PORTFOLIO MILESTONE

Jason Min-Liang Kang

SUID 894583556

Table of Contents

Introduction

I started the Applied Data Science program in 2018. I have been taking it slowly and carefully because I found it to be a broad and intricate subject. It took me awhile for all the pieces of knowledge to sink in and combined into a holistic understanding. In the following, I will talk about my journey to data science, the projects I have done, and how are they related to the overall data science process.

Journey to Data Science

Although statistical and data analytic have been studied for a long time, the term data science was first coined in 2008 by data researchers at LinkedIn and Facebook (Hu, 2017). The popularization was helped by the explosion of available data (big data) and by the need for teams of people to analyze that data in corporations and governments.

I became aware of the term data science around 2015. I came from software engineering background, so I had a foundational understanding of data structure and algorithm. However, it is a very different game when it comes to big data and drawing insight from data. I had a vague idea that data science comes from statistics so I started the initiative to take some online statistic classes. Those classes were interesting and helped to expand my knowledge in statistics. However, they seemed very remote from machine learning and big data.

Two years later, in 2017, I attended a local conference called LA Data Con. It was an embarrassing and defeating experience for me because I could understand every word spoken in the talks yet I could not connect them to my own experience/understanding. After the conference, I decided that I was going take some professional program/studies specific in data science. In April 2018, I started doing Applied Data Science program at Syracuse.

What is Data Science

Even before I started the Applied Data Science program, I heard all the great things about machine learning, such as autonomous driving vehicle, real-time language translation, and chess playing at grand master level. I thought machine learning was a new way of computer programming. Instead of code driven, it is now data driven. In fact, when I talked about data science, I meant machine learning. I carried that misconception when I started school. I was getting ready to learn all about machine learning from the get-go, but soon I realized that is not what data science is all about. In fact, machine learning is only a small part of data science. As depicted in figure 1, data science is a process that consists of data acquisition, data exploration, data analytic, and interpretation. Data analytic is further divided into descriptive analytic, predictive analytic, and prescriptive analytic. Machine learning is one application (perhaps arguably the most important application) of predictive analytic. This holistic understanding of the data science process is one of the most important knowledge I got out of the program. Instead of committing to a specific analytic technique, I now have the freedom and wisdom to pick the best strategy and technique to solve a specific problem domain.

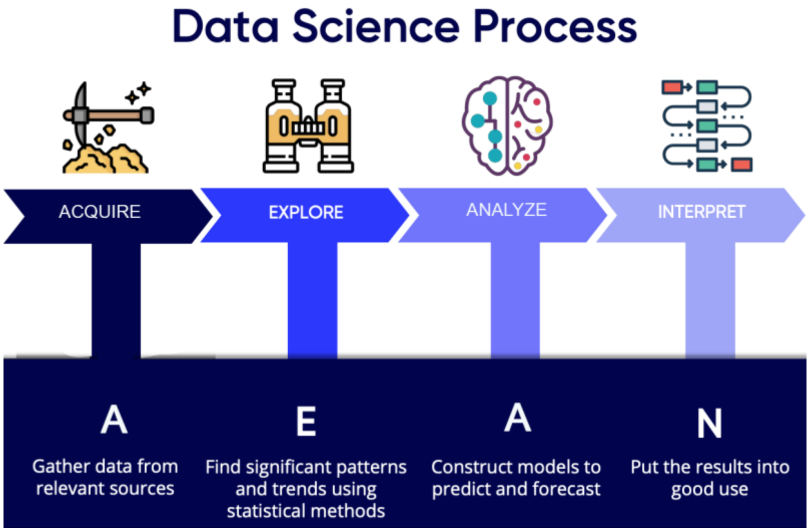

Data Science Process
ACQUIRE
EXPLORE
ANALYZE
INTERPRET
A
Gather data from relevant sources
E
Find significant patterns and trends using statistical methods
A
Construct models to predict and forecast
N
Put the results into good use

Figure 1. Data Science Process

Problem Formulation

Definition

Data science process is a problem-solving process. Before getting into the process, one has to have a clear problem to solve. Sometimes the problem is given and trivial. Other times, the problem may be buried deep in the business requirements and processes. When it comes to problem formulation, one course called IST 659 Data Admin Concepts really standout because it incorporates the formulation of real live problem.

IST 659 Data Admin Concepts

IST 659 Data Admin Concepts was taught by Professor Chad Harper. It was the first course I took in the program. At the time I was studying this course, I was also working on a side project idea. As a software engineer, I noticed how often I encountered obscure coding errors (i.e. error messages outputted by compiler to signal possible errors in the codes) during my daily coding works. The idea was to create a web-based platform for frustrated software engineer, like myself, to ask for and to share solutions to coding errors. When Professor Harper suggested to develop a real-life idea in the final project, I decided to give my little side project idea a try.

Under the direction of Professor Harper, I quickly developed a narrative. I named the platform CodeXchange. I designed three types of users – site curator, author, and wander; I developed detailed functions for each user. Based on the narrative, I extracted necessary business rules needed to develop logical models. Finally, with the insight derived from the logical models, I created the basic UI prototype (figure 2) designed to catch inputs from users. Detailed project documentation can be found in IST 659 Data Admin Concepts directory under Syracuse Data Science Portfolio GitHub repository (Kang, IST 659 Data Admin Concepts, 2020).

