DATA603 Project Update

Yun-Zih Chen, Pooja Kangokar Pranesh, Elizabeth Cardosa

Abstract

The goal of this project is leverage big data technologies to train a model using the UCI ML Drug Review dataset to predict the star rating of drug based on the sentiment of the review. This model will then perform inference in a streaming manner on ‘real-time’ reviews coming in. This data can then be used to help potential customers understand the overall sentiment towards a drug and if it might be useful for them.

Dataset:
https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29

Introduction

The Drug Review dataset contains over 200,000 patient reviews on specific drugs along with related conditions and a 10 star patient rating reflecting overall patient satisfaction, and is made available through the UCI ML Repository. This dataset is primarily used for research purposes and can be used to answer a variety of questions about patients conditions and the relationship between their reviews and rating of a drug.

	uniqueID	drugName	condition	review	rating	date	usefulCount
0	206461	Valsartan	Left Ventricular Dysfunction	"It has no side effect, I take it in combinati...	9	20-May-12	27
1	95260	Guanfacine	ADHD	"My son is halfway through his fourth week of ...	8	27-Apr-10	192
2	92703	Lybrel	Birth Control	"I used to take another oral contraceptive, wh...	5	14-Dec-09	17
3	138000	Ortho Evra	Birth Control	"This is my first time using any form of birth...	8	3-Nov-15	10
4	35696	Buprenorphine / naloxone	Opiate Dependence	"Suboxone has completely turned my life around...	9	27-Nov-16	37

Big Data business Problem

Potential Problems

- Seeking out reliable and effective medicines is a nerve-racking experience for people needing drug treatment.
- Most of the time consumers learn about drug treatments, their effectiveness, and side effects from physicians or ads.
- If it is possible to learn from the people who have taken specific drug treatments, consumers can obtain experiences and genuine feedback about these drugs.

Potential Solutions

- To investigate reviews about drugs to predict the star rating of the drug
- To know what elements of a review make it more helpful to readers
- Whether a review is positive, neutral, or negative

Big Data business Problem

Potential Outcomes

- To develop the drug star rating, consumers could quickly spotlight excellence in drugs as a reference.
- Consumers may bring the information and discuss it with their primary care providers.

