Humana-Mays 2023 HealthCare Analytics Case Competition:

Targeting At-Risk Patients

Table of Contents

1. Executive Summary

This report conducts predictive analytics on Humana member data to identify individuals at high risk of prematurely discontinuing drug A therapy due to adverse side effects. Drug A has proven efficacy as a targeted treatment for non-small cell lung cancer patients, substantially improving survival outcomes. However, frequently occurring side effects like nausea, fatigue, and elevated blood glucose often cause patients to discontinue treatment ahead of schedule.

Rigorous statistical modeling was undertaken to develop machine learning algorithms that predict patients prone to stopping Drug A treatment within the first six months. The modeling dataset incorporated medical and pharmacy claims records for Humana members undergoing Drug A therapy. Following comprehensive data preparation, six sophisticated predictive models were constructed: Quadratic Discriminant Analysis, Logistic Regression, Random Forest, LightGBM, XGBoost, and CatBoost.

The LightGBM model exhibited superior predictive performance, with an AUC of 0.7515 and accuracy of 91.54% on the holdout validation set. Analysis of feature importance revealed that side effect diagnoses and cost factors were most influential in predicting early Drug A discontinuation.

To assist high-risk members in continuing their essential therapy, we recommend Humana implement:

- Supportive care programs to help members effectively manage side effects
- Financial assistance initiatives to mitigate prohibitive Drug A treatment costs

With proactive patient outreach and support systems in place, Humana can empower members to complete full Drug A regimens, enabling improved cancer survival outcomes. The predictive modeling supplies a data-driven approach to identify and assist vulnerable members most in need.

2. Case Context

In this section, we outline the background and problem statement. Drug A is an effective lung cancer treatment but side effects often cause patients to prematurely discontinue therapy. Humana aims to identify at-risk members and help them adhere to full treatment regimens.

2.1 Case Background

Cancer remains one of the leading causes of mortality, with approximately 600,000 annual deaths attributed to cancer in the United States alone. Although novel therapeutic agents are steadily emerging, many present side effects that may deter patients from adhering to their crucial medication regimens.

Drug A has demonstrated efficacy, nearly doubling survival rates compared to absence of treatment. Moreover, adhering to the Drug A regimen reduces cancer recurrence risk by 80%.

However, frequently occurring side effects of Drug A often prompt patients to prematurely discontinue therapy. While many side effects can be managed with counseling and mitigation strategies, patients may cease treatment rather than seeking guidance on addressing them. Approximately one quarter of Humana members taking Drug A experience adverse side effects resulting in therapy cessation within the first 6 months.

2.2 Problem Statement

Humana, a leading national health insurance provider, is dedicated to assisting members in living their healthiest lives. They provide advanced healthcare analytics, pharmacy solutions, and primary care services to optimize member health outcomes.

This analysis will assist Humana in identifying members most likely to discontinue Drug A therapy within the initial 6 months due to adverse side effects. Additionally, it will propose solutions Humana can implement to improve member adherence to essential cancer medications.

3. Data Preparation

In this section, we detail the data preprocessing steps taken, including exploratory analysis, missing value imputation, feature encoding and selection, strategic data partitioning, and adversarial selection. These steps optimized the data for modeling.

3.1 Exploratory Data Analysis (EDA)

Private information, cannot be disclosed

3.2 Data Cleaning and Imputation

For the medical claim data, the columns 'medclm_key' and 'clm_unique_key' are unique for each records. For the pharmacy claim data, the columns 'document_key' and 'ndc_id' are unique. After dropping two unique columns for each of data, we dropped all the duplicates to avoid data redundancy.

We analyzed the percentage of null values in each column for both medical claims and pharmacy claims. We chose to drop the columns with over 60% of null values. Since we have the therapy start dates of the patients, we chose to analyze the records which are after the start of therapy. For the data left, we fill all the null values in categorical columns with “missing”. For numerical columns, we used KNN to impute the null values.

The most difficult part of this task is to merge three datasets into one. In this context, patients may have multiple medical and pharmacy claim records. Some may have no claim records. We had to merge all records into one row and add the row to the training data. We chose different strategies for different columns. For categorical data, we chose the most frequent value except missing value. For numerical data, we chose median, mean and sum for specific columns according to their context and meaning. We also created features of number of medical and pharmacy records for each patient.

