Ronen Reouveni
Portfolio Milestone

Table of Contents

Introduction 2

Natural Language Processing 4

Rotten Tomatoes - Deep Learning and Sentiment ---------------------------*4*

Overview ---4

Description---*4*

Reflection --*5*

Learning Goals ---*6*

Trip Advisor - Web Scrapping and Machine Learning-----------------------*7*

Overview ---7

Description---*7*

Reflection --*8*

Learning Goals ---*9*

Computer Vision 10

Chest X Rays — Transfer Learning in Medicine ----------------------------*10*

Overview---10

Description ---*10*

Reflection---*11*

Learning Goals --*12*

Fashion MNIST - Crushing Accuracy with CNN's---------------------------*13*

Overview---13

Description ---*13*

Reflection---*14*

Learning Goals --*14*

Time Series 15

Zip Codes - Predicting the future with RNN's ------------------------------*15*

Overview---15

Description ---*15*

Reflection---*16*

Learning Goals --*17*

Regression and Classification 18

Bias in Machine Learning - How code effects people ----------------------*18*

Overview---18

Description ---*18*

Reflection---*19*

Learning Goals --*19*

Introduction

My goal as a Data Scientist is to tell unique and deeply insightful stories through data. The Masters of Applied Data Science program at Syracuse has given me the skills and knowledge to solve complex problems with data-driven solutions. I have excelled in the program, finished with a perfect 4.0, and gained a deep understanding of machine and deep learning. In doing so I have mastered the skills needed to succeed in a data science career. Furthermore, my time in the program has allowed me to develop a portfolio of wide-ranging topics in Data Science. My portfolio has extensive work in natural language processing, computer vision, time series, traditional regression, classification, and database architecture projects. In this paper, I will highlight a selection of them. These research areas have allowed me to implement deep convolutional neural networks for computer vision and recurrent neural networks for NLP and time-series tasks. My background in object-oriented programming allowed me to push my Python skills and gain expertise in Tensorflow and Keras. My portfolio shows that I have a grasp of the primary practice areas of Data Science. This portfolio overview document has a brief explanation of each project followed by a discussion of its impacts on me and relevance to the learning outcomes directed by the program. Each piece of work discussed contains links to the code and a detailed report or presentation associated with that project. The reports contain technical information but are generally understandable to non-technical people, such as managers or investors. This outline of my portfolio serves to demonstrate how I achieved mastery of the MSADS learning goals. Each project discussed in this narrative will demonstrate achievement of one or more of the following learning goals.

1. Describe a broad overview of the major practice areas in data science
2. Collect and organize data
3. Identify patterns in data via visualization, statistical analysis, and data mining.
4. Develop Alternative strategies based on the data.

5. Develop a plan of action to implement the business decisions derived from the analyses.
6. Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization.
7. Synthesize the ethical dimensions of data science practice (e.g., privacy)

Each of these learning goals is discussed in at least one project. In their entirety, my projects shown in this portfolio fully demonstrate mastery of these learning goals. Throughout this narrative, each learning goal is directly connected to at least one project and combined these learning goals provide a well rounded education in the field. This document is split into sections based on major practice areas of data science. In doing so, I cover computer vision, natural language processing, traditional regression, and classification, as well as time series problems. Although learning goal #1 is referenced throughout this document, this introduction and narrative demonstrate achievement of learning goal #1; *describe a broad overview of the major practice areas in data science*. This introduction alone discussed computer vision, NLP, time series, and traditional classification and regression. This document is broken down by field, followed by the projects within that field. Each project section contains an overview, description, reflection, and learning goals. The overview contains which learning goals will be discussed, the GitHub link, and technologies used in that specific project. In the description and reflection, the project and results are outlined. Finally, the learning goals section formally connects each learning goal to that project. This structure, and wide range of topics covered show an ability to work with many areas of data science. Finally, it highlights my expertise and research. It clearly portrays an understanding of significant areas of interest within data science, and therefore, helps to demonstrate my achievement of learning goal #1. My main goal professionally is to work as a Data Scientist using Tensorflow or Pytorch and implement neural networks.

