Portfolio Milestone

Syracuse University: M.S. Applied Data Science

Matthew Laken

SUID: 226482612

January 2022

https://github.com/mlaken/Matthew-Laken-Applied-Data-Science

Table of Contents

Introduction

The Applied Data Science Master's program at Syracuse is designed to help students further their education by introducing them to the fundamental principles of Data Science and Analytics. Throughout the various courses in this program we students are encouraged to utilize programming languages such as Python, R, and SQL via their IDE as well as software such as R Studio, SQL Management Studio, and the Microsoft Suite (Access, PowerBI, Excel). My focus in the program was to develop a more comprehensive understanding of how I utilize the tools at my exposure to collect, analyze, interpret, and communicate information within any sector of business. This goal prompted me to enroll in classes such as IST 659 Database Administration, IST 652 Scripting for Data Analysis, IST 687 Introduction to Data Science, and IST 707 Data Analytics. Each of these courses required me to utilize information and data science to provide an in-depth analysis of real-world situations to best prepare me for what lies ahead in my career.

I have prepared a portfolio containing the projects I completed in my courses, which represent the School of Information Science's learning objectives established for students in the Master's in Applied Data Science Program.

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).

IST 659: Database Administration

Project Completed Fall 2020. For full details on the project please see the following: IST 659

Project: UFC Relational Database

At the culmination of the course on Database Administration, I had the opportunity to demonstrate my knowledge of both creating a database using SQL Management Studio and developing various scripts to execute procedures using SQL. The purpose of my database was to create a predictive model to be applied for sports with an individual competitor. For this assignment, I generated a relational database that contains the UFC (Ultimate Fighting Championship) roster as well as all fight cards data tracing back a decade. This was done by collecting the data from the UFC official website, generating tables for fighter information and statistics, and writing the scripts to put it all together. My intentions in starting this project were for one to be able to utilize this database to make informed predictions on the result of an upcoming match based on a fighter's history and the history of the UFC.

I established a set of rules which define the integrity of the database. Then I generated a conceptual model to show high-level relationships between tables. This was later refined when creating the logical model shown in *Figure 1*. Once the tables had been created, I generated functions to serve as a search engine within my dataset and help one better navigate the data. Also, I integrated Microsoft Access to update my tables/ queries efficiently while using R studio and the SQL query() function in R to analyze the questions I set out to answer in my assignment. For example, I analyzed the duration of match time across weight classes, shown in *Figure 2*.

