Masters Portfolio Requirement Exit Requirement Spring 2020

Data Science Program Overview Paper

Benjamin Allen

Data science is a multidisciplinary field which I have been preparing myself to be well versed in. From my program the major skills learned included how to explore data; which involved using visualization and descriptive tools. There is an exhaustive list of tools we used, but the ones that stand out are, filters in excel, and creating pivot tables of the data, using charts and graphs. In Python, using functions to understand the shape and size of data, and commands for knowing if there is missing data. Making regression models to predict outcomes for categorical and continuous variables. Overall, this paper explains these processes.

Upon entering my IST 618 (Introduction to Data Science) class I quickly learned new ways to process data. Cleaning data and determining insights is often an 80/20 process. So, in class we spent a lot of time understanding how to solve for missing data. The basics include, removing rows and columns where the data is missing. But a more robust method is to impute the data. Fill it using mean, median, mode or a more complex manner of using a technique like K-nearest neighbor that imputes missing values in a way that considers the correlation between other columns in the data. You can see how I filled missing data in both my Microsoft Malware Project and IBM Attrition Analysis.

Our program taught us analytical approaches to interpreting our data. We often had the goal of gaining insight from the data. In my Airbnb Exploratory Analysis in New York City project, I used plots and graphs to understand price differences between different boroughs. From this data, I made inferences and explained the models output in a non-technical way. Also, in my First Mini Project, I used python pandas and sci-kit learn to create regression models that predicted housing prices. What makes our program unique was our evaluation in class was based on our ability to answer difficult problems using complex algorithms and explain our process and results to non-technical audiences.

When working on my final Microsoft Malware Project in Data Analytics our goal was to predict what machines would get a virus on them. In doing this project we ran into technical issues when certain analysis could not be performed that were previously thought to be applicable to the problem. Those algorithms were K-Means and Bayesian Regression. These algorithms limitations for our use case was the outcome variable should be a probability of the detection of a virus while these algorithms dealt with clustering and a continuous variable outcome. Upon more research into our data problem, the type of algorithm we thought to use was logistic regression. This turned out to be a good plan being that we could have a prediction outcome, but one last step was necessary which was determining the outcome for each machine, not just all machines. This is where using a gradient boosting algorithm was preferred.

When using the logistic regression model on this dataset, we encountered problems with choosing which features to keep and eliminate in the model. The first model gave us

results that could determine which predictors would not add value to the model simply by looking at the p-value of the model. The p-value results show when variables are above 0.05, they would not be statistically significant and should be removed from the model. Once we removed the variables with high p-values and ran the model again, this approach yielded a 0.59 accuracy score for determining machines with viruses. From here, we wanted to improve the score, so we tried the Light Gradient Boosting algorithm. It is important to note that the problem with eliminating features based on p-values is that our data had many categorical columns which needed to be turned into a numerical representation of the categories. To handle this, we used one hot encoder, and this vastly increased the number of columns in our dataset, making our logistic regression model more susceptible to overfitting and increasing the run time when scoring the algorithm.

Despite that, when using the Light GBM algorithm, we had a 12 percent increase in the accuracy of the model. Reason being, the algorithm was able to process what features are important through a method called gain. The feature selection by using gain is an improvement from feature selection by using p-values, because we additionally consider the measure of strength each variable has toward predicting the outcome variable. This more robust process of feature elimination gives us the ability to select a smaller number of features that are the best at predicting the outcome.

Data is in three forms: structured, semi-structured, and unstructured. Structured data are distinguished as tables which contain columns and rows. While, unstructured data have a format that does not contain a columnar structure. Semi-structed data does not have the columnar structure but has something like a tag system to identify relationships in the data. An example would be a content management system like Apple Music that stores different media formats in a tagging system thus transforming unstructured data into a semi-structed format. The tagging system gives meaning to data like an album contains many songs or an artist can have many albums and/or songs.

Using Mongo DB in my IST 652 (Scripting for Data Analysis) class I collected and manipulated semi-structed data into a structured format. Json data was imported from a popular website that stores information on baby names. In order to import the parts of the json that is needed for analysis, I inspected the raw text, and upon inspection, I made a connection to the website and retrieved the data from the server and stored it into a local mongo database. To store the data properly I needed to transform the json strings into a python dictionary, and then create a list of dictionaries.

In my code, you can see how I did this:

