Predict Medical Events

Problem Statement:

The goal is to predict the next 10 events in 2014 for each patient in order of occurrence.
The “train.csv” file contains historical patient information from Jan 2011 to Dec 2013.
The “test.csv” file contains a list of Patient IDs for which we aim to predict the next 10 events for in the year 2014.

Error: There is error in sample_submission.csv where file mentions to find Event1 to Event9 which makes 9 events predictions. But problem statement requires it to be 10 events.

Data:

Given data, UID (Unique Patient ID), Age of Patient, Gender, Date of the event (it’s actually month and year) and our target variable Event Codes in ICD-9 format.

Approach:

According to me the data given for this challenge is very insufficient so as to make any predictions for future medical events for the patients. Any additional data would have been very useful to create a good predictive model.
As the test data contains only UID of 3000 patients, goal is to build a model which takes UID of patients as an input and gives back 10 most likely events in the order of occurrence.
We build a supervised model after doing an EDA on data which is described below.

EDA:

This is critical part of modelling as we get to know various characteristics of the data.
First we separated the date into two columns of month and year. Following this we sorted the data according to the Year, Month and UID to get an ascending order of events occured for each patient in year 2011 to 2013.

We see that there are growing number of events each year with highest in 2013.
We see that there are 3000 unique patients with “Id_e45ad2db” having highest number of events equal to 1401.

There is a lot of disproportion in distribution of number of medical events which is our target variable.

We will do an in-depth analysis on this variable.
There are total of 6472 unique events during the year 2011-13.
Among the 6472, there are 770 events with examples of 1 patient, 1740 with less than 5 patients.
After plotting a cumulative distribution of event codes, we see that 227 events comprise of 50% of the data with event code 9921 having the highest proportion.

Having **decided on number of events** to be considered for our predictive model.
It was time to split the data into train and validation sets.
This part also proved to be computationally tricky as we had to predict 10 events into the future.

So, we divided the each patient's events into 2 sets, with last 10 events into validation set and all remaining events into validation set.
After removing all patients with less than 10 events, we decrease the number of patients by 19.
This can very much hurt our generalisation of model as we will have to omit these number of patients from our test data.
This number is different depending on number of events we choose for training (**critical hyperparameter**).

Having done the split properly, it was time to use our data for training the model.

Model:

We use various classification models like DecisionTree, RandomForest, AdaBoost, Gaussian NB, KNeighbours Classifier, etc. and compare our model based on a new metrics called mean NDCG (Normalized discounted cumulative gain)@K, where K=10. NDGC is calculated as:

$$nDCG(k) = DCG(k) / IDCG(k)$$

where $DCG(k)$ is Discounted Cumulative Gain@K.