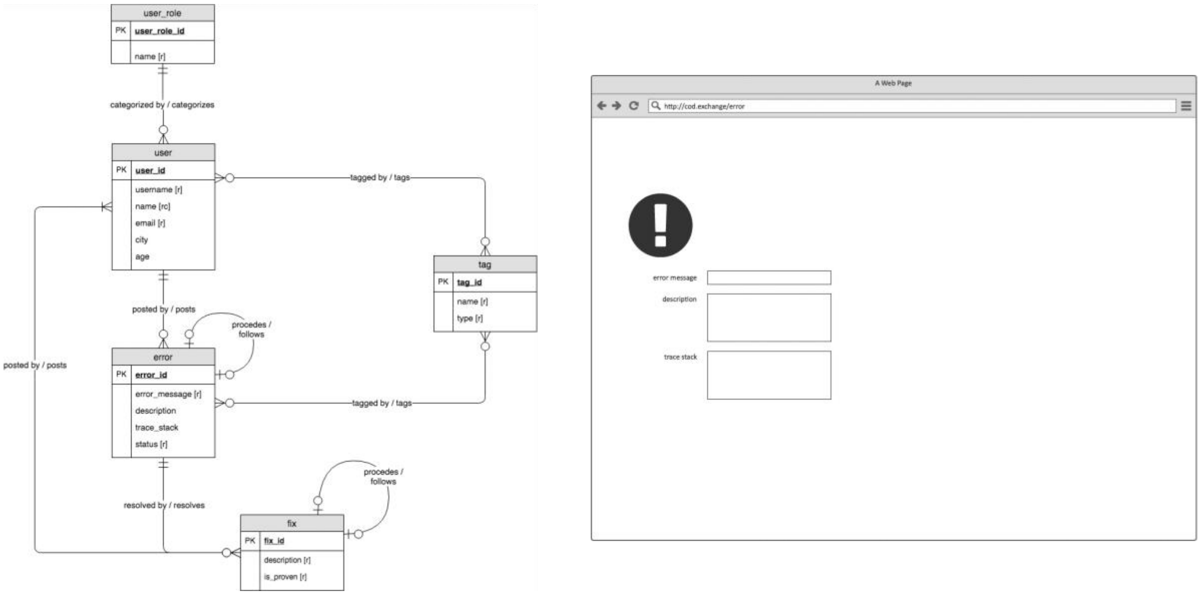


Figure 2 Conceptual Model and UI Prototype

Reflection

The most important skill that I learnt from this course is to write a narrative of the problem that I am about to solve. Writing it down forces me to think through exactly what I am trying to solve and what data I need to collect. This, in essence, is problem formulation. One takeaway from the course is that spending some extract time in problem formulation will pay off down the road.

Data Acquisition

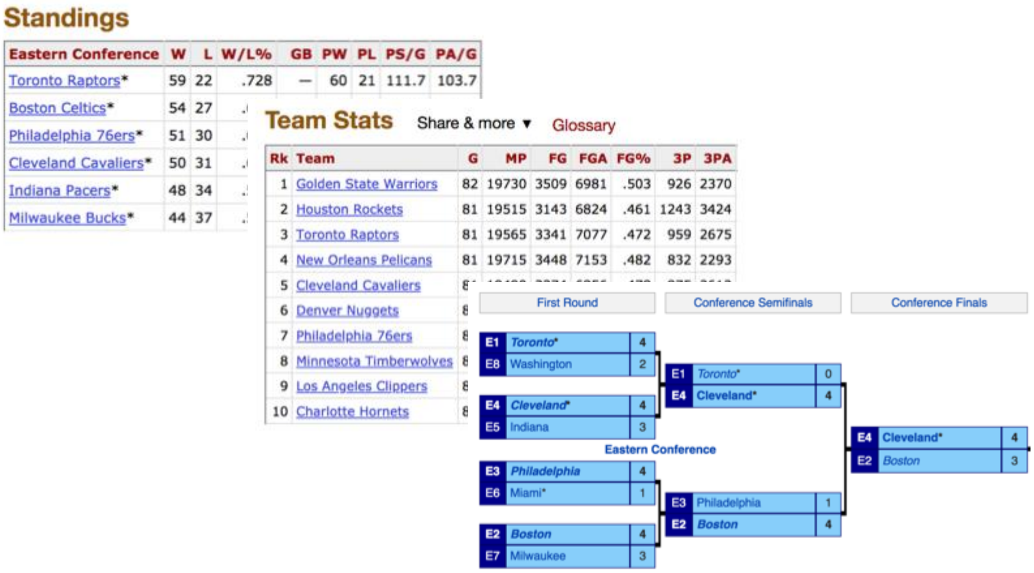
Definition

Data acquisition (i.e. collecting data) is the first step of the data science process. Data scientists use data to solve problems. Without good data, the rest of the data science process is fruitless. In software engineering, there is a concept called "garbage in garbage out". This is saying that during an algorithm computation, flawed input produces nonsense output. This same concept applies to data science as well. In most courses that I took, data was either provided or was readily available for download online. The most brutal data acquisition experience that I had was during the course IST 687 Introduction to Data Science.