Natural Language Processing

Rotten Tomatoes - Deep Learning and Sentiment

Overview

Learning goals demonstrated below:

- Collect and organize data (#2)
- Identify patterns in data via visualization, statistical analysis, and data mining (#3)
- Develop Alternative strategies based on the data (#4)
- Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization. (#6)

Code:

https://github.com/RonenReouveni/RottenTomatoe

Technologies used:

- Python
- NLTK
- Tensorflow
- Keras
- SKlearn

Description

The Rotten Tomatoes data set contains 156,000 phrases from movie reviews and their accompanying sentiment label. These labels range from 0-5, or very negative to very positive. In this project, I implemented various machine and deep learning models to predict sentiment. I also built a custom sentiment analyzer to improve upon the VADER and TextBlob sentiment analyzers. Features were extracted from the 156,000 phrases in order to fit the various machine learning models. These features included unigrams, bigrams, trigrams, and my custom sentiment polarity score. My custom sentiment analyzer uses TextBlob to get the sentiment polarity of each individual word in the desired string. The analyzer then iterates over the sequences and keeps a running sum of the polarity. While calculating this sum, the values are augmented based on several rules I defined. For example, words like ‘although’ will flip the sign of the previous set of words summed sentiment polarity score. Furthermore, it uses negation words such as ‘neither’ to augment its running polarity state. In the project report, I show that my custom analyzer is far superior on sentences and phrases than TextBlob is on its own. I used this custom analyzer on each phrase in the data set and add it as a feature used to predict sentiment. Both Naive Bayes and Random Forests were implemented using all of the features mentioned above. My custom sentiment feature was the most important among the unigrams, bigrams, and trigrams. However, the best results were from using deep learning. I implemented multiple LSTM’s using Tensorflow and Keras. These models all differed in-depth, the number of units, use of dropout, and window size. Within this paradigm, the original structure of the sequences of words is not degraded; between this fundamental concept and the sheer number of trainable parameters in the model, 505k, the results are far superior.

Reflection

This project allowed me to clean and preprocess text to fit machine and deep learning models. It contributed significantly to my growth as a data scientist and forced me to think differently about modeling data. Traditional machine learning paradigms destroy the order of words. This degradation is evident in the conversion of the sequenced text into a feature set of n-grams. However, an LSTM will keep the sequence of words intact. Modeling this specific data using RNN’s is a much better fit and therefore yields higher results. This project is an excellent example of how to develop alternative strategies based on data. It also shows a deep understanding of how to implement RNN’s in Tensorflow and create the pipeline necessary to clean and process the data as needed. Furthermore, this project shows my ability to identify patterns in data through analysis. The RNN’s resulted in excellent results on the validation set. This is excellent preparation for more research and work in the field of NLP. There is an extremely high demand for Data Scientists who can work in NLP and have the ability to fit RNN's in Tensorflow and create pipelines. This project has allowed me to express these skills.

Learning Goals

In this project, I showed mastery of various learning goals. One of these was learning goal #2: *collect and organize data*. I did this by creating the necessary data pipeline to use the models I did. Intensive formatting and preprocessing were needed to accomplish this because of Tensorflow's specific requirements. Often, these requirements are distant to the format data comes in. Without an ability to organize data this research would fail. This too shows achievement of learning goal #2.

The analysis in this project also demonstrates learning goal #3: *identify patterns in data via visualization, statistical analysis, and data mining*. I was able to mine for various patterns and gained insight from the models deployed. This included, analyzing where sentiment differences are not separable, such as closely related ideas and emotion. Uncovering these patterns was best achieved by classification algorithms. This demonstrates the multitude of information that can be extracted using data mining and therefore helps to support mastery of learning goal #3.

Finally, the work also clearly showed a mastery of learning goal #4: *develop alternative strategies based on the data*. In the report, I showed that specific strategies outperformed others, given the data. For example, using an LSTM will be a better strategy than using a traditional machine learning algorithm. This is because an LSTM can maintain the order of words where the latter model will destroy the sequences. It demonstrates mastery of the learning goal #4 in terms of trying multiple approaches to achieve optimal results, and developing alternative and individualized strategies based on the data.

I also synthesized the findings of the research into a report to communicate the results. This demonstrates learning goal #6: *demonstrate communication skills regarding data and its* *analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization*. In this report, I was able to communicate the ideas and findings in a manner that was digestible for programmers and non-technical management. This means that both technical and non technical people can gain insight and understanding of the data and its trends.

