Portfolio Milestone
Learning Goals

MS Applied Data Science

Ryan Timbrook
NetID: RTIMBROO
SUID: 386792749
EMAIL: rtimbroo@syr.edu

Course: Portfolio Milestone
Term: January/Winter, 2020

Syracuse University School of Information Studies, in collaboration with the Whitman School of Management

Table of Contents

1 Introduction

As a Software Development Manager with 20 years of industry experience, continuous learning and self-growth in the latest technology are a necessity for remaining relevant in your job and seen as a Leader in your chosen domain. Longevity comes from choosing a path of life-long learning. Since graduating with my undergraduate degree in 1998, I've continued my education through formal and informal channels. It's included a diploma in Computer Programming, graduate certificates in both Natural Language Technologies and Data Science from the University of Washington, industry-recognized technical certifications such as Oracle Certified Java Programmer, Certified Python Programmer, Certified Agile Scrum Master, Certified Agile Product Owner, as well as specialized domain certifications, as in Contact Center Technologies, Genesys Certified Professional - Systems Inbound Voice.

For any technical leader, the knowledge and skills learned in Applied Data Science is a natural fit for the ever-increasing responsibilities placed on them. With the technology boom in Big Data, AI, and advancements in Data Mining techniques and tooling, most Companies are adopting the Data-Driven approach for running their organizations. In my field of technology management, to be a Leader, you have to be a hands-on technician, able to architect, design, and write code when necessary.

Graduates of the MS Applied Data Science program at Syracuse University School of Information Studies, in collaboration with the Whitman School of Management, must be able to demonstrate that they have been able to master each fundamental aspect of this discipline, while also being able to synthesize their individual ability to analyze, interpret and recommend actions to stakeholders in organizations when challenged with new operational problems to solve.

The intent of this Portfolio Milestone is for the student to assemble evidence and reflect on how each course they have taken has contributed to their acquisition of the cognitive strategies defined in the program learning outcomes, and how this has enabled them to become professionally prepared in their chosen area of specialty. Within the Learning Objectives section of this document, the student will use references from their course projects or assignments and reflect on how their work has demonstrated their mastery of these concepts.

1.1 Scope

The Learning Outcome Objects for this program are:

1. Describe a broad overview of the major practice areas in data science
    a. Project Alignments:
        i. All Projects listed in section 4, References Applied Data Science Courses
2. Collect and organize data
    a. Project Alignments:
        i. IST 736 Text Mining - Public Sentiment Toward NFL Teams, Coaches, and Players
        ii. IST 707 Data Analytics - Real Estate Property Investment
        iii. IST 659 Database Administration Concepts and Database Management - A4B KPI BizOps Organization Hierarchy
        iv. MBC 638 Data Analysis and Decision Making - Purchase Order Process Improvement
3. Identify patterns in data via visualization, statistical analysis, and data mining

   a. Project Alignments:
      i. IST 718 Big Data Analytics - IEEE-CIS Fraud Detection
      ii. IST 736 Text Mining - Public Sentiment Toward NFL Teams, Coaches, and Players
      iii. IST 707 Data Analytics - Real Estate Property Investment
4. Develop alternative strategies based on the data
   a. Project Alignments:
      i. SCM 651 Business Analytics - Recruiting Advertising Strategy
      ii. IST 707 Data Analytics - Real Estate Property Investment
5. Develop a plan of action to implement the business decisions derived from the analysis
   a. Project Alignments:
      i. MBC 638 Data Analysis and Decision Making - Purchase Order Process Improvement
      ii. IST 659 Database Administration Concepts and Database Management - A4B KPI BizOps Organization Hierarchy
6. Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization
   a. Project Alignments:
      i. MBC 638 Data Analysis and Decision Making - Purchase Order Process Improvement
      ii. IST 736 Text Mining - Public Sentiment Toward NFL Teams, Coaches, and Players
      iii. IST 718 Big Data Analytics - IEEE-CIS Fraud Detection
7. Synthesize the ethical dimensions of data science practice (e.g., privacy)
   a. Project Alignments:
      i. IST 659 Database Administration Concepts and Database Management - A4B KPI BizOps Organization Hierarchy
      ii. IST 718 Big Data Analytics - IEEE-CIS Fraud Detection

2 Learning Objectives

2.1 Major Practice Areas in Data Science

2.1.1 *Overview*

Learning Objective: Describe a broad overview of the **major practice areas in data science.**

Below is a reference list breakdown of some of the major practice areas in data science that will be referenced by the course project summaries.

- Data Engineering
- Data Mining and Statistical Analysis
- Database Management and Architecture
- Business Intelligence and Strategy
- Machine Learning / Cognitive Compute Development
- Data Visualization and Presentation
- Operations - Related Data Analytics
- Marketing - Related Data Analytics
- Industry Domain-Specific Data Analytics

2.2 Collect and Organize Data

2.2.1 *Overview*

Learning Objective: Collect and organize data.

Data collection is the process of **gathering** and measuring **data**, information, or any variables of interest in a standardized and established manner that enables the collector to answer or test hypotheses and evaluate outcomes of the particular **collection**. - https://en.wikipedia.org/wiki/Data_collection

I manage a large DevOps organization for the third-largest U.S. telecommunications company. The primary function of the systems my teams build and maintain is for the CARE organization. Every customer who calls the customer service line is first met by the Voice Self-Service IVR and Call routing applications we build. These systems process one million phone calls daily, servicing a customer base of 90 million users. As could be expected, gigabytes of data are generated by these transactions daily. Data comes in the typical forms of systems operations, application logs, network traffic, as well as software delivery pipeline metrics that measure and automate code build and deployments. It also comes in forms unique to speech recognition, natural language systems, such as ASR Grammar Utterances (everything a user speaks into their phone) and audio .wav file recordings of users speaking with the IVR system and Agents. To better service both our internal customers and external customers, this data is collected, organized in many different ways, and analyzed to enhance both our systems' performance as well as our DevOps teams code delivery.

Below are a few characteristics regarding the collection of data that will support our understanding of these learning objectives' unique challenges.

Three Key dimensions to assess data:

- Provenance - do you trust that source, what level of quality can we expect in the data?
- Legality - essential to understanding what is and isn't allowed
- Sensitivity - breaching some ethical boundaries

Big Data - the new natural resource:

- Observations captured one by one and entered manually on paper or in the computer
- Human activity on the web leaves traces captured by various entities
- Internet of things - streams of data automatically obtained from sensors or human activity into databases or sophisticated graphs

Many 'mountains' of data:

- Cost of storage low, # of devices/sensors higher every year
- Reasons to store: financial vs. other (competitive advantage)
- Creating Data + Metadata (data about data)

Big Data characteristics:

- Volume: Data at Rest, Terabytes to exabytes of existing data to process
- Velocity: Data in Motion, Streaming data, milliseconds to seconds to respond
- Variety: Data in Many Forms, Structured, unstructured, text, multimedia
- Veracity: Data in Doubt, Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Who produces/collects data and for what purposes:

- Users/individuals: They like to keep/share what they create. "Sentimental value" vs. "make money." Data mostly shared
- Businesses: To optimize/grow the business - make more money. Data mostly NOT shared except when the primary function is the collection and sale of data and or insights.
- Government: To govern/protect/serve citizens. Mostly NOT shared, except sometimes for information purposes and public transparency.
- Education institutions: For research purposes. Data shared to the extent allowed by research and community scrutiny.

Anywhere from 60% to 80% of a data scientist's time is spent performing data preparation activities. Arguably, this is the most essential part of a data scientist's job.

Predictive analysis results of a data scientist can be only as good as the data they assemble.

Data comes from many sources, leaving it up to the data scientist to have to join all this data and make sure that the resulting combinations make sense for further analysis. This data may come from structured sources like organizations' internal database systems, or unstructured sources like application logs or social media feeds. All of this data, undoubtedly, has formatting inconsistencies.

