Drug Recommendation System using WebMD

I. Introduction

Due to how the human brain processes information, as stated in the blog [1], it is easier for it to process charts or graphs, to visualize large amounts of complex data, rather than poring, pondering over spreadsheets or reports. This is a quick, easy way to convey fundamental concepts, relationships and trends in a universal manner – and you can change things drastically, albeit dramatically in different scenarios by making minor adjustments or changes.

Data visualizations makes it easier for the human brain to understand, and visualize, big, multi-dimensional data. As shown in [2], it significantly reduces the amount of work done by the brain and painstaking efforts expended by a person to detect patterns, trends, and outliers in groups of data.

There are many applications of data visualizations. There are a plethora of its sub-disciplines. We, in the course of this project, try to implement a drug recommendation system. This is an application, which is by far, one of the most important uses of technology in improving the quality of life of individuals and populations. This is because simply making a recommendation system may be something of little importance or commonplace, but making one by providing background information and displaying proofs of why a particular drug was chosen could make all the difference in influencing and accepting decisions.

II. Explanation of the Solution

The Dataset

We use the [3] WebMD Dataset for our project. This dataset has 362807 records of different drug uses. Our dataset has the following dimensions:

Column	Description
Age	Age of the user
Condition	Symptom experienced
Date	Date of drug use
Drug	The name of the drug
DrugId	A unique ID of a drug
EaseOfUse	The ease of use of drug
Effectiveness	How effective a drug was
Reviews	Reviews about a drug
Satisfaction	Satisfaction of the drug
Sex	Male/Female
Sides	Side effects of the drug.
UsefulCount	Usefulness of the review

	A	B	C	D	E	F	G	H	I	J	K	L
1	Age	Condition	Date	Drug	DrugId	EaseofUse	Effective	Reviews	Satis	Sex	Sides	Useful
2	75 or o	Stuffy Nos	####	25dpl	146724	5	5	I'm a reti	5	Male	Drowsi	0
3	25-34	Cold Symp	####	25dpl	146724	5	5	cleared r	5	Female	Drowsi	1
4	65-74	Other	####	warfa	144731	2	3	why did	3	Female		0
5	75 or o	Other	####	warfa	144731	2	2	FALLING	1	Female		0
6	35-44	Other	####	warfa	144731	1	1	My granc	1	Male		1
7	55-64	Other	####	warfa	144731	4	4	help hea	4	Male		0
8	25-34	Birth Cont	####	wyma	163180	5	5	Haven't g	2	Female	Nausea	0
9	45-54	Disease of	####	wyma	163180	5	5	I have ta	5	Female	Nausea	0
10	25-34	Acne	####	wyma	163180	4	2		2	Female	Nausea	1
11	55-64	Stuffy Nos	####	12 ho	9800	4	2	The 12 h	2	Male	Tempo	0
12	65-74	Other	####	pyrog	12112	5	5	Excellen	5	Male		0
13	19-24	Birth Cont	####	lyza	164750	5	5	Taking Ly	2		Nausea	0
14		Birth Cont	####	lyza	164750	2	1	This	1	Female	Nausea	0
15	19-24	Birth Cont	####	lyza	164750	2	3	I usually	1	Female	Nausea	0
16	35-44	Birth Cont	####	lyza	164750	5	5	I was cor	5		Nausea	0
17	25-34	Birth Cont	####	lyza	164750	2	2	The birth	1		Nausea	1
18	25-34	Birth Cont	####	lyza	164750	1	1	LYZA	1	Female	Nausea	1
19	25-34	Birth Cont	####	lyza	164750	4	4	IÃ¢??ve	1		Nausea	1
20	25-34	Birth Cont	####	lyza	164750	5	5	I have be	4	Female	Nausea	1
21	25-34	Birth Cont	####	lyza	164750	2	5	I have	1	Female	Nausea	3
22	35-44	Birth Cont	####	lyza	164750	5	5	I took thi	1		Nausea	3
23	25-34	Birth Cont	####	lyza	164750	5	5	My OB/G	4	Female	Nausea	10
24	19-24	Birth Cont	####	lyza	164750	5	5	I started	3		Nausea	5
25	25-34	Birth Cont	####	lyza	164750	5	5	I have be	5	Female	Nausea	5
26	19-24	Birth Cont	####	lyza	164750	5	5	I switche	3		Nausea	1

