-
Syracuse University School of Information Studies
M.S. Applied Data Science

Portfolio Milestone

Courtney Hrdy
555365718

Table of Contents

Introduction

The M.S. in Applied Data Science is a practitioner's degree founded upon firm theoretical underpinnings. The project-based work reinforces the professional underpinnings of data science and gives opportunity for creativity. I have had an incredibly positive learning and professional development experience from this program. I am most grateful for the supportive faculty and peers.

In my portfolio, I will demonstrate achievement of the learning goals as follows:

1. Collect, store, and access data by identifying and leveraging applicable technologies.

I have leveraged a variety of applicable technologies for primary and second-party data collection. In Accounting Analytics, I accessed data from the Securities and Exchange Commission using Inline XBLR filings available in the Electronic Data Gathering, Analysis and Retrieval (EDGAR) system. In Database Administration and Concepts, I applied schema and data modeling techniques to database management system products using Microsoft Azure to store community bank information. For my internship with Thales Group, I collected, stored, and processed sensor data directly from a drone.

2. Create actionable insights across a range of contexts.

I worked with complex contexts of data and enjoyed seeing the universal application of data science techniques to ubiquitous domains including accounting, engineering, human resources, banking, and risk analysis. The CRISP-DM framework was foundational to every project's lifecycle. Financial application included collecting and analyzing loan data to identify risks and adjust algorithms for grading peer-to-peer lending customers. An application to human resources/people analytics included evaluating and identifying key areas of a workplace through survey data mining to help enterprises be more informed of drivers of their workforce productivity. Application to an aerospace context included conducting geospatial analysis for UAS Instrument Landing System (ILS) calibrations to allow military operations to land planes in GPS-denied environments.

3. Apply visualization and prediction models to help generate actionable insight.

I used Tableau to create data visualizations from a UAS ILS Calibration in Groningen Airport Eelde, Netherlands and from a flight conducted at Griffiss International Airport in New York. These visualizations were used to validate flight path coordinates and sensor data output in a format compatible to FAA ILS operations. In Accounting Analytics, Tableau was used to visualize and simplify long-term solvency ratios. Leveraging Tableau for financial data allowed easy calculations and flexible analytics. In terms of prediction modeling, both unsupervised and supervised learning methods were used for recent projects in Applied Machine Learning and Big Data Analytics. Unsupervised learning methods were used for analyzing features of risk in Lending Club's borrowers and used to better understand survey questions associated with productivity in the Australian workforce. Supervised learning models included linear regression, logistic regression, random forest, support vector machines, K-nearest neighbors, and gradient boosting algorithms to generate actionable insight. These models were subsequently evaluated by looking at F1 score, accuracy, recall, and precision outcomes as metrics of their performance.

4. Use programming languages such as R and Python to support the generation of actionable insight.

I used Python packages Sci-Kit Learn, Pandas, NumPy, Matplotlib, Seaborn, and Plotly written in Jupyter Notebook to accomplish our analysis in Applied Machine Learning. For Big Data Analytics, PySpark Machine Learning pipelines and Plotly were used for modeling and evaluation. I used Plotly, Matplotlib, and Seaborn libraries for initial data pre-processing and geospatial analysis of sensor data throughout my internship.

5. Communicate insights gained via visualization and analytics to a broad range of audiences (project sponsors and team leads).

I presented visualizations made in Tableau to the Federal Aviation Administration (FAA) and for meetings with the CTO of Canard Drones. The visualizations enabled the engineers and project managers to pinpoint expected values and better understand the output from the flights. These presentations and facilitated discussions with the experts allowed me to learn and visualize data in a format compatible to how a pilot would look at the data.

6. Apply ethics in the development, use, and evaluation of data and predictive models (fairness, bias, transparency, privacy).

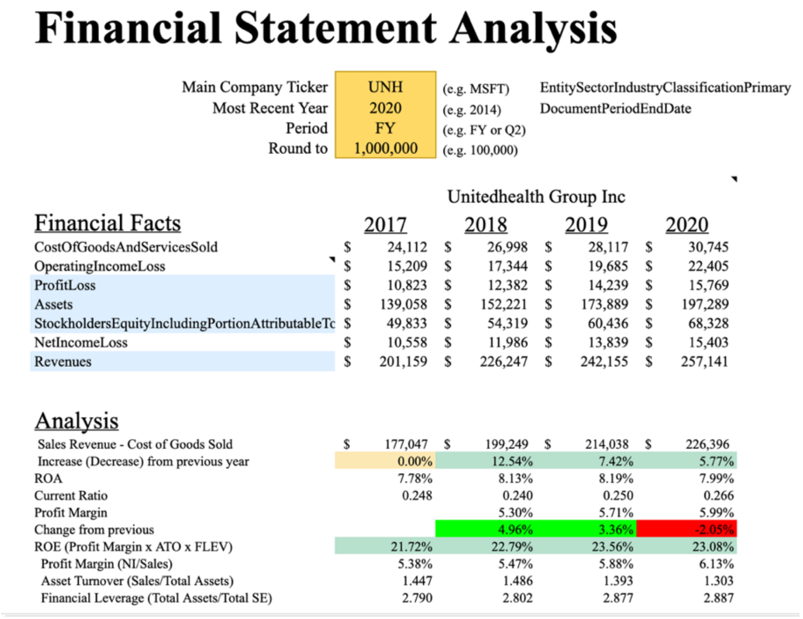
Ethical implications of AI-based lending were evaluated in my Applied Machine Learning project. Bias is an inherent risk to lending algorithms, as they can correlate certain features of the dataset that could result in unintentional discrimination. The disparate impact of a new algorithm could disproportionally exclude borrowers based on terms that violate discrimination laws. Data scientists must be fully aware of the trade-off between legal bias concerns and algorithmic bias concerns—if the wrong data is used for classification, the algorithm could pick up on certain attributes as a deciding factor which ultimately could constitute discrimination. On the other hand, if certain data is excluded from the modeling the results technically risk an exclusion bias without that data. Lending Club is currently attempting to create algorithms that would induce fairer lending and reduce income inequality in financial services, and recently sought guidance from federal agencies. Algorithmic lending does not solely rely on credit score for scoring borrowers, and the use of alternate data in such large quantities must be regulated for algorithmic bias.