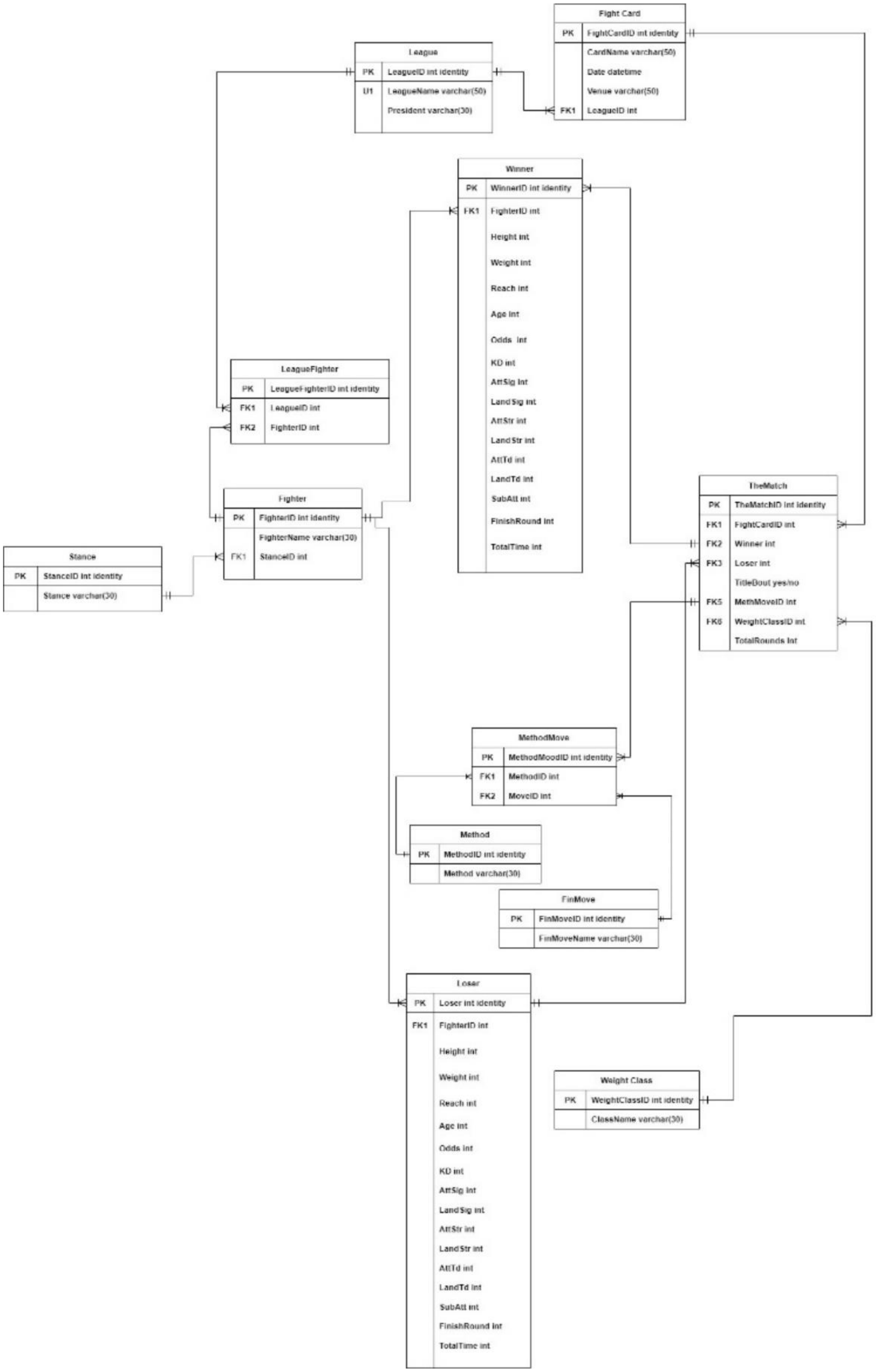

Fight Card
PK FightCardID int identity
CardName varchar(50)
Date datetime
Venue varchar(50)
FK1 LeagueID int
League
PK LeagueID int identity
U1 LeagueName varchar(50)
President varchar(30)
Winner
PK WinnerID int identity
FK1 FighterID int
Height int
Weight int
Reach int
Age int
Odds int
KD int
AttSig int
Land Sig int
AttStr int
Land Str int
AttTd int
LandTd int
SubAtt int
FinishRound int
TotalTime int
LeagueFighter
PK LeagueFighterID int identity
FK1 LeagueID int
FK2 FighterID int
Fighter
PK FighterID int identity
FighterName varchar(30)
FK1 StanceID int
Stance
PK StanceID int identity
Stance varchar(30)
TheMatch
PK TheMatchID int identity
FK1 FightCardID int
FK2 Winner int
FK3 Loser int
TitleBout yes/no
FK5 MethMoveID int
FK6 WeightClassID int
TotalRounds int
MethodMove
PK MethodMoodID int identity
FK1 MethodID int
FK2 MoveID int
Method
PK MethodID int identity
Method varchar(30)
FinMove
PK FinMoveID int identity
FinMoveName varchar(30)
Loser
PK Loser int identity
FK1 FighterID int
Height int
Weight int
Reach int
Age int
Odds int
KD int
AttSig int
Land Sig int
AttStr int
Land Str int
AttTd int
LandTd int
SubAtt int
FinishRound int
TotalTime int
Weight Class
PK WeightClassID int identity
ClassName varchar(30)

