Machine Learning in Oncology for predicting Drug Approval

Pranav Vijay
Pranav Raveendran
Pranavi Shekhar

SSN College of Engineering, Chennai

Mar 3, 2021

Abstract

- Drug development involves comprehensive **clinical trials**.
- **Clinical trials** - research that studies new tests and treatments and evaluates their effects on human health outcomes - done in **3 phases**
- **Phase 1** - Safety of drug evaluated and major side effects studied.
- **Phase 2** - Testing drug effectiveness and further safety analysis.
- **Phase 3** - Confirm effectiveness, understand all possible side effects and compare to similar drugs.
- If a drug clears Phase 3, it is deemed fit for usage and **approved for release**.
- The drugs used to treat cancerous tumours/neoplasms are called **Anti-neoplastic agents(ANAs)**.

Scope

- **Objective**
  The objective of this work is to develop a machine learning model to predict the likelihood of FDA approval of an Anti-Neoplastic agent from Phase 1 clinical trial data. The presence or absence of certain agents in the drug being tested in Phase 1 is exploited to predict the likelihood of approval.

- **Motivation**
  Drug development in Oncology is a time and cost intensive process. Phases 2 and 3 are long drawn and expensive, with a high attrition rate for final approval.

User Inputs/Outputs

- **Input:** The user provides the presence or absence of certain chemicals/enzymes in the Phase 1 drug being tested as input.
- **Output:** The model classifies the input drug into likely to be approved for release or unlikely to be approved for release.
- The classification is based on an approval score which is also provided to the user.
- A web interface is available to the user which provides a realistic overview of the prediction process. Users can enter the presence or absence of certain chemicals in the drug under consideration and obtain the approval score for the drug which represents whether the drug would be approved or not.