Trip Advisor - Web Scrapping and Machine Learning

Overview

Learning goals demonstrated below:

- Describe a broad overview of the major practice areas in data science (#1)
- Collect and organize data (#2)
- Identify patterns in data via visualization, statistical analysis, and data mining (#3)
- Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization (#6)

Code:

https://github.com/RonenReouveni/TripAdvisor

Technologies used

- Python
- BeautifulSoup
- MongoDB/pymongo
- NLTK
- SKlearn
- Scipy
- XGBoost

Description

In this project, I demonstrate the ability to build a custom web scraper and use it to build my own data set from Trip Advisor. I chose 20 different hotels worldwide and used the web scraper to obtain every individual review and rating, average price, and address. This resulted in around 13,593 reviews. The web scraper then parsed the data and inserted it into a local mongo database. This program was built so that the only thing needed was to choose a hotel, and it would be scraped of the relevant information, parsed, and inserted into the mongo database. In a separate script, I queried this local mongo database, implemented various machine learning models, and mined for multiple patterns. Using XGBoost and Random Forests, I was able to predict a user's rating based on the text of their review. Furthermore, I was able to show which words were associated with good reviews given a hotels' price. This unlocked interesting trends such as cheaper hotels being given five stars for cleanliness while more expensive hotels received five stars for location or amenities. I was also able to show that more expensive hotels are not necessarily associated with better reviews.

Reflection

This was an essential project in my progression as a data scientist because I successfully scraped the web for my own data, parsed and cleaned it, and implemented multiple machine learning models. This project demonstrates my ability to collect and organize data. Not only does this project scrape the web to build my own data set, it also cleans this data and stores it in a MongoDB for later use. This shows a deep understanding of collecting data, organizing it, and then storing it for easy access at any point. Structuring my program like this allowed me to store interesting data that was ready to be used for any relevant modeling or mining tasks. In real-world Data Science roles, there are not pre-made ready datasets. There is an expectation of being able to procure one's own data. This project helped prepare me for a position as a Data Scientist by demonstrating that I can collect, mine, and gain insights into raw internet data. I was also able to scrape nearly 14k reviews in a matter of minutes. It would not be challenging to add hotels and increase the size of this dataset. In this project, I was also able to mine interesting patterns and visualize them. For example, a goal of the research was to discover what type of words described a 5-star hotel (in terms of reviews) given price and geographic location. Wordclouds were used to compare highly-rated hotels in different places and price points. Mining data scraped from the web gave me confidence in my ability to work in NLP and Data Science professionally. Furthermore, I can develop a business plan for any hotel that wants to increase its ratings. With this program, I can easily scrape data from similar hotels and compare why one does better than another according to their online reviews. This is a pure example of using a data-driven approach to implement business decisions.

Learning Goals

By scrapping the web for my data I showed mastery of learning goal #2,: *collect and organize data*. Scraping the data and not using a pre-made dataset allowed me to demonstrate this achievement. Furthermore, this project highlighted the ability to mine patterns and gain insight via data.

This also shows achievement of learning goal #3: *identity patterns in data via visualization, statistical analysis, and data mining*. I was able to extract patterns in why hotels with different price points received good or bad reviews. For example, service became more of a concern as the price increased, and as the price dropped, cleanliness did. Mining these patterns via data shows achievement of learning goal #3: *identify patterns in data via visualization, statistical analysis, and data mining*. An ability to identify meaningful patterns is at the cornerstone of providing useful results from data, and critical to data science success.

Finally, this project also included a technical report that conveyed the ideas and methods used to conduct the research. This report displays achievement of learning goal #6: *demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization*. In my report, I communicated the analysis such that it was understandable to people across fields. Although one did not need programming skills to understand, it also included code and technical pipelines. Mastery of this learning goal requires acknowledging that different types of readers will need to be able to view and extract useful information from the technical analysis. This facet of providing results is critical in ensuring that analysis can be grasped and its conclusions acted upon.

Furthermore, this report also demonstrates learning goal #1: *describe a broad overview of the major practice areas in data science*. NLP is a major area of research within data science. It is an important skill to be able to mine large data sets for feedback from language reviews, and to find ways to analyze that data for meaningful conclusions and trends. This project covers those areas, albeit that they represent only a slice of the NLP world. The report associated with this project describes an overview of the practice area and inspires the problem at hand.