IST 687 Introduction to Data Science

IST 687 Introduction to Data Science was taught by Professor Mohammed Syed. Since it was one of the earlier courses I took in the program, my data science skills were still very raw at that time. The final project was a group project. The goal was to find out whether there exists a correlation between the stats of NBA teams and their winnings. Intuitively, there must exists a correlation. The question, therefore, is to find out the data fields that do correlate.

Unfortunately, we could not find any prefabricated dataset that would fit our purpose. Our last resort was to scrape the data directly off official web sites. From Basketball Reference, we scraped team stats, and team standings; From Wikipedia, we scraped playoffs data. The web scarping took most than a week to finish. Figure 3 shows what the data looked like in their original web site. Once completed, we then stitched the data together to form the final dataset with 31 data fields. After that, we ran correlation tests on the dataset and were able to identify the fields that should be include in the final linear regression model. Detailed project documentation can be found in IST 687 Introduction to Data Science directory under Syracuse Data Science Portfolio GitHub repository (Kang, IST 687 Introduction to Data Science, 2020).



Standings

Eastern Conference	W	L	W/L%	GB	PW	PL	PS/G	PA/G
Toronto Raptors*	59	22	.728	—	60	21	111.7	103.7
Boston Celtics*	54	27						
Philadelphia 76ers*	51	30						
Cleveland Cavaliers*	50	31						
Indiana Pacers*	48	34						
Milwaukee Bucks*	44	37						

Team Stats

Rk	Team	G	MP	FG	FGA	FG%	3P	3PA
1	Golden State Warriors	82	19730	3509	6981	.503	926	2370
2	Houston Rockets	81	19515	3143	6824	.461	1243	3424
3	Toronto Raptors	81	19565	3341	7077	.472	959	2675
4	New Orleans Pelicans	81	19715	3448	7153	.482	832	2293
5	Cleveland Cavaliers	8						
6	Denver Nuggets	8						
7	Philadelphia 76ers	8						
8	Minnesota Timberwolves	8						
9	Los Angeles Clippers	8						
10	Charlotte Hornets	8						

First Round

E1	Toronto*	4
E8	Washington	2
E4	Cleveland*	4
E5	Indiana	3
E3	Philadelphia	4
E6	Miami*	1
E2	Boston	4
E7	Milwaukee	3

Eastern Conference

Conference Semifinals

E1	Toronto*	0
E4	Cleveland*	4
E3	Philadelphia	1
E2	Boston	4

Conference Finals

E4	Cleveland*	4
E2	Boston	3

Figure 3 Data from Official Web Sites

Reflection

One takeaway from this course is that useful data does not always come cheap. This course prepared me to take on extreme measures to acquire useful data. In hindsight, instead of scraping the data manual, I should had invested the time to automate the web scraping process by using tools like Scrapy and BeautifulSoup.

Data Exploration

Definition

Once the dataset is acquired, a series of data exploration techniques are applied/performed on the dataset. There are two part of data exploration. The first is data cleaning. In this part, data records that contain missing or invalid data should be removed; those that contain incorrectly formatted data should be converted to the correct format. Some data types are prone to formatting issues such as dates, address, and phone number. The second part of data exploration is feature identification. In this part, useful or meaningful data fields are identified using correlation techniques and are selected to move on to data analytic phrase. Almost all projects I have done so far require some sort of data exploration. The final project in IST 718 Big Data Analytics course is the most challenging among them.

IST 718 Big Data Analytics

This course was taught by Professor Jillian Lando. It is an advanced course in data analytics that covers a wide range of data analytic techniques. The final project for this course was a group project. Since Kaggle just released its 2019 data science competition problem around that time, we tried to be fancy and decided to use the competition problem for the final project. The goal of the competition problem was to gain insights into how gaming can help children learn important skills for success in school and life. Dataset was raw data collected by an app called PBS KIDS Measure UP. It was an app that helped children ages 3 to 5 to learn early math concepts through completing levels and assessments in a game setting. Learning materials are presented to children as games. At the end of each game level, children had to pass an assessment test in order to gain access to next level. The success of learning is therefore measured by how many attempts children had to take to complete the assessment. In another words, the number of attempts is the dependent variable that we want to predict in the models. That was not given in the raw data, however. So, the first thing we did was to derived the number of attempts by aggregating data records by game session ID and assessment ID. We also removed invalid data during the process. Once we had the dataset cleaned up, we then proceeded to visualize the dataset for any correlation and pattern by plotting it out. We ran it through a couple of common plots, such as histogram, boxplot, heat-map, as depicted in figure 4. For modeling, we chose CatBoost, XGBoost Classifier and Decision Tree. The Decision Tree model yields the most accurate result. Detailed project documentation can be found in IST 718 Big Data Analytics directory under Syracuse Data Science Portfolio GitHub repository (Kang, IST 718 Big Data Analytics, 2020).

