**Data Generating Process / Bias Template**

**Student: Srivishnu Maganti**

**Date Submitted:** October/11/ 2020

**Task Selected:** *Response time (in seconds)*

Before we get to working with the data and building a predictive model, in this document you will provide an assessment of what you, the model builder, perceive to be the data generating process for your data sources and potential sources of bias that could result from either the data or the model building process.

**Data Generating Process**

Since you are role-playing a data analyst responding to your supervisor’s request to build a predictive model, you will be working with data that you didn’t have a hand in originally collecting, and you may not fully understand. This is common in many practices as a data analyst, and you have to build a healthy skepticism of **any** dataset you work with, even if you created it yourself. Before you can build your predictive model, you have to first understand how the data was generated, and what potential sources of bias might be present that could skew your results. This first section will focus on the data and its potential sources of bias; the next section will focus on issues that could arise with the modeling process itself.

Apart from the ambulance dataset I want to consider the temperature data of new York city so, by using few internet resources I found out two data sources which I have converted to csv format, one file consists of average temperature in New York from 2008 to 2016, and other file consist of amount of snowfall in new York city from 2008 to 2016.

I will mail both datasets for a clear understanding.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Item*** | ***Ambulance Calls for Service*** | ***Average temperature*** | ***Snow fall*** | ***{Additional Dataset 3}*** |
| ***Dates of Coverage*** | 01-01-2008 to 31-12-2016 and there are two entries for 01-01-2017(first action date) | 01-01-2008 to 31-12-2016 | 01-01-2008 to 31-12-2016 | … |
| ***Frequency of data collection*** | Data is collected per every incident, for instance data is being noted down for every action that they have taken in all those 8 years. | Data is collected every year.  Basically, the data that I collected was not in JSV format and I have converted in to excel sheet, it was collected year by year | Data is collected every year, it is collected from same previous data source, converted to jsv format. | … |
| ***Agency / Organization collecting the data*** | Basically, this dataset is published by data.cityofnewyork.us, And it was collected by the emergency medical services department and kept in NYS open data. | The data is published by national weather service broadcast office, and the data is available on the internet, it was available as time bound service. | national weather service broadcast. Taken from national weather service broadcast. | … |
| ***Original Unit of Analysis*** | For each factor that I consider there is different unit for analyses, some of the columns that I consider have code unit, and few columns such as borough has float type expression, and there are few columns which has date type format.  Response time which I am going to predict are in seconds. | Here the data average temperature is in Fahrenheit. | Snowfall is given in inches. | … |
| ***Transformed Unit of Analysis*** | One column that I am considering borough is in the float format like they are expressing in area, and I am going to convert it in to numerical (such as Brooklyn-1, Manhattan-2) like that, as numerical formats are very easy while modelling. And there are many columns in date format I don’t want to change or transform those formats. | No, I don’t want to convert this unit of analyses into any other unit. | No, I want to use this data in inches unit. | … |
| ***Data Generation Description*** | As per my research I can say that the data is being generated as a database storage, that every call that the ems response time receive is being stored and recorded and using natural language processing tools the initial incident details were recorded and final tools will be recorded as per the statement and dispatcher. | Data is being generated by survey, all the measurements have been taken from automated network of scientific instruments. | There are snow telemetry stations located and snow pack indicators uses these stations to record and automatically update the central server every day or every hour, depending upon the location | … |
| ***Data Collector*** | Few columns are being collected and recorded by dispatcher and few are machine oriented and they are computerized processed (such as date, time, time taken to act etc.) | Automated computerized instruments programmed for collecting the temperature. | Snowpack indicators at every stations records the data. | … |
| ***Triggering Process*** | The ems call in the situation can trigger the process. | Average Temperature change in a particular month can trigger the process. | Change in snowfall triggers the process. | … |
| ***Process Alignment*** | Data from this call service comes from call done to 911.  Firstly, when the call arrives to the service the call receiver enters the severity of the situation by code words to the CAD system and then the available dispatcher will record the final severity and time taken etc. | When the temperature is collected every day, It will be stored in automated systems, further it will be averaged every month and then it will be provided in the table. | Every subsidiary of new York city has different telemetry station, and that records all the snow fall data at respective station and reports the record automatically to central system. | … |

**Potential Sources of Bias**

In this project, you are using ambulance calls for service data from New York City during the years 2008-2016 to build your models. While the data is administratively collected from Computer-Aided Dispatch records, it doesn’t mean that it is free from differences in data scope or collection that could bias its results from one or more subgroups. In addition, any additional datasets you plan on bringing in also suffer from potential biases that need to be acknowledged at the outset. In this section, for the datasets mentioned above, we will consider some key questions to help flesh out those potential biases.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Item*** | ***Ambulance Calls for Service*** | ***{Additional Dataset 1}*** | ***{Additional Dataset 2}*** | ***{Additional Dataset 3}*** |
| ***Representativeness*** | {What is this dataset attempting to capture data about? Is the dataset designed to be representative of the underlying phenomenon that it is attempting to cover?}  Yes, the data set is actually covering all the emergency services in NYC from 2008 to 2016. | Basically, this dataset is attempting to cover average temperature of each month per year. Yes, it is underlying the phenomenon that is required to cover. | Yes, it is underlying snow fall for each year from 2008 to 2016 which measures each and every month. | … |
| ***Geographic Coverage*** | I want to cover all the New York city geographical area and the dataset all parts of it | It is collected average temperature of New York City; this is my specific area of interest | Yes, it covers all the areas of new York city | … |
| ***Demographic Coverage*** | {Does this dataset cover all demographics of interest for your model? Are there individuals of a specific age / race / ethnicity / ability that would not be included in this dataset? If so, what is missing? Why is it missing?}  All the demographics that I am considering are given in the dataset, I want to predict the overall response time of ambulance. | Yes, it is considering all the demographics that are needed,  Basically, it covers all the regions of the New York city. | Yes, It is considering all the demographics of the data in the respective years | … |
| ***Temporal Coverage*** | {Does this dataset cover all time periods of interest for your model? Are specific times, days, months, weeks, years missing from the data effort? If so, what is missing? Why is it missing?}  I am considering the model from 2008 to 2016, it covers all time periods of interest for my analyses. | It covers all the periods from 2008 to 2016 listed by each and every month. | Yes, it covers as required from 2008 to 2016 | … |
| ***Comprehensiveness*** | {Does the dataset capture all of the relevant features about your subject of interest that, you think, would be relevant for building your model? If not, what is missing and why is it missing? What limitations will those missing features have on the model you want to build?}  In this dataset there are few relevant features that I consider (listed in my initial plan) and there are few other conditions such as traffic situation, weather etc. So, for those features we need to consider other datasets. | I want to get more information regarding area wise, because New York city is a wide ranged city with many numbers of subareas such as Brooklyn, Manhattan, long island etc. I want the average temperature for each and every area, but I couldn’t find that listing over the internet, so I need to widen my research and I will try to include that ones I found out. | Same like previous average temperature snow fall is also specified for whole New York city, I want to get that for each subarea. | … |
| ***System Drift*** | This dataset as per my analyses, the dataset is perfect, and its ready for implementing a model on that. | IN this dataset I feel that more information is needed such as listing average temperature for each area etc. so like data need to be collected more appropriately based on the few other factors. | Other factors such as area,and there should be more information of this data that I needed for my analyses. |  |