System Architecture

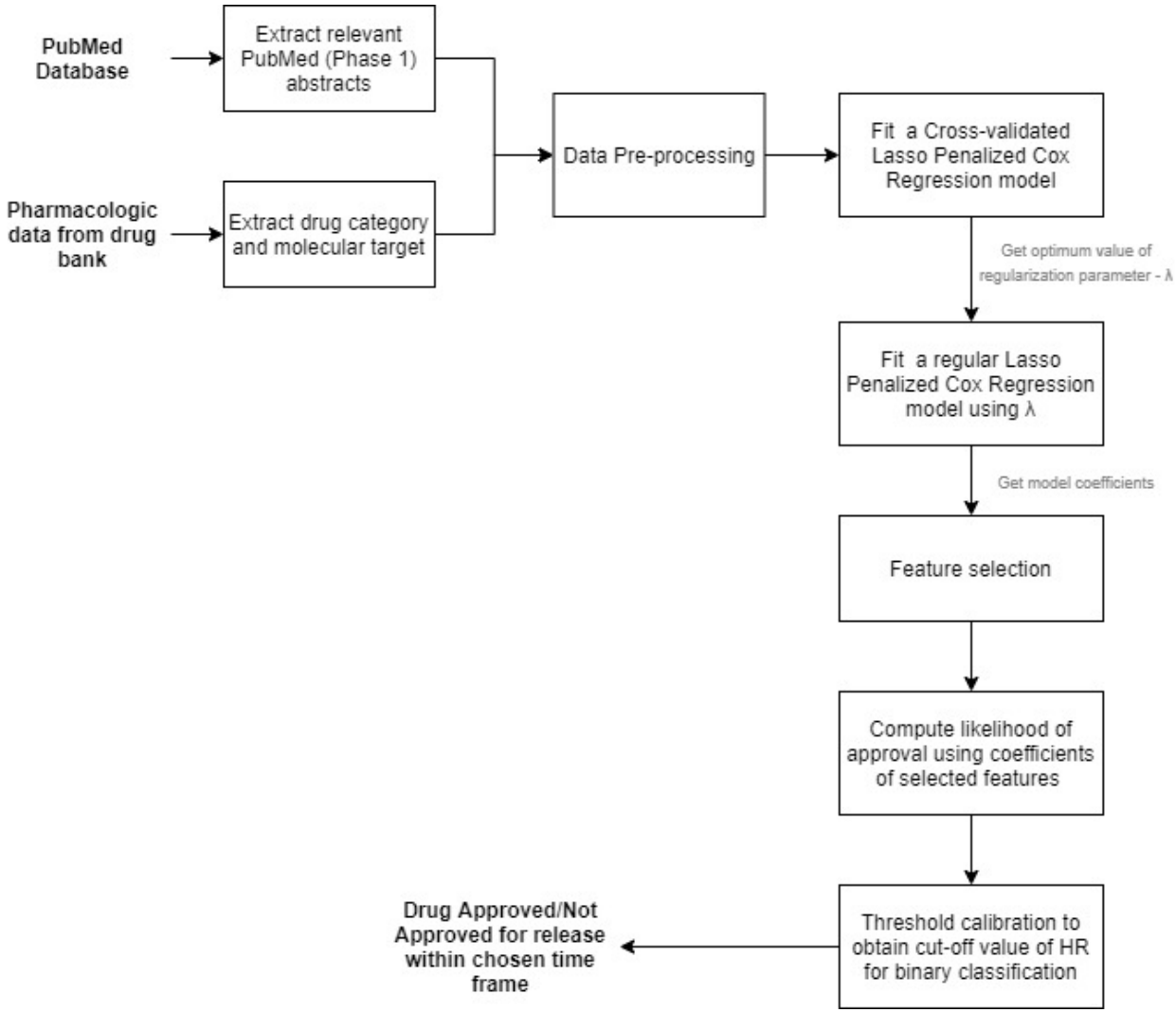

Figure: System Architecture

Dataset

- The data used for this work comes from the **DrugBank** and **PubMed** databases. These are the most comprehensive and exhaustive sources for all drug related data and also have dedicated resources to facilitate easy data extraction.
- The **DrugBank** database is a comprehensive, freely accessible, online database containing information on drugs and drug targets. DrugBank enables us to extract the drug categories and targets regarding the Anti-neoplastic agents of choice.
- **PubMed** is a free search engine accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics. PubMed can be used to obtain details about the drugs Phase 1 clinical trials of ANAs developed between 1972-2017.
- The data from these sources is combined and cleaned to obtain a final, usable dataset consisting of 462 rows and 1415 features.

Data Pre-Processing

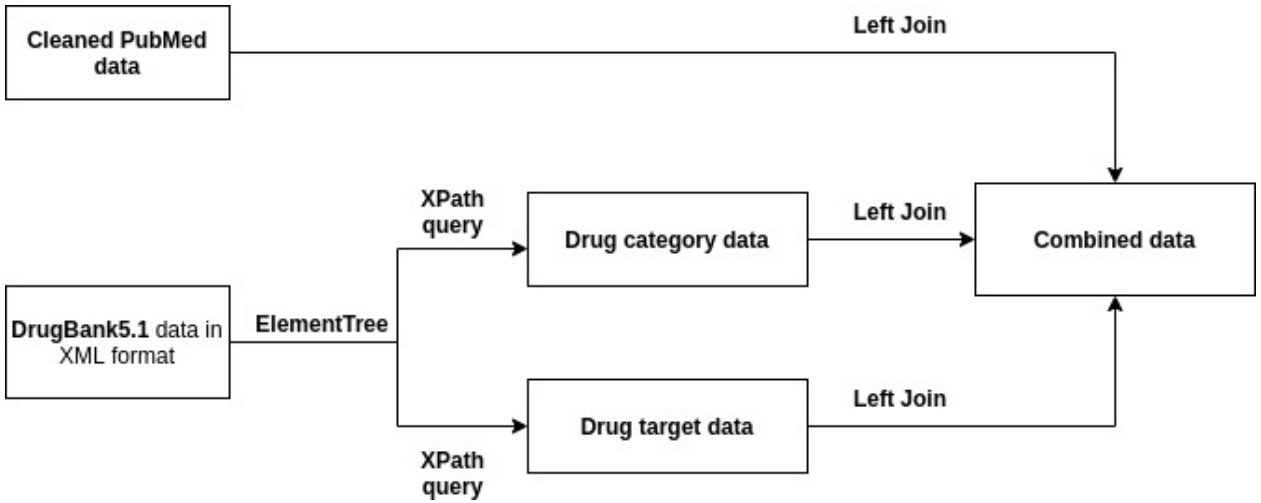

Figure: Data Pre-processing

Techniques/Algorithms

Survival analysis

- Survival analysis corresponds to a set of statistical approaches used to investigate the time it takes for an event of interest to occur.
- Survival analysis methods are usually used to analyse data collected prospectively in time, such as data from a prospective cohort study or data collected for a clinical trial.
- In the context of this project, the event of interest is FDA approval and the subjects of the clinical trial are the drugs for which survival data is available.
- The data we use describes the time taken for FDA approval of various drugs whose composition during Phase 1 clinical trials is considered. *All the methodologies used to build and calibrate the model are based on survival analysis.*

Techniques/Algorithms

Cox Regression

- Cox regression (CR) is method for investigating the effect of several variables upon the time a specified event takes to happen.
- The purpose of CR is to evaluate simultaneously the effect of several factors on survival. In other words, it allows us to examine how specified factors influence the rate of a particular event happening (e.g., infection, death, drug approval) at a particular point in time.
- Cox Regression has been chosen for this project because the data being used was collected over many decades (from 1970-2017) and the objective is to understand the effect of the predictors on drug approval.