Figure 2.3 : Snapshot of the original dataset

Analysis and Cleaning of the Dataset

The dataset cleaning is done in a separate python program. We use the Pandas library to read the csv file. This is because there is a use of a lot of commas in the reviews and side effect columns of the dataset; so traditional libraries struggle with this format. After we do the following to clean the dataset, we save the file as a csv file; only this time it is separated by a '~' instead of a traditional ',':

1. Any blank cell is replaced by 'NULL'.
2. If there are any cells where the Condition, Drug or DrugID is 'NULL', these rows are replaced.
3. Any row with Condition - 'other' is dropped, to get a high quality dataset.
4. The Reviews column is dropped as it has a lot of non UTF format characters and also, the reviews are quite long.
5. The Sex column is dropped.
6. Since we decide not to use the dataset as a time-series dataset.The date column is dropped.
7. UsefulCount column is normalised and set to be between 0 and 1.
8. We add a column SentimentScore to the dataset. This has the values of scores of the sentiment analysis of the reviews.
9. Any Side effects with value 'NULL' is replaced by 'There are no known side effects'.
10. In the ages column, all the records are changed from 'a-b' to '[a-b]' and from 'a or over' to '[a+]'.

The idea is to use the satisfaction value, effectiveness value and much more from the dataset to get a value for the drug for a particular set of symptoms. This value would be a

direct indication of what drug will be recommended. All of the information such as the effectiveness, satisfaction etc is given to us well in advance in the dataset.

We perform sentiment analysis on the reviews column of the dataset in order to get new metric information about the drugs, , by using a library called Vader [4]. We believe that different users might rate different drugs according to different parameters but their reviews would tell a real, consistent story. This will be something that, if we quantify, could give us a true drug rating.

So, we decide to display the satisfaction value, effectiveness value and the reviews' sentiment score as some charts and graphs for the users to see for themselves why a particular drug is in use. We believe that this will build confidence in the user about our recommendations.

III. Implementation

We decide to make a simple search box where a user will select various symptoms. Figure 3.1 displays this search box.

Figure 3.1

We use the dataset to show a drop down menu containing all the symptoms that a user can select. We make a sophisticated auto complete system that will provide suggestions to the users as he/ they type.

A user may select upto 8 symptoms. He has to limit this number because:

1. There is a limitation of space in the web app.
2. The major reason is that we believe that in a real life scenario there will not be a lot of symptoms simultaneously. And if they are there, a single drug may not be able to cure them all.

The symptoms stack next to each other on the top as and when the user adds them. Figure 3.2 shows a snapshot of this.

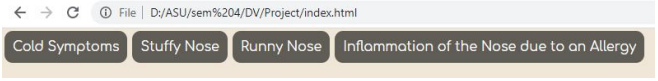

Figure 3.2

It is to be noted that we allow a user to select only symptoms that have at least one curable-drug in common. Otherwise, obviously no single drug will be able to cure all the symptoms.

As and when the symptoms are added, the drug panel is populated which is on the left side of the page. (Figure 3.3) The drug panel (Figure 3.3) , which is on the left side of the page is populated as and when symptoms are added.

Simultaneous to symptom addition and drug panel population, we draw a stacked bar chart in the centre of the page which shows the overall satisfaction of a drug for a particular condition using the d3js library.