ACC 652: Accounting Analytics

Project Description: **DuPont Analysis: Analyzing XBLR Financial Statements in Tableau**

Professor Karen Kukla's Accounting Analytics class introduced analytic techniques used in decision-making by accounting professionals. The focus was the examination of big data involving account information. The course demonstrated how both financial and managerial accountants can benefit from using data analytics and be able to solve accounting and business-related problems with the appropriate data modeling. The assigned case studies required students to deliver tangible insights by analyzing and visualizing patterns in company data with the aid of specialized software such as Tableau and Excel.

The coursework taught how to apply financial ratios to compare company performance by accessing data from the Securities and Exchange Commission using Inline XBLR filings available in the Electronic Data Gathering, Analysis and Retrieval (EDGAR) system. XBLR tags financial data and is a SEC filing requirement for public companies. This assignment required a financial ratios template in Google Sheets connected to the XBLRAnalyst plugin to retrieve customized financial statement data by inputting the company ticker to the template. Some of the XBLR tags differ, making them hard to find (e.g. StockholdersEquityIncludingPortionAttributableToNoncontrollingInterest is the tag for Total Shareholder Equity).



Financial Statement Analysis

Main Company Ticker	UNH	(e.g. MSFT)	EntitySectorIndustryClassificationPrimary
Most Recent Year	2020	(e.g. 2014)	DocumentPeriodEndDate
Period	FY	(e.g. FY or Q2)	
Round to	1,000,000	(e.g. 100,000)	

Unitedhealth Group Inc

Financial Facts	2017	2018	2019	2020
CostOfGoodsAndServicesSold	$ 24,112	$ 26,998	$ 28,117	$ 30,745
OperatingIncomeLoss	$ 15,209	$ 17,344	$ 19,685	$ 22,405
ProfitLoss	$ 10,823	$ 12,382	$ 14,239	$ 15,769
Assets	$ 139,058	$ 152,221	$ 173,889	$ 197,289
StockholdersEquityIncludingPortionAttributableTc	$ 49,833	$ 54,319	$ 60,436	$ 68,328
NetIncomeLoss	$ 10,558	$ 11,986	$ 13,839	$ 15,403
Revenues	$ 201,159	$ 226,247	$ 242,155	$ 257,141

Analysis				
Sales Revenue - Cost of Goods Sold	$ 177,047	$ 199,249	$ 214,038	$ 226,396
Increase (Decrease) from previous year	0.00%	12.54%	7.42%	5.77%
ROA	7.78%	8.13%	8.19%	7.99%
Current Ratio	0.248	0.240	0.250	0.266
Profit Margin		5.30%	5.71%	5.99%
Change from previous		4.96%	3.36%	-2.05%
ROE (Profit Margin x ATO x FLEV)	21.72%	22.79%	23.56%	23.08%
Profit Margin (NI/Sales)	5.38%	5.47%	5.88%	6.13%
Asset Turnover (Sales/Total Assets)	1.447	1.486	1.393	1.303
Financial Leverage (Total Assets/Total SE)	2.790	2.802	2.877	2.887

Figure 1*:The XBLRAnalyst add-on in Google Sheets establishes connectivity to the XBLR API to request and retrieve data from XBLR US public database*

The DuPont framework disaggregates company performance and can help answer the important question of what happened. The DuPont Ratio decomposes return on equity into three parts: profitability (profit margin), activity (asset turnover), and solvency (financial leverage) ratios. Comparing United Health and Cardinal Health showed how highly capital-intensive industries have low asset turnover ratio and need much higher profit margins to achieve competitive ROEs.

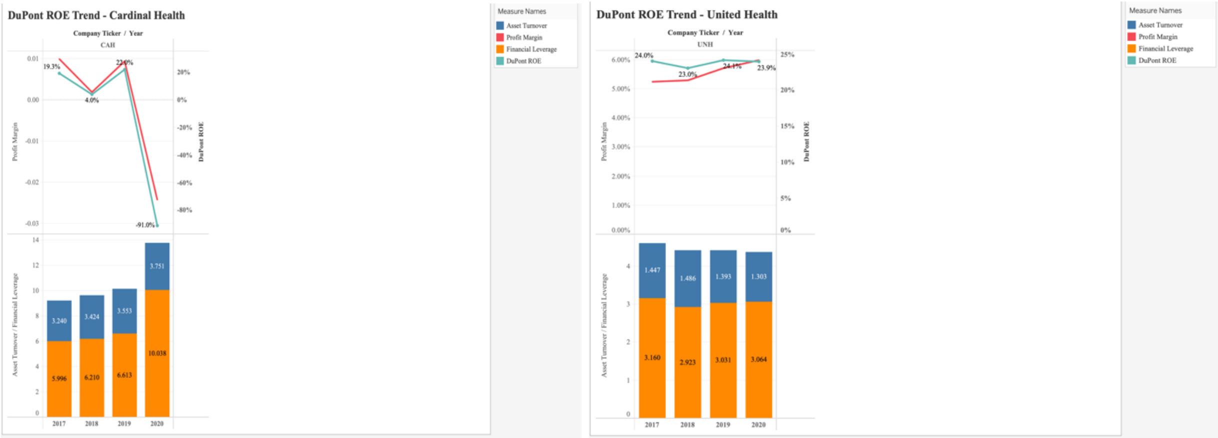


Figure 2*:The DuPont Analysis model would allow an investor to see that although United Health has lower asset turnover and financial leverage, its higher profit margin results in a better ROE than Cardinal Health.*

Reflection & Learning Outcomes

Being a data science student among intermediate accountants, I learned most from my classmates and professor sharing their experiences and knowledge with me. Working in groups with accounting students gave tangible insight to managerial accounting. I used Python and statistical techniques I had learned in other classes to supplement the requirements, and finally leveraged Excel for data cleaning and analysis. The framework of the class introduced four categories of analytics (descriptive, diagnostic, predictive, and prescriptive) and the category of questions they answer. Tableau was used to examine datasets and create calculations to produce long-term solvency ratios and profitability ratios to compare companies.

The various assignment and project deliverables introduced accounting calculations and analytics through visualizations. The Tableau proficiency takeaway from this class contributed to preparation for work in data visualization with financial data, but I have used it with all types of data since learning it. I love working with Tableau and I want to continue learning and developing dashboards to streamline solutions that provide easy to understand insights to operational activities for businesses. The visualization skills I learned in class enabled me to quickly look and get information from messy data in my internship and I will continue to apply these techniques as an analyst. The calculations that were used in evaluating and comparing company performance in Tableau will be a lifelong skill set.

