# PharmAnalysis

BAX 422 - Data, Design and Representation

Winter 2023

UC Davis School of Management


Submitted By:

Arpita Mangal

Rahul Rajput

Rishika Chaudhary

# Table of Contents

# Executive Summary

With small pharmacies facing increased competition from multinational pharmacy chains and online pharmacy platforms, to remain competitive they need to be highly organized and use resources more efficiently. To support this we have developed several Proof-of-Concept algorithms which would allow small pharmacies to improve inventory management and develop competent pricing strategies. These algorithms have shown promising results and can be significantly improved upon with more resources. We have provided recommendations for these models to be improved upon and be deployment ready.

# Background

In today's medical world there are thousands of medicines, healthcare, and wellness products available over-the-counter or via prescription to customers. In large parts of the world the role of dispensing these medicines lies with small pharmacies. However, organizing the sheer quantity of medicines is a labor-intensive and time-consuming task for these small pharmacies. A poor organization system leads to misplaced inventory and puts the small stores at a disadvantage against larger chain pharmacies or e-pharmacy portals with more resources.

In light of this scenario we aim to:

1. Build an algorithm which can classify medicines and other wellness products into categories by their manufacturer provided descriptions and details. We then aim to classify these medicines and healthcare products according to the list of 40 general categories of medicines as listed by the FDA.
2. Build a pricing model which would allow small pharmacies to accurately predict market prices of medicines and decide on a competitive selling point.
3. Build an algorithm which can provide substitute recommendations for medicines based on their description and features, to enable better inventory organization.

# Data

Data for this exercise was scraped from 1mg.com, India's largest online retailer of medicine and healthcare products. This enabled us to work with a more representative dataset of products available. Another advantage of scraping from 1mg was the availability of medicine information provided by the manufacturer in one place.Due to limited time and compute power, a sample set of 500 medicines and healthcare products were scraped from each alphabet. In total we built a dataset of 12,900 products. Information scraped for each medicine includes:

1. Name
2. Price
3. Manufacturer
4. Active Ingredients in the product
5. Dosage Form
6. Product Description
7. Mechanism of Action
8. Uses
9. Benefits
10. Side Effects
11. Other medicine facts
12. Substitutes

We also worked with a list of the 40 general categories of medicines and their descriptions as provided by the Food and Drug Administration, USA.

# Analyses

The analysis was conducted on Jupyter Notebooks and Google Colaboratory environment. The algorithms were built using Python via multiple libraries available for use including, but not limited to, Pandas, Numpy, SkLearn, and Gensim. For our analyses we did not have access to labeled 'True' data which could be used to guide out algorithms, therefore we could only use Unsupervised Learning methods with results being analyzed qualitatively. However, the exercise provides a POC and a platform for future models to be built upon and explores multiple methods of classification, price prediction, and substitution.

## Data Preprocessing and Exploration

1. We started by loading the CSV datasets for Medicines and Drug Categories and checking for duplicates, removing if any. The price column within the Medicines dataset was converted to numeric form, and several other columns modified to remove unwanted symbols. The column 'Therapeutic Class' was extracted from the 'Factbox' column while the text from columns 'Product Info', 'Benefits', and 'Description' were combined to a new column 'Combined_desc' providing a higher density of information within one embedding. Subsequently, the columns 'Factbox', 'Manufacturer', 'Description', ''Product Info', 'Side Effects', and 'Benefits' were dropped from the medicines dataset as these were not expected to be used further.
2. Next, we explored the 'Uses' and 'Therapeutic Class' columns to gain a high level understanding of how the medicines are currently distributed. The 'Uses' columns is a compilation of the categories mentioned by manufacturers and there are over 360 unique values. The top 10 uses, accounting for 40% of medicines listed, are related to treatment of bacterial infections, diabetics, pain, blood pressure, mental health, and acid reflux. Similarly, we explored the distribution of medicines by 'therapeutic class' and noticed that the top 5 out of the 22 total relate to Infection, Respiratory, GI, Pain, and Mental Health. We then explored the average prices of medicines per 'Uses' category and observed that the more common categories of medicines are cheaper.

## Aim 1: Categorisation of Medicines

To generate word embeddings for our analyses, we used the BioWordVec model which was trained on the PubMed text corpus and provides more accurate outputs for Biomedical data. The embeddings created were of dimensions 200 each. Each embedding was preprocessed using the simple_preprocess() function from genesis, removing stop words and stemming words.

