Applied Data Science

Project Portfolio DRAFT

Syracuse University

Toby Anderson

7/29/2023

# Data Science Overview

Data Science is a multidisciplinary field which unites methods from statistics, algorithms, data structures, and application domains. While Data Science is a relatively new emerging field, it draws upon long standing professions in Statistics, Data Analysis, and Computer Science. The goal of Data Science is to turn data into actionable insights, typically in a business setting. By collecting and structuring data, using exploratory data analysis, modeling, and visually communicating results, a Data Scientist can turn raw data into information, then into knowledge which provides business value.

In the Applied Data Science Masters program at Syracuse university, students are taught several topics, and create several projects which build understanding in these key areas. In this portfolio, I demonstrate how several of these projects have accomplished a well-rounded education in Applied Data Science.

# Collecting Data

A well-rounded data scientist should be able to collect both structured and unstructured data from various sources. Unstructured data would include images, video, audio, and raw text, while structured data would include databases, APIs, or tabular data.

In the course IST-659, we learned about database design and administration. In this course, I collected data from a sci-fi film ranking website, and transformed it into structured data in a local MS SQL database. I then designed several queries to gain insights from this data.

The project can be found here: [Sci-fi Film Project](https://github.com/tobot-hub/syracuse_portfolio/tree/master/1.%20Collecting%20Data/IST_659%20DB%20Admin)

Unstructured Data:

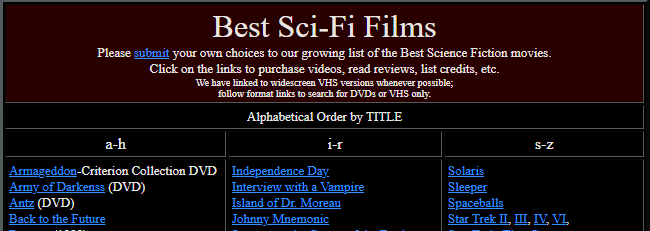


Figure 1 – Sample of unstructured data from the web.

Database Logical Model:

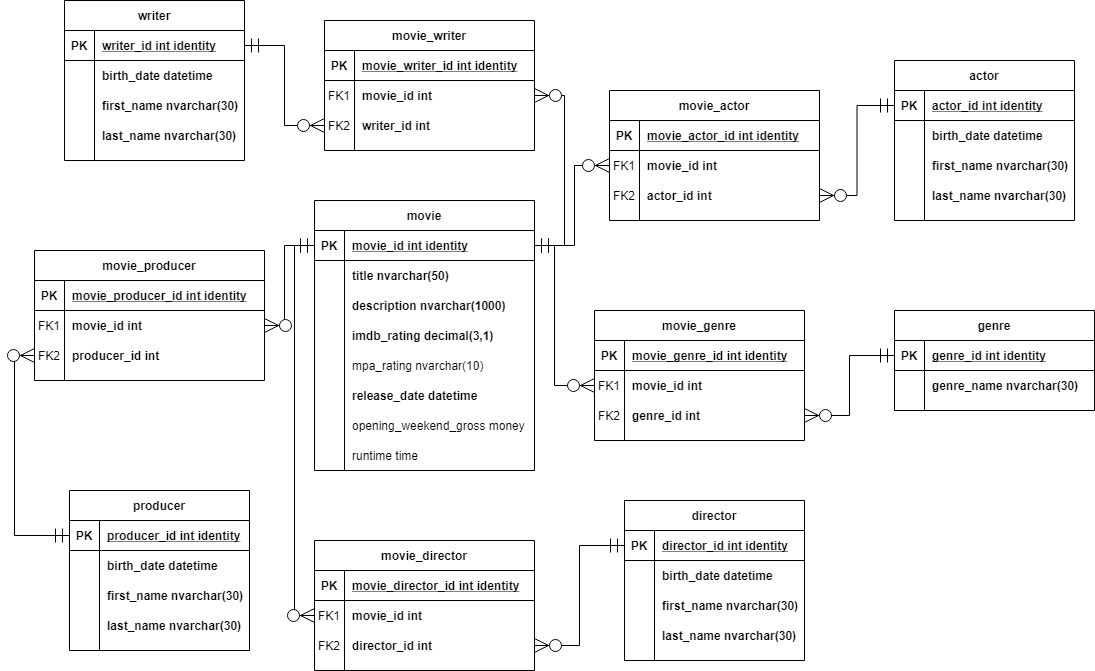


Figure 2 – Scifi Movie Database Logical Model diagram

The raw data was in HTML form, and consisted of a nested list of movies which included the list of actors, directors, genres, writers, and producers. I parsed the data and structured it into multiple data objects. After performing Second Order Normalization on the data, using linking tables to preserve unique primary keys, I created the database according to the Logical Model above.