Computer Vision

Chest X Rays — Transfer Learning in Medicine

Overview

Learning goals demonstrated below:

- Describe a broad overview of the major practice areas in data science (#1)
- Collect and organize data (#2)
- Identify patterns in data via visualization, statistical analysis, and data mining (#3)
- Develop Alternative strategies based on the data (#4)

Code:

https://github.com/RonenReouveni/chest_x_ray

Technologies used

- Python
- Tensorflow
- Keras
- SKlearn
- Cv2
- Lime

Description

This Kaggle dataset is 2.14 GB of chest x rays. These images are labeled as either covid19 pneumonia, classic pneumonia, or normal. My goal with this research was to compare different transfer learning techniques and leveraging of pre-trained models. I used these pre-trained models as a feature transformer that was then fed into my multi-layer convolutional neural network. This was done primarily with Tensorflow and Keras. I used ResNet, VGG16, and MobilNet as my base layers. The structure of the trainable portion of my architecture on top of the base layers included multiple layers of pooling, dropout, batch normalization, and convolution. In my implementation of this in Tensorflow, I included early stopping and data augmentation. By using both pre-trained models as base layers and data augmentation, I overcame the issue of having a small dataset. Data augmentation expanded the size of the dataset by rotating and zooming into each picture. Furthermore, the classification rates of these models far outperformed my baseline neural networks without the pre-trained base layers. However, with more investigation, it became apparent that some of what was being learned were non-medical features. For example, none of the X - rays of covid19 pneumonia were of children. Unfortunately, many of the observations in the other classes are of small children. The models were classifying an X-ray based on the size of the human in the image, not the presence of a medically related feature.

Reflection

Achieving strong classification rates on this dataset meant I needed to develop alternative strategies based on the data. For computer vision, this dataset is relatively small. This meant I had to rely heavily on transfer learning and data augmentation. Both of these were implemented in Tensorflow. This project shows my growth and development in the field of Data Science. Computer vision is a massive field, and applying it to the medical space is very interesting to me. While working on this task, I leveraged Tensorflow's ability to create an iterator to load data in batches. The images are not loaded into memory until they are passed into the neural network. Similarly, data augmentation is only applied once the image is loaded into memory and then thrown out after the network trains on that batch. This demonstrates an ability to organize data specific to the task and technological options. I was also able to use visualizations to compare misclassified images. This illuminated the fact that the networks were not always learning to notice the medical feature. They were often learning the fact that one class had smaller lungs, i.e., children's X rays vs. adult X rays. This shows an ability to identify patterns in the data and the results. My goal is to work professionally in medical computer vision, and this project has given me a great introduction. I want to further develop in the field and work on a 3-dimensional medical imaging problem.

Learning Goals

Working with multidimensional numpy arrays is critical in using transfer learning and TensorFlow in general. This project demonstrates achievement in learning goal #2: *collect and organize data*. Without being able to collect and organize data, this project would not have been possible. Using TensorFlow's ability only to load the data when needed also demonstrates mastery of this learning goal. This is important in the field because it makes training faster and is less computationally expensive. Although analysis is only as good as data involved there are ways to remedy deficiencies. It is always important to keep in mind the goal of the analysis and if the data lends itself well to that goal.

Furthermore, this project also allowed me to display my mastery of learning goal #4: *develop alternative strategies based on the data*. The limited amount of data does not allow for enough training to achieve strong convergence. Therefore, I needed to develop alternate strategies based on the data. This included using data augmentation to artificially add more data and using transfer learning. Using transfer learning allowed me to leverage pertained models as linear transformers. Both of these strategies were essential for approaching this problem. This clearly shows achievement of learning goal #4: *develop alternative strategies based on the data.* In real world applications data is often imperfect. Part of meeting this learning goal was to learn strategies for extracting information when data is less than ideal and specifically suited for a certain type of modeling. This project provided me with that opportunity.

Furthermore, the report associated with this project underscores computer visions' place within data science and how deep learning treats the problem differently from traditional learning. Discussing the broad overview contributes to the achievement of learning goal #1: *describe a broad overview of the major practice areas in data science*. Computer vision is a massive field that is of the utmost importance to many large companies. Whether it be google with search engines, Facebook with user uploads, or Tesla with self driving cars, understanding the basics of computer vision is essential for any data scientist.

Fashion MNIST - Crushing Accuracy with CNN's

Overview

Learning goals demonstrated below:

- Identify patterns in data via visualization, statistical analysis, and data mining (#3)
- Develop Alternative strategies based on the data (#4)

Code:

https://github.com/RonenReouveni/fashionMnist

Technologies used

- Python
- Tensorflow
- Keras
- SKlearn

Description

Famous datasets such as Fashion MNIST serve as a baseline to compare classification rates. Using various convolutional neural networks, I achieved a classification rate of over 94% on the validation set. In this project, I implemented many different types of neural networks to classify the images and a random forest to serve as a baseline. The advantage of convolutional neural networks over traditional machine learning and feed-forward artificial neural networks is the ability to maintain the structure of the image. In traditional learning models, the pixels in the images are flattened into a single one-dimensional vector; this is different from convolusion, where the image structure remains multi-dimensional. The most successful CNN I implemented had roughly 7.2 million trainable parameters. It used multiple layers of convolution, dense layers, batch normalization, pooling, and dropout. I also used learning rate decay and early stopping to optimize results.