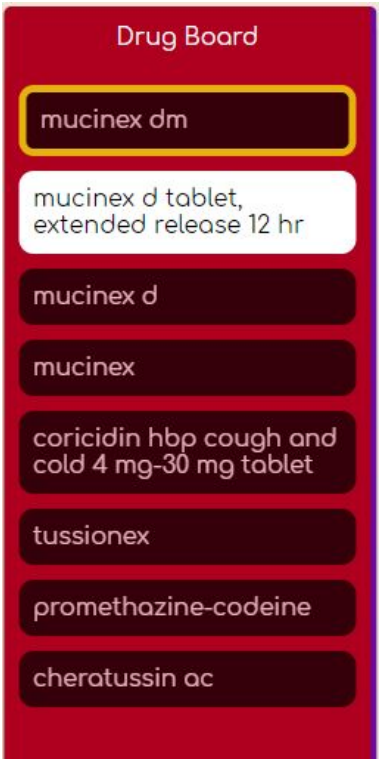

Figure 3.3

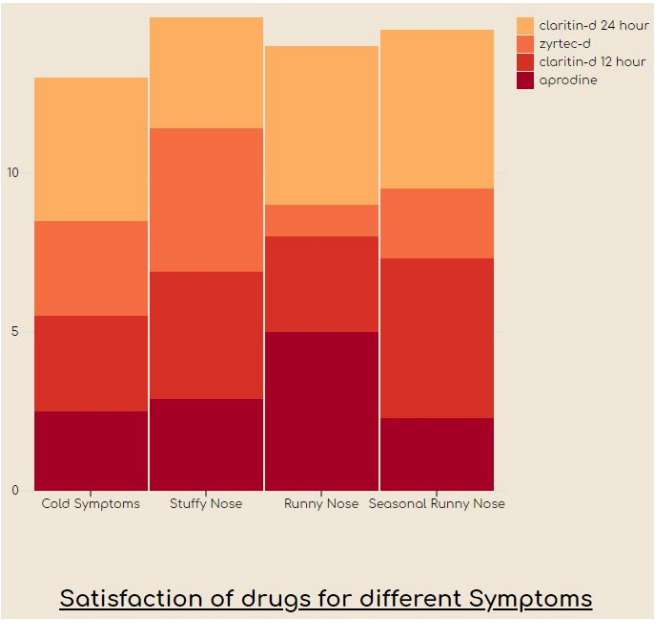

Figure 3.4

We choose to use a stacked bar chart for this because we can easily show 3 dimensions of data with easy potentially comparative studies. A user can simply hover over any area to view the value of satisfaction in a tooltip.

There is yet another panel on the right of the page which shows further information about the drug for a particular condition, when a user hovers over a particular area of the stacked bar graph. Figure 3.5 shows this.

We use a simple bar chart: the one in blue in Figure 3.5 to show the effectiveness of a drug for a particular condition for various age groups. We believe in showing as much

information as possible but at the same time we believe that we can't put a price tag on the advantages of simplicity. This was the motivation behind using a simple bar chart.

Figure 3.5

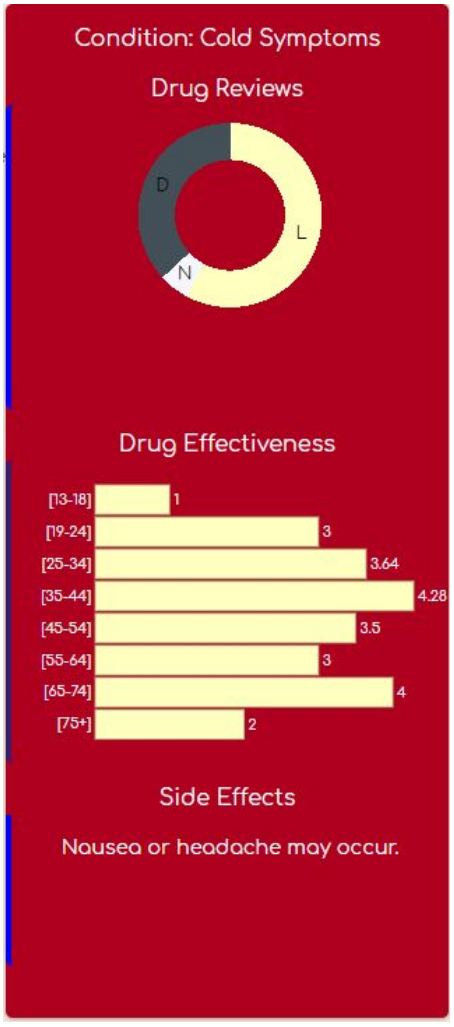

Figure 3.6

We also see a donut chart in Figure 3.6. This shows the number of positive, negative and neutral reviews that different drug users give to this drug for particular symptoms. In Figure 3.5 we see that if a user hovers over any area of donut chart corresponding to a sentiment, he will see the number of reviews for that particular sentiment.

We use a plain off-white background for the web app, with red panels to draw attention to them. The colors of the stack in the stacked bar graph gradient from red to orange to yellow to blue as the number of drugs increase. We use a background contrasting yellow color on the right panel for the donut chart and the bar chart depicting the sentiment analysis of reviews and effectiveness respectively.