Structuring the data allowed me to answer several questions, such as:

|  |  |
| --- | --- |
| Which actors starred in the most sci-fi movies? |  |
| What genres overlap the most with the sci-fi genre? |  |

This clearly demonstrates the value of collecting and organizing unstructured data to provide deeper insights.

# Actionable Insights

In order to provide business value, a well-rounded Data Scientist should be able to use analysis to connect data insights into recommendations and actions. In the course MBC-638, we learned about data driven decisions making, and it’s applications in process improvement. The course took us through the Six Sigma process called DMAIC; Define, Measure, Analyze, Improve, and Control. By following the DMAIC process, I was able to create meaningful improvements to my Data Science resume which will improve my chances of landing interviews in the job market.

The project can be found here: [Data Science Resume Project](https://github.com/tobot-hub/syracuse_portfolio/tree/master/2.%20Actionable%20Insights/MBC638%20Decision%20Making)

By utilizing the API for a Large Language Model, I was able to generate focused feedback when comparing my original resume against a curated list of Data Science job postings.

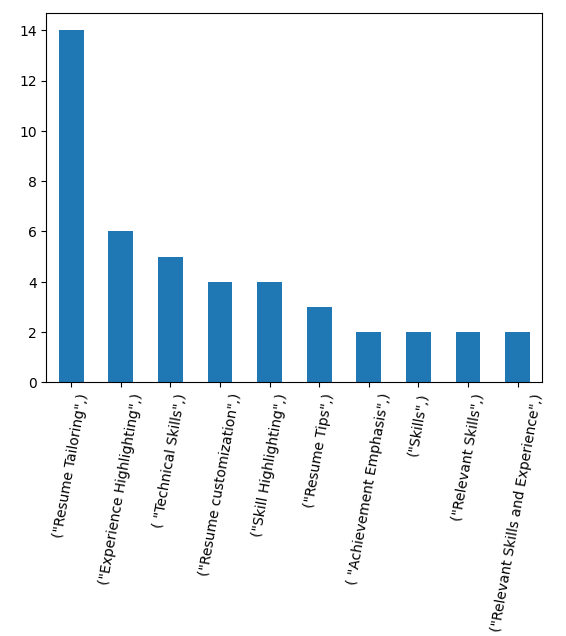


Figure 3 – Pareto chart of resume improvement advice

By categorizing the feedback, and organizing it into a pareto chart, I was able to make the most effective changes to my resume to improve fit with the job postings. While some of the advice, such as “Resume Tailoring” and “Resume Customization” could not be generalized into overall improvements, I was able to pick up on 1 major generalized improvement: Expanding Technical Skills.

Figure 4 – Sample of resume skills expansion

Before:

* Python, C++, VBA

After:

* Python: NLTK, Scipy, Numpy, Pandas, Openai, Matplotlib, Scikit-learn, Gensim, Seaborn, Pyodbc

As shown in figure 4, I went through the Skills section of my resume and expanded out a much higher level of detail, including modules, packages, and toolkits that I’ve used in various development settings. After updating my resume, I was able to use the same Large Language Model to confirm improvements.

|  |  |  |
| --- | --- | --- |
| **Resume** | **Yield %** | **SQL** |
| Original Resume | 16.8% | 0.55 |
| Resume with expanded technical details | 47.4% | 1.45 |
| Resume with technical details and professional summary. | 66% | 1.9 |

Table 1 – Validated improvements from resume updates

The updated resume was simulated to be selected for an interview 66% of the time, compared to the original resume, which was only simulated for interview 17%. This is a huge improvement, and demonstrates my ability to use data analysis to create actionable insights and increase business value.

# Modeling

A well-rounded Data Scientist can not only draw insights into existing data, but use modeling to extrapolate results to new situations, and potentially provide automation and real-time decision assistance. By using predictive modeling, a Data Scientist can leverage insights into much larger business value.