IST 659: Database Administration and Management

Project Description

The course taught the definition, development, and management of databases and their different business applications across industries. The coursework applied schema and data modeling techniques to database management system products using Microsoft Azure. Fundamental data and database concepts included conceptual and logical entity-relationship modeling, Structured Query Language (SQL) programming, and web-based database applications. The database development lifecycle is the framework for building database applications, starting with feasibility and ending with the maintenance of the completed application. The data model is the driver of modern data-oriented applications. Other functional requirements are captured based on user interactions with the data model, defined by user-stories as a high-level functional requirement which focus on the capability of a user within the system to be built. For my final project, the implementation of a credit card portfolio database application for community banks formalized a functional requirement into a formalized user story. PowerApps was used to improve accessibility to assessments and enhance dataflow to staff experts and enhance customer service.

Attribute	**Description**
Bank Name	Community bank name
City	City of branch location
State	State of branch location
Contact Person	First and last name of community bank representative
Phone Number	Phone number of community bank representative
Email	Email of community bank representative
Total Number of Consumer Checking Accounts	Number of consumer checking accounts with the community bank
Total Number of Business Checking Accounts	Number of business checking accounts with the community bank
Cost of Funds Rate	The weighted average of interest rates that banks pay on savings accounts held by their customers and money borrowed from other institutions. That percentage is then divided by 12 and multiplied by the outstandings to show an annualized amount. (Cost of Funds % / 12) x Outstandings = Annualized Cost of Funds in $
Annual Personal Expense Related to Allocated Credit Cards	The expenses entered by your bank are allocated equally by the number of card plans in your portfolio, except for the cost of funds figure.
Promotion participation	Does your bank participate in promotions from Visa, MasterCard, Internal or Employee Incentives

Figure 3*: The ICBA Automated Credit Expert portfolio analysis assessment data requirements*

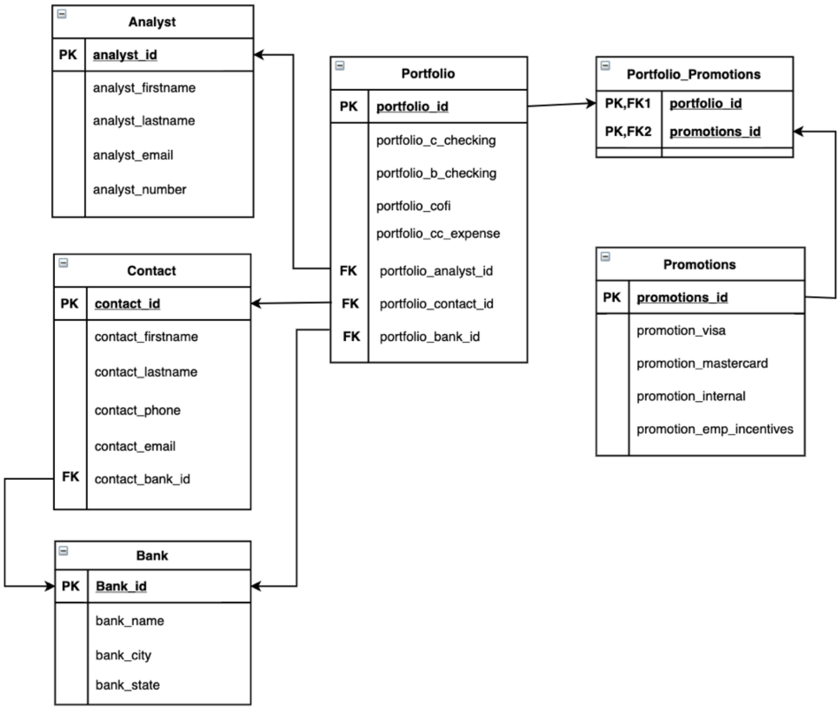


Figure 6*: The logical model*

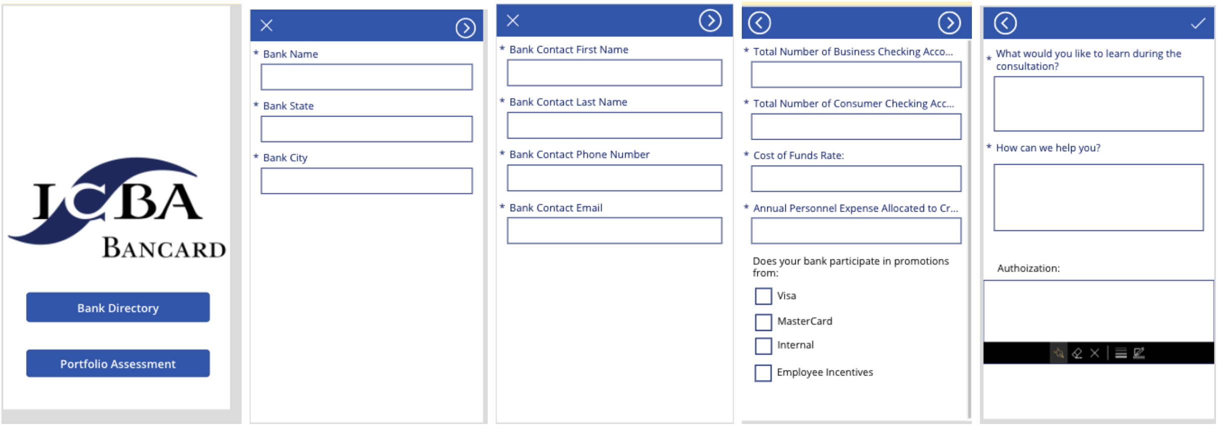

ICBA
BANCARD
Bank Directory
Portfolio Assessment
* Bank Name
* Bank State
* Bank City
* Bank Contact First Name
* Bank Contact Last Name
* Bank Contact Phone Number
* Bank Contact Email
* Total Number of Business Checking Acco...
* Total Number of Consumer Checking Acc...
* Cost of Funds Rate:
* Annual Personnel Expense Allocated to Cr...
Does your bank participate in promotions from:
□ Visa
□ MasterCard
□ Internal
□ Employee Incentives
* What would you like to learn during the consultation?
* How can we help you?
Authoization:

Figure 7*: The PowerApp*

- https://apps.powerapps.com/play/247b6213-b48a-427f-afde-d6c03502599e?tenantId=4278a402-1a9e-4eb9-8414-ffb55a5fcf1e&source=portal&screenColor=rgba%280%2C+176%2C+240%2C+1%29&skipAppMetadata=true
- https://vimeo.com/635581888/4ebbd7b3f4