2.2.2 *Course Projects Alignments*

2.2.2.1 IST 736 Text Mining - Public Sentiment Toward NFL Teams, Coaches, and Players

This is a text mining project of data found in social media and through web scraping. It focuses on using sentiment analysis machine learning techniques to understand the popular opinion of a specific NFL Team, Coach or Player. See the project description for details.

Complete details of this project and the IST 736 Text Mining course learning objectives can be found in section 4.1, Couse: IST 736 Text Mining.

2.2.2.1.1 Data collection and organization techniques/steps summary:

- Web scraping
- Twitter API Live Stream Capture
- Twitter API Historical Search Capture
- NFL type object modeling
- Local file system, text corpus, storage and retrieval
- Data engineering pipeline

Details of the data mining techniques used for collecting and organizing this dataset are in the course project report, **section 2 - Initial Data Mining**. This section of the project report details each of the methods used in collecting, analyzing, transforming and organizing the data which also includes data modeling along with file system storage and version persistence. Each of the module components referenced in the document consists of the source code file name, the location of which can be found here, along with a visual design diagram of the core process steps of each module component.

2.2.2.2 IST 707 Data Analytics - Real Estate Property Investment

This is a data mining project that has structured and unstructured components. Its business focus is real estate property value prediction. See the project description for details.

Complete details of this project and the IST 707 Data Analytics course learning objectives can be found in section 4.3, Course: IST 707 Data Analytics.

2.2.2.2.1 Data collection and organization techniques/steps summary:

- **Internet search for real estate datasets**:
  - Base Real Estate data provided by: Zillow
    - Zillow Data: Timeseries Real Estate data by U.S. Zip Code
      - SingleFamilyResidence
      - AllHomes
      - MedianRentalPricePerSqft
      - MdeianRentalPrice_AllHomes
      - MedianListingPrice_AllHomes
- **Internet search for U.S. Economic datasets**:
  - Base Economic datasets provided by: datahub.io
    - U.S., National Yearly Economic Reports
      - interest_rates
      - inflation-consumer
      - inflation-gdp
      - education_budget_data
      - population
      - investor_flow_funds_monthly
      - housing_price_cities
      - household-income
      - employment
      - cpi
      - cash-surp-def
      - bonds_yields_10y
      - gdp_quarter
      - gdp_year

Dataset Info: Economic

- The time series date ranges for modeling and analysis were from 1997 through 2018. All of the Real Estate datasets achieved this desired range, however some of the Economic datasets did not. To achieve parity with the real estate data and have a fuller dataset for baseline testing, time series future prediction forecast methods were applied to the below three features.
    - GDP Year: Forecasted from 2016, 2017, 2018
    - Inflation: Forecasted for 2017, 2018
    - Interest Rates: Forecasted for 2016, 2017, 2018

Dataset Info: Real Estate

- This data is the base datasets and provides the core insights into predictable housing market trends given prior knowledge of price-performance coupled with economic fluctuations. Time series prediction models are created for each type of housing dataset mentioned above by zip code and it's monthly price value from 1997 to 2018. For this initial analysis, zip codes' were limited to the U.S. State of Washington. This represents 351 unique zip codes that were modeled with a five-year future price prediction. These zip codes were then combined with the economic features to create a dataset for identifying and or predicting events that could have a positive or negative impact on housing prices given a zip code.

Details of the data mining techniques used for collecting this dataset can be found in the course project report, **section 1.2 - About the Data**. This section of the project report details each of the data sources used in the analysis along with a description of the source data. Multiple sources of data were used in this project such as Zillow Single Family Residence data sets and U.S. National Yearly Economic Reports pulled from Datahub.io, fourteen distinct datasets in all. All of these datasets had to be analyzed, transformed, and merged for usability in model experimentation. These datasets are also in time series format to enable forecast predictions. Thorough details of the data organization steps can be found in the project code which includes descriptive comments, visualizations and graphs.

2.2.2.3 IST 659 Database Administration Concepts and Database Management - A4B KPI BizOps Organization Hierarchy

This is a relational database design and implementation project that focuses on a companies organization's hierarchy structure of people and systems of application ownership. See the project description for details.

Complete details of this project and the IST 659 Database Administration Concepts and Database Management course learning objectives can be found in section 4.4, Course: IST 659 Database Administration Concepts and Database Management

2.2.2.3.1 Data collection and organization techniques/steps summary:

- Collection of data for this project came from internal operational data sources of company X along with project team interviewing of organization leads and resources who own applications or have a job role need where they should have access to the application KPIs.
- Data organization is in the form of a relational database system.

Details of the data organization techniques/steps performed in modeling this system can be found in the project report, section 2: **Conceptual Model**, section 3: **Normalized Logical Model**, and section 4: **Physical Database Design**.

2.2.2.4 MBC 638 Data Analysis and Decision Making - Purchase Order Process Improvement

This project is an internal organizations' process improvement initiative. It focuses on cost savings based on perceived excessive Purchase Order (PO) validation and approvals cycle-time. See the project description for details.

Complete details of this project and the MBC 638 Data Analysis and Decision Making course learning objectives can be found in section 4.5, Course: MBC 638 Data Analysis and Decision Making.

2.2.2.4.1 Data collection and organization techniques/steps summary:

- Collection of data came from internal business operational systems such as SharePoint, Rally, Ariba, Emails
- Data are organized in spreadsheets, and SharePoint input forms exported as .csv files.
    - The data elements collected were defined during the Define phase of the DMAIC framework process.
        - For a complete listing of the data attributes, see page 5, Data Collection Attributes of the project report.

2.3 Identify Patterns

2.3.1 *Overview*

Learning Objective: Identify patterns in data via visualization, statistical analysis, and data mining.

Pattern recognition is concerned with the automatic discovery of regularities in data through the use of computer algorithms and with the use of these regularities to take actions such as classifying the data into different categories. In machine learning, pattern recognition is the assignment of a label to a given input value. - https://en.wikipedia.org/wiki/Pattern_recognition

2.3.2 *Course Projects Alignment*

2.3.2.1 IST 718 Big Data Analytics - IEEE-CIS Fraud Detection

This project is a big data problem that uses machine learning algorithms to identify possible fraudulent e-commerce transactions. The data is provided from VESTA, a payment service company, and is part of a kaggle competition. See the project description for details.

Complete details of this project and the IST 718 Big Data Analytics course learning objectives can be found in section 4.2, Course: IST 718 Big Data Analytics

2.3.2.1.1 Patterns

Performing Exploratory Data Analysis (EDA) on large datasets like the VESTA e-commerce transactions requires the use of useful visualization techniques to highlight trends in the data and gain the meaning of its attributes. This dataset contained transaction attributes and identity attributes. Each had both categorical features and numeric, totaling 434 features. Throughout the EDA process, as shown in the EDA project code Timbrook_Fraud_ExploreCleanTransform, visualizations and statistical analysis are used to identify interesting patterns on specific attributes that could be associated with fraudulent transactions.

In the project presentation, slide 6: **Data Exploration Feature Correlation & Class Imbalance**, visualized **correlation heatmaps'** of various dataset features are highlighted to represent how they relate to fraud transactions. This representation enabled us to narrow our focus with further analysis of those interesting

features and structure the data for machine learning model development in a more efficient way. Another visualization on the same slide shows a clear picture of the isFraud **class imbalance** of the labeled dataset and the 'Transaction Amount' feature distribution over the entire dataset. The transaction amount is a key feature, when combined with other attributes, in helping to identify fraudulent transactions.

2.3.2.2 IST 736 Text Mining - Public Sentiment Toward NFL Teams, Coaches, and Players

This is a text mining project of data found in social media and through web scraping. It focuses on using sentiment analysis machine learning techniques to understand the popular opinion of a specific NFL Team, Coach or Player. See the project description for details.

Complete details of this project and the IST 736 Text Mining course learning objectives can be found in section 4.1, Couse: IST 736 Text Mining.