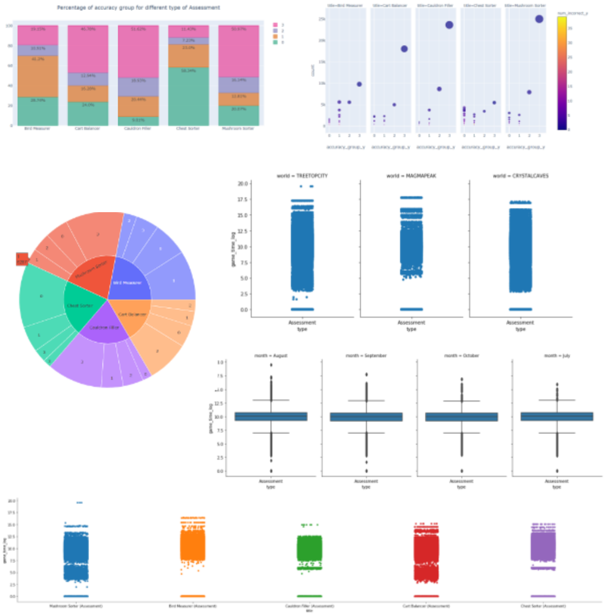


Figure 4 Plots used in Data Exploration

Reflection

The raw dataset was almost 4 GB in size. This makes it really slow to perform data exploration routines from a laptop computer. It would even crash our computers from time to time. In hindsight, we should had stored the dataset in a cloud database and perform data exploration routines directly from the cloud for better performance and response time.

Data Analytic

I used to think that data analytic is the longest and most complicated part of data science. In practice, it is actually easier and shorter comparing to data acquisition and exploration because all commonly used analytic algorithms are well studied and implemented. The hardest part is to use the right analytic algorithm and to understand the results. Data analytic is divided into three types: descriptive, predictive, and prescriptive.

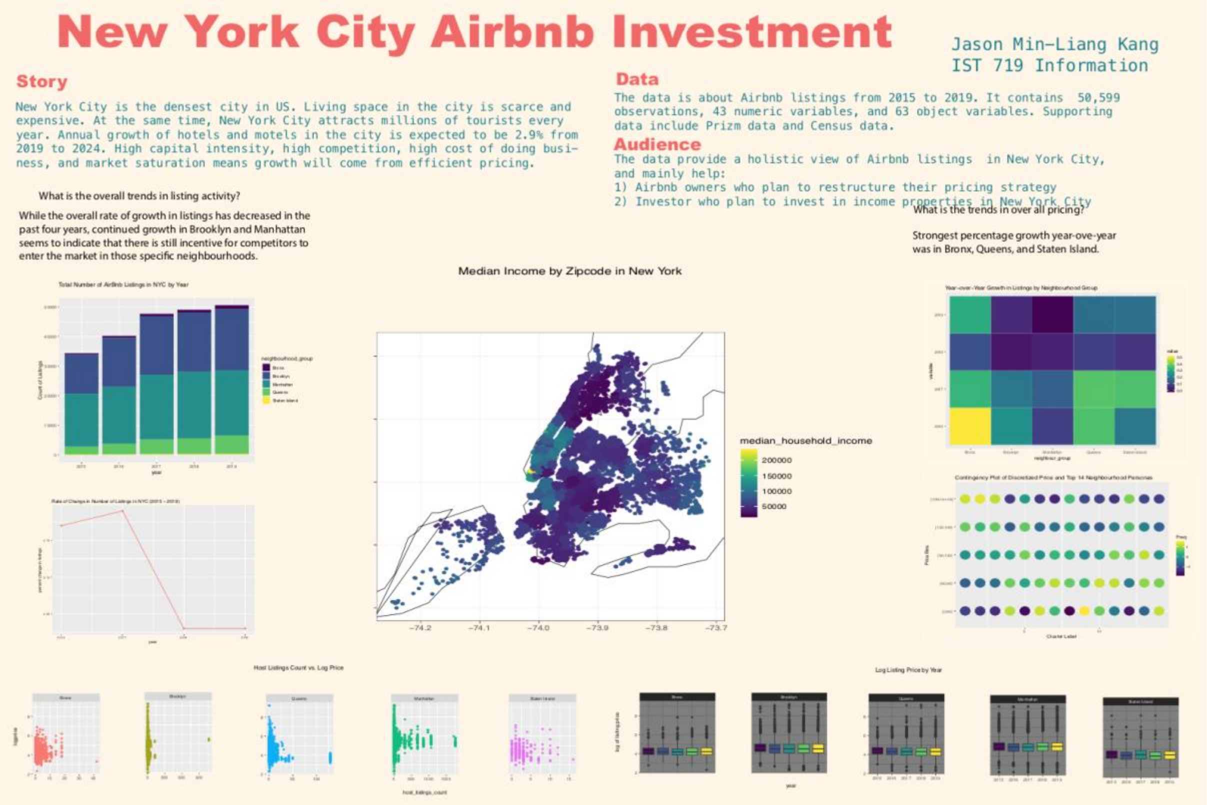
Descriptive

Definition

Descriptive analytic looks at what is in the most current data. It reflects what is in the past and what is in the present. Descriptive analytic is further divided into four types: measures of frequency, measures of central tendency, measures of dispersion, and measures of position. The resulting measurements are commonly plotted for visualization purposes. For this reason, descriptive analytic techniques are often used in data exploration phrase as well. The final project in IST 719 Information Visualization employs the most descriptive analytic techniques.