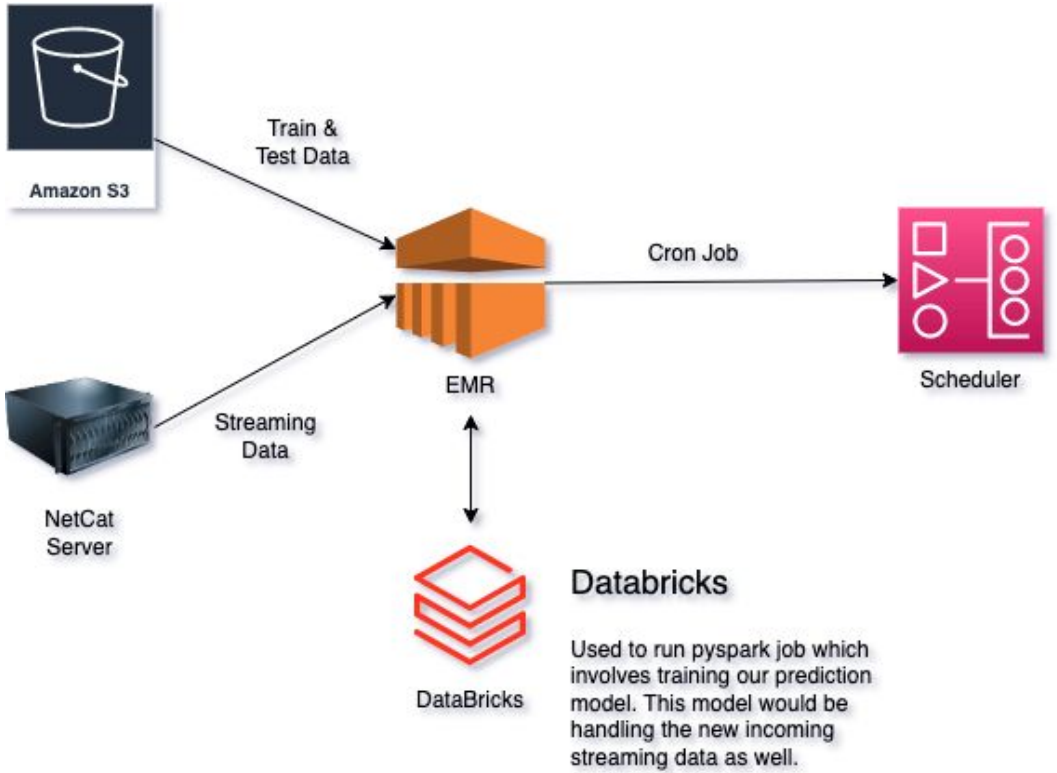
Potential Challenges

- Despite having drug star ratings, consumers still need to consult professionals about if they can take this type of drug due to drug interaction.

Technology Stack

1. S3 Bucket
2. EMR Cluster
3. Netcat Server
4. Pyspark
5. Cron job (Scheduling)

Proposed Solution - Architecture



- S3 bucket is used to store the data
- EMR cluster with 2 datanodes runs the spark job which predicts the result of streaming data that comes from netcat server.
- The process is scheduled to run every 30 minutes using Cron job.

Exploratory Data Analysis - Schema

Spark read:

```
[ ] # Read in training data file
    customschema = StructType([
      StructField("UniqueID", IntegerType(), True)
      ,StructField("drugName", StringType(), True)
      ,StructField("condition", StringType(), True)
      ,StructField("review", StringType(), True)
      ,StructField("rating", DoubleType(), True)
      ,StructField("date", StringType(), True)
      ,StructField("usefulCount", IntegerType(), True)
      ])

    df = spark.read.format("csv")\
              .option("delimiter", "\t")\
              .option("header", "true")\
              .option("quote", "\"")\
              .option("escape", "\"")\
              .option("multiLine","true")\
              .option("quoteMode","ALL")\
              .option("mode","PERMISSIVE")\
              .option("ignoreLeadingWhiteSpace","true")\
              .option("ignoreTrailingWhiteSpace","true")\
              .option("parserLib","UNIVOCITY")\
              .schema(customschema)\
              .load(working_folder + "Data/drugsComTrain_raw.tsv")
```

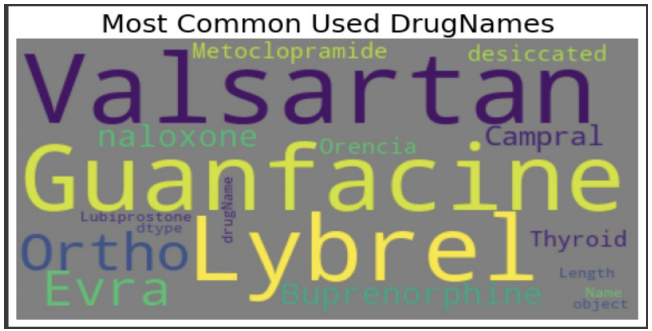
EDA - Continued

Top 5 Rows:

UniqueID	drugName	condition	review	rating	date	usefulCount
206461.0	Valsartan	Left Ventricular ...	"It has no side e...	9.0	May 20, 2012	27
95260.0	Guanfacine	ADHD	"My son is halfwa...	8.0	April 27, 2010	192
92703.0	Lybrel	Birth Control	"I used to take a...	5.0	December 14, 2009	17
138000.0	Ortho Evra	Birth Control	"This is my first...	8.0	November 3, 2015	10
35696.0	Buprenorphine / n...	Opiate Dependence	"Suboxone has com...	9.0	November 27, 2016	37