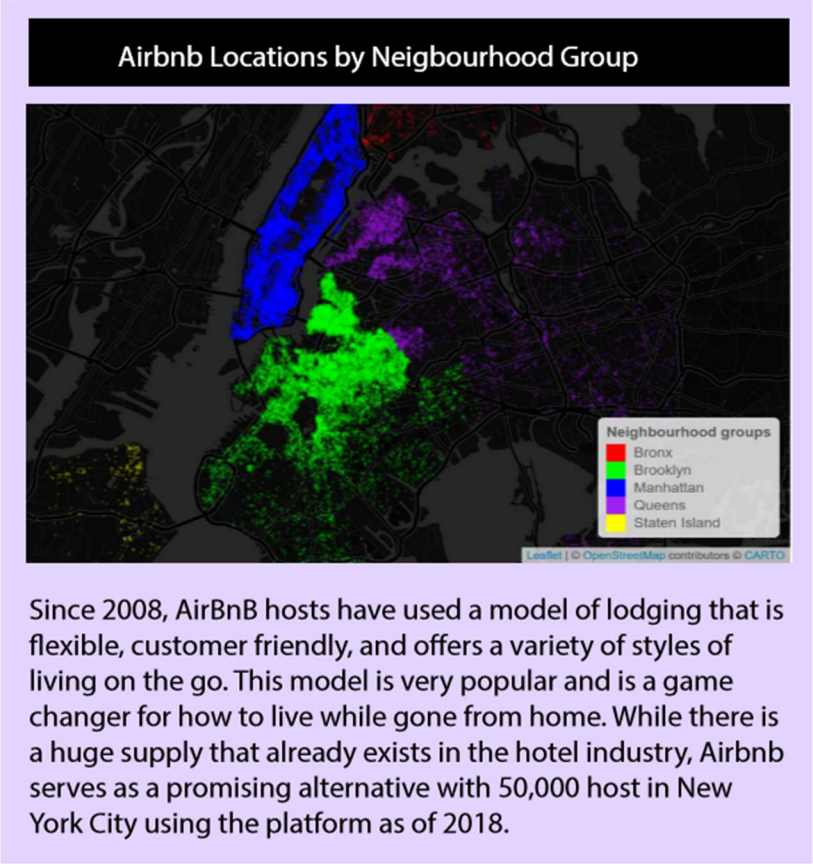
```
name = df.values.tolist()
dict_list = ["Year","Gender","Ethnicity","Name", "Instance", "Rank"]
name_dict = []

for elem in name:
    dict_temp = {}
    for i in range(len(elem)):
        dict_temp[dict_list[i]]=elem[i]
    name_dict.append(dict_temp)
name_dict
```

```
[{'Year': '2011',
  'Gender': 'FEMALE',
  'Ethnicity': 'HISPANIC',
  'Name': 'GERALDINE',
  'Instance': 13.0,
  'Rank': 75.0},
 {'Year': '2011',
  'Gender': 'FEMALE',
  'Ethnicity': 'HISPANIC',
  'Name': 'GIA',
  'Instance': 21.0,
  'Rank': 67.0},
 {'Year': '2011',
  'Gender': 'FEMALE',
```


Once transformed, I can look at the names across different years, gender, or race, and get an understanding of my data. This is a crucial step.

At the end of my fall semester in 2019, I recorded my data visualizations project and I choose to record using Prezi (a digital and interactive PowerPoint presentation). This assignment was a very detailed presentation showing an exploratory analysis of Airbnb Data in New York City. This project's goal was to use an exploratory analysis and link together interesting facts found in the data, thus telling a data story of why Airbnb is popular in certain areas of New York City. Using the leaflet package in R, I created a map that distinguished each borough in New York City and the density of Airbnb units in the respective boroughs.



This data visualization I created displays the demographics for Airbnb's listing types in NYC and provides insight into why certain areas have more Airbnb units.

When companies understand the motivation behind employee retention, they can properly make plans that make their job offerings more beneficial. In my IBM Attrition project, I used a logistic regression model to determine what leads to attrition. This problem is associated with a need to understand people's motivations for staying in a specific role. This action is considered an ethical application of the algorithm because the company wants to minimize people leaving a job, which helps the company and employees. But one could say that this algorithm should be used to determine what prospects are more likely to leave, therefore we should avoid hiring/retaining people who are more at risk.

Now, hiring is a risk for both parties involved because there is an exchange of money, promises, and expectations. Considering that, the algorithm can try to identify characteristics of people and know whether they are good candidate to stay with a company. The algorithm used in this project inherently could suggest women between the age 18-35 are more likely to leave a job, because they are in childbearing years. Should that person not be hired? Should more men be hired than woman to account for this potential issue? What we are facing now is an ethical dilemma where data scientist must decide how to handle data that has the potential to discriminate against woman. As a data scientist, these patterns exist in the data and I understand that decisions cannot be made

based on strong relationships along. In interpreting and understanding data there exist a responsibility to not be bias and make moral choices.

In practice, I have worked with data enough to know that you can spend countless hours trying to manipulate data to get the desired outputs you want. Common enough, data scientist must manage time resources and people when completing projects. In our program the data science projects given are usually within a team, and teams must decide the path for completing the project. This includes choosing our data of interest to work on and following a roadmap to get some desired result. Our time windows to explore the data are very important because topics chosen must align with the current skill level of the individuals in the group, and the data must meet certain size and complexity based on the class requirements. I mention this because there are a lot of people skills and organizational skills that I have learned in school, and they are valuable to becoming a data scientist.

Overall this program has taught me how to collect data and organize it, which I learned in my scripting class. It also provided ways to display visualizations and interpret them. In most of my classes I was challenged to rethink strategies I once thought could work. From there, I had to implement models that lead to actionable insights in a business context. The purpose of the final project reports in class and including this paper is to be able to communicate my understanding to others. Overall, I learned that it is very important to communicate effectively and hone in on my skills as I pursue a career in this field and do that with integrity, as I did in my school work.