In the course IST-718, we used python to wrangle large datasets and train multiple predictive models. I was able to use feature engineering and modeling to predict fraudulent transactions in banking data. By leveraging the Featuretools Python package, I was able to create relationships in the data which allowed for sophisticated feature synthesis, which created powerful modeling opportunities.

The project can be found here: [Fraudulent Transaction Project](https://github.com/tobot-hub/syracuse_portfolio/tree/master/3.%20Modeling/IST718%20Big%20Data)

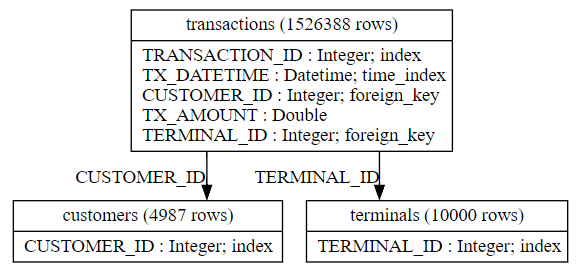


Figure 5 – Featuretools data model

The bank transaction data from Kaggle included a single table of transactions, which listed the customer and terminal IDs, as well as specific information about the transaction. Similar to database modeling, I extracted customers and terminals as separate objects from the transaction, then related them together within the Featuretools framework, see Figure 5.

With the data objects related, I was able to customize deep feature synthesis, to create time-based aggregated features across transactions, customers, and terminals. While featuretools created over 100 unique features for me, the most significant features are shown below.

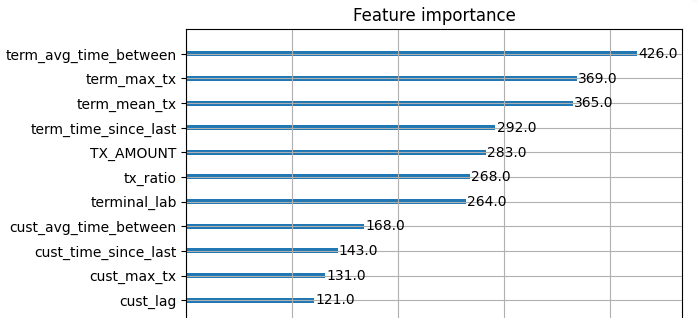


Figure 6 – Ranked features by XGBoost feature importance

After training an XGBoost model on the data, I was able to interrogate the model and visualize the most important features in Figure 6. By viewing the important features, I was able to make better progress in continued feature synthesis to improve the modeling accuracy.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Parameters** | **F1 Score** | **AUC** |
| **Decision Tree** | ccp\_alpha = 0.01  class\_weight={0:1,1:10} | 0.939 | 0.949 |
| **Naïve Bayes** | Gaussian Kernel | 0.416 | 0.933 |
| **XGBoost** | subsample=0.9  early\_stopping = 20 | 0.945 | 0.999 |

Table 2 – Model results for fraudulent transaction detection

Developing strong features allowed me to create a model with a 99.9% AUC, which is a very powerful model for detecting fraudulent transactions. Additionally, after analyzing the most important predictive features, we can see that transaction amount and terminal transaction frequency are the most influential factors. This could mean that attackers target unused ATMs to extract the maximum money under the transaction limit. This analysis demonstrates the ability to create predictive models and to visualize the results to provide business insights.

# Leveraging Programming Languages

A well-rounded Data Scientist should be familiar with the most prevalent tools used in the industry, including programming languages such as Python and R, and the query language SQL. In the course IST-652, we learned how to use Python to scrape and transform data from various sources, including APIs and JSON, and mine insights out of the data.

The project can be found here: [Data Science Salary Project](https://github.com/tobot-hub/syracuse_portfolio/tree/master/4.%20Programming/IST652%20Scripting)

I was able to use Python to mine insights out of Data Science salary data.