only showing top 5 rows

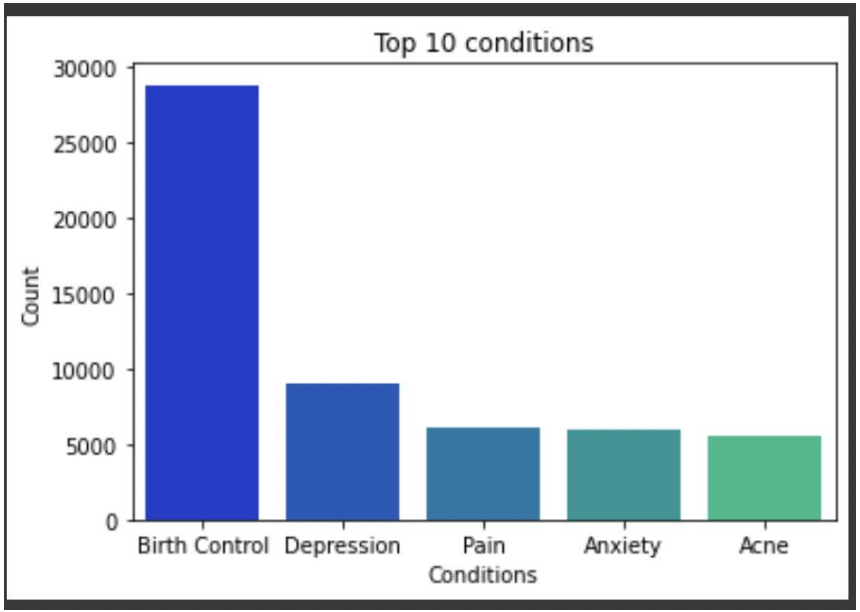
Common Drug Names:



