THE ERDŐS INSTITUTE

Data Science Bootcamp, May 2022.

Members: Diego Galdino and Wayne Uy

Mentor: Kritika Singhal

Problem statement

This project is concerned with predicting the copayment that is expected of a patient for a medication using claim billing data that records information on the transaction date, pharmacy name, diagnosis, drug details, insurance plan details, claim status, and patient copay. The goals of this project are 3-fold: predict the expected copay of a patient for a particular medication, predict whether the claim will be accepted or rejected by the insurance plan, and cluster medications according to formulary status and copayment requirements.

Proposed solution

Our proposed predictive model has 2 stages. In the first stage, a classifier determines whether the claim will be accepted or rejected. If the claim is rejected, the patient copay is $0. Otherwise, a regression model is then used to compute the payment that is expected of a patient.

The data set was pre-processed before it was used to train machine learning models. Duplicate rows were removed as well as rows for which pcn and group were both missing. In the training data, we investigated computing the maximum or the average copay among data samples with identical features. New features were created from the data set which served as predictors for the machine learning models. In particular, we extracted the month and day of the week from the transaction date, merged bin, pcn, and group to identify the insurance plan specific to a patient, classified whether the drug is branded or generic, extracted the drug name, and extracted the first letter of diagnosis.

We investigated the use of various machine learning models to build classifiers and regressors, however, we mainly focused on the gradient boosting algorithm provided by LightGBM. Gradient boosting offers a compromise between computation time (unlike neural networks), accuracy, and interpretability (by way of decision trees). More importantly, LightGBM is able to natively handle categorical data without the need of one-hot encoding. This is important since all features in the data set are categorical in nature. In training the LightGBM model, cross-validation was utilized to tune the hyperparameters which include the number of decision trees in the model, the number of leaves in the decision tree, the depth of each tree, and the learning rate.

Results

For our regression model, we achieved 6.7921 MSE and 0.0192 MAPE on the test data set. The low MAPE suggests strong predictive capability of our model given that the copay ranges from $0 up to $500. For our classification model, we tried various undersampling and oversampling strategies to overcome the unbalanced class problem. Our classifier achieves -7.6304 negative log loss, 0.7791 Accuracy, 0.2743 Precision, 0.9922 Recall, 0.4298 F1, and 0.8970 AUC on the test set.

For practical use of our machine learning models, we created an online application (URL: https://share.streamlit.io/diego-galdino/erdos_bootcamp_may22_cover_my_meds/main) that addresses the main goal of this project which is to anticipate the patient copay and the formulary status to help the patient in decision making.