IST 719 Information Visualization

This course was taught by Professor Gary Krudys. It focused exclusively on visualization features in R. The final project was a poster presentation. Since I was also working on New York City Airbnb dataset for another course at that time, I decided to repurpose that dataset for the poster. The goal is to inform Airbnb investors which neighborhood has the highest potential for rental pricing growth. I created plots to compare pricing growth between neighborhoods and pricing growth overtime. I organized the plots according to visualization principles learnt from the course. Final poster is shown in figure 5. Detailed project documentation can be found in IST 719 Information Visualization directory under Syracuse Data Science Portfolio GitHub repository (Kang, IST 719 Information Visualization, 2020).



New York City Airbnb Investment

Jason Min-Liang Kang
IST 719 Information

Story

New York City is the densest city in US. Living space in the city is scarce and expensive. At the same time, New York City attracts millions of tourists every year. Annual growth of hotels and motels in the city is expected to be 2.9% from 2019 to 2024. High capital intensity, high competition, high cost of doing business, and market saturation means growth will come from efficient pricing.

Data

The data is about Airbnb listings from 2015 to 2019. It contains 50,599 observations, 43 numeric variables, and 63 object variables. Supporting data include Prizm data and Census data.

Audience

The data provide a holistic view of Airbnb listings in New York City, and mainly help:
1) Airbnb owners who plan to restructure their pricing strategy
2) Investor who plan to invest in income properties in New York City

What is the overall trends in listing activity?

While the overall rate of growth in listings has decreased in the past four years, continued growth in Brooklyn and Manhattan seems to indicate that there is still incentive for competitors to enter the market in those specific neighbourhoods.

What is the trends in over all pricing?

Strongest percentage growth year-ove-year was in Bronx, Queens, and Staten Island.

Median Income by Zipcode in New York
median_household_income
200000
150000
100000
50000
−74.2
−74.1
−74.0
−73.9
−73.8
−73.7
Total Number of AirBnb Listings in NYC by Year
Year-over-Year Growth in Listings by Neighbourhood Group
Contingency Plot of Discretized Price and Top 14 Neighbourhood Personas
Host Listings Count vs. Log Price
Log Listing Price by Year

Figure 5 Final Poster

Reflection

Descriptive analytic is often expressed through data visualization because the best way for human to recognize pattern in measurements is to visualize it. One takeaway from the course is that viewers can only digest so much information at a time. Therefore, clean and simple visuals can convey the message/information far better than eye catching yet complex visuals.

Predictive

Definition

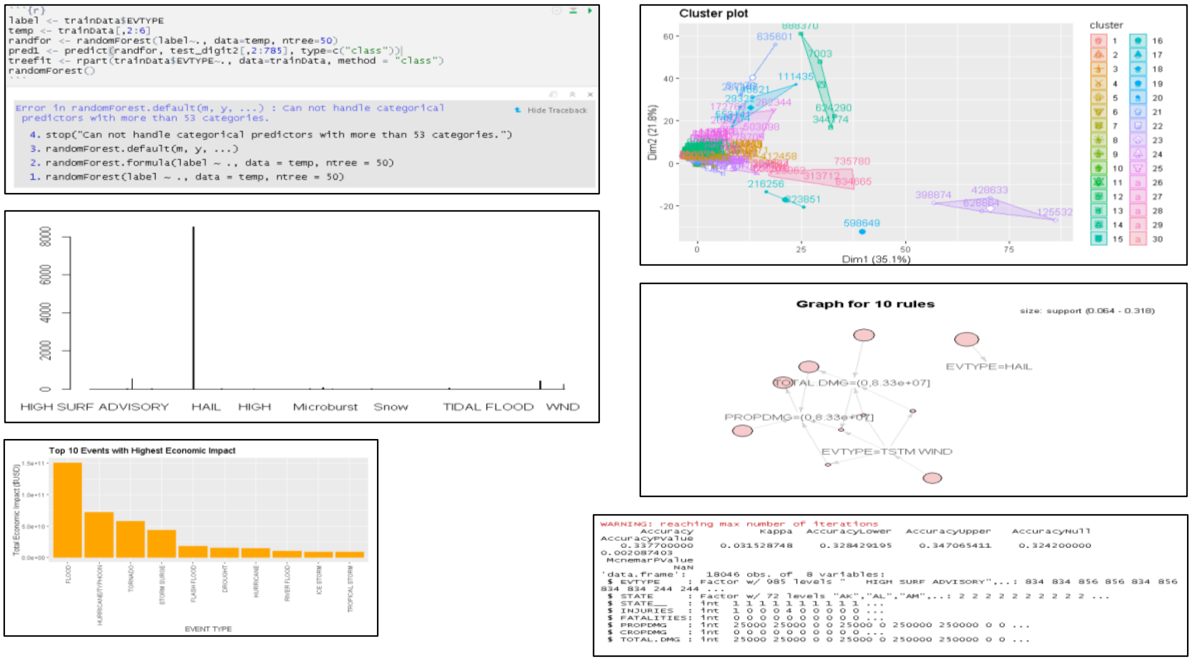
Predictive analytic looks at what is likely to happen in the future given what has happened in the past and present. There are roughly four types of predictive analytic models: regression, classification, clustering, and neural network. Each type is has advantages in different situations. The final project in IST 707 Data Analytics was an interesting one because requires the use of different predictive analytic models to analyze the same dataset.

