DATA 603 Final Report

Yun-Zih Chen, Pooja Kangokar Pranesh, Elizabeth Cardosa

Problem Statement

- Seeking out reliable and effective medicines is a nerve-racking experience for people needing drug treatment.
- The most time-consuming activities are learning about drug treatments, their effectiveness, and side effects from physicians or ads.
- It becomes more challenging for users to review all textual comments.
- An efficient structured algorithm is needed to explore the textual reviews and generate a new drug rating for consumers before making purchase decisions.

Objective

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.

1. Predict the rating of drug based on the sentiment analysis of the review
2. Examine this model in a streaming data of ‘real-time’ reviews
3. Help potential customers understand the overall sentiment towards a drug

Literature Review

- Garg (2021) developed a drug recommendation system that predicted the sentiment using patient reviews and tested different vectorization method and classification algorithms.
- Vijayaraghavan and Basu (2020) built a model using ANN algorithm on Count Vectorizer method to predict the sentiment of the drug review for three conditions such as birth control, depression, and pain.
- Shiju and He (2021) compared the traditional machine learning with transformed-based neural network models to classify drug ratings based on textual data.
- Colón-Ruiz and Segura-Bedmar (2020) compared various deep learning tools to use sentiment analysis on drug reviews and examined various models of word embedding.
- Gujjar and Kumar (2021) demonstrated a method that understood customers' opinion using TextBlob API in python and proposed this technique to help decision making of product and service benchmarking.
- Bose et al. (2020) investigated customary text on twitter about the application of drugs for COVID- 19 treatment using sentiments analysis (TextBlob) across 8 countries.

Dataset

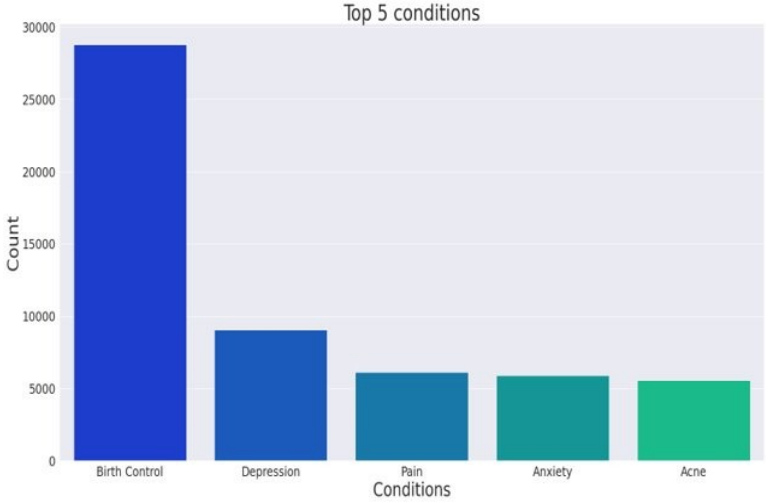
The Drug Review dataset contains over 200,000 patient reviews on specific drugs along with related conditions and a 10 star patient rating reflecting overall patient satisfaction, and is made available through the UCI ML Repository. This dataset is primarily used for research purposes and can be used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

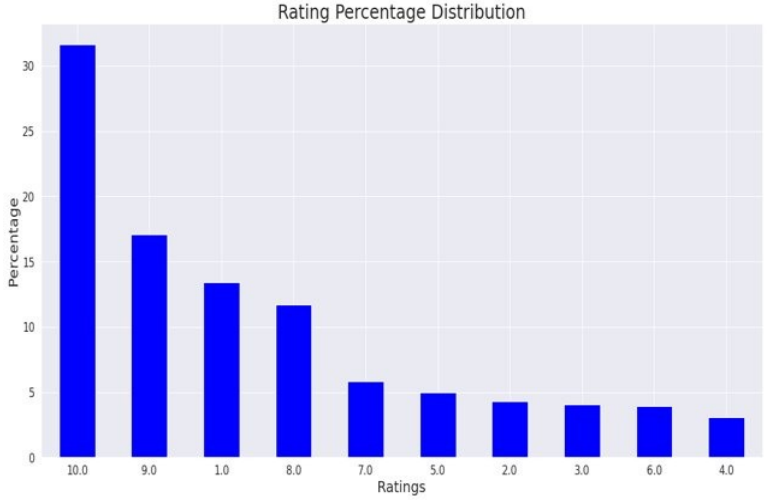
	uniqueID	drugName	condition	review	rating	date	usefulCount
0	206461	Valsartan	Left Ventricular Dysfunction	"It has no side effect, I take it in combinati...	9	20-May-12	27
1	95260	Guanfacine	ADHD	"My son is halfway through his fourth week of ...	8	27-Apr-10	192
2	92703	Lybrel	Birth Control	"I used to 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

Exploratory Data Analysis

Most Common Used DrugNames

WordCloud of Reviews





Big Data Machine Learning Pipeline

Sentiment Extraction

Initially set out to use SparkNLP by John Snow Labs, but faced a handful of challenges:

- Their “Drug Reviews Classifier”, based off the BioBERT Transformer, was not within the free-tier
- Instead we tried their “Sentiment Analysis of Tweets” pipeline, but it only classifies sentiments into “Positive”/”Negative” and the inference proved to be too slow on the free-tier GCP colaboratory instance.