Figure 1

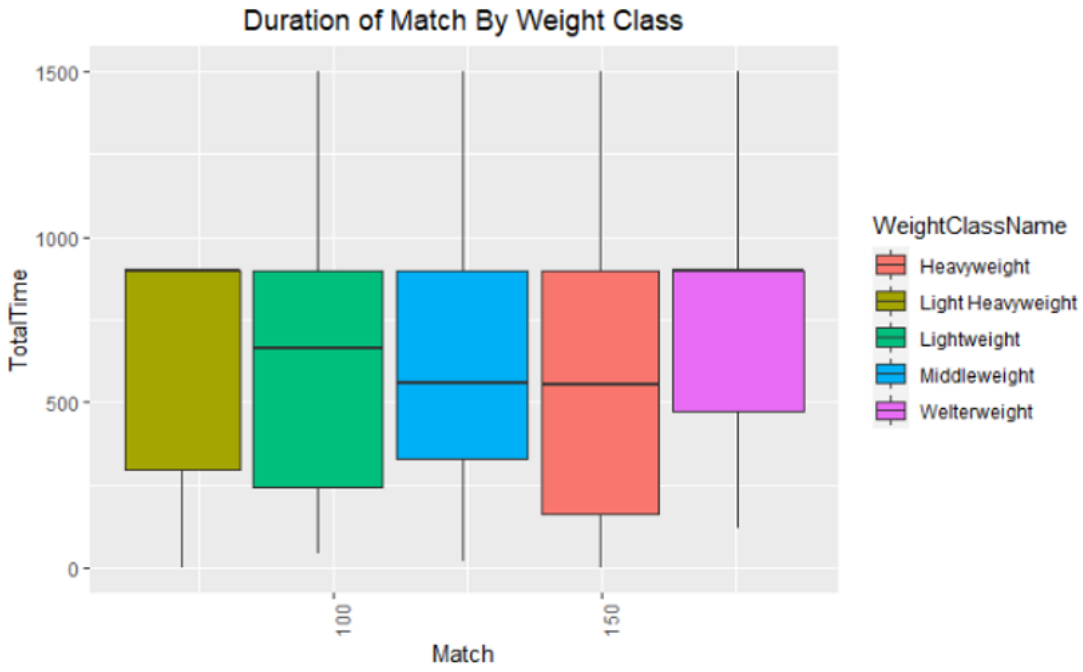


Figure 2

Learning Objectives

IST 659 Database Administration was the first course I took in the system of Applied Data Science and from the start of the program, we began discussing the fundamentals of data science and how they are applied to database management. Specifically, this course introduced the concepts of database design and the hierarchal model regarding the relational database. We had the opportunity to learn about these topics while utilizing Microsoft Access as a means for collecting data which we retrieved from various sources and then organized into tables. Additionally, we constructed complex problems that could be solved by using Structured Query Language (SQL) on SQL Server Management Studio. We reviewed the basics of query writing, the database development lifecycle, and how to properly store, save, and back up one's data. Additionally, we discussed the means of utilizing multiple programs to access our database with the ODBC driver that allowed us to generate our database using SQL and create visualizations using the packages in R language learned in other courses.

The IST 659 course provided exposure to software and skills that will be necessary for maintaining company databases and solving problems. The emphasis on data mining and statistics provided us with the tools to be successful in the field. A major component of the course was understanding the foundation for the various assignments we completed, which were essential in communicating the findings and results of my work.

IST 687: Introduction to Data Science