IST 707 Data Analytics

The course was taught by Professor Amy Gates. It was a comprehensive study of different predictive analytic models. For the final group project, Professor Gates wanted us to use different analytic models to analyze the same dataset, and compare the results and tradeoffs. We decided to use severe weather events data from National Oceanographic and Atmospheric Administration (NOAA). The goal is to predict which events are most harmful to population health, and have the greatest economic consequences. Since this is a classification problem, we experimented with a couple of classification models, such as Decision Tree, Random Forest, Support Vector Machine, Naïve Bayes, K Means, and Association Rule Mining. The results from some of the models are shown in figure 6. We had to make slight adjustments to the dataset to fit each model of course. Our findings shown that Tornados as the leading cause of injuries, and that flooding had the greatest impact on U.S. economy. Detailed project documentation can be found in IST 707 Data Analytics directory under Syracuse Data Science Portfolio GitHub repository (Kang, IST 707 Data Analytics, 2020).

Reflection

One takeaway from the course is the realization that a dataset can be rearranged to fit almost any analytic models. A good practice is to run the dataset through different models then compare the results. However, to know which model works best with a dataset and why take practice and experience.



~~~
```{r}
label <- trainData$EVTYPE
temp <- trainData[,2:6]
randfor <- randomForest(label~., data=temp, ntree=50)
pred1 <- predict(randfor, test_digit2[,2:785], type=c("class"))
treefit <- rpart(trainData$EVTYPE~., data=trainData, method = "class")
randomForest()
```
~~~

```
Error in randomForest.default(m, y, ...) : Can not handle categorical
 predictors with more than 53 categories.
  4. stop("Can not handle categorical predictors with more than 53 categories.")
  3. randomForest.default(m, y, ...)
  2. randomForest.formula(label ~ ., data = temp, ntree = 50)
  1. randomForest(label ~ ., data = temp, ntree = 50)
```

```
WARNING: reaching max number of iterations
      Accuracy          Kappa  AccuracyLower  AccuracyUpper   AccuracyNull 
AccuracyPValue
   0.337700000    0.031528748    0.328429195    0.347065411    0.324200000 
0.002087403
 McnemarPValue
           NaN
'data.frame':   18046 obs. of  8 variables:
 $ EVTYPE    : Factor w/ 985 levels "   HIGH SURF ADVISORY",..: 834 834 856 856 834 856 
834 834 244 244 ...
 $ STATE     : Factor w/ 72 levels "AK","AL","AM",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ STATE__   : int  1 1 1 1 1 1 1 1 1 1 ...
 $ INJURIES  : int  1 0 0 0 4 0 0 0 0 0 ...
 $ FATALITIES: int  0 0 0 0 0 0 0 0 0 0 ...
 $ PROPDMG   : int  25000 25000 0 0 25000 0 250000 250000 0 0 ...
 $ CROPDMG   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ TOTAL.DMG : int  25000 25000 0 0 25000 0 250000 250000 0 0 ...
```