2.3.2.2.1 Patterns

In this text mining project, finding patterns in the public opinion toward NFL characters on a weekly bases was the primary objective. After the data had been collected, organized, and engineered, polarity scoring was performed on every twitter tweet document collected. In the project report, section 2.1.7.3 VEDAR Results, polarity scores **descriptive statistics** are shown along with a **scatter plot** and **bar plot visualizations** showing the polarity scores of neutral, positive, or negative. In Figure 2.1: Vedar Sentiment Scoring - NFL Trend over 13 Game Weeks, visualizes a **time series trend** for each of the three categories of Players, Coaches, and Teams, tweet sentiment classification for each of the 13 weeks game schedules collected. This trend showed an apparent emotional reaction from the public that correlated with how the NFL team performed each week.

After labeling the corpus of tweet documents using the VADAR API, further experimental classification modeling were performed on the dataset. The purpose was to find the best classification algorithm and its configurations for this dataset using two of those learned in the course objectives. The algorithms experimented with were, **Multinomial Naive Bayes (MNB)** and **Support Vector Machine (SVM)**. Details of the experimentation can be found in the project report section 3 - Sentiment Classification Modeling and in the project code classification_modeling_mnb_svm. Each of the experiments' output accuracy results is shown along with its accompanying confusion matrix visualization. In section 3.1.1.5.1 Model Accuracy Comparison Summary, there are visualizations that show the accuracy score for each of the experiments. This visual allows for a quick and straightforward way of showing how each test performed in comparison to each other. The top-performing classifier was the SVM Unigram Model with a prediction accuracy result of 93%.

2.3.2.3 IST 707 Data Analytics - Real Estate Property Investment

This is a data mining project that has structured and unstructured components. Its business focus is real estate property value prediction. See the project description for details.

Complete details of this project and the IST 707 Data Analytics course learning objectives are found in section 4.3, Course: IST 707 Data Analytics.

2.3.2.3.1 Patterns

The Real Estate investment project had a great deal of data visualization throughout each stage of the project. This was partially due to the number of datasets that had to be analyzed and reformatted to fit into a valuable master data source. And also to identify which economic factors had the most notable impact on real estate property value over time. In the project report, section 1.4 Data Exploration - Scrub - Clean -

Transform, there are several **correlation heatmaps**, attribute **distribution scatter plots**, **time series trend** graphs, and **linear regression scatter plots**. All of these visuals added in understanding how features were related to each other or not. Which lead to how the final dataset used for predictive modeling was engineered.

Other notable data visualization graphs include the **elbow chart** used in identifying the optimal number of k clusters to be used in the **K-Means clustering** algorithm. This is shown in section 3.1.3 - KMeans of the project report. And the **Decision Tree** algorithm results, **feature importance** bar graph, shown in section 4.4 of the project report.

2.4 Alternative Strategies

2.4.1 *Overview*

Learning Objective: Develop alternative strategies based on the data.

A strategy is a thoughtful plan focused on changing the current state to reach a vision for the future. In other words, the right strategy needs to start with a vision, and the strategy is a way of making a series of changes, usually requiring innovation to achieve the vision. Every enterprise must have a business vision and a business strategy in place before having a data strategy. A data strategy should go hand-in-hand with a business strategy and serve to realize the business vision. Data lives with technology while providing value to businesses and customers. Data strategy, therefore, is also the strategy of both data and technology. - https://towardsdatascience.com/how-to-create-a-successful-data-strategy-1293bacf463c

Data is a representation of either a social or physical reality. Any data source is ever only a sample of the fullness and complexity of the real world. Information is data imbued with context. The raw data collected from reality needs to be summarized, visualized and analyzed for managers to understand the reality of their business. This information increases knowledge about a business process, which is in turn used to improve the truth from which the data was collected.

In speech recognition, the continuous collection and analysis of users' spoken words are needed for tuning the recognition engines, as well as discerning if users understand the context of the dialog. Based on this analysis, I work with my business partners regularly on designing more efficient voice user interfaces our customers understand and enjoy.

2.4.2 *Course Projects Alignment*

2.4.2.1 SCM 651 Business Analytics - Recruiting Advertising Strategy

This project is an advertising analysis of an internet marketing campaign, using google analytics, to identify patterns and opportunities in order to establish a strategy for the coming year. See the project description for details.

Complete details of this project and the SCM 651 Business Analytics course learning objectives can be found in section 4.6, Course: SCM 651 Business Analytics

2.4.2.1.1 Strategy

Based on the historical data and insights gained from analyzing the google analytics platform, using the following metrics: Cost per Click, Users, New Users, Bounce Rate, Pages per session, the most effective campaign identified was the Whitman.syr.edu. Details, including graphs, can be found in the project report, section 2 - Measurements of Prior Campaigns Effectiveness to Cost

Given those results, next year's campaign recommendations were:

- The 2020 Campaign recommendations include continued use of Google Ad Marketing, with the additional recommendations to incorporate Facebook Ad Business and GMASS Targeting marketing. Taking advantage of these platforms should be tested. Budge costs and strategy are described section 3.1 Recommendations. Supporting data for these recommendations is provided in this section's Tables and Graphs reference.

The recommendations included in the report detail the advertising regions to focus their ad campaigns to, keywords to use in search analytics, the best days of the week and time of the day for the advertising to be published, the expected advertising costs which includes a budget plan, a post-implementation success measurement plan, and a list of other factors we found essential to the success of the upcoming campaign.

2.4.2.2 IST 707 Data Analytics - Real Estate Property Investment

This is a data mining project that has structured and unstructured components. Its business focus is real estate property value prediction. See the project description for details.

Complete details of this project and the IST 707 Data Analytics course learning objectives can be found in section 4.3, Course: IST 707 Data Analytics.

2.4.2.2.1 Strategy

The original data strategy for this project was collecting Real Estate property value datasets from Zillow and U.S. Economic reports datasets from Kaggle. During exploratory data analysis, it was found that the Economic dataset was incomplete and would not be usable for our study. Additional internet searches found, relatively complete, national economic report datasets on Datahub.io.

The time series date ranges for modeling and analysis were from 1997 through 2018. All of the Real Estate datasets achieved this desired range, however some of the Economic datasets did not. To achieve parity with the real estate data and have a fuller dataset for baseline testing, time series future prediction forecast methods were applied to the below three features.

- GDP Year: Forecasted from 2016, 2017, 2018
- Inflation: Forecasted for 2017, 2018
- Interest Rates: Forecasted for 2016, 2017, 2018

The prediction models generated for state zip code, median home price value, influenced by economic factors should have further experimentation using state-level economic metrics versus the national level used in this study.

2.5 Plan of Action

2.5.1 *Overview*

Learning Objective: Develop a plan of action to implement the business decisions derived from the analysis.

2.5.2 *Course Projects Alignment*

2.5.2.1 MBC 638 Data Analysis and Decision Making - Purchase Order Process Improvement

This project is an internal organization process improvement initiative. It focuses on cost savings based on perceived excessive Purchase Order (PO) validation and approvals cycle-times. It's designed from a real-world scenario, one that I had direct control over as a Software Development Manager at the company and who's responsibilities included validating and approving vendor POs. See the project description for details.

Complete details of this project and the MBC 638 Data Analysis and Decision Making course learning objectives can be found in section 4.5, Course: MBC 638 Data Analysis and Decision Making.

The business problem statement was defined as:

- As a Software Development Manager who augments my teams' resource pool with a Managed Vendor Service agreement, it is my responsibility to approve and sign-off all Purchase Order agreements my vendors submit monthly for the work they've completed.

  Due to recent system changes and financial forecasting requirements, all POs are submitted three days prior to the start of a two-week Sprint; The details of which must include specific User Stories and their associated Project Finance Codes needed to bill back to the various projects the work is associated to.

