**Initial Look at the Dataset**

A first thing is to look at the dimensions of the dataset. The training set has 23,814 rows, while the testing set (public leaderboard or public LB) has 3,982 rows. The private leaderboard (or private LB) has 4\*3,982=15,928 total rows. So, the final test set size is about 66% of the training set size (###use simple graphic to explain how this works out).

There are 876 total columns in the data. The first column is a unique row identifier called ‘sig-id.’ There are three categorical variables: ‘cp\_type’, ‘cp\_time,’ and ‘cp\_dose.’ The first category, ‘cp\_type’ identifies whether a subject is given either some placebo or the actual treatment. The second category, ‘cp\_time,’ tells the duration of the treatment as either 24, 48, or 72 hours. The last category, ‘cp\_dose,’ indicates whether the treatment was given in either a high or low dosage. It can be seen in the below figure that both ‘cp\_time’ and ‘cp\_dose’ are roughly balanced throughout the training set. However, the ‘cp\_type’ is quite imbalanced, with a relatively low proportion being given the placebo.

Chart, bar chart

Description automatically generated

Figure

The other regressors are quantitative variables that can be split into two groups. These are the gene expression data and the cell viability data. The former is denoted by ‘g-0,’’g-1,’…, ‘g-771.’ The latter is denoted by ‘c-0,’ ‘c-1,’…, ‘c-99.’ Therefore, there are 772 variables related to gene expression and 100 variables related to cell viability. Combined, that makes up 872 quantitative variables to examine.

Looking towards the target variable, there are 206 different classes. Each subject in the dataset can potentially have zero to seven total labels. A histogram of the number of classes per subject can be seen in the figure below. Around 39.33% have no label, 52.62% have one label, and the remaining (about 8.05%) have multiple labels. This makes the problem a multilabel classification problem, different from either binary classification or multiclass classification as traditionally seen. This is not a new phenomenon, but it has become more common recently for example in computer vision tasks, where the goal is to label the separate objects in a video or image.

Chart, bar chart

Description automatically generated

Figure

From the above, it can be seen that this dataset is a rather high-dimension dataset with many features. Furthermore, these are features that are difficult to interpret for a layperson. Without sufficient medical knowledge, it would be difficult to make too many interpretations about the dataset itself. This is good in terms of learning experience. In the real-world, there are many situations for a data scientist where there are some or many variables that are difficult to understand and cannot be approached logically. The approach then will to be more systematic, and to utilize what is apparent and obvious with needing a background in medicine.

**Targets**

**Features**

**Modeling**

Given that there are 206 classes for the target variable, an approach first will be to use the one-vs-rest (OVR) method utilizing traditional classifiers (e.g., logistic regression, naïve Bayes, etc.). It has been seen on Kaggle already, that many of the best performers are implementing neural networks of some sort to achieve high scores. However, it could be wiser to first develop a baseline model somehow and change the approach carefully towards these more advanced techniques. The evaluation for the score is based on the log-loss metric, so the goal is for competitors to minimize this value for the final testing set.

To determine a suitable model, the approach will be to first develop a pipeline that can be improved over time. This pipeline has multiple phases. The idea behind using a pipeline is that it is required that many models and approaches are tested first before a final model is settled on to submit to the competition to be judged in the final private LB where it is compared with all the other competitors. The pipeline will be such that the dataset is first preprocessed, where categorical variables can be one-hot-encoded (OHE), numeric variables can be normalized, and new features can be added or removed from the dataset through some feature engineering or feature selection process. This preprocessed dataset will then be split into k-folds for k-fold cross-validation (CV) based on a special multilabel stratified approach that takes into account both the fact that classes can be multilabel and imbalanced (this is done by a special library automatically). After splitting the data into separate folds, it will be modeled, and the average log-loss will be analyzed. After a model or set of models have been tuned, the optimized parameters will then be applied to the given test set so that it can be submitted to the public LB for comparison.

This process needs to be improved over time as modeling develops. It will be iterative, starting from a baseline logistic regression model. From there, more complex models and approaches can be added on to try and slowly improve performance via CV and comparison to the public LB.