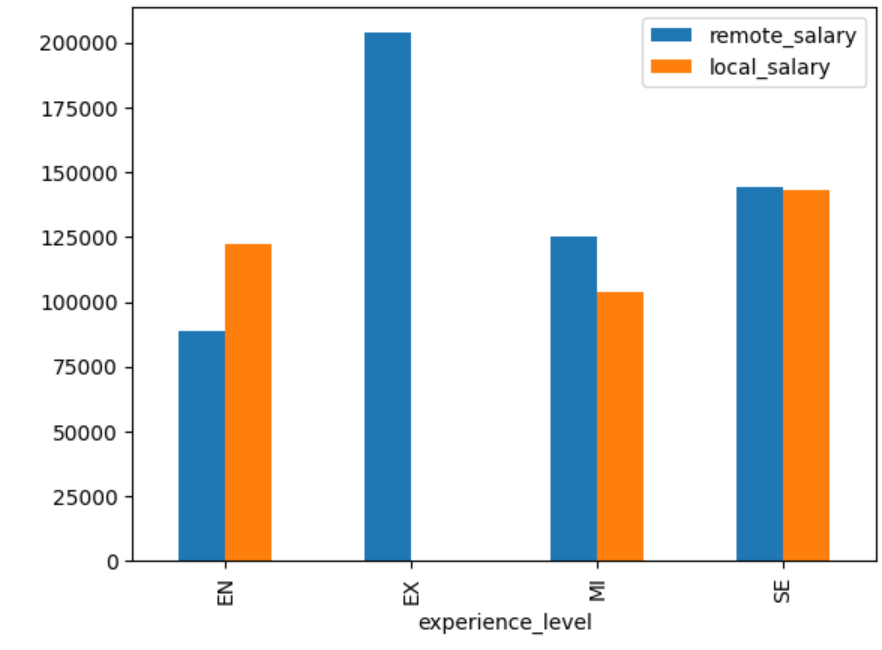
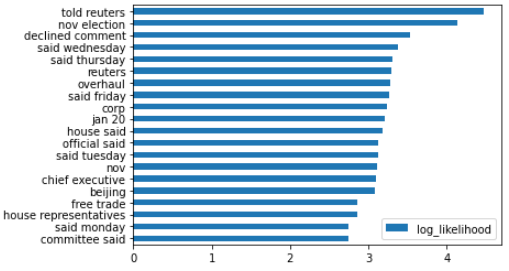


Figure 7 – Data Science salary by experience level and remote vs local

The main finding was that Entry Level (EN) employees should look for medium sized locally companies, while more experienced employees should look for larger companies offering remote jobs in order to maximize their salary growth potential.

Additionally, in the course IST-736, we used scripting to perform Text Mining to gain insight out of unstructured text. I was able to use curated Fake News data to determine which terms were more likely to indicate a True or False article.

The project can be found here: [Fake News Article Project](https://github.com/tobot-hub/syracuse_portfolio/tree/master/4.%20Programming/IST736%20Text%20Mining)

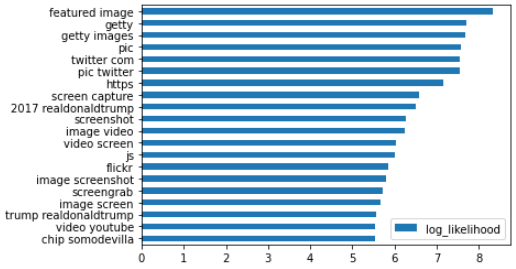
Figure 8 – Phrases indicating a genuine news article

Figure 9 – Phrases indicating a fake news article

By using Natural Language Processing (NLP) techniques within Python, I was able to extract key phrases from news articles. I was then able to “Vectorize” the data into a more machine-readable format, which was then used to train a Naïve Bayes model to predict genuine/fake news. While the text processing required a good deal of sophisticated programming, the results are very satisfying. This demonstrates my ability to leverage Python to perform advanced data analysis.

# Visual Communication

A well-rounded Data Scientist needs to not only gather data, mine insights, and create predictive models. They also need to communicate results to both technical and non-technical audiences. In IST-719, Information Visualization, we learn how to design visualizations around human psychology. By creating visual hierarchies, removing chart junk, and showing contrast in interesting data, we can communicate much more effectively. In this class, I was able to take several data sources and create a wide range of visualizations.