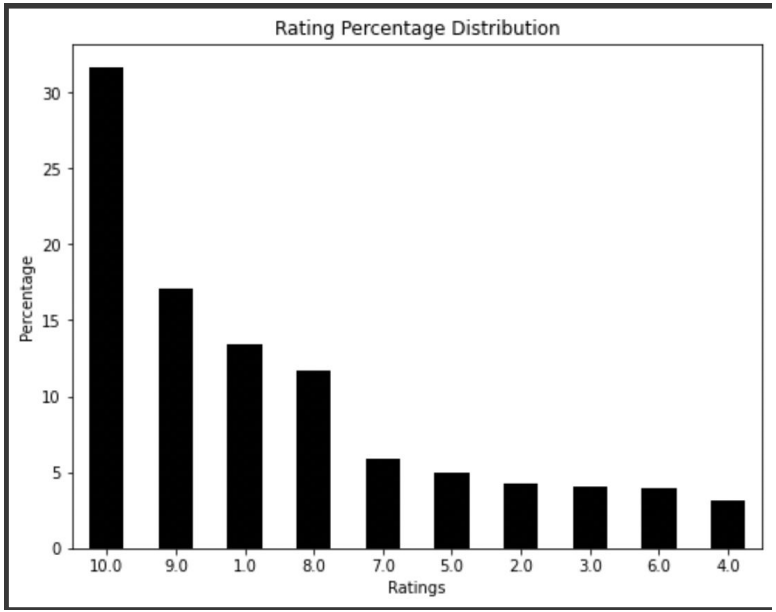
Common words used in reviews

EDA - Continued

Top 5 Conditions in our dataset

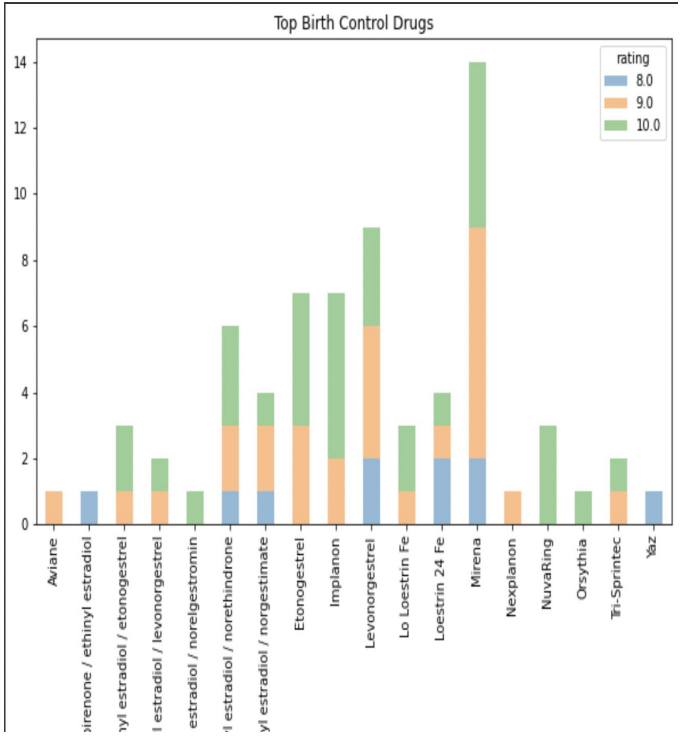


Rating Percentage Distribution

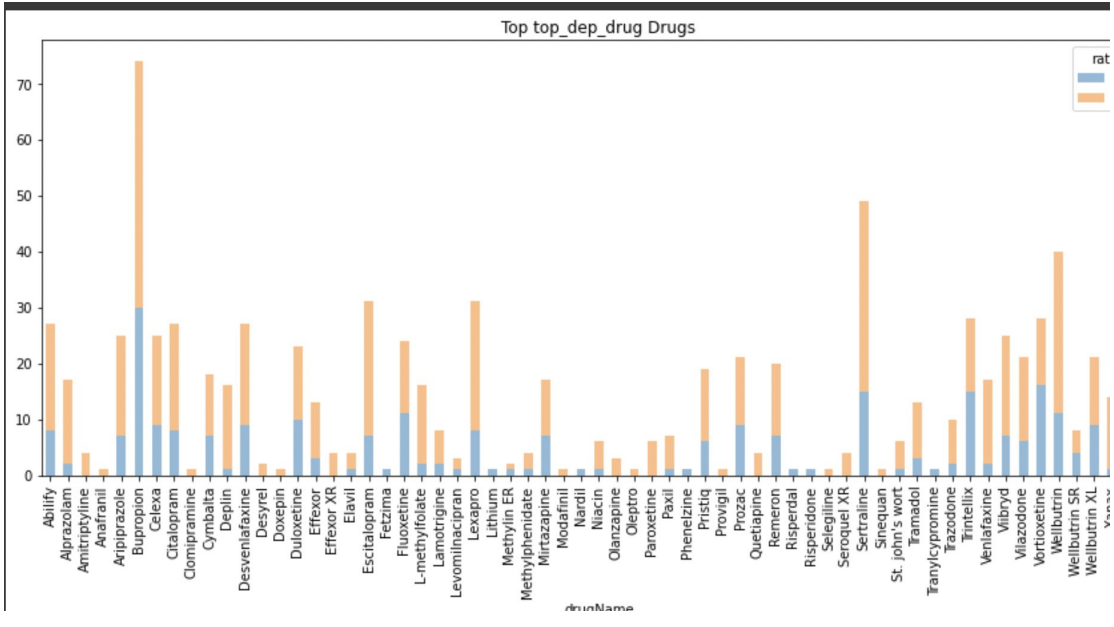


EDA - Continued

Top birth control drugs based on ratings



Top depression drugs based on ratings



Sentiment Analysis on Review Column

```
[ ] # https://medium.com/analytics-vidhya/sentiment-analysis-with-sparknlp-couldnt-be-easier-2a8ea3b728a0
    # https://nlp.johnsnowlabs.com/
```

```
[ ] # https://colab.research.google.com/github/JohnSnowLabs/spark-nlp-workshop/blob/master/jupyter/quick_start_google_colab.ipynb#scrollTo=tyMMD
```

```
[ ] pipeline = PretrainedPipeline('analyze_sentimentdl_use_twitter', 'en')
```

```
    analyze_sentimentdl_use_twitter download started this may take some time.
    Approx size to download 935.1 MB
    [OK!]
```

```
[ ] # TODO: CREATE MORE ROBUST PREPROCESSING PIPELINE FOR THE REVIEW COLUMN
```

```
    result = pipeline.fullAnnotate(["im meeting up with one of my besties tonight! Cant wait!!  - GIRL TALK!!", "is upset that he can't update h
```

```
[ ] result[0]['sentiment']
```

```
    [Annotation(category, 0, 71, positive, {'sentence': '0', 'positive': '1.0', 'negative': '0.0'})]
```

```
[ ] # rename the text column as 'text', pipeline expects 'text'
    df_result = pipeline.transform(df.withColumnRenamed("review", "text"))
```

