR
* Timeseries Analysis
  + Autocorrelation
  + Auto Regressive Models, AR1
  + Moving Average Models
  + Exponential Smoothing Models

The final process improvement project allowed me to define the problem using my data.

# Marketing Analytics | Prof. Raghavshyam Ramamurthy

The objective of this course is to develop marketing strategies and resource allocation decisions driven by quantitative analysis

Professor Ramamurthy took us through the complete journey with marketing analytics starting with Customer Segmentation, Marketing mix models, Choice, and Churn models, and Conjoint analysis. We are exposed to all these models already, but the practical application of the models to business problems proved to be very tough. Below are the topics covered during the course

* Marketing Resource Allocation
  + Opportunity
  + Resource Allocation
  + Optimizing Sales
  + ROI
* Cluster Analysis
  + Segmentation using k-means
  + Profiling Segmentation
* Regression
  + Omitted Variable Bias
  + Price Elasticity
  + Statistical vs Economic significance
* Customer Lifetime Value
  + Base CLV
  + Cohort
* Propensity models
  + Market Share Prediction
  + Hit Rates
  + Association Rule Mining
* Churn Prediction
  + pAlive Calculation
  + Logistic Regression
* Choice Models
  + Utility Theory
  + Conjoint Analysis
  + Cross-Selling
* Marketing Experiments
  + Marketing-Mix models
  + Test & Control groups
* Collaborative Filtering

Our final project gave us a chance to explore all the marketing analytics techniques we learned, and the simulation gave us an idea of how hard it is to run a multi-year marketing campaign.

# Big Data Analytics | Prof. Jonathan Fox

The objective of the course is to complete the entire lifecycle of data science, from collecting data to delivering insights.

Professor Fox made the class, competitive and fun. I created the highest number of models in this class. He took us through the full OSEMIN (Obtain, Scrub, Explore, Model & Interpret) cycle. Below are the topics covered during the course

* OSEMiN
* Python & Spark
* Data Review
* Model Review
* Code Review
* Consumer preference
* Evaluation Methods
* Forecasting – ARMIA
* Poisson & Negative Binomial Distributions
* Decision Trees & Random Forest
* Neural Networks
* Convolution Neural Networks
* Presentation
  + Elders Rule

Each homework assignment pushed the limits on what we could do. The Football dataset homework made us crawl the web to augment external data. The time-series Zillow dataset made us run 30 thousand models, and finally, the Fashion MINIST data set made us compare all the classical approaches against Neural Networks.

The Final Project, “COVID Detection from X-Ray Images,” gave us a chance to produce meaningful work and showcase the power of transfer learning in neural networks.

# Information Visualization | Prof. Raghavshyam Ramamurthy

The objective of the course is to learn and practice advanced visual storytelling.

Professor Ramamurthy is hands-on in transforming the basic SVG plots created in R to beautiful data stories in Adobe Illustrator. Below are the topics covered during the course

* Graphic Design Principles
  + Colors
  + Layouts
  + Element Design
* ggplot
* Social Network Analysis & Graphs

The final Netflix analytics Project allowed me to convey my data story in a visually appealing way.

# Natural Language Processing | Prof. Michael Larche

The objective of the course is to understand the levels of linguistic analysis and apply them using the NLTK library

Professor Larche made the class interactive by using the async assignments in class and asking questions to explain the concepts. He methodically introduced regex, Parts of Speech tagging, context-free grammar, sentiment analysis and finally ended in deep learning. Below are the topics covered during the course

* Corpus Statistics
  + Unigrams & Bigrams
  + Mutual Information
* Language Modeling
* Morphology
* Parts-of-Speech Tagging
* Context-Free-Grammar
* Semantics
* Discourse & Dialogue
* NLTK
* Sentiment Analysis & Classification

The final project was structured in a way we get to use the classical NLP methods along with deep learning. Although we learned the classical NLP, working on deep learning made us question the need for the classical approaches.

# Quantitative Reasoning | Prof. Gregory Block

The objective of the course is to advance our understanding of the inference of the data by comparing Frequentist and Bayesian methods.

Professor Block's had a hands-on approach in explaining the frequentist and the corresponding Bayesian equivalents like High-Density Interval, Markov chain - Monte Carlo methods, and Bayes Factor. Below are the topics covered during the course

* Population & Samples
* Law of large Numbers
* Frequentist vs Bayesian approaches on
  + ANOVA
  + Cross-Products/Pearson Product-Moment Correlation
  + Chi-Square Distribution
* Linear Multiple Regression
  + Least Square Criterion, Sum of Squared Errors
  + Multicollinearity
  + Bayesian Inference
* Logistic Regression
  + GLM & Link Functions
  + Bayesian Estimates
* Timeseries
  + Changepoint Analysis
  + Repeated Measure ANOVA
  + Trend Removal
  + Diagnosing and Testing Stationarity
  + ARIMA

The Final exam allowed us to apply all the Bayesian techniques and compare them with the corresponding frequentist methods.

# Conclusion

The program increased my knowledge and confidence in the data science practice. Introduced me to unique tools to help me understand the data communicate effectively. Changed my perspective on how to approach the business problem.

All the work I have done in the program is consolidated in the below GitHub repository. The repository is organized by the order I took the course. The individual work is organized as subfolders, and the contents are self-explanatory.

<https://github.com/tamtam0/datascience_portfolio>