

ZS ASSOCIATES – YOUNG DATA SCIENTIST CHALLENGE 2017

APPROACH DOCUMENT: SECTION A

SUBMITTED BY: NISHANT RAJ

IIT ROORKEE

nishant95.raj

**Approach:** There were 6472 unique events in dataset. This was a clear cut indication that normal machine learning techniques like regression and classification would not apply. Also it was not a time series regression problem wherein you can take some previous windows to predict the outcome of the next window but what I would like to call it as a “time based sequence classification” problem. I thought to use the available columns and generate classification with top 10 softmax probabilities but that approach was clearly not going to work (even if we used top algorithms like XGBoost) given we had only 3000 unique PIDs and 6472 events. Clustering of events was one possibility but still there would be very little information to learn and predict. I tried using markov models but the result was not good(for higher orders too). I then tried some rule based approach as we do in time series similar to concepts like moving average that gives more weightage to recent observation and hence I tried giving weights to all the previous events in various possible ways like linear weights, logarithmic weights, exponential weights and various other increasing weight combinations and finally selected a model. Also, as there was higher probability for first events, I replaced last 3 events with most recent 3 events as I had observed that events were repetitive in last. Also the position of the events was an important parameter in the NDCG scoring parameter.

**Quality Checks and Errors found:** The first and foremost error that I found in the dataset was that it had ICD-10 values. If you see Wikipedia and other references, you would observe that ICD-09 had only E and V first letter encoding.

Another issue I found was that some of the codes were not found both in both types of encodings like 7087. Some of the patient ids had very few occurrences of events like 10001508, 1016210, 1006395, 10006122 etc.

Just to confirm about the other letter encoding existence and to see if that was being evaluated I made a dummy submission with all values being H100 for all PIDs and events. I expected a flat zero score but got 0.0001 score that left me in doubt if I can replace these variable with their ICD 09 codes or not.

**Data processing steps:** I extracted the first letter of the codes (like “H” from “H100”) and saw their value count and also observed the value counts of these alphabet starting code apart from E and V and observed that their number was very little to influence the score and also my dummy submission left me in doubt so I left the ICD-10 encodings as it is in the dataset because their number was very less in prediction file.

The dataset was grouped according to Patient IDs and Events as well as Dates were generated as sequences (please refer code file).From this data frame further processing was done.

**Feature Generation**: The first feature that I generated was “Year” that I had extracted from Date column and the second feature being “Month”. Also to generate weightage value I had written a code to allocate different weights based on increasing mathematical functions like linear weights, exponential weights, logarithmic weights and their arithmetic various combinations.

Apart from this I created another feature by replacing the above generated month column by “Month Count” from starting from beginning like for February 2012 the value would be 14 for Month Count. This would ensure that the value that I generated had effect of both month and year while I generate weights for the events and score them by adding their weights.

**Key Trends / Observations:** The most striking observation that I made in the dataset was that the most recent events affected the next 10 events that is the events which occurred in the past long ago (even if they were very repetitive in beginning) had little to no effect in the end. Thus prompted me to give excess weight to most recent events.

Some events were found to co-exist together like for example whenever some code occurred it was always followed by another particular code only.

**Model Choice Explanation :** No generic machine learning algorithms like Random forest, XGBoost etc could be used as the number of events were very large with a few PIDs to train upon(model would not be able to learn). Also an approach similar to starter script in Python would generate very sparse dataset.

So, I was left with the choice of two models that could work for sequence predictions: HMM and LSTMs. I worked on HMMs but could not improve their accuracy beyond a limit even for higher order values. I tried working with LSTM which I think must give best results but my processor being i3 could not handle that much amount of sequence data on the whole and hence I scrapped this approach. I feel that if LSTMs are trained on individual patient IDs with sufficient epochs and memory units they can create best results.

The motivation for choosing my present model came from time series modelling where we use time windows as features and also in some approaches we give more weights to recent events. Since, I had observed that recent events had more profound effect I decided to use time based weighting of different event sequences with recent events getting more weight. The problem was to decide what should be the correct function for our case. So, I tried various weights like linear, logarithmic and chose the one that gave me the best results in both cross-validation and leaderboard which was multiplication of exponential of months(from 1 to 36) & natural logarithmic value of months. The weight for each event was added and whichever events had highest 10 scores were selected as potential top 10 candidates. I replaced last three candidates with first three to ensure that a repeating trend was observed in the dataset. This worked for me well.

**Expected Error / Expected Gain :** I feel that there might be some chances of overfitting as the function I have selected for weight assignment is based on trials but since I have cross-validated to some extent I feel that that won’t affect NDCG score much and I expect my NDCG score to be around a value that is similar to the public leaderboard score(0.105 approx).

**Top5 Most Significant Variables in the model :** I believe an approach is more valuable than using insignificant variables. Some of the most significant variables in my approach were :

1. Events : Events that were nearby were more helpful in prediction than the events far away in sequence.
2. Month : Later months had more weightage than the current months.It used Date variables to extract month from it.
3. Year : Recent medical history in present year were more valuable than the medical history in previous years in general cases. Hence I extracted Year variable from Date variable in the dataset.
4. Modified Month : Months were mapped initially from 0 to 12 for each year. To take into effect of year, for January 2012 mapping used was 12+1 = 13 and for March 2013 mapping used was 24+3 = 27
5. Weighted\_factor\_expolog : Weights were obetained by :

Weight = exponential (Modified Month) \* logarithm (Modified Month)