Prediction of Therapeutic Classifications based on a Medication’s Reported Side Effects

By Brady Bauer

Introduction

Whether you are a doctor, data scientist or expert in a field somewhere between the two, there is one consensus that is agreed upon by all parties: success in the medical field is entirely dependent on how well we can statistically analyze the relationships between experimental input factors and their results. In order to greenlight the usage of new medications or procedures, countless hours of data exploration and inquiry is required such that we are able to fully understand the tools, prescriptions, systems, etc., to improve the health and wellbeing of the world’s populous. Triumph in meeting the standards in place for the approval of these devices requires rigorous inspection; avoiding the repercussions of releasing unrefined or harmful products onto a nation is arguably one of the most important undertakings in the modern age. Failure to do so has dire consequences.

Granted, while the gravity of contents of this paper isn’t nearly as much a matter of life and death as some other endeavors for our current medical professionals, it is still necessary to stress the importance of medical data analysis. Understanding human health is a challenging subject and we require, now more than ever, talented and driven people to help create solutions to many of the problems presented in this field. This was the reasoning used when choosing a medical dataset for this project. Therapeutic classification of certain medications as a function of their common side effects is what will be examined and interpreted for this analysis. The goal being to answer the following questions: is there some kind of observable pattern that exists within the reported side effects of listed medications, and, if so, can it be used to predict how that medication is used?

Data Processing and Hypothesis

The data examined during this research assignment is titled “250k Medicines Usage, Side Effects and Substitutes” by authors Vishal Thakur and Vivek Tiwari. This is a publicly available data set found on Kaggle.com and is updated monthly. This dataset contains information on approximately 250,000 medications collected over a 23-year period just past the 21st century. The following is the authors description of exactly what kind of data is contained within their dataset:

* + 1. **Drug name**
    2. **Adverse reactions and side effects**
    3. **Drug interactions**
    4. **Drug class**
    5. **Substitute drugs**
    6. **Active ingredients**

This data was used for predicting the drug classification as a function of the respective medications’ 21 most common side effects. A 22nd predictor is used as binary factor for whether the drug is considered habit-forming which is also considered as a side effect for the purposes of this analysis. The following is the exact list of the most common side effects chosen to be used as predictors for the models:

**Nausea, Vomiting, Diarrhea, Headache, Rash, Bleeding, Sleepiness, Dizziness, Stomach pain, Dryness in mouth, Flatulence, Abdominal pain, Indigestion, Constipation, Fatigue, Loss of appetite, Increased liver enzymes, Allergic reaction, Heartburn, Weakness , Insomnia, Itching, Prone to Habit Formation**

The response variable Therapeutic Classification is a way of identifying what a drug does generally or how the medication is used. The following is a list of the possible therapeutic classes that a drug can fall under:

**Anti Diabetic, Anti Infectives, Anti Malarials, Anti Neoplastics, Blood Related, Cardiac, Derma, Gastro Intestinal, Gynaecological, Hormones, Neuro CNS, Ophthal, Ophthal Otologicals, Others, Otologicals, Pain Analgesics, Respiratory, Sex Stimulants Rejuvenators, Stomatologicals, Urology, Vaccines, Vitamins Minerals and Nutrients**

The raw data contained in this file is quite difficult to work with as there are 41 columns for possible side effects of each medication containing many holes with empty values. Binary predictor variables were constructed and allocated a column in the set for whether a certain side effect is or is not present for each observation. Initially, all possible side effects reported in the data set were planned on being used for every model (this totaled around 1,500 unique instances). However, simply constructing a data frame of that size took considerable time and computing power, let alone running multiple regression models using the data. As a result, side effects were ordered by how many times they occurred within the dataset and the 22 most common occurrences were selected as predictors. 250,000 observations is also far to large an amount of data to process at once using available computing means. To rectify this a random sample of the medicine data file was made without replacement after dropping all empty values from the original dataset. 5,394 medications were used as training data for each model. ChatGPT was also utilized to help with processing data. The code for the initial data cleaning was written in the Python programming language. ChatGPT was able to help translate the written code for making a function and applying that function in Python to a data frame in the R programming language.