  Because of the manual steps and system limitations in capturing details of each PO it requires multiple people to collect, organize and categorize each of the Sprint teams' work and manually validate the deliverables and costs specified in the PO are in fact, what was delivered at the end of each Sprint. Due to the manual nature of the process and the high volume of work the team handles it's often observed, after the fact, that POs have been billed to wrong projects and their actual costs haven't been reconciled against their estimated costs given at the start of the Sprint. All of which causes delays and errors invalidating the PO information which is submitted by our vendors is correct.
- Business Impact:
  - I spend on average 8 hours per bi-weekly Sprint cycle, manually collecting and organizing data for five Sprint Teams and reconciling it with the vendors' PO submissions. This does not include comparing estimates to actuals or loading the collected data into a system of record for future analysis. At a standard bill rate of 100/hr. this cost of time is $800 per Sprint, or $20,800 per year.
    Due to the nature of this manual process, it is also observed that 80% of the POs take greater than 15 days to be received in the Ariba financial system. By going over the 15-day threshold we lose the opportunity of saving 2% on the total PO cost our vendor offers as a discount.
    At $120K per Sprint, per Sprint Team (5), this is a loss of 24K in savings every two weeks, annually $624K.

2.5.2.1.1 Action Plan

During the '**Measure'** and '**Analyze'** phases of the DMAIC process improvement lifecycle, it was found that: Managers have a low-frequency rate of approving PO Invoices under the vendor discount time period; PO Invoices having to be corrected lead to longer approval cycle times; PO Invoices with mid-to-high User Story counts increase the likelihood an invoice will need to be adjusted; The more time a team and manager have to spend validating an invoice the more time it takes to approve a PO. Details can be found in the project report, sections **Measure** and **Analyze**.

In the **'Improve'** phase, a plan was executed that included the design and pilot of a new process workflow that incorporated automation through Microsoft SharePoint tooling. By designing a standard user interface for

entering all of the necessary attributes needed for an accurate and timely PO validation and approval process along with a detailed cross-functional workflow diagram for role/resource training, an improved cycle time efficiency of 70% was observed during the pilot of the new process. Details can be found in the project report, section **Improve**.

To measure on-going success and determine the business value of pushing this new process to other teams within the organization, controls were implemented that would continue to measure and report cycle-times along with POs being approved within the given discount thresholds. Details can be found in the project report, section **Control**.

Additional implementation planning artifacts can be found in the Appendix of the project report.

2.5.2.2 IST 659 Database Administration Concepts and Database Management - A4B KPI BizOps Organization Hierarchy

This is a relational database design and implementation project that focuses on an organization's hierarchy structure of people and systems of application ownership. See the project description for details.

Complete details of this project and the IST 659 Database Administration Concepts and Database Management course learning objectives can be found in section 4.4, Course: IST 659 Database Administration Concepts and Database Management

2.5.2.2.1 Action Plan

After completing this project, as per the course requirements, I ran a pilot initiative internal to my team at T-Mobile, which leveraged this new database in building DevOps, operational Amazon Alexa 4 Business Skills that answered the hierarchical system KPI relationship questions described in the project report, section 1.1 Purpose. The pilot was executed as an Agile, 2-week sprint cycle, an initiative following all of my team's standard DevOps procedures for running sprint projects. This includes having a Product Owner writing User Stories, a Scrum Master coordinating the Agile DevTeam's scrum ceremonies along with Business Stakeholder demo's and CI/CD (Continuous Integration / Continuous Development) pipelines enabling rapid prototyping and customer feedback. Additionally, being a greenfield application, this system was fully deployed to AWS Cloud leveraging their cloud managed services and Amazon RDS to host the SQL Server.

2.6 Communication Skills

2.6.1 *Overview*

Learning Objective: Demonstrate communication skills regarding data and its analysis from managers, IT professionals, programmers, statisticians, and other relevant professions in their organization.

My role as a Software Development Manager requires that I'm presenting to leadership and customers regularly. It includes everything from day-to-day operation metrics to new product innovation sales pitches to technical all-hands town-hall discussions on things like digital transformation. Communicating to a wide range of people and topics takes a great deal of thought and preparation if you're to get your message across effectively. Continuous practice of this skill is needed more than most due to the unique nature of how people hear and respond to messages.

2.6.2 *Course Projects Alignment*

2.6.2.1 MBC 638 Data Analysis and Decision Making - Purchase Order Process Improvement

This project is an internal organization process improvement initiative. It focuses on cost savings based on perceived excessive Purchase Order (PO) validation and approvals cycle-time. See the project description for details.

Complete details of this project and the MBC 638 Data Analysis and Decision Making course learning objectives can be found in section 4.5, Course: MBC 638 Data Analysis and Decision Making.

2.6.2.1.1 Communication Plan

As a deliverable for this project, a presentation slide deck was created to communicate executive summary results through a single slide dashboard for each of the DMAIC phases along with support slides that provide the low-level details of how the results were achieved. A live-video presentation to the class of my findings was held on the last day of class.

2.6.2.2 IST 736 Text Mining - Public Sentiment Toward NFL Teams, Coaches, and Players

This is a text mining project of unstructured data found in social media and through web scraping. See the project description for details.

Complete details of this project and the IST 736 Text Mining course learning objectives can be found in section 4.1, Couse: IST 736 Text Mining.

2.6.2.2.1 Communication Plan

This project had both a technically detailed final report, Final_Project_Timbrook_Ryan, along with a final presentation, Team_5_Final_Project_Presentation, held via live video demonstration. The final report document, along with the automated model summary report .csv output files located in the 05-Report subdirectory, represent communication artifacts for all levels of technical resources. The presentation deck is a good representation of communicating our results to Business Owners and Stakeholders at a Sprint Demo or Retrospective.

The main features and functions of the coding for this project was built using Jupyter Notebooks in Python. These notebooks are a great way of communicating low-level programming to data scientists, programmers, and technical leadership.

2.6.2.3 IST 718 Big Data Analytics - IEEE-CIS Fraud Detection

This project is a big data problem that uses machine learning algorithms to identify possible fraudulent e-commerce transactions. The data is provided from VESTA, a payment service company, and is part of a kaggle competition. See the project description for details.

Complete details of this project and the IST 718 Big Data Analytics course learning objectives can be found in section 4.2, Course: IST 718 Big Data Analytics

2.6.2.3.1 Communication Plan

The final deliverable for this project was a presentation deck and demo conducted video-live. The presentation highlighted our summarized results through the Obtain, Scrub, Explore, Model, Interpret, execution lifecycle. A communication presentation such as this would be held during a Sprint Demo and or

Retrospective with our Business Partners and Stakeholders. Following the demo, the next sprint planning session would kickoff, taking lessons learned and applying them to the next sprint cycle.

The main features and functions of the coding for this project was built using Jupyter Notebooks in Python. These notebooks are a great way of communicating low-level programming to data scientists, programmers, and technical leadership.

2.7 Synthesize Ethical Dimensions

2.7.1 *Overview*

Learning Objective: Synthesis of the ethical dimensions of data science practice (e.g., privacy)

Over my career, the majority of my years (14) have been designing, developing and managing tier-1, customer-facing, business-critical systems for Fortune 50 companies like T-Mobile and Anthem Blue Cross and Blue Shield Inc. These systems are voice self-service applications where customers can do things like make payments to a health insurance claim, hear their health benefits and other claims information. And for the T-Mobile system, they could do similar things like making payments or scheduling payments for their phone service. To perform these functions sensitive information from the customer has to be collected through our technology that authorizes them and processes the transaction. This data is highly regulated at the government level which required my systems to be regularly audited for security compliance. These security compliances included: **Payment Card Industry Data Security Standard (PCI DSS**), **Personally Identifiable Information (PII),** and **Health Insurance Portability and Accountability Act (HIPAA).**
Designing systems such as these require careful and thorough planning where the security of our customers' private data is the focus of every layer.
A few examples of the considerations that had to be designed and included in our systems architecture were:

- **Encryption**: Sensitive data had to be encrypted while traveling through the network or when residing on storage targets.
- **Application Logging**: No PII or sensitive data was allowed to be written to logs or other non-authorized targets.
- **Secure credentials and endpoints**: Service credentials and source/target endpoints were kept outside of memory. Native tokenization services with minimized privileges were utilized to minimize potential blast radius. With human operators created credentials, identity and access management rules with segregated and specific policies were a standard.
- **Audit Logs**: Automated alerting and notification was established to detect non-authorized personal from accessing the systems along with tracking the actions of possible hundreds of users who may be interacting with our cloud environment.
- **Testing**: No real-customer PII data was to be used for any type of customer experience testing. Data were generated and mockup host systems created to replicate real-world transactions.