Instead, we decided to use TextBlob’s sentiment polarity extraction which returns a continuous polarity score between [-1,1] and only took about 30 seconds to run when applied to the training data.

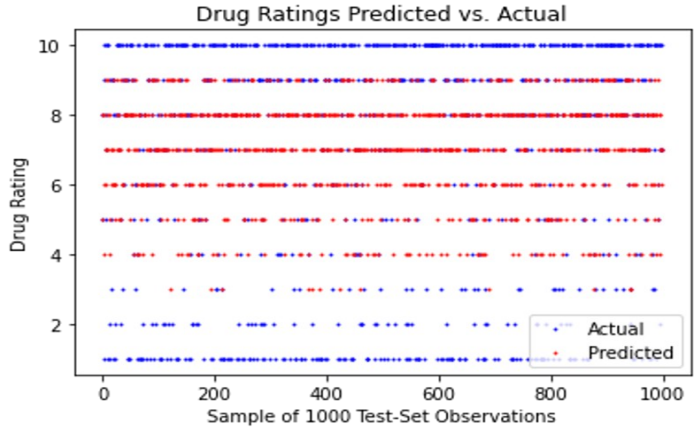
The distribution of the sentiment scores extracted from the training data’s reviews were largely neutral and mostly around zero, but there were sentiments at either end of the polarity range.

```
pd_df_train.sentiment.describe()
```

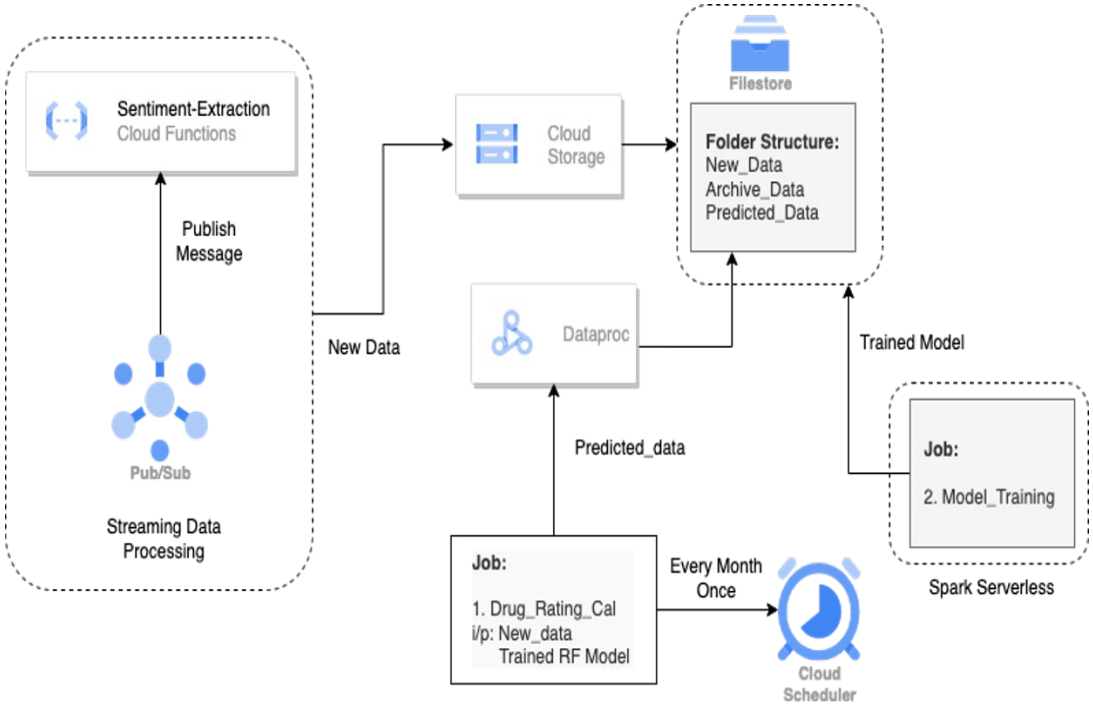
count	159498.000000
mean	0.064715
std	0.222336
min	-1.000000
25%	-0.043333
50%	0.057975
75%	0.173611
max	1.000000

Final Model

We trained our final model using MLlib's Random Forest Regressor with the target variable being the rating from 1-10, and the features being the useful count and sentiment polarity score extracted from the reviews. Our best model had a RMSE of 2.94719. The image below shows the predicted drug ratings overlaid on the actual ones from the test set. This shows that our model does not predict ratings will be a perfect 10 or between 0-3. It also is skewed towards predicting higher rating, potentially due to a right-tail skew in the training data.



Architecture Diagram & Technology Stack



GCP services:

- Pub/Sub
- Cloud Functions
- Cloud Storage
  - File structure: new_data, archive_data, predicted_data
- DataProc
  - Jobs - Drug_Rating_Calculation_RF.py
- Spark Serverless
  - Model_Training.py

Solution

Step 1: Creating Pub/Sub Topic on GCP

- *gcloud pubsub topics create DRUG-REVIEW-TOPIC --schema=drug-reviews --project=Drug-Analysis*

Step 2: Create the schematic format of incoming data.