The hypothesis for this project is that the prediction accuracy for the fitted models **will not exceed 60%**. This is because, while there is likely to be some kind of pattern observed by a medication’s classification and its side effects, it is not yet certain if the therapeutic class can be predicted well purely based on a medication’s reported side effects.

H0: Ω < 60% HA: Ω ≥ 60% where Ω is the maximum of all Mean Model Accuracies

Exploratory Data Analysis

Exploratory anlyses for this project was quite difficult to perform because all data was either categorical or Boolean. Because of this, the graphed data was produced in a very similar manner to one another, but with ultimately separate inferences. Below is a bar graph that summarizes the total observations and their categories for each of the sampled drugs.

Chart

Description automatically generated

Here we can see that the number of drugs used as an anti-infective is very high and makes up approximately 25% of the entire sample. However, at the same time, drugs classified as being used as vaccines, otologicals and stomatologicals make up less than a percentage point of the sample. In fact, there are only 18 observations from the original 250,000 that are able to be classified as stomatologicals. Coming from this, there was a need to ensure that each of the classes were represented by at least 18 observations within the sampled data because, when taking a completely random sample from the entire set, some samples failed to include observations representing all possible classes. It should be noted that, because of how nonuniform the observations were to certain classes, the models generated using this data could less accurately predict the classes of medications such as vaccines, otologicals and stomatologicals when compared to a class such as anti-infectives.

Chart, bar chart

Description automatically generated

The above bar graph shows the percentage amount of data that has listed a certain side effect. Again, when selecting which side effects to analyze, each of the possible parameters were ordered based on how frequently they occurred in the data. From there, the top 22 most frequent side effects were selected. A parameter for whether the drug was habit-forming was also included as a side effect. From this graph, it is apparent the data is, again, not uniform; some side effects are much more common than others. Nausea is a side effect that is present in over 60% of medications sampled; however, bleeding and heartburn were reported as side effects in under 1% of sampled medications. It could be possible that the presence of less common side effects might be influential in helping signal to statistical model that this medication belongs to a very specific class. On the other hand, side effects such as nausea could help a model filter out some possible therapeutic classes. Regardless, the ultimate prediction will likely be made using a wide combination of listed side effects.

Model Fit and Performance

Originally, there were plans to fit 7 models (1 LDA, 2 Tree Algorithms, 3 SVMs and 1 Neural Network); unfortunately, however, the computing power necessary to run all 7 at once in a reasonable timeframe was not available. Because of this, a few sacrifices were made in order to properly train on the amount of data that was able to be processed by the machine used for this analysis.

Firstly, only a single algorithm for each model type (1 LDA, 1 Tree, 1 SVM and 1 Neural Network) was run in order to improve upon the time taken to compile the code. Secondly, for all of these models, some individual runs on smaller data were performed in order to gauge where optimal values for tuning parameters fell most frequently. The range over which those parameters were tested during cross-validation was reduced. Lastly, 5-fold was implemented instead of 10-fold cross validation in order to reduce time complexity. Because of these sacrifices, more complex models were able to be fitted on a larger amount of test data while accounting for many more possible side effects as predictors. This, in turn, would produce results that more accurately reflect the data at large resulting in a clearer understanding of just how small the misclassification rate of each model could be when fitted using more computing power.