Figure 6 Data Analytics Model Results

Prescriptive

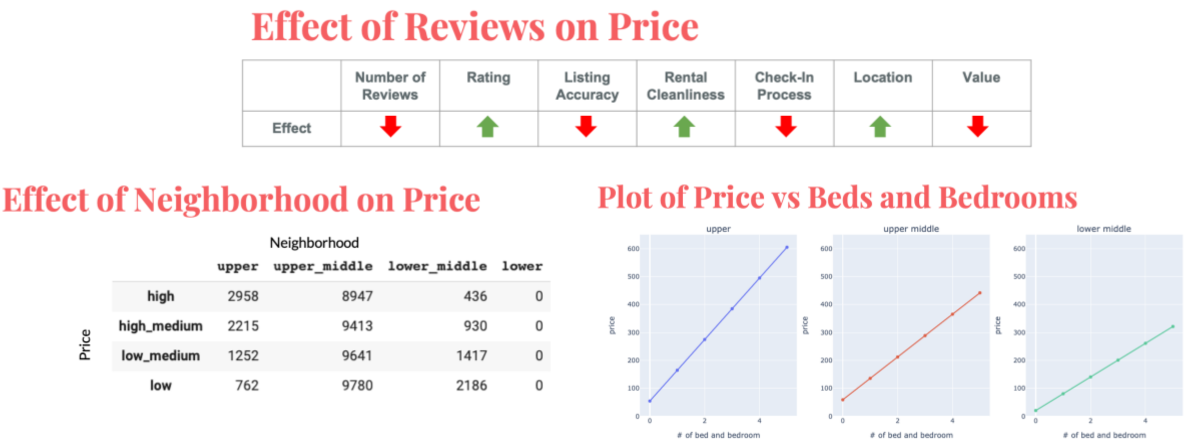
Definition

Prescriptive analytic employs the same kind of algorithm found in predictive analytic. However, rather than looking for the most accurate predicted outcome, prescriptive analytic focuses on actionable predictor. That is, feature/variable that can be controlled/manipulated at present time. The goal is to find out which variable is an actionable predictor and by how much it should be changed. In some cases, the more accurate but unactionable predictors need to be dropped from the model in order to see the effect of lesser accurate but actionable predictors.

Most of the courses I took only talk about predictive analytic. This is understandable because data scientists concern more with the overall accuracy of the model rather than improving the dependent variable. My only experience working with prescriptive analytic modeling is in MAR 653 Marketing Analytics course.

MAR 653 Marketing Analytics

The course was taught by Professor Shaam Ramamurthy, and was focusing on analytic algorithms geared toward business and marketing researches. For the group final project, we decided to play with New York City Airbnb dataset. Our goal was to find out what can Airbnb investors do to improve their rental income (i.e. increase the rental price of their Airbnb property). We ran the Airbnb dataset through a series of data exploration analytics, and narrowed down to a handful of actionable predictors — reviews, neighborhood, number of beds, and number of bedrooms (depicted in figure 7). For each predictor, we evaluate how it would affect rental pricing. Detailed project documentation can be found in MAR 653 Marketing Analytics directory under Syracuse Data Science Portfolio GitHub repository. Detailed project documentation can be found in MAR 653 Marketing Analytics directory under Syracuse Data Science Portfolio GitHub repository (Kang, Mar 653 Marketing Analytics, 2020).



Effect of Reviews on Price

	Number of Reviews	Rating	Listing Accuracy	Rental Cleanliness	Check-In Process	Location	Value
Effect	⬇	⬆	⬇	⬆	⬇	⬆	⬇

Effect of Neighborhood on Price

	Neighborhood			
Price	upper	upper_middle	lower_middle	lower
high	2958	8947	436	0
high_medium	2215	9413	930	0
low_medium	1252	9641	1417	0
low	762	9780	2186	0

Plot of Price vs Beds and Bedrooms

Figure 7 Actionable Predictors

Reflection

One takeaway from this course is to discern actionable predictors from predictors. In the project, we were able to get the actionable predictors and to evaluate their effects. This would give Airbnb investors some idea when they need to make their investment decisions. However, further analytics are needed to find out the optimal values that can yield the best output from these actionable predictors.

Interpretation

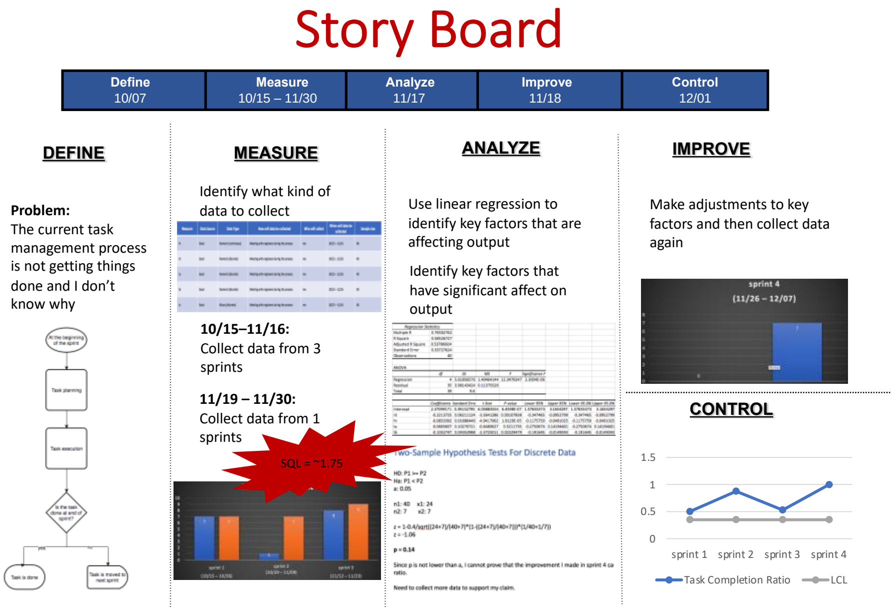
Definition

As mentioned before, data science process is also a problem-solving process. The previous steps focus on finding and analyzing evidences/data. The final step, interpretation, focuses on finding a solution and presenting the solution with evidence. Every data science course I took involves interpretation in one form or another. Data science is a discipline of making recommendation, after all. Of all the presentations I have done in the program, the one from MBC 638 Data Analysis and Decision Making is most memorable.