1. Method 1: Categorisation by comparing Medicine and Drug Category Descriptions.

For the first method we generate word embeddings for Drug Category Descriptions offered by the FDA by averaging the vectors for all important words present in the description. Embeddings for the 'Combined_desc' variable for each medicine were generated similarly. This method simply identified the closest Drug Category embedding to the medicine embedding, by creating a cosine similarity matrix between the Drug Category embeddings and Combined_desc embeddings, and assigning the medicine to that particular category. The Cosine distance measures the angle between two vectors and is a good metric to identify similar contexts between documents. A higher cosine distance value implies a smaller angle between two vectors, and therefore a more similar contextual meaning.

2. Method 2: Categorisation through K-Means Clustering and Drug Category Embeddings.

For the second method we decided to use more features - 'Active Ingredients', 'Therapeutic Class', along with Combined Description, for each medicine to generate the embeddings. Instead of averaging the embeddings we decided to concatenate them and create a single vector of dimension (600,1). This would allow more weightage for the different features and allow for better clustering rather than averaging them out. We then performed K-Means Clustering with K=40, mirroring the number of Drug Categories. The medicines were then grouped by their respective cluster assignments and for each cluster we calculated the average embeddings of corresponding 'Uses' values present in it. A cosine distance matrix was created for these embeddings and Drug Category medicines, with each cluster being assigned to the

closest Drug Category. Involving the ‘Uses’ column allowed us to incorporate more information about medicines.

3. Method 3: Categorisation through Topic Modeling and Drug Category Embeddings.

The final method for classification involved extracting the key information present in the combined description of each medicine through Topic Modeling and then calculating distance between the average embeddings for each medicine’s key words and the Drug Category embeddings. Topic Modeling is a method used on unlabelled data to identify common themes within and between documents and then cluster them. It assumes that each document references multiple topics defined by a certain probability distribution and that each word belongs to a particular theme with a particular probability, given the corpus of data. Or, in other words, it assumes that each document is composed of multiple topics and each topic is composed of multiple words. We used Latent Dirichlet Analysis, which assumes that topics in each document and words in each topic are distributed via a Dirichlet distribution. The algorithm creates a vocabulary via the ‘Bag-of-Words’ model and then identifies hidden structures in the data. For our case, we performed LDA over each medicine’s description and extracted the topic which had the highest probability of representing the document and subsequently generated embeddings for the top 20 words for each topic. And similar to the previous methods, categories were assigned to each medicine using a cosine similarity matrix.

## Aim 2: Building Pricing Models

For the pricing models we started by splitting the dataset into train and test sets with a 80/20 ratio. The pricing models were built on the concatenated embeddings of ‘ActiveIngredients’, ‘Therapeutic_Class’, and ‘Combined_Desc’ features, with the final embedding dimensions representing each medicine equal to (600,1). After generating embeddings for each medicine we then performed Principle Component Analysis on them and extracted the first 100 components. The reduced dimensionality increased the prediction performance of our models by reducing noise. Each model was trained on the train dataset and

then tested by comparing predictions from the test set with the actual values listed in the test set. The final performance benchmark used was Out of Sample R-squared.

1. Method 1: Linear Regression Model

The first model involved a simple Linear Regression model with Price regressed on the PCA Embeddings. This particular model assumes that there is a linear relationship between the price and the variables being used to predict. The R-squared value came out to be 0.04.

2. Method 2: Random Forest Regression Model

The first model involved a simple Linear Regression model with Price regressed on the PCA Embeddings. A Random Forest model is a Bootstrapped Aggregation of the simple Decision Tree method. The R-squared value for this method came out to be 0.10.

3. Method 3: Neural Network Model

The final model was a Neural Network with a single hidden layer that had 300 neurons and used a ReLu activation function, as price can't be negative. We decided to test out a simple neural Network as the data appears to be highly non linear. A Neural Network is able to handle differently scaled data and outliers well by virtue of constant Loss Optimisation. The R-squared value of the Neural Network came out to be approximately 0.20.

## Aim 3: Finding Medicine Substitutes

A key aspect of the pharmacy business is keeping track of the multitude of medicines available for the same ailment. We developed a model using the same cosine similarity matrix method above, with the embeddings for Combined Description of each medicine, and then comparing the distance of each medicine with others. The top 6 medicines were then shortlisted as substitutes.