Project Completed Fall 2020. For full details on the project please see the following: IST 687

Project: Analysis of Cost and Salary Potential of Attending College

The Introduction to Data Science course final project challenged students to generate a thorough business presentation on a collection of data of their choosing while working in a team of peers to communicate findings. For my group project, we analyzed a Kaggle dataset on college tuition cost across the United States and the average starting salaries upon earning a degree to investigate the Cost and Salary Potential of Attending College. We utilized a basic linear model as well as a Kernel Support Vector Machine to determine which is superior in finding an expected value for mid-career salary when comparing the errors for the two models. Both models consider each school's total cost and early-career salary values.

This project was done in R Studio using R packages such as dplyr to aggregate the data by region, for example, and ggplot to visualize it. We found the mean values of the cost of tuition, financial aid, and salary after grouping the university data by region. This allowed us to see which part of the country provides students with the best value for their degree as shown in Figure 3. Also, we were able to narrow down the specific schools that have the greatest ratio of salary to cost of tuition, shown below in Figure 4. Overall, this project showed me how to conduct a more technical analysis using various calculations and models while being able to clearly illustrate my results and communicate findings.

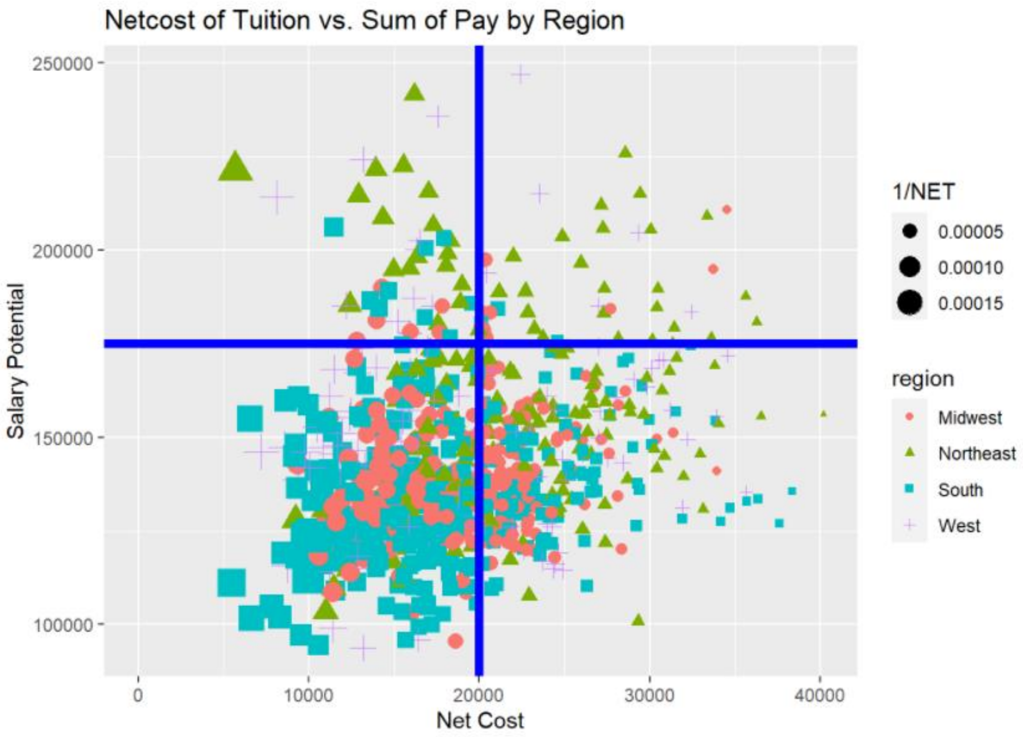
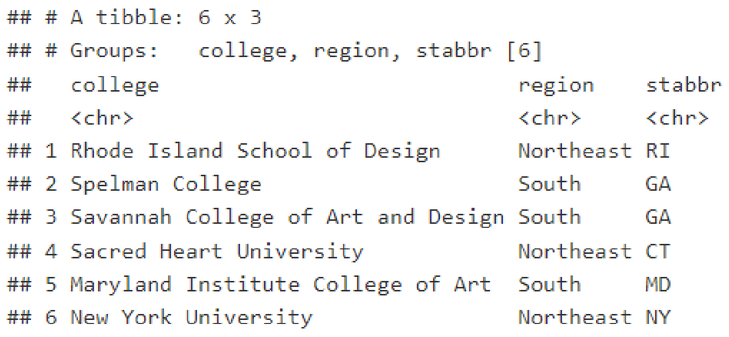