3.3 Feature Encoding

For later model training, we encoded the categorical data into numerical data (float numbers). We tried a few techniques such as one-hot encoding and LightGBM. We found LightGBM suited our model better so we used it.

3.4 Feature Selection

We used several different techniques for features selection. We checked these techniques separately and made different experiments whilst iteratively dropping features within a fast LGBM pipeline.
We dropped features using a mix of the following methods:

- Features with a high percentage of missing values (60% and up)
- Collinear (highly correlated) features
- Features with zero importance and zero influence (like 'therapy_id')
- Sensitive features (like sex and race)

3.5 Data Selection

As we stated before, patients may have multiple medical and pharmacy claim records, while some may not. Thus, we decided to train our model based on the 1103 patients who had at least one claim record. For the 129 patients who don’t have any record, we plot the distribution of their target. Only two of them had unsuccessful therapies. Most of the patients with no records had information loss in other columns, so their information is too little for us to predict their therapy states. Thus, we chose to assign all the patients with no records in the holdout data with 0 (“all other therapies”).

3.6 Adversial Selection

Usually, we would randomly choose 70% of data as training data and 30% of data as test data. However, we employed adversarial selection to help improve our model's AUC and accuracy. Instead of randomly selecting 70% of data for training, we chose the 70% of data that are most similar to the holdout data.

We labeled the training data (without the target column) as 1 and holdout data as 0. Then we trained a CatBoost model to help classify whether the data is from training or holdout. The model assigned the probability of a record coming from training data. We sort the data ascendingly as the data with lower probability are more similar to the holdout data. By training the model with these data with similar distribution to the holdout's, we can better predict our target.

4. Model Building

In this section, we present the machine learning algorithms utilized in our analysis to predict which patients are more likely to successfully complete their medical procedures without quitting. We employed a diverse set of models, each with unique strengths and characteristics, to maximize the robustness of our predictions.

1. Quadratic Discriminant Analysis (QDA)

QDA is a discriminative model that makes use of the distribution of predictor variables for both classes (patients who complete and those who quit) to make classifications. It assumes that the features follow a Gaussian distribution and estimates the mean and covariance of each class. QDA has been chosen for its ability to capture complex relationships between predictors and the target label.

2. Logistic Regression

Logistic Regression is a fundamental model for binary classification problems. It models the log-odds of the probability of a patient quitting as a linear combination of predictor variables. It's interpretable and well-suited for understanding the impact of individual features on the likelihood of quitting.

3. Random Forest

Random Forest is an ensemble learning method based on decision trees. It is adept at handling high-dimensional data and capturing complex interactions between features. By combining multiple decision trees, it provides a robust predictive model while controlling overfitting.

4. Light Gradient Boosting Machine (LightGBM)

LightGBM is a gradient boosting framework that leverages a histogram-based learning method to efficiently build decision trees. It excels in handling large datasets and performs well in terms of speed and predictive accuracy. Its ability to work with categorical data is particularly valuable in healthcare analytics.

5. XGBoost

XGBoost is another gradient boosting library that has become a staple in machine learning competitions. It optimizes model performance by using an ensemble of decision trees and is known for its regularized objective function, reducing overfitting and improving prediction accuracy.

6. CatBoost

CatBoost is a gradient boosting algorithm designed to handle categorical features naturally. It employs a powerful technique called ordered boosting to train the model efficiently. This algorithm is well-suited for scenarios where the dataset contains a mix of numerical and categorical features.

The selection of these models is motivated by the need to explore a wide spectrum of techniques and their varying ability to handle different types of data and relationships. The diversity of the algorithms ensures that we have considered various aspects of the data, enabling us to make more informed predictions. We also employed robust cross-validation (10 fold) and hyperparameter tuning techniques to optimize the performance of each model, especially for the tree-based ensemble models.