2.7.2 *Course Projects Alignment*

2.7.2.1 IST 659 Database Administration Concepts and Database Management - A4B KPI BizOps Organization Hierarchy

This is a relational database design and implementation project that focuses on an organization's hierarchy structure of people and systems of application ownership. See the project description for details.

Complete details of this project and the IST 659 Database Administration Concepts and Database Management course learning objectives can be found in section 4.4, Course: IST 659 Database Administration Concepts and Database Management

2.7.2.1.1 Privacy

This project had multiple layers of privacy security. In the cloud, all of the IAM rules were set to 'least privilege' to protect the integrity of the data and not allow unauthorized persons from having access to restricted organizational KPI metrics. The hierarchical relationship model design was structured so that only users who had the proper level of ownership and seniority in the organization could access the KPI information. And in the SQL Database role-based permissions were set up to restrict users capabilities based on their individual role needs. Details of these designs can be found in the project report, section 3: Normalized Logical Model and section 4: Physical Database Design.

2.7.2.2 IST 718 Big Data Analytics - IEEE-CIS Fraud Detection

This project is a big data problem that uses machine learning algorithms to identify possible fraudulent e-commerce transactions. The data is provided from VESTA, a payment service company, and is part of a kaggle competition. See the project description for details.

Complete details of this project and the IST 718 Big Data Analytics course learning objectives can be found in section 4.2, Course: IST 718 Big Data Analytics

2.7.2.2.1 Privacy

The objective of this project was to use advanced machine learning algorithms to identify real-time fraudulent e-commerce transactions so that automated applications could be designed and built to alert and block highly suspicious activities. This would help hundreds of thousands of businesses reduce their fraud loss and increase their revenue; while securing consumer's peace of mind and wallets.

This problem is important because just about everyone uses e-commerce technology and modern banking systems are at risk of being a victim of fraud. It cost both the individual as well as the merchant who offers refunds for fraudulent transactions; and not all scenarios are covered, leaving many individuals to pay.

Company's who have had data security breaches put everyone who uses electronic forms of payment at risk. A few of the larger breaches such as eBay in 2014 with 145 million user accounts compromised, Heartland Payment Systems in 2008 had 134 million users credit cards stolen, Target in 2013 had up to 110 million customers credit/debit card, and contact info was taken and Yahoo in 2014 with it's 1 billion user accounts and passwords compromised, represent how crucial early warning fraud detection systems are to everyone.

Details of the project outcomes can be found in the project report.

3 Conclusion

I have specialized in Customer Service, Contact Center Technology, most of my career. As I continue to grow with the changes in technology, I aim to apply the knowledge and skills I've learned in the Applied Data Science program to the daily business problems I face. Having these skills enables me to be a stronger leader who can more effectively use data to drive the vision and roadmap of my organization.

Business Key Performance Indicators (KPI) are at the heart of most organizations operating models. Those I've belong to have used KPIs as their primary means of communication, both upward and outward. Data Science techniques, such as those learned throughout this program, are core in defining and measuring all forms of business metrics that make up the language of these KPIs. As an example, the most communicated and measured KPI for the systems I manage is Call Handle Rate (CHR). Simply put, it's a calculation of the number of callers who enter the voice system, perform a self-service transaction, and hang-up rather than being transferred to a Customer Service Agent. At $5.40 cost per call that reaches an Agent, companies with a large membership base and or churn leverage a voice self-service system as a cost-savings tool above all else. At my company, our voice system handles approximately 1 million calls per day with an average CHR of 40%, which is roughly $400,000,000 in annual cost savings.

A great deal of resource time and money goes into analyzing the data generated from these systems and in understanding our customers in order to design the most efficient applications our customers find friendly and straightforward to use versus needing to speak with a live agent. Data Science along with Speech Science, are critical skills needed in my organization to be able to provide the most effective data-driven recommendations to my business partners and stakeholders.

The learning objectives outlined in this portfolio, along with the course project samples provided, reflect new skills and knowledge learned in this program that has prepared me to execute the strategic vision I've set for myself and team for 2020. It is as follows.

VXPD Speech Services Strategy / Vision:
Expand speech interaction capabilities to new channels while managing CARE costs and KPIs with focused speech IVR investments.

- EASY ACCESS: (CONNECT)
  - Make it 'SIMPLE' for customers to reach their team (TEX) with voice.
- SIMPLE-RELIABLE-SECURE (SERVE)
  - Provide key self-service experiences that bring sufficient volume, savings & customer satisfaction.
- SEAMLESS TRANSITIONS (SHIFT)
  - Transition callers to appropriate strategic channels that better service their experience needs.

4 References: Applied Data Science Courses

4.1 Course: IST 736 TEXT MINING

4.1.1 *Description*

Introduces concepts and methods for knowledge discovery from large amounts of text data and the application of text mining techniques for business intelligence, digital humanities, and social behavior analysis.

The main goal of this course is to increase student awareness of the power of large amounts of text data and computational methods to find patterns in large text corpora. This course will introduce the concepts and methods of text mining technologies rooted in machine learning, natural language processing, and statistics. This course will also showcase the applications of text mining technologies in:

- Information organization and access
- Business intelligence
- Social behavior analysis
- Digital humanities

4.1.2 *Learning Objectives*

- Describe basic concepts and methods in text mining, for example, document representation, information extraction, text classification and clustering, and topic modeling;
- Use benchmark corpora, commercial and open-source text analysis and visualization tools to explore interesting patterns;
- Understand conceptually the mechanism of advanced text mining algorithms for information extraction, text classification and clustering, opinion mining, and their applications in real-world problems; and
- Choose appropriate technologies for specific text analysis tasks and evaluate the benefit and challenges of the chosen technical solution.

4.1.3 *Portfolio Project*

4.1.3.1 Final Project - Public Sentiment Toward NFL Teams, Coaches, and Players (Team Project)

4.1.3.1.1 Project Reports

- **Root Directory:** .\Timbrook_Portfolio_Milestone\ist736_text_mining\final_project\00-Requirements
  - **Project Requirements**: project-instructions.pdf
  - **Final Project Report**: Final_Project_Timbrook_Ryan.pdf
  - **Final Project Live Presentation**: Team_5_Final_Project_Presentation.pdf

4.1.3.1.2 Project Code

- **Root Directory:**
  .\Timbrook_Portfolio_Milestone\ist736_text_mining\final_project\03_Build
  - **Subdirectory:** \text_mine_stats
    - text_mine_nfl_players_list.ipynb
    - rtimbroo_utils.py
  - **Subdirectory:** \text_mine_twitter_streaming
    - stream_mine_tweets_nfl_team.ipynb (or .pdf)
    - stream_mine_tweets_nfl_coach.ipynb (or .pdf)
    - stream_mine_tweets_nfl_player.ipynb (or .pdf)
    - rtimbroo_utils.py
  - **Subdirectory**: \text_mine_coaches
    - search_twitter_nfl_coach.ipynb (or .pdf)
    - search_twitter_nfl_coach_premium.ipynb (or .pdf)
  - **Subdirectory**: \text_mine_teams
    - search_twitter_nfl_team.ipynb (or .pdf)
    - search_twitter_nfl_team_premium.ipynb (or .pdf)
  - **Subdirectory**: \text_mine_players
    - search_twitter_nfl_player.ipynb (or .pdf)
    - search_twitter_nfl_player_premium.ipynb (or .pdf)
  - **Subdirectory**: \data_engineering_pipeline
    - format_raw_twitter_data.ipynb (or .pdf)
    - process_raw_twitter_data_from_file.ipynb (or .pdf)
    - merge_datasets_to_master.ipynb (or .pdf)
    - nfl_sentiment_analsysis_data_merge_master.ipynb (or .pdf)
  - **Subdirectory**: \classification_modeling
    - classification_modeling_mnb_svm.ipynb (or .pdf)
    - rtimbroo_utils.py
  - **Subdirectory**: \topic_modeling
    - nfl_tweets_topic_modeling.ipynb (or .pdf)
    - rtimbroo_utils.py

4.1.3.1.3 Objective

The objective of the project is to use the main skills taught in this class to solve a real text mining problem. Find or create a dataset that suites the projects business problem to solve. Experiment design, define a problem on the dataset as a classification and/or clustering problem, and describe it in terms of it's real-world organizational or business application. The investigation is to include aspects of experimental comparison: depending on the problem, that may be with different types of algorithms or techniques.