Figure 3



```
## # A tibble: 6 x 3
## # Groups:   college, region, stabbr [6]
##   college                            region    stabbr
##   <chr>                              <chr>     <chr>
## 1 Rhode Island School of Design      Northeast RI
## 2 Spelman College                    South     GA
## 3 Savannah College of Art and Design South     GA
## 4 Sacred Heart University            Northeast CT
## 5 Maryland Institute College of Art  South     MD
## 6 New York University                Northeast NY
```

Figure 4

Comparing the plot of errors from both models in figure 5, we can see that the KSVM model tracks closer to the true data than the linear model. Both models have about the same overall error with -0.08% for the linear model and -0.07% for the KSVM. The linear model has a wider range of error (-9%,10%) compared to the KSVM values (-7%,5%). The narrower range of error makes the KSVM more accurate, but the linear model runs much faster, seconds compared to about ten minutes. Depending on the relative importance of speed and accuracy to a user, either model seems sufficient to predict mid-career salary.

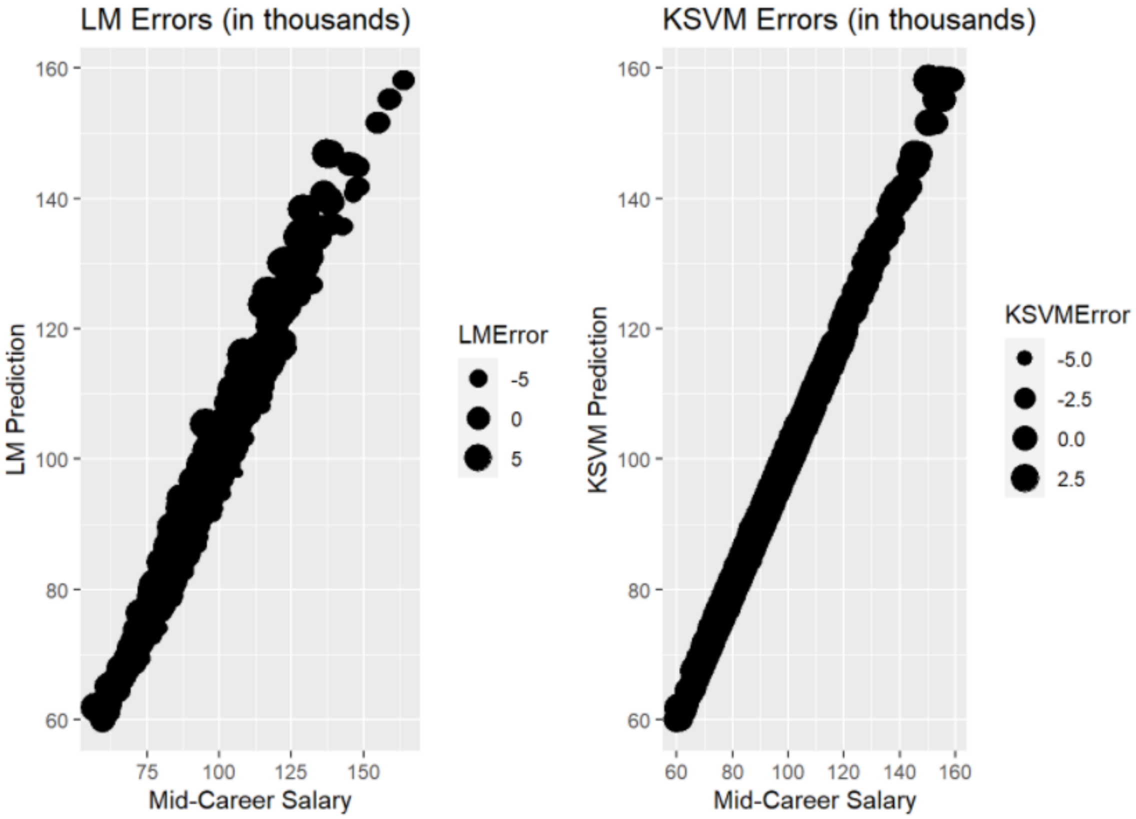


Figure 5

Learning Objectives

In completion of the Introduction to Data Science course, I was challenged to utilize the fundamentals of organizing and reporting data to collect, evaluate, and then compare different predictor models to gain a more in-depth understanding of the dataset as well as when to use specific models. The Introduction to Data Science course reviewed the essential concepts of data science while utilizing R and R-Studio to write code to solve the course assignments. From a statistical perspective, this course encourages students to identify problems as well as the data and methods needed to investigate the problem. Students were taught to clean and configure datasets through the processes of data transformation such as aggregation, then determine which methods of analysis were best to utilize in solving specific problems.

The assignments and projects in this course provided knowledge of how to take high-level problems that one would encounter in the workplace and develop a plan of action in terms of analyzing, visualizing, and applying statistical models like the linear or KSVM models depending on the conditions of the problem. It is important to know how to take a project through its lifecycle, thus being able to conduct a proper investigation with modeling and interpretation. This is something that I know will be utilized in my future work.

IST 707: Data Analytics

Project Completed Winter 2020. For full details on the project please see the following: IST 707

Project: Marvel vs DC Analysis

The Data Analytics course IST 707 shared various tools and methods of data mining while introducing its real-world application to solving problems in business. Essentially, we utilized a variety of tools such as R, Weka, Orange, and Alteryx which all are vehicles for conducting data mining algorithms. For my final project in this course, I explored the age-old debate on Marvel vs. DC Comics. Using data about various superheroes/villains of both Marvel and DC from Kaggle, I combined two datasets. One of the datasets, the Score set is comprised largely of numeric variables, and the other, the Power set, has a majority of categorical variables. I joined the data on the character name and then began my analysis to predict whether a character was from Marvel or DC as well as their alignment (good vs. evil). The algorithms I utilized were Random Forest, Naïve Bayes, kNN, and SVM. I evaluated the results using 3-fold cross-validation to determine which dataset variables were most effective in predicting both publisher and alignment. I configured and ran these models using a software introduced through this course called Orange which is a data mining software that uses python. Figure 5 shows the layout of my project in Orange.