Reflection & Learning Outcomes

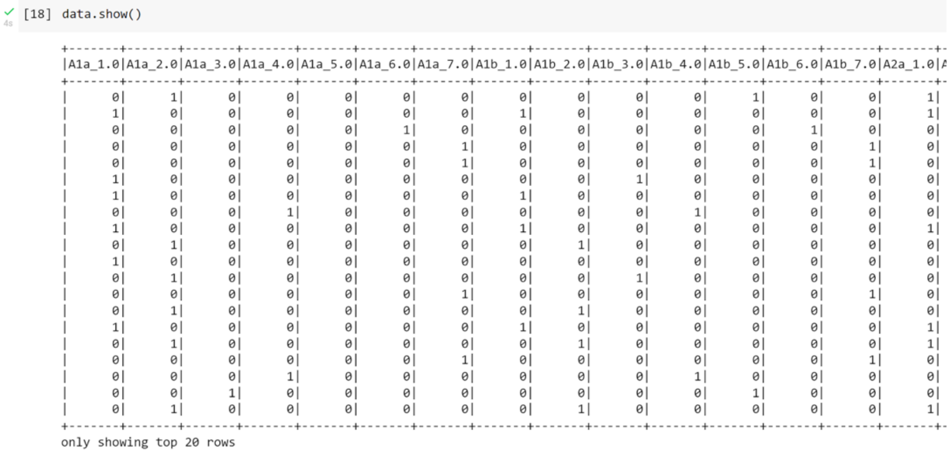
Learning PowerApps was the most challenging yet the most rewarding project in the program. Configuring, designing, and modeling the database requires the right balance between creativity and knowledge. The various assignments throughout the course showed applications of fundamental database concepts and required critical thinking about the user story and business requirements. I liked the creative potential in designing a PowerApp and building a database specifically for banking as a client-based industry, given the opportunity for much needed innovation in banking data management and easy application to PowerApps.

IST 718: Big Data Analytics

Project Description

Our project aims to explore the topic of human resources analytics, specifically on how employees interact with their workplace. There are many existing projects currently in the topic that explore employee attrition, employee absenteeism, employee retention and how financial incentive motivates employees' productivity. However, there is a lack of projects that offer a more holistic view and are people-centric in terms of workplace evaluation. For example, looking at how overall happiness or satisfaction of employees can directly link to increases in productivity. The objective for this project is to evaluate and identify key areas of a workplace, beyond financial incentive, that can increase employee productivity. Productivity is the main driver of profitability since employees are those working towards key performance indicators. By understanding how these key areas boost productivity, human resources departments at enterprises can make better internal investment decisions. In our model, we are expecting to see indirective incentives that improve the work environment such as opportunity for position advancement, advocacy for diversity and inclusion, offering of mentorship by senior leaders, better accommodations, etc. to have a positive impact on employee productivity. With COVID-19 and how the workforce is more attracted to workplaces that are more accommodating and embracing of diverse backgrounds, enterprises would have to lean into these insights to attract and retain their prospective and current employees. Our project was completed in Jupyter Notebook using Pyspark.

To accomplish our project, we used survey data from the Public Sector Commission of Western Australia. The survey's objective is to bring light to feedback on how to enhance integrity, effectiveness, and efficiency to the public sections. The survey was completed by employees of 11 public sector organizations. In 2016, there were a total of 3883 valid responses. The questionnaire has 109 questions. The questions are structured on a scale of 1-8 corresponding to responses between strongly agree - strongly disagree, very satisfied - very dissatisfied, never - very frequently, yes/no. In addition, we have both a question key and a response key that allows us to view each question and each response. Prior to our analysis, we created dummy variables for each question. After this transformation our dataset contained either a 0 or 1 value for each response for each question.

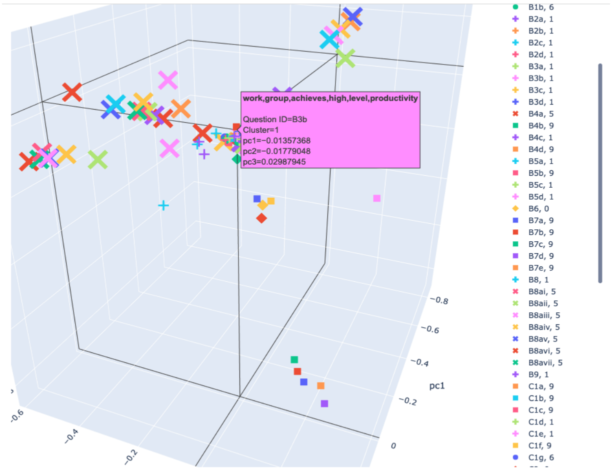


```
[18] data.show()
```

```
+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-
|A1a_1.0|A1a_2.0|A1a_3.0|A1a_4.0|A1a_5.0|A1a_6.0|A1a_7.0|A1b_1.0|A1b_2.0|A1b_3.0|A1b_4.0|A1b_5.0|A1b_6.0|A1b_7.0|A2a_1.0|A
+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-
|      0|      1|      0|      0|      0|      0|      0|      0|      0|      0|      0|      1|      0|      0|      1|
|      1|      0|      0|      0|      0|      0|      0|      1|      0|      0|      0|      0|      0|      0|      1|
|      0|      0|      0|      0|      0|      1|      0|      0|      0|      0|      0|      0|      1|      0|      0|
|      0|      0|      0|      0|      0|      0|      1|      0|      0|      0|      0|      0|      0|      1|      0|
|      0|      0|      0|      0|      0|      0|      1|      0|      0|      0|      0|      0|      0|      1|      0|
|      1|      0|      0|      0|      0|      0|      0|      0|      0|      1|      0|      0|      0|      0|      0|
|      1|      0|      0|      0|      0|      0|      0|      1|      0|      0|      0|      0|      0|      0|      0|
|      0|      0|      0|      1|      0|      0|      0|      0|      0|      0|      1|      0|      0|      0|      0|
|      1|      0|      0|      0|      0|      0|      0|      1|      0|      0|      0|      0|      0|      0|      1|
|      0|      1|      0|      0|      0|      0|      0|      0|      1|      0|      0|      0|      0|      0|      0|
|      1|      0|      0|      0|      0|      0|      0|      0|      0|      0|      0|      0|      0|      0|      0|
|      0|      1|      0|      0|      0|      0|      0|      0|      0|      1|      0|      0|      0|      0|      0|
|      0|      0|      0|      0|      0|      0|      1|      0|      0|      0|      0|      0|      0|      1|      0|
|      0|      1|      0|      0|      0|      0|      0|      0|      1|      0|      0|      0|      0|      0|      0|
|      1|      0|      0|      0|      0|      0|      0|      1|      0|      0|      0|      0|      0|      0|      1|
|      0|      1|      0|      0|      0|      0|      0|      0|      1|      0|      0|      0|      0|      0|      1|
|      0|      0|      0|      0|      0|      0|      1|      0|      0|      0|      0|      0|      0|      1|      0|
|      0|      0|      0|      1|      0|      0|      0|      0|      0|      0|      1|      0|      0|      0|      0|
|      0|      0|      1|      0|      0|      0|      0|      0|      0|      0|      0|      1|      0|      0|      0|
|      0|      1|      0|      0|      0|      0|      0|      0|      1|      0|      0|      0|      0|      0|      1|
+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-
only showing top 20 rows
```