Reflection

This project is a fantastic example of how to change strategies based on the data. The more convolutional layers added, the more susceptible the model is to overfitting. On this problem, it is very easy to get close to 100% on the training set but then below 93% on the validation set. This gap between the two accuracy scores indicated that the model would benefit from regularization and dropout. I applied L2 regularization to every dense and convolutional layer. This, in conjunction with adding dropout between the layers, finally pushed my accuracy score to above 94%. This shows an ability to reason about my results and use them to improve. It also offers an ability to change strategies based on patterns found in the data. Finally, this project is an excellent example of implementing many neural networks with different tunings using Tensorflow and Keras.

Learning Goals

The biggest takeaway from this project is combatting overfitting and focusing on generalization while using advanced neural networks. By analyzing confusion matrices and visualizations of resulting accuracies, I recognized the overfitting and combated it. This demonstrates achievement of learning goal #3: *identify patterns in data via visualization, statistical analysis, and data mining*. Without recognizing these patterns, I would not have been able to achieve the results I did. The learning goal once again demonstrates that the data scientist has to evaluate results on an ongoing basis and focus on generalized results.

This project also exemplifies learning goal #4: *develop alternative strategies based on the data*. The data itself drove the approach of using CNN's. Furthermore, using dropout and regularization improved results dramatically. This shows an ability to develop alternate strategies based on results and data. It also shows the importance of a flexible mindset in approaching the type of analysis being requested. This is demonstrated clearly in that I noted the overfitting and amended my approach to ameliorate the issue.

Time Series

Zip Codes - Predicting the future with RNN's

Overview

Learning goals demonstrated below:

- Describe a broad overview of the major practice areas in data science (#1)
- Develop Alternative strategies based on the data (#4)
- Develop a plan of action to implement the business decisions derived from the analyses (#5)

Code:

https://github.com/RonenReouveni/ZipCodes

Technologies used

- Python
- Tensorflow
- Keras
- SKlearn
- FB prophet

Description

In this time-series dataset from Zillow, there are 20 years of monthly median home values for each zip code in the USA. The goal is to recommend to investors which zip codes to invest in over the next year. To handle a dataset of this size a few heuristics to filter observations are needed. It is not feasible to train and test an LSTM for every single zip code. By filtering the dataset and reordering it by historic median returns, I could lower the number of zip codes to make projections. The two primary modeling techniques are Facebook's prophet, and a Bidirectional LSTM implemented using Tensorflow. Both of these models were fully implemented with a functional approach. This meant that I could seamlessly filter the 14k zip codes by historical returns and fit a Bidirectional LSTM or prophet on each zip code. In the said functions is code that splits the data into train and test, where the test set is the final year of the dataset. Additionally, each model is trained on all relevant data, and then makes an out-of-sample prediction into the future. Finally, all of this information is printed to show the predicted returns and error rates for each zip code using both models. The deep RNN's far outperformed Facebook's prophet. Another significant achievement in this investigation was implementing functions to visualize any zip code, city, county, or state. Multiple locations can be passed into these functions to visualize them together. For example, I showed the historical returns for California, San Diego, and two more counties within San Diego, all on the same graph. These visualizations allow for flexible insight into the data.

Reflection

One of my passions in data science is modeling and deep learning. I believe neural networks are some of the most beautiful models ever discovered. Treating a time series problem as a sequence and using a Bidirectional LSTM to make predictions is often far superior to traditional time series models. In the analysis above, it even far outperformed Facebook's prophet. This shows an ability to develop alternate strategies based on the data. Furthermore, it demonstrates the ability to implement different kinds of neural networks in Tensorflow and Keras. By taking an object-oriented approach, my implementations are all functional . They can be fit on any subset of data in the dataset. This means that I can fit the network seamlessly on specific zip codes, counties, or cities. The entire point of this analysis was to inform a financial recommendation on which areas to invest in. I was able to mine for patterns using many techniques. These included visualizations, traditional time series techniques, and, most importantly, deep learning. I was able to uncover zip codes in the US with projected growth of up to 36%. There were also investment opportunities with 16% returns and far less variance or risk. An essential aspect of this project is how it directly suggests a business choice.