4.1.3.1.4 Description

This project focused on how data science techniques such as text mining and sentiment analysis could be used to gain insights into public opinion on the lucrative NFL industry.

Specifically, its goal was to identify public sentiment toward NFL teams, and it's coach and players that could help fans, who play fantasy football, choose teams and players to play each week in their leagues.

4.1.3.1.5 Purpose

Identify public sentiment toward NFL teams and its players that could help fans choose teams and players to play in their fantasy leagues.

4.1.3.1.6 Scope

- Gather public data from the web and social media services such as Twitter and Facebook using text mining techniques to mine for:
  - Public opinion toward NFL teams, coaches, and players.
    - Reduce data gathering to one of each for initial POC.
    - Data is in time-series format from the first week of the NFL 2019 session to current schedule week
  - Players weekly Fantasy Football performance stat forecasts on a daily time scale.
- Perform sentiment analysis modeling ML techniques on data set to determine model prediction accuracies.
- Perform unsupervised topic modeling ML techniques on document text to gain insights into public opinion and primary opinion drivers.
- Evaluate weekly sentiment trends aligning with Fantasy Football performance stat forecasts.

4.1.3.2 Tools and Technologies:

- **Programming Languages:** Python - Anaconda 3 (Jupyter Notebook)
- **APIs**: Twitter Premium Search, Twitter Standard Streaming
- **Machine Learning Algorithms:**
  - **Classifiers**: Multinomial Naive Bayes (MNB), Support Vector Machine (SVM)
  - **Topic Modeling**: LatentDirichletAllocation
- **Machine Learning Packages**: sklearn.svm.LinearSVC, sklearn.naive_bayes.MultinomialNB, sklearn.decomposition.LatentDirichletAllocation
- **Other Notable Packages:** nltk, pandas, numpy, sklearn.feature_*extraction.text.CountVectorizer, sklearn.feature*_extraction.text.TfidfVectorizer, bs4.BeautifulSoup, seaborn, matplotlib.pyplot, requests, tweepy.OAuthHandler, tweepy.Stream, tweepy.streaming.StreamListener

4.2 Course: IST 718 BIG DATA ANALYTICS

4.2.1 *Description*

A broad introduction to analytical processing tools and techniques for information professionals. Students will develop a portfolio of resources, demonstrations, recipes, and examples of various analytical techniques

4.2.2 *Learning Objectives*

- Obtain data and explain data structures and data elements.

- Scrub data by applying scripting methods, to include debugging, for data manipulation in Python, R or other languages.
- Explore data by analyzing using qualitative techniques including descriptive statistics, summarization, and visualizations.
- Model relationships between data using the appropriate analytical methodologies matched to the information and the needs of clients and users.
- INterpret the data, model, analysis, and findings. Communicate the results in a meaningful way.
- Select an applicable analytical methodology for real problems in areas such as business, science, and engineering.

4.2.3 *Portfolio Project*

4.2.3.1 Final Project - IEEE-CIS Fraud Detection (Team Project)

4.2.3.1.1 Project Reports

- **Root Directory:** .\Timbrook_Portfolio_Milestone\ist718_big_data_analytics\final_project\05-Report
  - **Final Project Live Presentation:** Team_AUQ42_Presentation_CourseProject.pdf
  - **Final Project Checkpoint 1:** AUQ-42_Week5ProjectCheckIn.pdf

4.2.3.1.2 Project Code

- **Root Directory**: .\Timbrook_Portfolio_Milestone\ist718_big_data_analytics\final_project\03-Build
  - **Obtain Data and Reduce Memory Consumption**: Timbrook_Fraud_ObtainAndReduceMemoryUse.ipynb (or .pdf)
  - **Exploratory Data Analysis:** Timbrook_Fraud_ExploreCleanTransform.ipynb (or .pdf)
  - **Model - Neural Network - Classifier:** Timbrook_Fraud_BaseNeuralNetwork.ipynb (or .pdf)

4.2.3.1.3 Objective:

For the final project, the student will identify a data-focused problem, bring together different data sources, conduct analysis, draw conclusions, and produce a report explaining the results. It's expected that the work demonstrates the student's ability to select the appropriate analytical methods to the chosen problem; interpret the data, model, analysis, findings; draw appropriate conclusions, and present the results in a meaningful way.

4.2.3.1.4 Business Question

Improve the efficacy of fraudulent transaction alerts, helping hundreds of thousands of businesses reduce their fraud loss and increase their revenue; while securing consumer’s peace of mind and wallets!

4.2.3.1.5 Problems to solve

Using the VESTA's dataset from Kaggle, identify real-time fraudulent e-commerce transactions, using advanced Machine Learning algorithms by automating alerts that block highly suspicious activities. The data come from Vesta's real-world e-commerce transactions and contains a wide range of features from device type to product features.

https://www.kaggle.com/c/ieee-fraud-detection

4.2.3.1.6 Why the problem is important

Everyone who uses e-commerce technology and modern banking systems are at risk of being a victim of fraud. It costs both the individual as well as the merchant who offers refunds for fraudulent transactions; and not all scenarios are covered, leaving many individuals having to pay.

Chargebacks area a growing costly burden for merchants. By eliminating chargebacks, fines, and fees related to third-party fraud and unauthorized charges, the client, VESTA , is able to significantly reduce the operational costs and resources associated with complex chargeback management solutions and the specialized staff necessary for rapid, scalable business growth. This leaves all the cost risk on the client. Improving automated fraudulent detection technology will greatly reduce this cost.

4.2.3.1.7 Scope:

- Given the large-scale Vesta data set, create advanced machine learning models that can detect potentially fraudulent transactions in real-time in order to provide automated alerting and prevention solutions for the customer.
    - Obtain the data set
    - Perform thorough exploratory data analysis
    - Perform experimental modeling on the data, reporting on the results and accuracy of each model
    - Interpret the results, providing recommendations and next actions

4.2.3.2 Tools and Technologies:

- **Programming Languages:** Python - Anaconda 3 (Jupyter Notebook)
- **Data Source**: VESTA e-commerce Credit Card Transactions - Kaggle
- **Machine Learning Algorithms:**
    - **Classifiers**: KerasClassifier (https://keras.io/models/sequential/)
    - **Neural Networks**: Keras Sequential - multiple dense layers
- **Machine Learning Packages**: sklearn.preprocessing, sklearn.metrics, sklearn.model_selection.GridSearchCV, sklearn.model_selection.cross_validate, sklearn.pipeline, keras.modles.Sequential, keras.layers, keras.layers.normalization, keras.wrappers.scikit_learn.KerasClassifier
- **Other Notable Packages:** pandas, numpy, matplotlib.pyplot, matplotlib.pylab, seaborn, pickle,

4.2.3.2.1 Subfolder: final_project

The final-project subfolder contains additional subfolders for each of the SDLC phases of the project. The 03-Build subfolder contains all of the code I built for the project.
Note-This was a team project. For the Portfolio purposes, I've only included my work and contributions to the project, leaving my partners work out. This subdirectory additionally contains:

Root Files:

- Readme File
- Grade and Feedback Received
- Final Project Presentation ('Team_AUQ42_Presentation_CourseProject.pdf')
    - Slides I created:
        - 1,2,3,4,5,10,11
        - My topics:
            - Data Acquisition / Transformation / Cleaning
            - Data Exploration
            - Neural Network

4.3 Course: IST 707 DATA ANALYTICS

4.3.1 *Description*

Introduction to data mining techniques, familiarity with particular real-world applications, challenges involved in these applications, and future directions of the field. Hands-on experience with open-source software packages. This course introduces popular data mining methods for extracting knowledge from data. The principles and theories of data mining methods will be related to the issues in applying data mining to problems. Students will acquire hands-on experience using state-of-the-art software to develop data mining solutions to scientific and business problems. The topics of the course include the key tasks of data mining, including data preparation, concept description, association rule mining, classification, clustering, evaluation, and analysis.