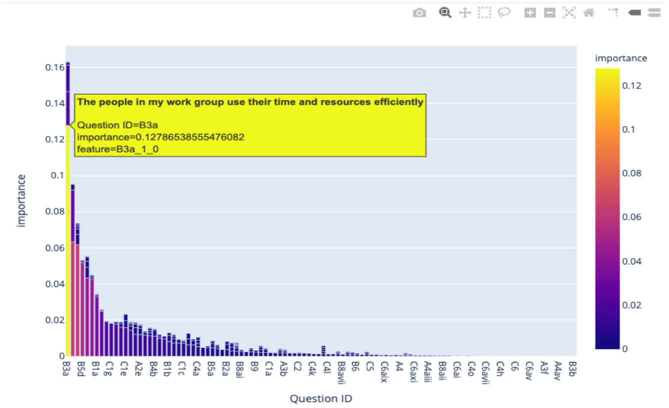
After creating dummy variables for each question response, our data frame contained 559 columns. We performed a correlation matrix as seen below. The highest correlation value was .44, telling us that our data is not highly correlated. Looking at the Agency Size attribute, we found that 63% of the responses were made by employees at large agencies with over 1,000 employees. Roughly 30% of responses were recorded from employees at agencies with 200-1,000 employees and 7% with less than 200 employees. Our target variable is B3b_1. This variable corresponds to the response, 'Strongly Agree', for the question: My work group achieves a high level of productivity. Most employees responded with strongly agree and agree.

For exploratory data analysis of the questionnaire, the normalized features of the filtered words from the TFIDF matrix of the questions were mapped into a low dimensional space using the Principal Component Analysis (PCA). The beauty of having the PCA is the ability to reduce the number of variables by combining them into meaningful features where they are orthogonal to each other. The questions clustered with the target variable would be expected to have high feature importance or correlation in the answers.



3D PCA

The Random Forest is used to look at features by their mean decrease in impurity to evaluate their importance in classification. As an ensemble method of decision trees, random forest removes correlation by fitting decision trees to subsets of the features and is a good way to assess feature importance without overfitting the data. Machine learning applications to survey research have recently been used to estimate response propensities and are able to process given responses to make predictions. Survey answers that have the highest feature importance in relation to the target variable had responded with being very satisfied to the following questions: “The people in my workplace use their time and resources efficiently”, “The people in my workplace are committed to providing excellent customer service and making a positive impact in the community”, “In the last 12 months my company has implemented innovative processes and policies”, and “Your coworkers treat employees from all diversity groups with equal respect.” Contrary to the clustering results, the questions with the highest feature importance are The colors in the following visualization represent the number of times the feature was within the topmost important features among all questions.



Link to Random Forest Feature Importance

Accuracy	Precision	Recall	F1 Score
0.81	0.85	0.70	0.77

From our results, we see that the Random Forest has fewer false positives than it does false negatives, meaning each tree can capture impurity but it is overfitting the data. However, evaluation of model performance transcends looking at the traditional metrics, as we are predicting a nuance of human behavior. The algorithm is making predictions using survey answers without understanding the sentiment of the question. In contrast, our unsupervised learning models used the transformed TF-IDF features of the questions without learning the correlation of responses. To holistically evaluate human factors like productivity, the combination of unsupervised and supervised learning methods needs to evaluate interaction sequences that reveal cognitive and emotional elements of the predictions to bridge the gap between computational algorithms and human reasoning. So, if we are only going to make decisions from this dataset using the basic metrics (accuracy, precision, recall, f1 score) we are missing out on understanding user responses.

Reflection & Learning Outcomes

I am most proud of my accomplishments from this class, as I really learned so much and feel very confident in my ability to apply different machine learning pipelines and evaluation metrics to any dataset. I most enjoyed applying deep learning frameworks for facial recognition, using unsupervised learning methods for different analyses (recommendation systems, risk analysis, natural language processing, etc.), and transforming messy data into high-value prediction models. I hope to never lose the skills I earned from this class and look forward to harnessing them in the future. I will continue to learn and leverage innovative applications using the coding, mathematics, and data science common sense that was taken from this class.

IST 707: Applied Machine Learning

Project Description

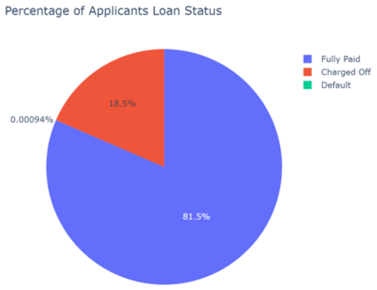
This course introduces popular data analytics methods for extracting knowledge from data. The focus of this course is understanding data and how to formulate data analytics tasks to solve problems. Topics of the course include the key tasks of data analytics, including data preparation, concept description, association rule analytics, classification, clustering, evaluation, and analysis. The exploration of the concepts and techniques of data analytics and practical exercises will provide skills that can be applied to business, science, or other organizational problems.

The objective of the Risk Scoring Loans assignment is to propose a new model that would potentially assign a risk score to loan applicants. The objective is to clean, merge, and analyze LendingClub's dataset of loan applications to identify target variables indicative of risk. LendingClub provisions 3-5 year loans between $1,000 and $35,000 to accepted applicants based on a risk score calculation. The assigned risk score to all loan applicants determines whether a loan is accepted, and the interest rate the approved loans will receive. The risk score that has been used requires evaluation and improvement, given the increase in defaulted loans.