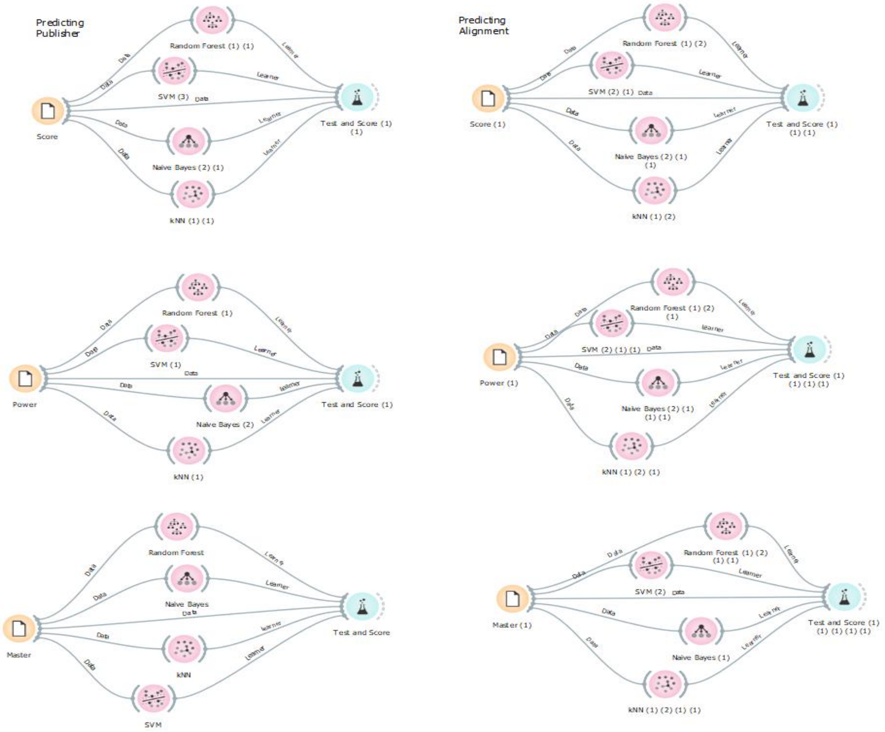
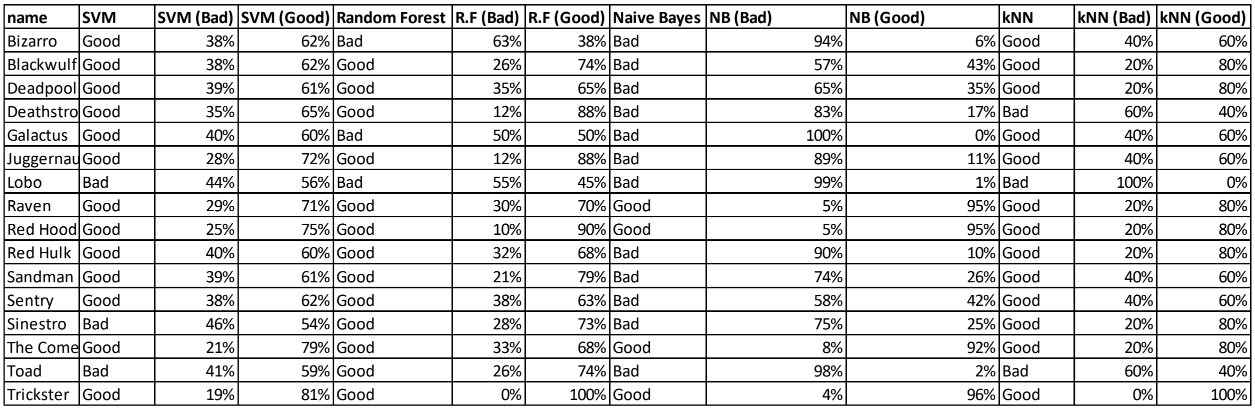


Figure 5

I began by training the models with the character’s data and then used a portion of my dataset with the removed target variables to generate predictions in my testing set. My findings showed me that using the Combined dataset was most effective in predicting both publisher and alignment, 69.1% and 68.2% respectively. I evaluated my testing group based on the percentages for each induvial test shown below in Figure 6.



name	SVM	SVM (Bad)	SVM (Good)	Random Forest	R.F (Bad)	R.F (Good)	Naive Bayes	NB (Bad)	NB (Good)	kNN	kNN (Bad)	kNN (Good)
Bizarro	Good	38%	62%	Bad	63%	38%	Bad	94%	6%	Good	40%	60%
Blackwulf	Good	38%	62%	Good	26%	74%	Bad	57%	43%	Good	20%	80%
Deadpool	Good	39%	61%	Good	35%	65%	Bad	65%	35%	Good	20%	80%
Deathstro	Good	35%	65%	Good	12%	88%	Bad	83%	17%	Bad	60%	40%
Galactus	Good	40%	60%	Bad	50%	50%	Bad	100%	0%	Good	40%	60%
Juggernau	Good	28%	72%	Good	12%	88%	Bad	89%	11%	Good	40%	60%
Lobo	Bad	44%	56%	Bad	55%	45%	Bad	99%	1%	Bad	100%	0%
Raven	Good	29%	71%	Good	30%	70%	Good	5%	95%	Good	20%	80%
Red Hood	Good	25%	75%	Good	10%	90%	Good	5%	95%	Good	20%	80%
Red Hulk	Good	40%	60%	Good	32%	68%	Bad	90%	10%	Good	20%	80%
Sandman	Good	39%	61%	Good	21%	79%	Bad	74%	26%	Good	40%	60%
Sentry	Good	38%	62%	Good	38%	63%	Bad	58%	42%	Good	40%	60%
Sinestro	Bad	46%	54%	Good	28%	73%	Bad	75%	25%	Good	20%	80%
The Come	Good	21%	79%	Good	33%	68%	Good	8%	92%	Good	20%	80%
Toad	Bad	41%	59%	Good	26%	74%	Bad	98%	2%	Bad	60%	40%
Trickster	Good	19%	81%	Good	0%	100%	Good	4%	96%	Good	0%	100%

Figure 6

Learning Objectives

The Data Analytics Course, IST 707, was essential in developing the skills needed to be a data storyteller in the real world as we learned how to develop solutions to our data mining problems using various techniques, models, and programs. This course focused on the various statistical modeling techniques that can be used in classification problems in addition to serving as the means for configuring and conducting such analysis. Essentially, we dove into models such as SVM, Random Forest, Naïve Bayes, and kNN to understand when it is appropriate to use each model and what information we can derive from the results. By developing alternative strategies, we are opening up the possibilities of how to analyze our data, thus generating a more thorough evaluation of the problem at hand.