MBC 638 Data Analysis and Decision Making

The course was taught by Professor Marc Miller. It focused on implementation of process improvement using data science techniques. The final project was an opportunity to identify issues in a real-world process, and to improve the process. As a software developer, I naturally decided to use software development process as an experiment.

Software development process works in sprints of two weeks. For each sprint, each software developer is given a set of tasks to complete in that sprint. The goal is to improve the task completion rate. After collecting relevant data/measurements and done analytics on them, I found out that total estimated hours and number of tasks have negative impact on task completion rate. The whole story is captured and summarized in one slide as shown in figure 8. Detailed project documentation can be found in MAR 653 Marketing Analytics directory under Syracuse Data Science Portfolio GitHub repository (Kang, MBC 638 Data Analysis and Decision Making, 2020).



Story Board

Define 10/07	Measure 10/15 – 11/30	Analyze 11/17	Improve 11/18	Control 12/01

DEFINE

Problem:
The current task management process is not getting things done and I don't know why

At the beginning of the sprint
Task planning
Task execution
Is the task done at end of sprint?
yes
no
Task is done
Task is moved to next sprint

MEASURE

Identify what kind of data to collect

10/15–11/16:
Collect data from 3 sprints

11/19 – 11/30:
Collect data from 1 sprints

SQL = ~1.75
sprint 1
sprint 2
sprint 3

ANALYZE

Use linear regression to identify key factors that are affecting output

Identify key factors that have significant affect on output

Two-Sample Hypothesis Tests For Discrete Data

HO: P1 >= P2
Ha: P1 < P2
a: 0.05

n1: 40 x1: 24
n2: 7 x2: 7

z = 1-0.4/sqrt(((24+7)/(40+7)*(1-((24+7)/(40+7)))*(1/40+1/7))
z = -1.06

p = 0.14

Since p is not lower than a, I cannot prove that the improvement I made in sprint 4 ca ratio.

Need to collect more data to support my claim.

IMPROVE

Make adjustments to key factors and then collect data again

CONTROL

Figure 8 One Slide Story Board

Reflection

This course taught me that story board is an effective way to communicate data science process and findings. I regret not able to use it often enough in my other projects due to time constraints, but it will remain a useful tool in my data science arsenal.

Data Security

Definition

According to Forbes, data is the most important currency used in commerce today (Evans, 2018). Data scientists have the responsibility not only to extract information/answer from data but also to protect the data against malicious use. There are two types of malicious use. The first type is malicious use by the data scientist. For example, the data scientist may have included sensitive or classified information in data analytic phrase. The second type is malicious use by an unauthorized person such as a hacker. In this case, the hacker usually gets to the data through network vulnerability. Both types are thoroughly discussed in IST 623 Introduction to Information Security taught by Professor Tyson Brooks.

Moving Forward

Last three years have been a wonderful journey into data science. As the program comes to an end, however, my journey does not end here. After the program, I plan to take a deep dive into data analytic algorithms. In addition to how to use them, I want to learn the ins and outs of how they work. I also want to gain more knowledge in data/information security as this will become a big challenge for the coming decades. More importantly, I want to apply what I have learnt in my workplace as well as the community.

Works Cited

Evans, M. (2018, March 12). *Why data is the most important currency used in commerce today*. Retrieved from Forbes: https://www.forbes.com/sites/michelleevans1/2018/03/12/why-data-is-the-most-important-currency-used-in-commerce-today/?sh=407b1e8e54eb

Hu, J. (2017, July 27). *The Development of Data Science*. Retrieved from Aponia: https://aponia.co/development-data-science-ny

Kang, J. (2020, December 6). *IST 659 Data Admin Concepts*. Retrieved from Github: https://github.com/jasonmlkang/syracuse_data_science_portfolio/tree/main/ist_659_data_admin_concepts

Kang, J. (2020, December 6). *IST 687 Introduction to Data Science*. Retrieved from Github: https://github.com/jasonmlkang/syracuse_data_science_portfolio/tree/main/ist_687_introduction_to_data_science

Kang, J. (2020, December 6). *IST 707 Data Analytics*. Retrieved from Github: https://github.com/jasonmlkang/syracuse_data_science_portfolio/tree/main/ist_707_data_analytics

Kang, J. (2020, December 6). *IST 718 Big Data Analytics*. Retrieved from Github: https://github.com/jasonmlkang/syracuse_data_science_portfolio/tree/main/ist_718_big_data_analytics

Kang, J. (2020, December 6). *IST 719 Information Visualization*. Retrieved from Github: https://github.com/jasonmlkang/syracuse_data_science_portfolio/tree/main/ist_719_information_visualization

Kang, J. (2020, December 6). *Mar 653 Marketing Analytics*. Retrieved from Github: https://github.com/jasonmlkang/syracuse_data_science_portfolio/tree/main/mar_653_marketing_analytics

Kang, J. (2020, December 6). *MBC 638 Data Analysis and Decision Making*. Retrieved from Github: https://github.com/jasonmlkang/syracuse_data_science_portfolio/tree/main/mbc_638_data_analysis_and_decision_making