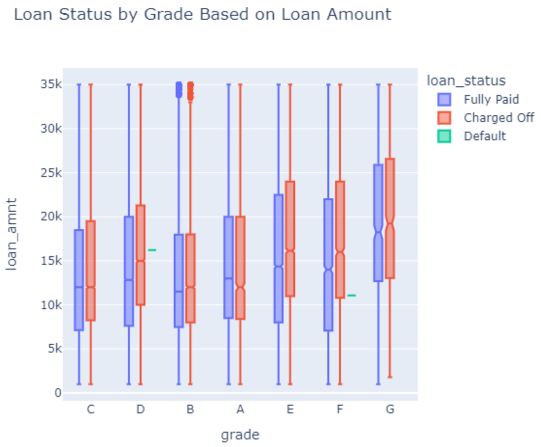
LendingClub is looking for a new proposed model that would help potentially assign a risk score to loan applicants. loans dataset contained 235,629 records with 107 metadata attributes.

To accomplish our analysis, we used Python packages Sci-Kit Learn, Pandas, NumPy, Matplotlib, Seaborn, and Plotly written in Jupyter Notebook. We completed the following models: SVM, KNN, Gradient Boosting, Random Forest, and Logistic Regression.

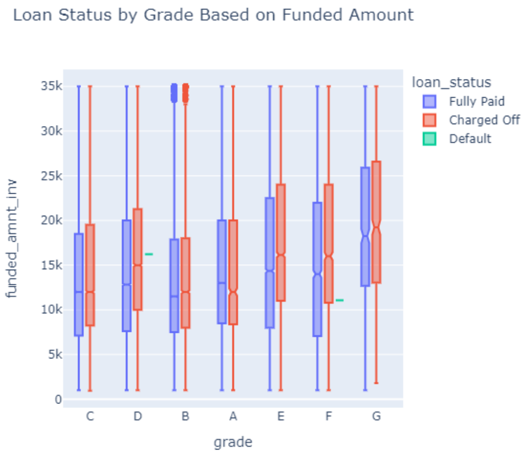
Most applicants have fully paid off their loan. However, there is a large percentage of applicants that have charged off their loan. When a loan is charged off, it means that Lending Club has given up on being repaid according to the original terms of the loan. This does not mean that the applicant is no longer responsible for the amount owed, but rather Lending Club has removed that loan from their balance sheet. We then removed the loans that have been classified as in grace period, late, and current to better visualize the percentage of loans that were fully paid and charged off. Our dataset contains just 2 records of applicants who have defaulted, so we are using ‘charged off’ as our target variable throughout our analysis. The following pie chart depicts the percentage of fully paid and charged off loans:



We wanted to look at the proportion of fully paid and charged off loans broken down by the applicant’s grade and loan amount. We found that the lowest grade has the highest loan amounts, interest rate and funded amount for charged off loans. Better grades have lower interest rates and have less funding. This is indicative of LendingClub algorithm optimizing net annualized return for investors without accounting for risk properly. We can see that they are prioritizing loans with high interest rates to receive higher loan amounts. The following boxplots depict these findings:



Loan Status by Grade Based on Loan Amount



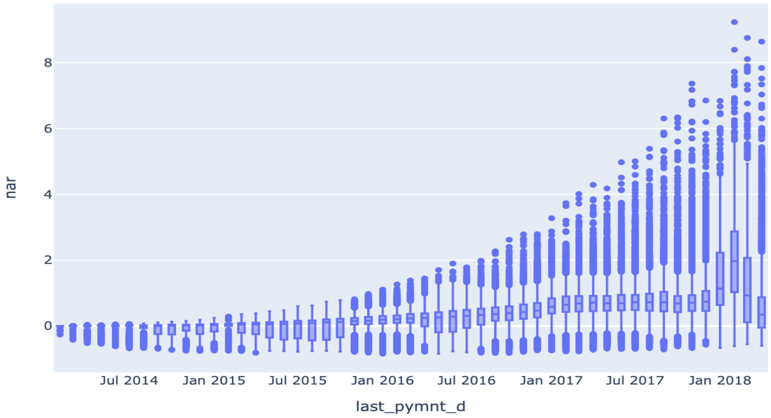
Loan Status by Grade Based on Funded Amount

Net Annualized Return Calculation

Net Annualized Return (NAR), according to LendingClub, is “an annualized measure of the rate of return on the principal invested over the life of an investment. NAR is based on the actual Borrower payments received each month, net of service fees, actual charge off amounts and recoveries.” We added a column to for NAR to our dataframe using the following calculation:

(Total Payment / Funded Amount) ** (1 / (365/ days)) - 1

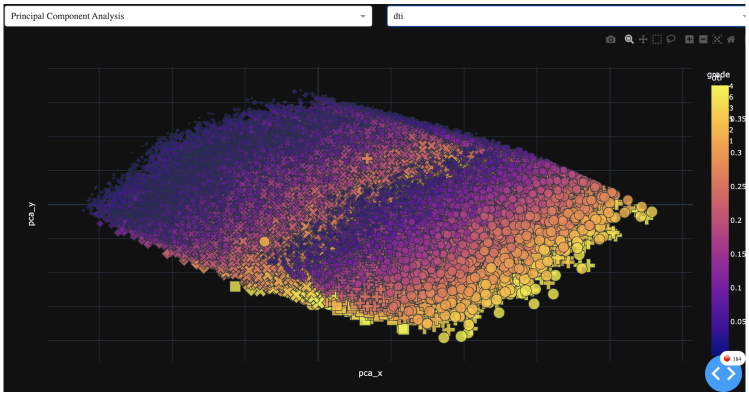
We then plotted NAR dependant on the last payment date. From the plot below, you can see that NAR began to decrease in April 2018



Model: PCA

The Principal Component Analysis is used to map standardized features into a low dimensional space where they are separated by maximum variance. Dimensionality reduction keeps relationships between variables but changes the dimensionality by producing a series of linear combinations of the data where each linear combination is uncorrelated with the others.

This allows us to look at the data in a space where they are orthogonal to each other. PCA is used extensively in business and finance for risk analysis. By creating a principal component representation of the portfolio of borrowers, an investor can analyze the weights of different borrowers to better understand the percentage return on a statistical risk factor. From the PCA constructed using all borrowers in the dataset, the negative values in the first principal component correspond to borrowers with lower interest rates and lower debt to income ratios, indicating lower risk. This makes sense as the grades with negative values range from A-C. The positive grades have higher interest rates, higher debt to income ratios indicating higher risk. This also makes sense, as the positive grades correspond to grades F and G. Applying PCA allows the data to be clustered by matching systematic risk and return characteristics without any prior knowledge of their fundamentals.