One must be able to effectively combine datasets as exemplified in the preliminary steps of the final project to develop an adequate sample to conduct the analysis. More importantly, understanding both the factors which are needed to adopt specific strategies for a project and what is appropriate to utilize in each situation are essential to carrying out an agenda are crucial to professional supporting business decisions. A common theme within the program is being able to observe patterns within the data and communicate one’s findings, which is only achievable by comprehending the fundamentals of data analysis and presenting it effectively.

IST 652: Scripting for Data Analysis

Project Completed Spring 2021. For full details on the project please see the following: IST 652

Project: MLB Strikeout Investigation

The Scripting for Data Analysis course employed a hands-on approach to utilizing python as a means for collecting, analyzing, and evaluating data. This course allowed me to explore the many packages and applications of python while developing a fundamental understanding of high-level problem-solving. For my final project, I utilized the Pybaseball package to dive into the sport of baseball and how the Major League Baseball game has evolved. Specifically, I worked with a peer on analyzing the rise in strikeouts over the past decade and how it has stayed consistent with home runs and batting average by observing the fastball location within the strike zone.

To accomplish this project, we began by calling specific fields from different data sources available within the pybaseball package and combining them into a dataset. Then we set out to answer 5 questions revolving around pitching and batting using our dataset. To properly communicate our findings, we generated visuals that show the trend of team batting average, strikeouts by game, team home runs, and team total runs per season. Figure 7 features a point for every game and team respectively for a given season. These scatter plots feature the season averages shown by the red lines.

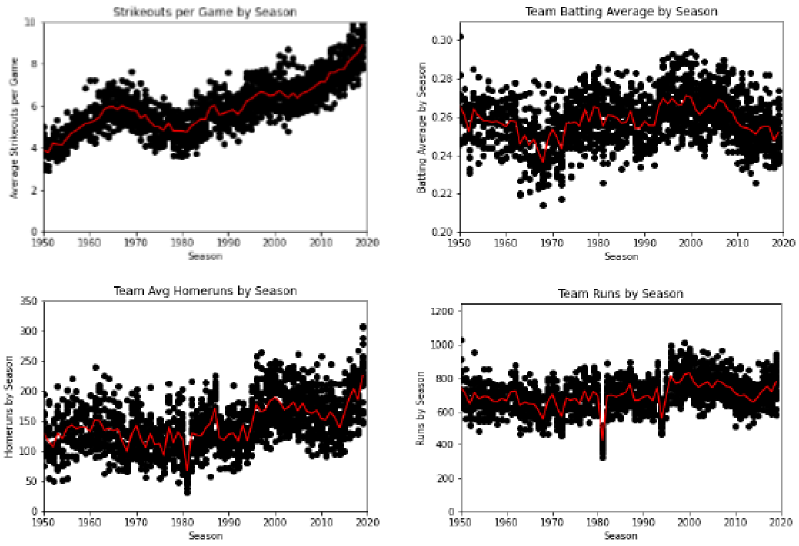


Figure 7

Additionally, we explored the evolution of the fastball regarding where it falls within the strike zone. We generated graphs to emulate the batter's box and plotted fastballs in 2010, 2015, and 2019 to show how the hot zone gradually moved up in the batter's box, shown in figure 8.

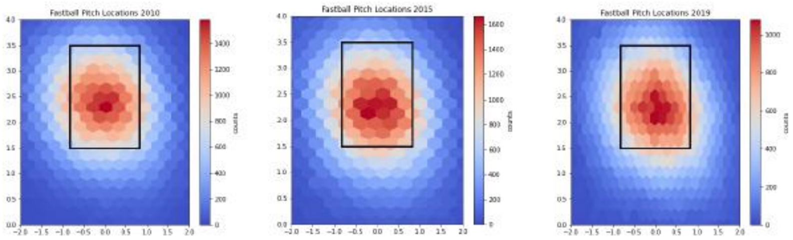


Figure 8

This can be explained by the homerun phenomenon where players are swinging for the fences and willing to risk the strikeout or pop out by swinging at the elevated fastball. We furthered our investigation by observing plate discipline as we understood that the overall batting average has decreased since the mid-2000s. Figure 9 shows how pitchers are forcing hitters to swing at more pitches outside of the strike zone thus increasing the number of strikeouts and lowering opposing hitters batting average. Specifically, the yellow line denotes that the hitters are focused on that elevated fastball rather than waiting for their pitch in the zone.