Techniques/Algorithms

Cox Regression

- The method does not assume any particular survival model but it is not truly non-parametric because it does assume that the effects of the predictor variables upon survival are constant over time and are additive in one scale.
- Essentially, this model states that risk of the event in any group is a constant multiple of the risk in any other.
- In other words, if a drug has a risk of non-approval at some initial time point that is twice as high as that of another drug, then at all later times the risk of non-approval remains twice as high.

Techniques/Algorithms

Hazard ratio

- The coefficients in a Cox regression are called **hazard ratio** - a positive coefficient indicates a worse prognosis and a negative coefficient indicates a protective effect of the variable with which it is associated.
- In our case, the hazard is defined by drug approval. Therefore, a HR$<$1 indicates reduced chance of approval whereas a HR$>$ 1 indicates an increased chance of approval.
- A linear combination of HRs for each ANA gives us the likelihood of FDA approval for said ANA. We call this the approval score. This score will be used to classify these ANAs into likely to be approved/non-approved for release.

Techniques/Algorithms

Lasso Penalty

- Lasso stands for *Least Absolute Shrinkage and Selection Operator*. Lasso shrinks regression coefficients toward zero by penalizing the regression model with a penalty term called L1-norm, which is the sum of the absolute coefficients.
- The penalty has the effect of forcing some of the coefficient estimates, with a minor contribution to the model, to be exactly equal to zero. It is also called the *regularization parameter*.
- In the context of this project, applying a Lasso Penalty to the Cox Regression model will enable us to remove unimportant features from the model and thereby reduce its complexity - essentially enabling us to carry out feature selection.

Techniques/Algorithms

Choosing the regularization parameter

- The strength of the lasso penalty must be tuned to an optimum level in order to obtain the correct features. There are two ways to select optimum λ - **minimize information criteria** or **cross-validation**.
- An information criterion is a measure of the quality of a statistical model and it's used to estimate goodness-of-fit of the model.
- Cross-validation is a used for assessing how the results of a statistical analysis will generalize to an independent data set and it's used to estimate how accurately a predictive model will perform in practice.
- For the purpose of this work, **cross-validation** is chosen to estimate optimum λ because the objective is to predict whether a drug is approved or not - this means that better predictive performance is more important to us than goodness of fit.

Techniques/Algorithms

Threshold calibration using Logrank test

- Threshold calibration is the process of identifying an optimum value to classify regression outputs into the desired buckets.
- The **Logrank** test is used to test the null hypothesis that there is no difference between the populations in the probability of an event (here - FDA Approval) at any time point.
- The analysis is based on the times taken for the event of interest to occur. It is most likely to detect a difference between groups when the risk of an event is consistently greater for one group than another.
- We apply the LR test to a list of possible thresholds and pick the one which gives the least p-value on the LR test. This threshold is used to classify the drugs into approved/non-approved for release based on their approval scores.

Techniques/Algorithms

Performance analysis using Concordance Index

- The **Concordance index** or **C-index** is commonly used to evaluate the predictive power of models which utilize survival data. In our case, it measures the model's ability to correctly provide a reliable hazard ratio.
- **IPCW estimator:** The Inverse Probability of Censoring Weighted (IPCW) estimator is used to account for the presence of informative censoring. The basic idea of this estimator is to correct for censored subjects by giving extra weight to subjects who are not censored.
- The results of performance analysis of our model gave a C-index of **0.898** and IPCW value of **0.890**. *It can be observed from the C-index and IPCW values obtained that our model is strong and provides reliable hazard ratios.*

References

1. Biotechnology Innovation Organization: Clinical Development Success Rates 2006-2015, 2016.
2. Jonsson B, Bergh J: Hurdles in anticancer drug development from a regulatory perspective. Nat Rev Clin Oncol 9:236-243, 2012
3. Paul SM, Mytelka DS, Dunwiddie CT, et al: How to improve R and D productivity: The pharmaceutical industry's grand challenge. Nat Rev Drug Discov 9:203-214,2010.
4. Chakiba C, Grellety T, Bellera C, et al: Encouraging trends in modern phase 1 oncology trials. N Engl J Med 378:2242-2243, 2018.
5. Wishart DS, Feunang YD, Guo AC, et al: DrugBank 5.0: A major update to the DrugBank database for 2018. Nucleic Acids Res 46:D1074-D1082, 2018.

Thank You