# Results

## Categorisation

As there was no labeled data, the results of the categorisation models were analyzed in a qualitative manner. Medicines were first aggregated by their respective assigned Drug Categories, for each of the three methods employed. We then looked at the proportional distribution of Drug Categories across all medicines. Out of the 40 Drug Categories, a total of under 30 unique categories were present in the output. The top 10 Drug Categories were ‘Cold-Cures’, ‘Anti-Inflammatories’, ‘Laxatives’, ‘Diuretics’, ‘Cough Suppressants’, ‘Antifungals’, ‘Analgesics’, ‘Antivirals’, ‘Antipsychotics’, and ‘Hypoglycemics (Oral)’. This output is extremely similar to the initial qualitative assessment made using information provided by Manufacturers. To better understand why certain Drug Categories were left out we analyzed the Euclidean distances between the embeddings for each of the 40 Drug Categories. We observe that the most common categories are quite close to almost all categories of medicines which may lead to misallocation of certain medicines to these common categories. Drug categories absent from the outputs are predictable extremely far from all other embeddings.

We also calculated the average Euclidean distance between the 3 different categorisations of each medicine. The distances were normally distributed with a large number of distances equivalent to zero, which implies certain medicines had the same outcome from each of the three methods. A normal distribution of these distances points towards an issue with the data used to build models, over a systematic error.

On further manual inspection of the dataset, it was noticed that several medicines were mis-assigned to categories which are extremely similar to their intended categories, for example cold cures and analgesics, antibiotics and antivirals. A few words could make the difference between correct assignment or not in such cases.

We can infer from these findings that the embeddings generated were not accurate enough to successfully distinguish between Drug Categories which appear similar. This is likely due to short descriptions of

medicines with too many general over specialized words and Drug Categories with similarly less detailed descriptions.

However, overall the classification models performed well enough for the majority of cases. Medicines were labeled within the correct, or adjacent, categories.

### Pricing Models

The pricing models created were able to achieve a maximum Out of Sample R-squared value of 10% which implies that the model could explain up to 10% of the variation in prices. This is a reasonable starting point.

From the results it can also be inferred that the relation between price of medicine and the features we used as independent variables, anomaly Active Ingredients in medicine, Therapeutic Class, and Description, is highly non-linear as the Out of Sample R-squared value increases as we move from Linear Regression to Neural Networks.

### Substitute Models

The study was unsupervised but the results from the model were satisfactory. We observed that the bio word2vec embedding was helpful in identifying medicines with similar description, product information and benefits. For ~60% of the medicines the first substitute was the medicine itself, and hence we recommended the second most similar medicine as a substitute. We observed that for ~98% of medicines the top substitute belonged to the same therapeutic class. For ~85% of the substitutes the dosage form (tablet, syrup, injection) is the same as the medicine. The type of dosage is really important as the speed with which the medicine affects the body differs. The elderly and kids have difficulty consuming tablets and prefer syrup dosage. Hence, keeping the dosage form the same in substitute is necessary. However, the model recommended a higher mg medicine as a substitute for a lower mg medicine, and vice versa. The model did not take into account the amount of dosage.

# Recommendations and Conclusions

Despite the initial successes of the models for each of the intended Aims, there remains a large scope for improvement. In this section we will provide some recommendations which can improve the performances of the models built in this report and be ready to provide business value.

1. Use a larger dataset of medicines so that the models have access to more information, enabling better classification results. Similarly look for more detailed descriptions of Drug Categories, increasing the distances between each category embeddings
2. Try out different pre-trained embedding models, eg ClinicalBERT, using different dimensions for encoding information. There is potential that the models might perform better if the embedding dimensions are lowered, especially if the text data are not distinguished.
3. Generating own embeddings on a larger dataset. Allows the model to build more accurate context for the data being used.
4. Using different similarity criteria such as Euclidean distance instead of Cosine Similarity.
5. For Pricing models, using more features without concatenating embeddings could provide better outputs while simultaneously reducing the number of neurons in the hidden layer to generate more meaningful relations. To improve the results that we got from Neural Networks, we could also implement GridSearchCV to get optimal parameters, which we were not able to execute this time due to computational issues.
6. For predicting the substitute we utilized the description, product information and benefits. The data was verified using the therapeutic class and the type of drug that was recommended. However, in the model we did not take into account the dosage information. The right amount of dosage is quite necessary. The model could be restricted to recommend medicines of equal dosage.