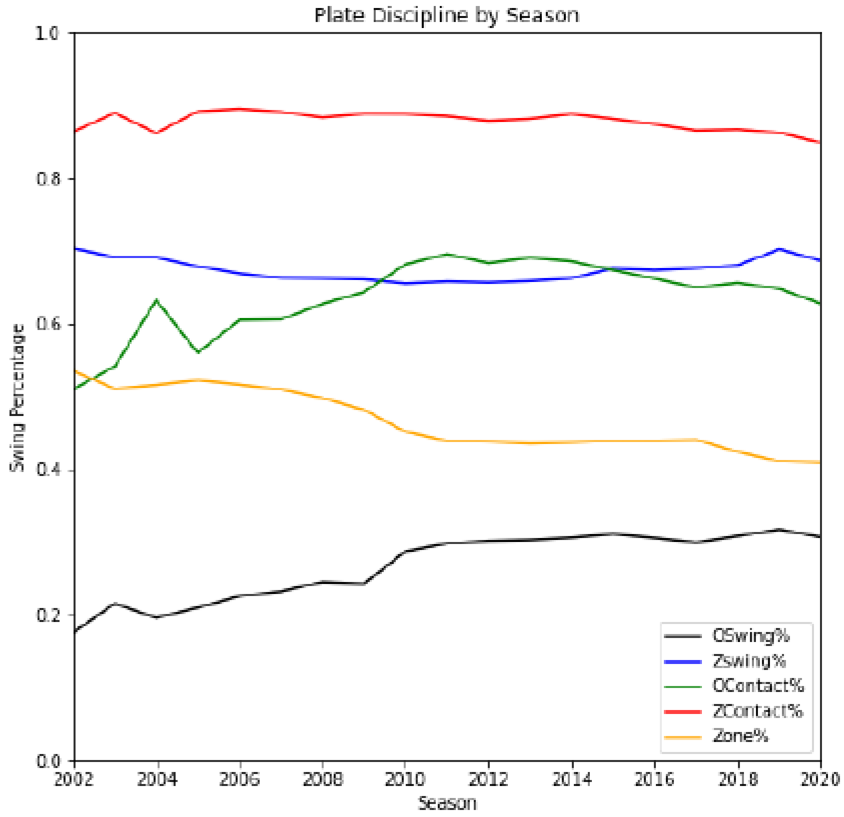


Figure 9

Learning Objectives

The Scripting for Data Analysis course was essential in developing an understanding of the methods to utilizing both structured and unstructured data to investigate complex problems. The assignments and projects allow for exploration into various python packages, as well as the development of the knowledge to make use of the proper ones based on the conditions of a given problem. It is important to understand the versatility of Python and that there are many applications that allow one to write and run scripts to process, clean and analyze data.

The accomplishment of the goals of the course provided students with the ability to identify patterns, analyze and solve problems. Additionally, the course challenged students to be able to comprehend and communicate findings, responding to real-world questions. Another aspect of this course was to securely access data stored in various formats and collaborate with colleagues using various software such as Google Colaboratory. It is essential to gain exposure to the tools and technologies used in a professional setting within the field.

Conclusion

The Master's in Applied Data Science program has taught the technical skills in SQL, R, Python, and many other software packages to enable students to be comfortable in conducting their investigations to solve real-world business problems and make informed decisions. There are endless possibilities of fields that one can impact by developing solutions with the foundational knowledge of analytics and managing different types of data.

The coursework has prepared me for both the analytical and interpersonal aspects of working in a data science role, since I have worked on projects where I acquired and cleaned data, established a database, utilized modeling and algorithms to run queries and visualize the data, and finally, communicated my findings in a professional, comprehensive manner. Also, throughout this program, I have taken on many team-oriented projects during my studies which have shaped the way I collaborate with peers and work within a team setting. Through the methodologies and teachings of data science, I plan to challenge myself to expand the applications of information and data science and believe that my impact in future ventures will exemplify the principles of the Applied Data Science Master's Program at Syracuse University.

References

Laken, M (2021). IST 659 – Work completed in SQL using SQL Management Studio project Retrieved from Github folder: https://github.com/mlaken/Matthew-Laken-Applied-Data-Science/tree/main/Projects/IST%20659

Laken, M (2021). IST 687 – Work completed in R using R Studio project Retrieved from Github folder:https://github.com/mlaken/Matthew-Laken-Applied-Data-Science/tree/main/Projects/IST%20687

Laken, M (2021). IST 707 – Work completed in Orange / Python project Retrieved from Github folder:https://github.com/mlaken/Matthew-Laken-Applied-Data-Science/tree/main/Projects/IST%20707

Laken, M (2021). IST 652 – Work completed in Python project Retrieved from Github folder:https://github.com/mlaken/Matthew-Laken-Applied-Data-Science/tree/main/Projects/IST%20652