4.3.2 *Learning Objectives*

- The document, analyze, and translate data mining needs into technical designs and solutions.
- Apply data mining concepts, algorithms, and evaluation methods to real-world problems.
- Employ data storytelling and dive into the data, find useful patterns, and articulate what patterns have been found, how they are found, and why they are valuable and trustworthy.

4.3.3 *Portfolio Project*

4.3.3.1 Final Project - Real Estate Property Investment (Individual Project)

4.3.3.1.1 Project Reports

- **Root Directory:** ./ist707_data_analytics/final_project/05-Report/
    - Ryan_Timbrook_Project_Report.pdf
    - Ryan_Timbrook_Project_Report.html

4.3.3.1.2 Project Code

- **Root Directory:** ./ist707_data_analytics/final_project/03-Build
    - Ryan_Timbrook_Project_Report.ipynb (or .pdf)

4.3.3.1.3 Objective:

The objective of the project is to use the main skills taught in this class to solve a real data mining problem. For the project, students choose their own dataset. The format of the business problem is in experimental design. The problem will use all of the data mining algorithms/techniques taught in the class, including **full cleaning**, **prep**, **EDA** and **visual EDA**, **clustering unsupervised machine learning**, **decision tree supervised machine learning**, **naive bayes classification**, and **support vector machine classification**.

4.3.3.1.4 Business Problem

Targeting low-risk property investment opportunities for property management firms and or individual investors who buy-rent-sell single-family homes throughout the United States. Given a base set of investment criteria, provide a predicted N-best list of U.S. geolocation regions by zip code that offers the best ROI. Utilizing Zillow's Timeseries Real Estate, Zillow Home Value Index (ZHVI), data by Zip Code in the U.S. combined with U.S. National Yearly Economic data, develop a prediction algorithm that provides investors with actionable insights that support their needs.

4.3.3.1.5 Problems to solve

- How to predict a low risk / high yield return on property investment in a volatile real estate market.
- Where and when to buy and sell that maximizes investment profits.
- Forecast future growth and decline of a region in order to guide investors with optimized, data-driven recommendations.

4.3.3.1.6 Scope

- Obtain datasets from Zillow and other sources that provide U.S. economic reports which have potential impacts on real estate home value prices.
- Perform exploratory data analysis on datasets to identify insightful patterns.
    - Clean, Transform, Merge individual datasets
    - Include:
        - **Feature Correlation** Analysis
        - **Linear Regression** Analysis
        - **Timeseries Analysis** and Future Price Prediction Analysis
    - Create **time series price prediction** models for each of the U.S. regions by ZipCode that covers two, three, five-year future price value predictions.
    - Perform **K-means**, **unsupervised** machine learning techniques, to classify zipcode models into categories of 'buy', 'hold', 'sell'.
    - Perform **Experimentation** using different supervised machine learning techniques to identify which has the best accuracy results for the given datasets.
        - Perform **Decision Tree**, **supervised** machine learning techniques, modeling on labeled data created based on 'Price Thresholds'.
        - Perform **Random Forest**, **supervised** machine learning techniques, modeling on labeled data created based on 'Price Thresholds'.
        - Perform **Naive Bayes**, **supervised** machine learning techniques, modeling on labeled data created based on 'Price Thresholds'.
        - Perform **Support Vector Machine**, **supervised** machine learning techniques, modeling on labeled data created based on 'Price Thresholds'.
    - Report experimentation outcomes including visualizations.
    - Provide recommendations and insights on dataset and business problem.

4.3.3.2 Tools and Technologies

- **Programming Languages:** Python - Anaconda 3 (Jupyter Notebook)
- **Data Source**: See project report, Ryan_Timbrook_Project_Report.pdf section 1.2 About the Data
- **Machine Learning Algorithms:**
  - **Time Series Analysis & Forecasting:** Prophet (Facebook)
  - **Unsupervised Clustering:** K-means (sklearn.cluster.Kmeans)
  - **Supervised Classifiers**: Decision Tree (sklearn.tree.DecisionTreeClassifier), Random Forest (sklearn.ensemble.RandomForestClassifier), Naive Bayes (sklearn.naive_bayes.GaussianNB), Support Vector Machine (sklearn.svm.SVC)
- **Machine Learning Packages**: sklearn.cluster.Kmeans, sklearn.tree.DecisionTreeClassifier, sklearn.ensemble.RandomForestClassifier, sklearn.naive_bayes.GaussianNB, sklearn.svm.SVC, fbprophet.Prophet
- **Other Notable Packages:** pandas, numpy, pickle, matplotlib.pyplot, seaborn

4.3.3.3 Subfolder: final_project

The final-project subfolder contains additional subfolders for each of the SDLC phases of the project. The 03-Build subfolder contains all of the code and abstracted data used for the project.
Root Files:

- Readme File
- Grade and Feedback Received
- Final Project Report

4.4 Course: IST 659 Database Administration Concepts and Database Management

4.4.1 *Description:*

IST 659 is an introductory course to database management systems. This course examines data structures, file organizations, concepts, and principles of database management systems (DBMS) as well as data analysis, database design, data modeling, database management, and database implementation. More specifically, it introduces hierarchical, network, and relational data models; entity-relationship modeling; basics of Structured Query Language (SQL); data normalization; and database design. Using Microsoft's Access and SQL Server DBMSs as implementation vehicles, this course provides hands-on experience in database design and implementation through assignments, lab exercises, and course projects. This course also introduces advanced database concepts such as transaction management and concurrency control, distributed databases, multitier client/server architectures, web-based database applications, data warehousing, and NoSQL.

4.4.2 *Learning Objectives*

- Describe fundamental data and database concepts
- Explain and use the database development lifecycle
- Create databases and database objects using popular database management system products
- Solve problems by constructing database queries using Structured Query Language (SQL)
- Design databases using data modeling and data normalization techniques

- Develop insights into future data management tool and technique trends
- Recommend and justify strategies for managing data security, privacy, audit/control, fraud detection, backup and recovery
- Critique the effectiveness of DBMS in computer information systems

4.4.3 *Portfolio Project*

4.4.3.1 Final Project - A4B KPI BizOps Organization Schema (Individual Project)

4.4.3.1.1 Project Reports

- **Root Directory:** ./ist659_database_admin_management/final_project/
  - **Final Report**: RyanTimbrook_IST659_Project_Final.pdf
  - **Project Requirements**: IST659_Project_Description.pdf

4.4.3.1.2 Project Code

- **Root Directory:** ./ist659_database_admin_management/final_project/sql
  - a4b_kpi_db_prototype.R
  - a4b_hierarchy_db_ddl.sql
  - a4b_hierarchy_db_ddl_functions.sql
  - a4b_hierarchy_db_ddl_inserts.sql
  - a4b_hierarchy_db_ddl_stored-procedures.sql
  - a4b_hierarchy_db_ddl_views.sql
  - a4b_hierarchy_db_ddl_exec-procedures.sql
  - a4b_hierarchy_db_ddl_qa-selects.sql
  - a4b_hierarchy_db_ddl_test-views.sql
  - a4b_kpi_hierarchy-DB-Create.sql

4.4.3.1.3 Objective

Design and implement a database that solves a data management problem. Throughout the process, the intent is to exemplify all of the above listed learning objectives for this course. The final deliverable will have two main parts; the first part is the design specifications detailing the data to be tracked and how all of the elements work together. It will include any business rules that dictate how the data are to be managed. Stakeholders will be identified and details of what data they will need to access and maintain the system. Data questions will be provided that represent how this database can answer the business-specific questions identified. The second part is the implementation of the design created in part one. This includes the SQL statements to create the tables and columns to hold the data and any constraints that implement the business rules. Also included in this section are representative statements for the basic Data Manipulation Language (DML) that implement the create, read, update, and delete statements (referred to as CRUD) used in maintaining the data.