7. The model predicts a medicine substitute as itself only for 60% medicines, which tells us that the model could be improved further as we expect the similarity of a vector with itself to be highest.
8. Since the dataset scraped here was limited and did not cover the entire medicine dataset on the 1mg website we could not verify our substitute predictions with the one on the website itself. The unsupervised model could be extended further and verified if we scrape information for all the medicines.

# References

BioWordVec: https://www.nature.com/articles/s41597-019-0055-0

FDA: https://www.fda.gov/drugs/investigational-new-drug-ind-application/general-drug-categories

# Appendix

Figure 1: Distribution of Medicines by top 100 ‘Uses’.

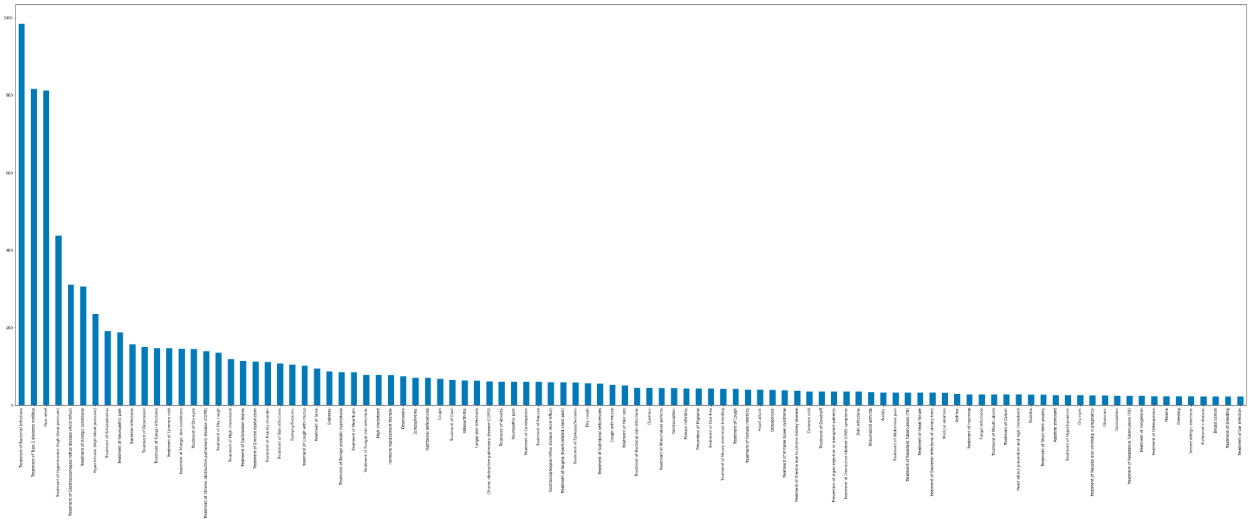

Figure 2: Distribution of Medicines by ‘Therapeutic Classes’.