5. Model Evaluation

In this section, we provide a detailed analysis of our model evaluation with regard to two performance metrics of AUC (area under the ROC curve). For our round 1 submission, we made a strategic decision to select LightGBM as our final model. LightGBM consistently demonstrated exceptional performance, particularly excelling in both the AUC and accuracy metrics. Although Logistic Regression also exhibited strong performance, we conservatively opted not to select it. Our decision was based on the concern of potential overfitting, given the utilization of target encoding for handling categorical data. Target encoding can make the logistic regression model more susceptible to capturing noise in the training data. To ensure the highest quality predictions for our prediction, we favored the robustness of LightGBM, which efficiently handled the complexities of our dataset while reducing the risk of overfitting.

5.1. Model Evaluation on Validation Set

Model	AUC	Accuracy
Random Forest	0.7107	0.9335
CatBoost	0.7177	0.9154
XGBoost	0.7226	0.9245
Light GBM	0.7515	0.9154
QDA	0.7606	0.7462
Logistic Regression	0.7613	0.9335
Meta Model	0.8590	0.9334

5.2 Future Work

In our pursuit of optimizing the predictive accuracy for patient procedure completion, we recognize that the true strength of ensemble learning lies in harnessing the collective wisdom of diverse models. To fully exploit this potential, we introduce a **meta-model** that will act as the conductor of our ensemble orchestra, harmonizing the predictions from various base models and seeking to produce a unified, superior performance. Each base model offers a distinctive perspective on the problem at hand, and while they have individual strengths and weaknesses, the concept of ensemble learning lies in merging their predictions to create a more robust, accurate, and reliable prediction.

The meta-model, also known as the stacking model, will serve as the centerpiece of this ensemble. It will take the predictions generated by the base models as its input and apply its own learning algorithms to fuse these predictions into a final, aggregated prediction. By doing so, it leverages the complementary strengths of the base models and aims to mitigate their weaknesses, resulting in a more resilient, high-performing model. Initially, we implemented the meta-model using Logistic Regression. However, due to the constraints of the competition's submission rules, we had limited opportunities to test its performance fully. Furthermore, there was a growing concern that the logistic regression-based meta-model might be prone to overfitting the training data, given its relatively simplistic nature.

As a result of these concerns and our commitment to producing the best possible predictive model, we decided to explore alternative meta-models. After careful consideration and extensive testing, we ultimately selected LightGBM as our meta-model. LightGBM offers a number of advantages, including strong predictive performance, the ability to handle high-dimensional data efficiently, and a built-in mechanism for mitigating overfitting, making it an ideal choice to lead our ensemble.

6. Analysis and Recommendations

In this section, we employed feature analysis to reveal side effect diagnoses and costs as top predictors of Drug A discontinuation. We also make recommendations including providing supportive care resources and financial assistance to high-risk members.

6.1 Model Analysis

From the results of feature importance calculated by our model, we can see that features related to side effects diagnosis and costs contribute higher importance in prediction; they are:

- Ade_diagnosis, diag_cd2, primary_diag_cd, diag_cd3 (side effects)
- Avg_rx_cost, rx_count, avg_tot_cost (costs)

Hypothetically, if patients experience ADE or have financial issues paying for Drug A, they are more likely to discontinue therapy ahead of time. Therefore, we propose recommendations in two major directions.

Feature importance graph omitted to protect privacy

6.2 Recommendations

6.2.1 Ensuring Supportive Care and Accessible Resources

The first and most important thing is that we need to inform patients of any potential side effects before the therapy, explain the outcome (improve survival), and encourage them to stay positive during the therapy.

It is also necessary to inform patients of online or in-person resources if they encounter any problems. Side effects such as diarrhea, nausea, or pain can be difficult to tolerate. We want to ensure that our patient can get help from our/public resources. Specifically, we can provide medications for those side effects and psychological consulting to our patients.

Moreover, for patients who experience nasty side effects, dosage adjustment may be applied to this situation. Consult the patient's oncologist to see if reducing the dose while maintaining efficacy is possible. It can help mitigate severe side effects.

6.2.2 Providing Financial Assistance

From our model analysis, we understand that the cost of taking a treatment is likely to be the cause of early quitting the therapy. Hence, we suggest providing financial assistance to patients in need.

We recommend that Humana collaborate with local government and banks and initiate a financial assistance program for cancer patients, so that patients with low-income can receive funds for the therapy.

Additionally, we suggest patients participate in hospital-based programs which may provide assistance with medical bills, co-pays, and other related expenses.