4.4.3.1.4 Business Problem

The Voice Services team at Company X is developing a new set of Amazon Alexa Skills that are intended for internal use by employees of the company. These new voice services will enable employees of different organizational levels to use their voice to ask Alexa for Key Performance Indicators (KPI) they are responsible for as a function of their role in the company. In order to make the skills simple, user-friendly, and fast, the Voice Services team understands that a level

of knowledge about the companies organizational hierarchy along with it's hundred's of KPIs is needed to be captured and stored in a relational database system for quiring as well as updating as people, systems, and organizations change regularly. There is no system like this at the company they could leverage, so they will design and implement a new one.

4.4.3.1.5 Scope

The Database Design will consist of the following:

- Conceptual Modle
- Normalized Logical Model
- Physical Database Design
- Data Creation
- Data Manipulation
- Answering Data Questions
- Implementation Screen Shots

4.4.3.2 Tools and Technologies

- **Languages:** SQL, R
  - **SQL DDL Programming:** Views, Functions, Stored Procedures
  - **SQL DML Programming:** Create, Read, Update, Delete
- **Database Tooling:** Microsoft SQL Server
- **Other Tools:** Microsoft Visio

4.4.3.3 Subfolder: final_project

The final-project subfolder contains additional subfolders for sql scripts, R coding, and images used in the final project report.
Root Files:

- Readme File
- Grade and Feedback Received
- Final Project Presentation
  - File Name: RyanTimbrook_IST659_Project_Final.pdf

4.5 Course: MBC 638 Data Analysis and Decision Making

4.5.1 *Description:*

This course will familiarize students with the assumptions underlying various statistical techniques and assist in identifying their appropriateness in a variety of situations. The student should be able to perform statistical analysis and interpret results in a meaningful way. Students are expected to relate the results of such analyses to become an information-based decision-maker.

4.5.2 *Learning Objectives*

- Help students understand the value of data collection and analysis in acquiring knowledge and making decisions in today's business environment.
- Students will be able to identify and apply the appropriate statistical technique for a given set of conditions in order to answer a particular question.

4.5.3 *Portfolio Project*

4.5.3.1 Final Project - Process Improvement Project (Individual Project)

4.5.3.1.1 Project Reports

- **Root Directory:** ./mbc638_data_analysis_decsion/final_project
  - **Final Project Presentation**: RTIMBROO-MBC638-ProccessImprovementProject.pdf
  - **Project Definition Statement**: Timbrook_FinalProject_ProblemDefinitionWorksheet.pdf
  - ***Subdirectory***: /00-Requirements
    - **Project Requirements**: ProjectRequirements.pdf

4.5.3.1.2 Objective

Take a business problem and apply the DMAIC framework to communicate and implement a data-driven solution to the problem. DMAIC stands for: Define, Measure, Analyze, Improve, Control.

DMAIC:

- **D**efine: Identify the problem and the team's scope.
- **M**easure: Develop data collection plan and implement it.
- **A**nalyze: Determine root causes; identify and verify critical variables.
- **I**mprove: Develop/select/pilot and then implement a solution.
- **C**ontrol: Put a control plan in place; ensure the problem stays fixed.

4.5.3.1.3 Business Problem

Company X is losing time and cost savings dollars that could be realized by approving Vendor Purchase Order payments within a discount timeframe. Only 20% of Vendors supplied Sprint Teams Purchase Order (PO) Invoices are being approved within the discount threshold of 19 days. Not achieving this discount benchmark costs the company, on average, $5,911 in lost revenue on each PO invoice it processes. Managers and Teams spend to much time validating Sprint Team PO invoices for correctness. This time not only costs in resource dollars, it as well as negative impacts on the amount of software delivery work they can complete each Sprint Cycle, leading to a decreased 'Time to Market' performance measure.

The overall goal of the project is to REMOVE WASTE, in terms of cycle time deficiencies, which will improve EFFICIENCY and PRODUCTIVITY, saving the company a great deal of money.

4.5.3.1.4 Scope

- Create an executive summary slide in a Storyboard format that highlights each of the DMAIC phases' key discoveries.
- Create backup slides for each of the DMAIC phases that outlines the details with which the executive summary is based.

4.5.3.2 Tools and Technologies

- **Statistical Tools/Techniques:** Microsoft Excel (Stats Package)

- Timeseries Analysis
- Linear Regression Analysis
- Chi-squared test for independence
- Correlation Analysis
- Descriptive Statistics
- Probability Distributions / Hypothesis Testing
- Process Control Chart Algorithms

4.5.3.3 Tools and Technologies

- **Workflow Tools:** Microsoft SharePoint
- **Design Tools:** Microsoft Visio

4.5.3.3.1 Subfolder: final_project

The final-project subfolder contains additional subfolders for each of the SDLC phases of the project.

Root Files:

- Readme File
- Grade and Feedback Received
- Final Project Presentation
  - RTIMBROO-MBC638-ProccessImprovementProject.pdf
  - Timbrook_FinalProject_ProblemDefinitionWorksheet.pdf

4.6 Course: SCM 651 - Business Analytics

4.6.1 *Description:*

This course is intended for the graduate student who is interested in developing a portfolio of skills in business analytics. The class discussions are based on case situations and on articles from business and technical publications. The class includes substantial hands-on work in data collection, analysis, and interpretation.

4.6.2 *Learning Objectives*

- Data collection: using tools to collect and organize data (e.g., Google Analytics)
- Data analysis: identify patterns in the data via visualization, statistical analysis, and data mining
- Strategy and decisions: develop alternative strategies based on the data
- Implementation: develop a plan of action to implement business decisions

4.6.3 *Portfolio Project*

4.6.3.1 Assignment - Recruiting Advertising Strategy (Team Assignment)

4.6.3.1.1 Project Reports

- **Root Directory:** ./scm651_business_analytics/assignments/
  - **Assignment Deliverable:** Timbrook_HW2_Team5_GoogleAnalytics.pdf

4.6.3.1.2 Objective

Use Google Analytics to examine the Whitman Graduate Programs Internet marketing campaigns. Perform an advertising analysis using the analytics collected through the Google Analytics platform, identifying patterns and opportunities, to be reported as data-driven recommendations.

4.6.3.1.3 Business Problem

The Whitman School of Management launched an Internet recruiting campaign in February 2011, using Google ads and Delta Airlines flight magazine advertisements. An assessment of opportunities and results is now necessary to establish the direction for next year. The goal of the assessment is to use data-driven insights to recruit the best United States students, measured by GMAT scores, but there's a marketing budget constraint of $100,000. The budget must cover advertising costs, but no Whitman administration costs.

4.6.3.1.4 Scope

- Identify time frames for each marketing campaign and the cost spent on each.
- Identify what the effectiveness was for each of the previous campaigns.
    - whitman.syr.edu
    - MBA Marketing - Full-time
    - MBA Marketing - iMBA
    - Delta
- Identify the key aspects of a United States campaign for next year.
    - which geographic regions should be advertised and why.
    - what key words would be the most effective.
    - which days of the week and what time of day should be focused for advertising and why.
- Identify the costs for the advertising campaigns.
    - By region
    - By degree program
- Specify how to best measure performance of the decisions made after implementation.
- Specify other factors or considerations that are important. And other data that would help in developing an Internet advertising strategy if it was something that could be collected.

4.6.3.2 Tools and Technologies

- **Analytic Service/Platform:** Google Analytics
- **Analytic Tools:** Microsoft Access Database, Microsoft Excel, R, Tableau