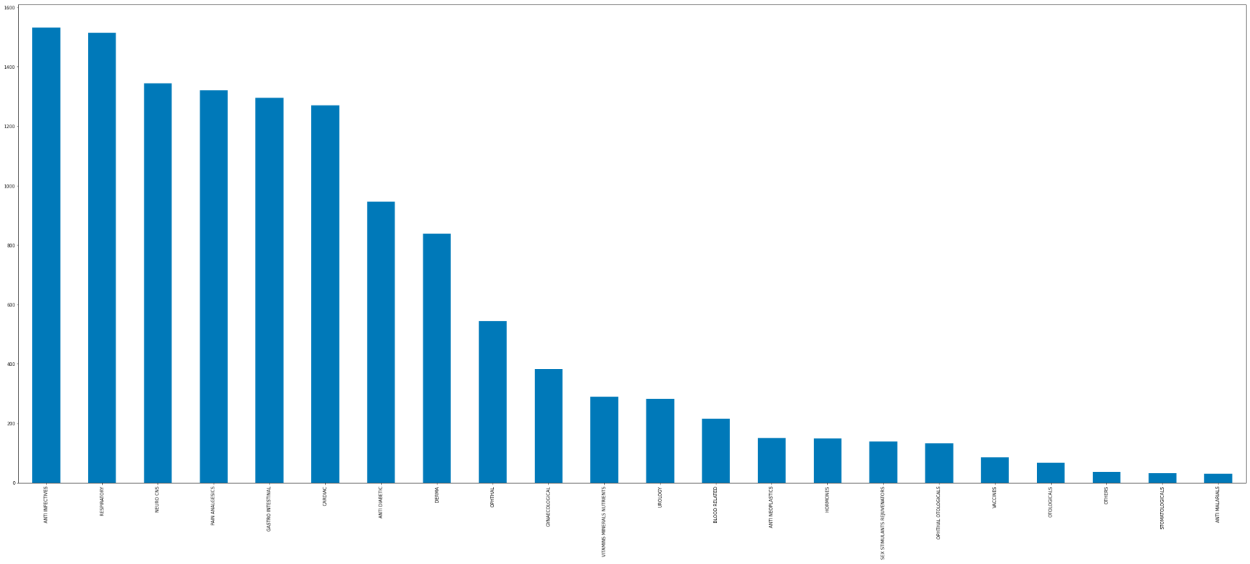

Figure 3: Distance between Drug Category Embeddings.

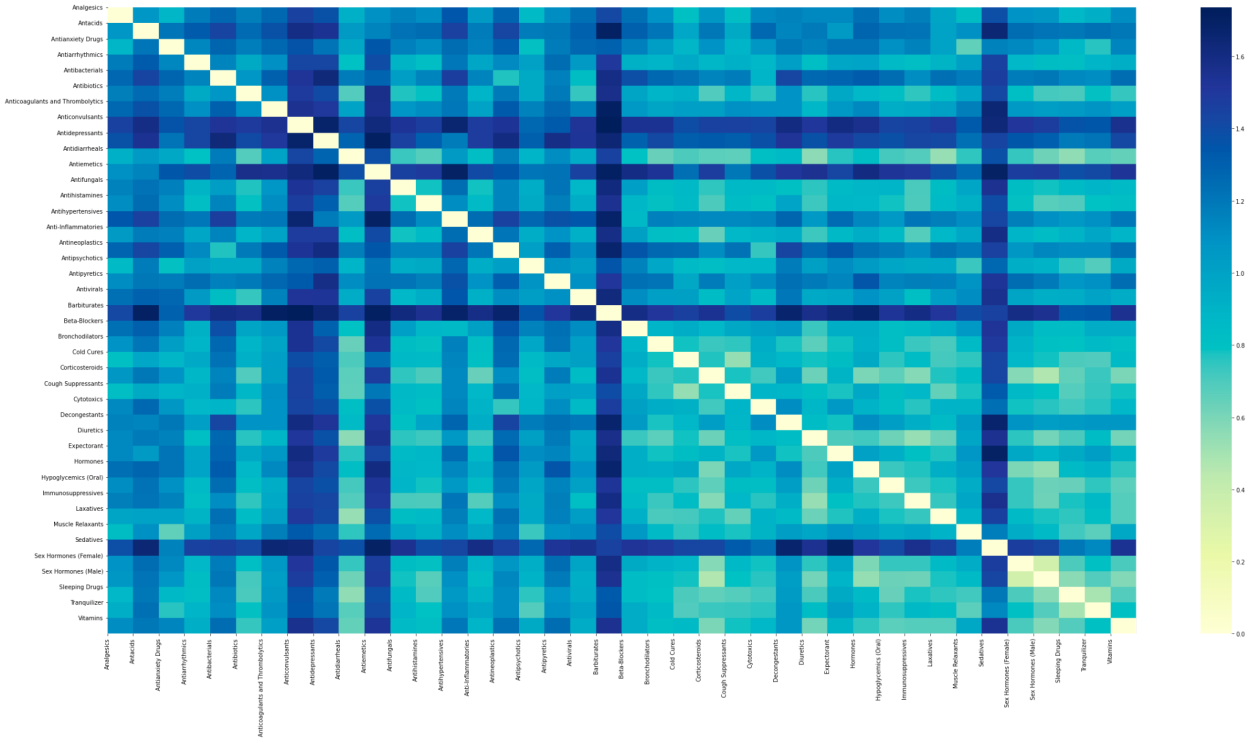

Figure 4: Distribution of medicines by Drug Category.