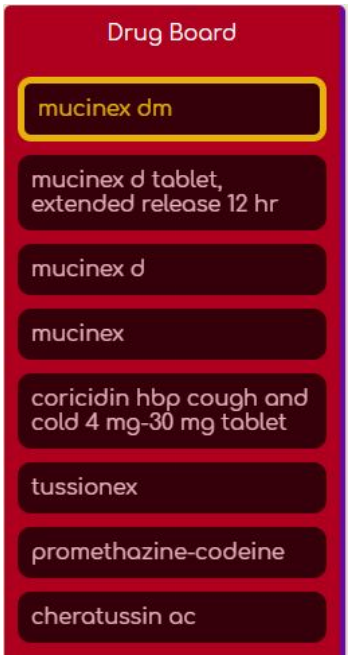

Figure 3.7

At last, we use the sentiment score, effectiveness, satisfaction score, and the usefulCount score to get which particular drug is best fit for the given symptoms. As shown in Figure 3.7, this drug is simply highlighted in the drugs panel. We have also placed the labels of the drugs of the same brand together in the legend, according to Gestalt's Principle of Proximity [5],

We try to use percentage values almost everywhere in the code. This ensures that the web app is equally responsive for different screen sizes.

IV. RESULTS

We were able to successfully:

1. Cleanse the data.
2. Import the data in json format into the script of the webpage.
3. Display the information as desired.
4. Use d3js to make several svg graphic components.
5. Do sentiment analysis of the reviews.
6. Get a pretty good statistical measure of the drugs, for comparison.
7. Make the recommendation system.

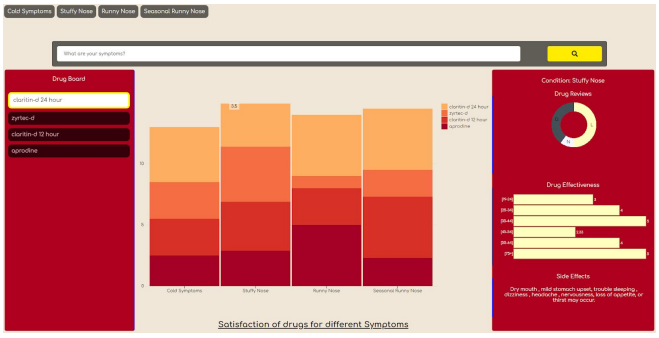

Figure 4.1 shows a picture of the entire web application.

V. Lessons Learned

We learned the impact a visualization plays over raw and even structured data but which is without visualization. Just informing the users what drug to use is a thing of the past now. We need to convince them that what the system is recommending is, rather, quite accurate and precise. For this we display various charts and graphs.

We learned the use of sentiment analysis tools to generate more and in some cases better data which is more suitable for our scenario or application..

Everyone has a different computer, a different mobile device, a different tablet etc. We can not expect the screen size of their device where they plan to view the recommendation system to be the same. Thus, one of the major lessons that we learned was of using proportions for values of width and height and at various other places instead of using fixed ones in the entire project. This helped us make the webpage responsive to different screen sizes.

One of the things that I learned from this project is that it is possible to overwhelm yourself with something, especially when dealing with data of this size and so to keep or develop an exit strategy. Thus it is very important to know how to tackle such a situation, in the course of one's work.

VI. Team Members

The following are the names of the people who made this project possible:

1. Aabhaas Gupta
2. Tanishka Singh
3. Sarthak Shetty
4. Sumitava Ghosh

VII. Personal Contributions

My contributions are as follows:

1. I was a major part of the initial planning process and deciding on how to carry it out.
2. I tried to help the group with whatever I could when we were stuck at some place.
3. I was responsible for implementing most of the code on the web page itself.
4. I helped partly with the documentation.

VIII. References

[1] https://www.sas.com/en_us/insights/big-data/data-visualization.html

[2] https://www.searchenginejournal.com/what-is-data-visualization-why-important-seo/288127/

[3] Rohan Harode. 2020. WebMD Drug Reviews Dataset. (Mar 2020). https://www.kaggle.com/rohanharode07/webmd-drug-reviews-dataset/metadata

[4] Rossen LM, Bastian B, Warner M, Khan D, and Chong Y. 2019. NCHS Data Visualization Gallery - Drug Poisoning Mortality. (Apr 2019).https://www.cdc.gov/nchs/data-visualization/drug-poisoning-mortality/index.htm

[5] Johan Wagemans, James H. Elder, Michael Kubovy, Stephen E. Palmer, Mary A. Peterson, Manish Singh, and Rüdiger von der Heydt. 2012. A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure-ground organization. Psychological bulletin 138, 6 (Nov 2012), 1172–1217. DOI: http://dx.doi.org/10.1037/a0029333 22845751[pmid].