Model 1: Linear Discriminant Analysis

This model was a struggle to fit for a very long time because whenever attempts were made to tune the shrinkage parameter of this model, computing errors would be left unresolved. It wasn’t until looking closely at the predictor variables that the issue became apparent. Shrinkage is a tuning parameter used when fitting a linear discriminant analysis model so that the covariance matrix of our variables is altered to generate a more stable model. But because all the predictors for this model are binary, the covariance matrix can’t be properly computed. There isn’t variation in the values the predictors can take on and whether or not a medication has a certain side effect is a solid yes or no. Therefore, it is not possible to tune specifically the shrinkage parameter for this dataset in an LDA model. As a result, the intended purpose of this model is to serve as some kind of basis of comparison for the other, much more complex, classification models that were able to be properly tuned.

The final linear discriminant analysis model had a mean model accuracy of **55.540%** for accuracies ranging between **53.099%** and **57.763%**. There were no tuned parameters used in this final model.

Model 2: Random Forest

Initially, the idea for predictive tree algorithms was to fit a gradient boosted tree along with the random forest model and compare the results of the two methods. However, for the sake of computational ease, the random forest algorithm was the algorithm chosen for implementation. The reason this model was included was to observe if the aggregation of random decision tree structures would be the best way to make predictions for the dataset. The tuning parameter “mtry” was constrained over the range [10,20] based on good performance in trial runs on smaller datasets. Because of this, the number of trees allowed to be made by the model is quite high and, timewise, was able to be computed reasonably.

The final random forest model had a mean model accuracy of **78.076%** for accuracies ranging between **77.273%** and **78.889%**. The chosen “mtry” for the final model was **17**. This means that 17 parameters were chosen randomly for each decision split.

Model 3: Support Vector Machine

Again, initial planning was to fit 3 SVM’s to the dataset; one with a linear kernel, one with a polynomial kernel and one with a radial kernel. However, there were 22 possible classifications for the medications in this dataset which means that fitting an SVM of any kind will be extremely computationally expensive due to the way SVMs classify categorical variables with more groupings than typical binary classifications. After a few trial runs on smaller data, the kernel that appeared to perform the best for this dataset was the radial kernel and was, therefore, the kernel chosen for the final fitted support vector model.

The final support vector machine had a mean model accuracy of **78.243%** for accuracies ranging between **77.593%** and **78.887%**. The chosen value of the cost parameter was **2.556** allowed error and sigma was held at a constant of **0.1** for computing performance reasons.

Model 4: Neural Network

For all trial runs of the neural network model for this project, optimal values for weight decay or L2 regularization were found to be extremely small values approximating 0 every time. Therefore, for increased computing performance, the penalty term was not considered for the final model. As a result, the number of layers that the neural network could test was very large. This value was maxed out at around 20 layers. It was found that, for this dataset, continued addition of neural network layers produced increasingly accurate model results which suggests a deeply complex relationship exists between the predictors and response of this data model.

The final neural network model had a mean model accuracy of **76.853%** for accuracies ranging between **75.185%** and **77.911%.**The chosen parameter value for the amount of layers was **20** weight decay was held at a constant of **0** for computing performance reasons.

Conclusion

Based on the findings of this analysis, the hypothesis for this project, that the prediction accuracy that the models fitted to the “250k Medicines Usage, Side Effects and Substitutes” dataset **will not exceed** **60%**, has been **rejected**. Even on the limited amount of randomly sampled data used to train each model, the prediction accuracy of every model exceeded 60% excluding the initial LDA model. The model that performed the best was the support vector machine which used a radial kernel to classify the medications, although, the random forest and neural network models were not very far behind in terms of accuracy.

What the reader should ultimately take away from the findings of this analysis is that machine learning models are very capable of predicting what category of illness a certain medication is used to treat. Based simply off a very minute subsample of medicine data, the support vector machine and random forest models could correctly classify nearly 80% of medications on average based solely on the patterns they observed in the medication’s reported side effects. That is ridiculous. Using the full training data, it would not be surprising if 99% accuracy could be achieved for this classification problem by certain models. Following the findings presented in the analysis of this paper in this, it should be suggested that future research is be done into the pattern of side effects that occur for these medications perhaps so that we may understand why and prevent them.