PCA Visualization Dashboard

Model: SVM

Support Vector Machines are a powerful supervised learning classification algorithm. The SVM finds a hyper-plane creating a boundary between the types of data and maximizes that boundary. We ran three SVMs using the Linear Kernel, Gaussian Kernel, and Polynomial Kernel. We first ran the SVMs using all of the numeric columns in our dataset with the target variable being ‘charged_off’. We used the train_test_split method from sklearn to split our data with a test size of .33. The Linear Kernel and Gaussian Kernel performed the same with the following results:

Accuracy	F1	Precision	Recall
.87	.53	1	.36

The Polynomial Kernel performed slightly worse:

Accuracy	F1	Precision	Recall
.84	.31	.99	.18

Model: Random Forest

Forest models are an ensemble of decision trees, each one able to predict its own response to a set of input variables. Each tree is created on a different sample selected with replacement, so the different combinations of cases and inputs for the splits are more varied. Results are combined to provide the final prediction, in which the strongest association with the target is used in the splitting rule from all available inputs. The feature importance is computed using the mean decrease in impurity and expressed relative to the maximum.

Accuracy	F1	Precision	Recall
.84	.31	.99	.18

Model: Gradient Boosting

Gradient Boosting combines weak ‘learners’ into a single strong learner iteratively. Each iteration is weighted based on whether misclassification increases in the current iteration, thus higher misclassification has higher weight. The weight indicates the likelihood each case is selected again in the iterative sample. The focus on the misclassification count is what ultimately optimizes model performance.

Accuracy	F1	Precision	Recall
.99	.96	1.00	.93

Model: KNN

K-Nearest Neighbor (k-NN) reads in all training examples, and immediately makes predictions based on the similarity between the test example and all training examples with the majority-voted category label in the k nearest training examples. The decision boundary has no predefined shape and there are no assumptions made about independence. Given that kNN is sensitive to noisy training data and works best when all attributes are relevant to prediction, the algorithm likely was unable to filter some of the inconsistencies to make predictions.

Accuracy	F1	Precision	Recall
0.83	.45	.61	.35

Model: Logistic Regression

Logistic regression performed the best of all of our models, indicating that probability algorithms work best with our data. Ranking features by highest coefficient weights descends from funded_amnt (total amount funded for the loan), total_rec_int (the interest received to date), revol_bal (the amount of credit the borrower is using relative to all available revolving credit), total_il_high_credit_limit (the total installment/high credit limit), and the installment (the monthly payment owed by the borrower if the loan originates). Logistic regression has a discrimination threshold (the cutoff imposed on the predicted probabilities for assigning observations to each class) that would allow Lending Club decision-makers to adjust cut-off values for installments based on this model. The logistic regression supports the initial analysis we made, in that higher loan amounts are given to lower grades without accounting for risk factors.

Accuracy	F1	Precision	Recall
1.00	.96	1.00	.93

Reflection & Learning Outcomes

This class provided foundational conceptual and quantitative underpinnings of machine learning strategies. The ability to experiment with algorithms and their applications to different data (given certain evaluation objectives in each assignment) allowed me to apply combinations of models using Spark and scikit-learn to achieve high-value predictions that would guide better decisions. The ability to evaluate the models and understand how to solve non-standard problems will contribute to all work I do with machine learning in the future. My final project has largely sparked interest in understanding the intersection of privacy laws and machine learning, and I would support any company trying to build trust as well as government accountability for machine learning decisions. Given the lack of coding requirements for the class, I applied all the Spark learned in IST 718 to complete projects and assignments. Spark is very useful for relatively simple machine learning with big data, but for the smaller datasets and more complicated frameworks used in assignments I built models using scikit-learn. I intend to leverage learning more sci-kit learn and continue to advance in Spark machine learning pipeline techniques in my program of life-long learning.

IST 974: Internship in Data Science

Project Description

Over the summer I had the opportunity to participate in a co-op through the Center of Advanced Systems and Engineering (CASE) at Syracuse University, where I worked as a data science intern for Thales Group. I cleaned, analyzed, and visualized output from an Instrument Landing System (ILS) receiver collected by drone flight. I learned the data requirements for an ILS inspection by an airplane in a US airport, a drone in a European airport, and for calibration by the USAF in a GPS denied environment. The internship allowed me to exercise the basic data cleaning and geospatial visualization tools in Python and Tableau that I have learned over the past year. On the operational side, I learned about differences and similarities in ILS calibration between drones and airplanes. I learned how the positioning of a drone can be tracked in a GPS denied environment using a Trimble (a surveying equipment used by civil engineers) to get the polar coordinates. I learned how to interpret the frequency, modulation, and signal strength of antennas from the positioning of a drone using longitude and latitude to calculate its angle and distance. Everything I learned came from asking the right questions to the subject matter experts. Making these connections not only helped me learn more about the dynamics of the project I was working on, but they also provided me with valuable resources and data. I had never worked with drones, trigonometry, Trimble, or aeronautical data and was learning everything for the first time. I performed extremely well and was offered a full-time position. By the end of the internship, I started to bridge the gap between UAS and airplane ILS calibration from ICAO to FAA standards.