- *gcloud pubsub schemas create DRUG-REVIEWS --type=AVRO --definition=SCHEMA_DEFINITION (json file)*

  Ex: {"UniqueID":1235, "drugName": "Dolo", "condition": "Fever", "review": "Worst didn't work", "rating": 0, "date": "10-30-2022", "usefulCount": 10}

Step 3: Create a Cloud Function with the above Pub/Sub Topic as the Trigger.

- *Use cloud_function.py in our code folder structure*

Step 4: Spark serverless batch creation for Model training

- *Create batch using model_training.py file*

Step 5: Create a DataProc Cluster to schedule the Rating_Prediction Job.

- # Following gcloud command cerates dataproc cluster with 2 worker nodes of 200GB disk size and 1 master node with boot disk of size 100GB.
  gcloud dataproc clusters create spark-drug-analysis-dataproc --enable-component-gateway --region us-central1 --subnet default --zone us-central1-b --master-machine-type n1-standard-4 --master-boot-disk-size 100 --num-workers 2 --worker-machine-type n2d-standard-4 --worker-boot-disk-size 200 --image-version 2.0-rocky8 --optional-components JUPYTER,ZOOKEEPER --scopes 'https://www.googleapis.com/auth/cloud-platform' --project newagent-bba27
- # Schedule the job using crontab
  *Crontab -e {Contents: 30 6 30 * * shell_invoke.sh >> /var/logs/cron.log 2>&1}*

Code Run

- Sentiment_Extraction Cloud Function is triggered automatically and dumps the new data with sentiment details to cloud storage.
- A spark serverless batch is created to run the Model Training and to save the trained model to cloud storage.
- Using the pre-trained model DataProc Schedules a Cron Job with shell script that runs the Rating_Exatration.py file on new streamed data.

Future Work

Model Perspective:

- Training the model on more number of specific reviews for each drug.
- Testing the models like Drug Reviews Classifier by John Snow Labs for more accurate analysis.

Architecture Perspective:

- Incorporating life-cycle policies on cloud storage bucket to maintain standard, near-line and cold-line. This will help in maintaining minimal cost.
- To orchestrate the jobs it’s ideal to use cloud composer (Airflow) instead of crontab.

Impact

- Standardized drug reviews rating can help users identify positive or negative reviews without going through all comments
- Consumers could quickly spotlight excellence in drugs as a reference
- Consumers may use the information to discuss with their primary care providers
- A drug review will often result in the identification of actual or potential medication-related problems and recommendations to optimise drug use.
- Helps in evaluating efficiency of a drug on diverse set of people as every region has its own food habits and living style.
- Drug business owners can use these ratings and reviews detail to improve their performance and raise profit.

References

•Bose, R., Aithal, P.S., & Roy, S. (2020). Sentiment Analysis on the Basis of Tweeter Comments of Application of Drugs by Customary Language Toolkit and TextBlob Opinions of Distinct Countries. *International Journal of Emerging Trends in Engineering Research*, 8(7), 3684-3696. https://doi.org/10.30534/ijeter/2020/129872020

•Colón-Ruiz, C., & Segura-Bedmar, I. (2020). Comparing deep learning architectures for sentiment analysis on drug reviews. *Journal of Biomedical Informatics*, *110*, 1-11. https://doi.org/10.1016/j.jbi.2020.103539

•*Drug Review Dataset (Drugs.com) Data Set.* UCI ML Repository. (n.d.). Retrieved September 30, 2022, from https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29

•Fan and Fuel. (n.d.). *No online customer reviews means BIG problems in 2017.* Retrieved November 29, 2022, from https://fanandfuel.com/no-online-customer-reviews-means-big-problems-2017/

•Garg, S. (2021). Drug Recommendation System based on Sentiment Analysis of Drug Reviews using Machine Learning. *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 175–181. https://doi.org/10.1109/Confluence51648.2021.9377188

•Gujjar, J.P., & Kumar, H. P. (2021) Sentiment Analysis: Textblob For Decision Making. *International Journal of Scientific Research & Engineering Trends*, 7(2), 1097-1099.

•*NLP Models Hub.* John Snow LABS.(n.d.). Retrieved October 13, 2022, from https://nlp.johnsnowlabs.com/models?q=explain_document_dl

•Shiju, A. & He, Z. (2022). Classifying Drug Rating Using User Reviews with Transformer-Based Language Models. *2022 IEEE 10th International Conference on Healthcare Informatics,* 163-169. DOI 10.1109/ICHI54592.2022.00035

•*TextBlob: Simplified Text Processing.* TextBlob. (n.d.). Retrieved October 12, 2022, from https://textblob.readthedocs.io/en/dev/

•Vijayaraghavan, S., & Basu, D. (2020). *Sentiment Analysis in Drug Reviews using Supervised Machine Learning Algorithms.* http://arxiv.org/abs/2003.11643

Project Code and Dataset

Github Link: https://github.com/cardosa1/UMBC-DATA603-Group3-Project

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