Sentiment Analysis Results

```
[ ]  df_result.show()
```

condition	text	rating	date	usefulCount	document	sentence_embeddings	sentiment
ft Ventricular ...	"It has no side e...	9.0	May 20, 2012	27	[{document, 0, 78...	[{sentence_embedd...	[{category, 0, 78...
ADHD	"My son is halfwa...	8.0	April 27, 2010	192	[{document, 0, 73...	[{sentence_embedd...	[{category, 0, 73...
Birth Control	"I used to take a...	5.0	December 14, 2009	17	[{document, 0, 75...	[{sentence_embedd...	[{category, 0, 75...
Birth Control	"This is my first...	8.0	November 3, 2015	10	[{document, 0, 44...	[{sentence_embedd...	[{category, 0, 44...
Opiate Dependence	"Suboxone has com...	9.0	November 27, 2016	37	[{document, 0, 71...	[{sentence_embedd...	[{category, 0, 71...
nign Prostatic ...	"2nd day on 5mg s...	2.0	November 28, 2015	43	[{document, 0, 40...	[{sentence_embedd...	[{category, 0, 40...
ergency Contrac...	"He pulled out, b...	1.0	March 7, 2017	5	[{document, 0, 14...	[{sentence_embedd...	[{category, 0, 14...
Bipolar Disorde	"Abilify changed ...	10.0	March 14, 2015	32	[{document, 0, 73...	[{sentence_embedd...	[{category, 0, 73...
Epilepsy	" I Ve had nothi...	1.0	August 9, 2016	11	[{document, 0, 19...	[{sentence_embedd...	[{category, 0, 19...
Birth Control	"I had been on th...	8.0	December 8, 2016	1	[{document, 0, 73...	[{sentence_embedd...	[{category, 0, 73...
igraine Prevention	"I have been on t...	9.0	January 1, 2015	19	[{document, 0, 72...	[{sentence_embedd...	[{category, 0, 72...
Depression	"I have taken ant...	10.0	March 9, 2017	54	[{document, 0, 45...	[{sentence_embedd...	[{category, 0, 45...
Crohn's Disease	"I had Crohn&#039...	4.0	July 6, 2013	8	[{document, 0, 40...	[{sentence_embedd...	[{category, 0, 40...
Cough	"Have a little bi...	4.0	September 7, 2017	1	[{document, 0, 59...	[{sentence_embedd...	[{category, 0, 59...
Birth Control	"Started Nexplano...	3.0	August 7, 2014	10	[{document, 0, 78...	[{sentence_embedd...	[{category, 0, 78...
Obesity	"I have been taki...	9.0	January 19, 2017	20	[{document, 0, 73...	[{sentence_embedd...	[{category, 0, 73...
inary Tract Inf...	"This drug worked...	9.0	September 22, 2017	0	[{document, 0, 67...	[{sentence_embedd...	[{category, 0, 67...
ibromyalgia	"I've been t...	9.0	March 15, 2017	39	[{document, 0, 71...	[{sentence_embedd...	[{category, 0, 71...
Bipolar Disorde	"I've been o...	10.0	November 9, 2014	18	[{document, 0, 76...	[{sentence_embedd...	[{category, 0, 76...
ronic Myelogeno...	"I have been on T...	10.0	September 1, 2015	11	[{document, 0, 47...	[{sentence_embedd...	[{category, 0, 47...

Proposed Solution - Analysis & Model Training

- PySpark on Google Colab to perform exploratory analysis on the data (Completed)
- John Snow Labs SparkNLP Models to get sentiment intensity and subjectivity scores for the text reviews (Completed)
  - Determine which model makes the most sense to extract sentiments for our review's (Upcoming)
  - Extract sentiment scores out of returned outputs (Upcoming)
- Spark's MLlib(Spark NLP) to train our model (Upcoming)
- Literature review, conclusions, and impact (Upcoming)
- Stand up infrastructure for proposed deployment (Upcoming)