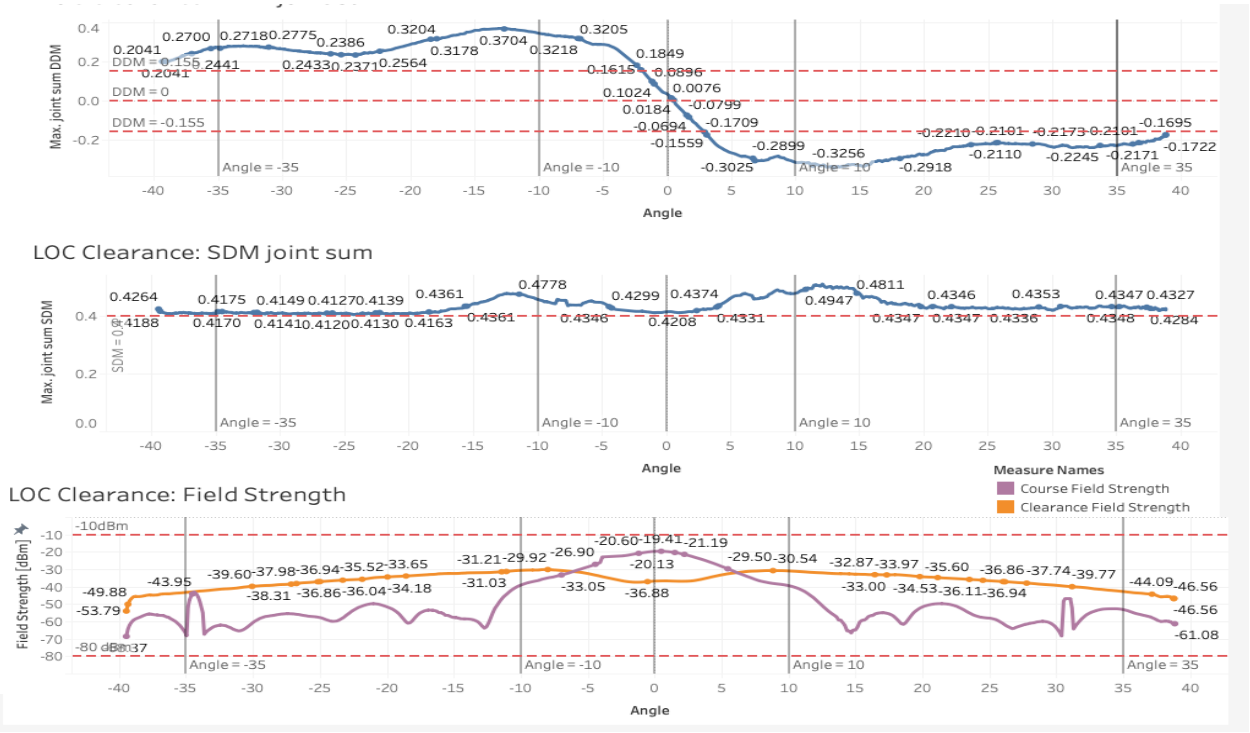


Figure 8: *Visualizations of sensor data collected from drone ILS inspection in the Netherlands*

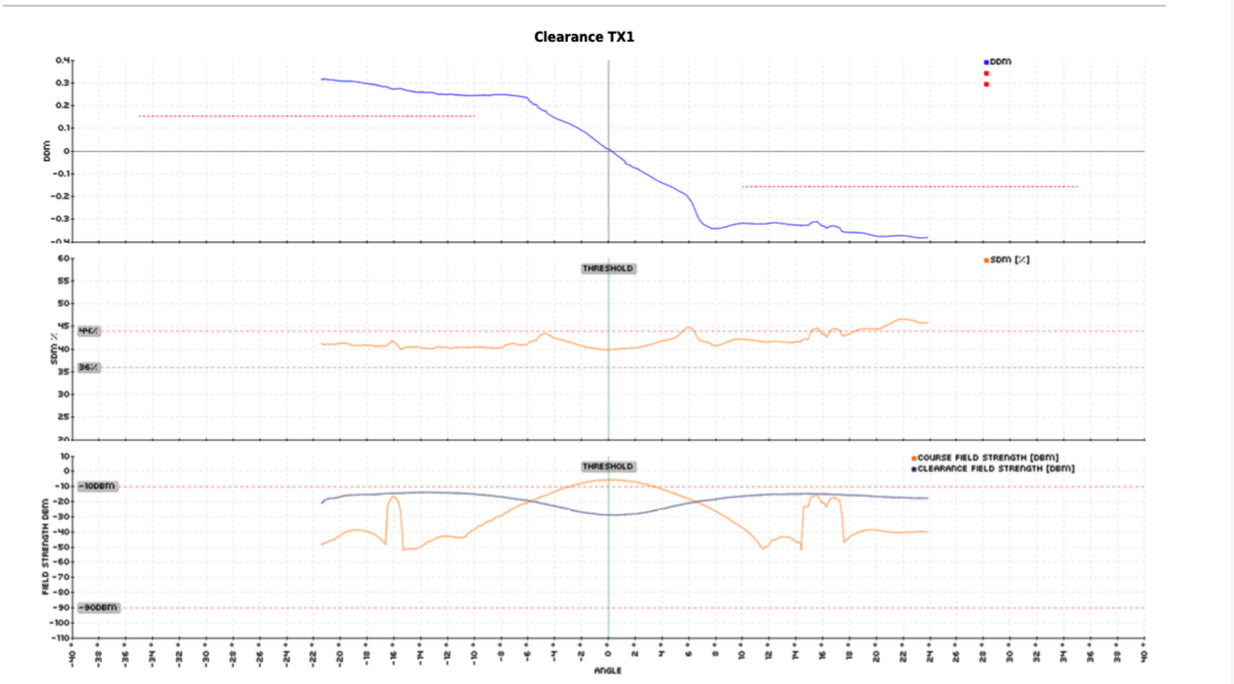


Figure 9: *Visualizations of sensor data collected from airplane ILS inspection in Spain*

Reflection & Learning Outcomes

I learned how hard work, eagerness to learn, commitment to solving a problem, and being personable with the people you work with makes a great impression. Being that this was an in-person internship, I found that making the effort to be present and meet with people in the office was important. I found the culture of the organization to be friendly and supportive. I learned to take initiative in communication, and that being prepared before asking questions will lead to better conversation and more long-term learning. I found the most important soft skills were showing up early, being prepared, and eagerness to learn. The leadership had high expectations for my performance. As I was learning the dynamics of the project for myself, I took initiative to discuss questions and progress regularly. Towards the end of the internship, I wanted to focus on using what I had learned to complete changes in the project requirements individually and I found that in working completely alone I struggled more to complete meaningful work. Collaboration and conversation are important, especially when you are learning something new. Outside of the organization, I worked with CANARD Drones and the FAA. These organizations were the sole insight and main characters to everything I was doing; they were the key subject matter experts I learned from. I found that being personable with these subject matter experts, demonstrating efforts to learn the data (e.g., showing them visualizations with prepared analysis questions) went a long way in building relationships, trust, and getting more resources to understand the data. I learned most from CANARD and the FAA and working with them was my favorite experience of the internship.

I have never studied anything related to what I was working on. I brought the Tableau skills I learned in Accounting Analytics. Learning the engineer approach to data science was completely new to me and a challenge, that is why I wanted to do this internship. I would like to learn more about business analytics in planning the operation. On the technical side of things, I am eager to learn more about autonomous analytics and advanced visualization tools. I would also like to improve my data science workflow and project management over the semester. During the internship, I reached out to the civil engineering professor who teaches the class on surveying. He met with me and showed me the basics of the Trimble in Link Hall, and I introduced him to the organization. He offered good advice to the project, and he even came to the flight at Griffiss. From this experience, I learned that sourcing subject matter experts to the organization is incredibly helpful. I am so grateful for the faculty support and wish I could take Professor Joyce's class.