Learning Goals

In this research, I had the opportunity to work with time-series data. Although Facebook's prophet works, I needed to develop alternate strategies to improve results. I treated the time series as a sequence and fit Bidirectional LSTM's to model the problem. This was a more sophisticated approach than just using prophet, and paid off with better results. This shows a grasp of learning goal #4: *develop alternative strategies based on the data*. Furthermore, I needed to subset the data based on historical returns to reduce the dataset's size. Otherwise, the computational cost would have been too expensive, which is a factor in the business world. This also shows achievement of learning goal #4: *develop alternative strategies based on the data.* An essential aspect of this research is the direct relation to a business objective.

The final results of the study are three unique zip codes to invest in. This shows achievement of learning goal #5: *develop a plan of action to implement the business decisions derived from the analyses*. The analyses provided projected returns of the top historical performing zip codes. Based on this, a plan of action was implemented to make an investment in the top 3 zip codes. This included analysis on not only projected returns but assessments of risk and volatility. I use this analysis in considering personal real estate investments. This research clearly shows an ability and aptitude in learning goal #5; *develop a plan of action to implement the business decisions derived from the analyses*.

The written portion of the report associated with this research impresses regression and time series modeling. This project, along with others discussed in this report that round out classification, computer vision, and NLP, contributed to the achievement of learning goal #1: *describe a broad overview of the major practice areas in data science*. I emphasized time series and financial forecasting in the research and described a broad overview of each field. This kind of analysis is a key factor in the decision making of many companies today, with far reaching implications for the decisions companies make. This project, which required an understanding of financial goals and the real estate market demonstrated my understanding of the challenges in the field and my achievement of learning goal #1.

Regression and Classification

Bias in Machine Learning - How code effects people

Overview

Learning goals demonstrated below:

- Describe a broad overview of the major practice areas in data science (#1)
- Synthesize the ethical dimensions of data science practice (e.g., privacy) (#7)

Code:

https://github.com/RonenReouveni/BiasInML

Technologies used

- R
- randomForest
- Caret
- e1071

Description

Data modeling and deployment have a genuine impact on people's lives. As Data Scientists, it is imperative to understand the impact that we have on society. In this project, I used banking data to illustrate bias in machine learning. The dataset is a classic financial problem in which one is tasked to predict defaulting or not on a loan. Once deployed, a model such as this would be used to decide an applicant's loan eligibility. The dataset included race, gender, family size, relationship status, job type, education, and more. A Random Forest was then used to predict if an observation would default or not on their loans. The point of interest is to calculate the probability that an observation will wrongfully be denied a loan given their race. It was very clear from the analysis that the probability of being wrongfully predicted to default was very inconsistent between races. For example, 'American Indians' were twice as likely to be wrongfully denied than 'American Asians.' Being wrongly denied loans is a significant barrier to generating and maintaining financial freedom. Across cultural and racial lines it is a major reason for certain groups to have significantly poorer economic success.

Reflection

This project highlights the connection between modeling and impacts on the world. The issue of bias in machine learning is a large area of concern. It is important to remember that while predictions on a computer seem like numbers, they often correspond to real people. Understanding this area in data science makes me a stronger candidate for these types of roles in the field. There are real repercussions to ignoring bias. Major banking institutions have paid hundreds of millions in fines for wrongful denial of loans. Furthermore, this is only one example of bias in machine learning. This issue is present in many fields of Data Science. Finally, conducting this analysis in my first quarter in the program set the stage for me to be aware of this throughout my graduate degree.

Learning Goals

This project was unique in that it focused on the ethics of data science and its impact on people and their lives. The research connected to learning goal #7: *synthesize the ethical dimensions of data science practice*. My research on bias in machine learning and how it relates to the financial industry highlighted the importance of learning goal #7. I understand that data science can provide results that fundamentally affect every aspect of people’s lives, and that financial decisions are only one small area of that. Ensuring that data is used in an ethical way and that the conclusions drawn from it omit bias as far as possible and provide a picture that is honest and fair are part of the responsibility of every person who analyzes and extrapolates that data. Understanding this learning goal is fundamental to a career in data science, especially when its analysis impacts social outcomes.

The report and presentation associated with this research demonstrate an understanding of this major area of practice within data science both from the financial analysis perspective and the social impact perspective. Along with the other projects mentioned in this report, this specific research project shows aptitude and completion of learning goal #1; *describe a broad overview of the major practice areas in data science*. It does this by rounding out the other major areas of practice discussed throughout my portfolio, time series, computer vision, NLP, and traditional regression and classification.