The project can be found here: [Adobe Illustrator Homework](https://github.com/tobot-hub/syracuse_portfolio/tree/master/5.%20Visual%20Communication/IST719%20-%20Visual%20Communication)

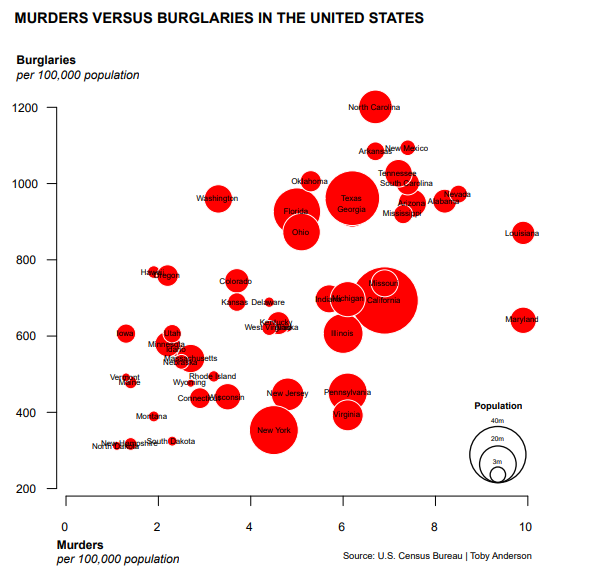


Figure 10 – Murders and Burglaries in the United States

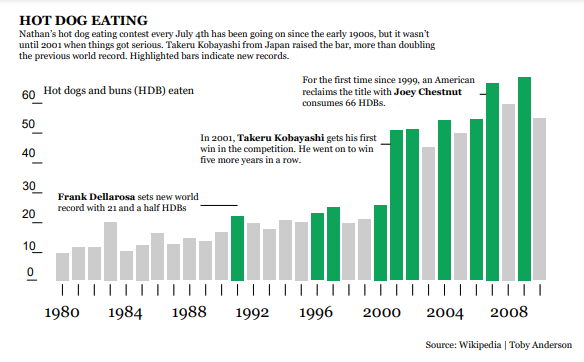


Figure 11 – Hot Dog Eating Records

Figures 10 and 11 are charts which were created using data analysis in RStudio, then imported into Adobe Illustrator for customization. In Illustrator, I was able to remove chart junk, like unneeded boxes and reference lines, which made the visuals much easier to read. Then I was able to customize the text, use color to highlight key elements, and add custom wording to highlight messaging behind the chart. Overall, transforming the charts in Adobe Illustrator is a powerful way to make a chart into a communication tool that can be easily absorbed by a wide audience.

# Ethics in Data Science

Data Science is a powerful toolkit capable of prediction on a mass scale, which can provide business value. However, with this massive predictive power also comes the risk of social harm. Data Science models are capable of codifying bias in training data, crossing personal privacy boundaries, and exploiting consumers. While the Syracuse program didn’t specifically emphasize the ethical aspect of Data Science, we touched on these concerns throughout many of the courses.

In the course IST-664, Natural Language Processing, I trained a model to detect spam emails. Additionally, in IST 736, I created a model to detect fake news articles. Both of these projects demonstrate an interesting dynamic. While Data Science has been used to create more convincing spam emails and take news articles, it can also be used to combat bad actors to protect consumers. In this way, Data Science is a tool that can both attack and protect, and it’s important to consider the harms your project might have on your customers and wider society.

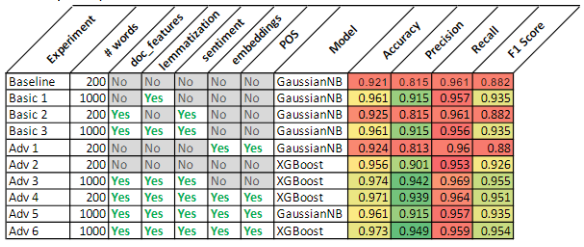


Figure 12 – Spam detection model results

As shown in the model results, Fig 12, advanced Data Science techniques, such as those found in Natural Language Processing, can be used to protect users from suspicious content. With each advance in Data Science capabilities, we must continue to look out for misuse, and to proactively consider how we can defend against bad actors.

# Conclusion

Throughout my Applied Data Science program at Syracuse, I’ve received a well-rounded Data Science education. I was able to demonstrate mastery of Data Collection, Creating Actionable Insights, Predictive Modeling, Programming, Visual Communication, and Ethics. These are core skills for a Data Scientist, and I hope that the skills I’ve demonstrated will make me more marketable when searching for lucrative Data Science careers. Regardless of my future career path, this portfolio is a milestone that testifies to my achievements and the opportunities for learning I’ve had throughout this Syracuse program.