**Big Data Competition**

**Health Equity Data Set**

Mike Roylance, David Crockett

**Introduction:** We have provided Claim data for our members and we want you to predict when our members need to add money to their HSA accounts.

You can include any other data sources you like to make a case to show why people should add money to their HSA accounts at a given time.  Using Age/location can make a difference in the prediction as well.

**Methods Used:** First, we decided to look how money came out of the system.

As members receive services, Health Equity receives information about what type of service was rendered (CPTCode), how much it cost the insurance company (RepricedAmount), how much it cost the member (PatientResponsibilityAmount), and when it ended (ServiceEnd). These tuples of information exist in an ordered sequence of time.

To predict when a person should money into their account, we wanted to predict what the most likely next service is that would be rendered and a rough estimate of how much it would cost.

First, we knew we needed to track the CPT codes. However, CPT codes are very granular in their service description and we needed to group together similar ones. We decided to group them together using the same groupings as described here:

<http://en.wikipedia.org/wiki/Current_Procedural_Terminology>

The exact set provided by The Healthcare Cost and Utilization Project (HCUP):

<http://www.hcup-us.ahrq.gov/toolssoftware/ccs_svcsproc/ccscpt_license.jsp>

We did some data wrangling with the list to produce a dictionary of CPT code to CCS code.

Next, we denormalized the Claim, ClaimDetail, Member and Dependent data into a single csv file containing the proper columns:

NewMemberID, DependentID, CPTCode, CCSCode, PatientResponsibilityAmount, RepricedAmount, BirthYear, Gender, Zip, State, ClaimType, ServiceStart, ServiceEnd

This information was ordered by NewMemberID, then DependentID, then ServiceEnd – all ascending.

With this new csv file, we then build a few dictionaries:

Transition – the transition dictionary looks at each person (which is a composite key of NewMemberID and DependentID) and records the probability that the transition happened. This starts at the first person’s service rendered and builds a transition to the next service rendered until there are no more services rendered. This also builds into each transition grouping elements BirthYear (3 groups, under 30, under 60 and over 60) and Location (state). Each record is considered 1 occurrence, the probabilities are calculated at the end. For example, a standard transition will look like this:

**Under60\_169 -> Under60\_147 : 0.013157894736842105**

The probabilities are calculated after all the transitions have been counted up.

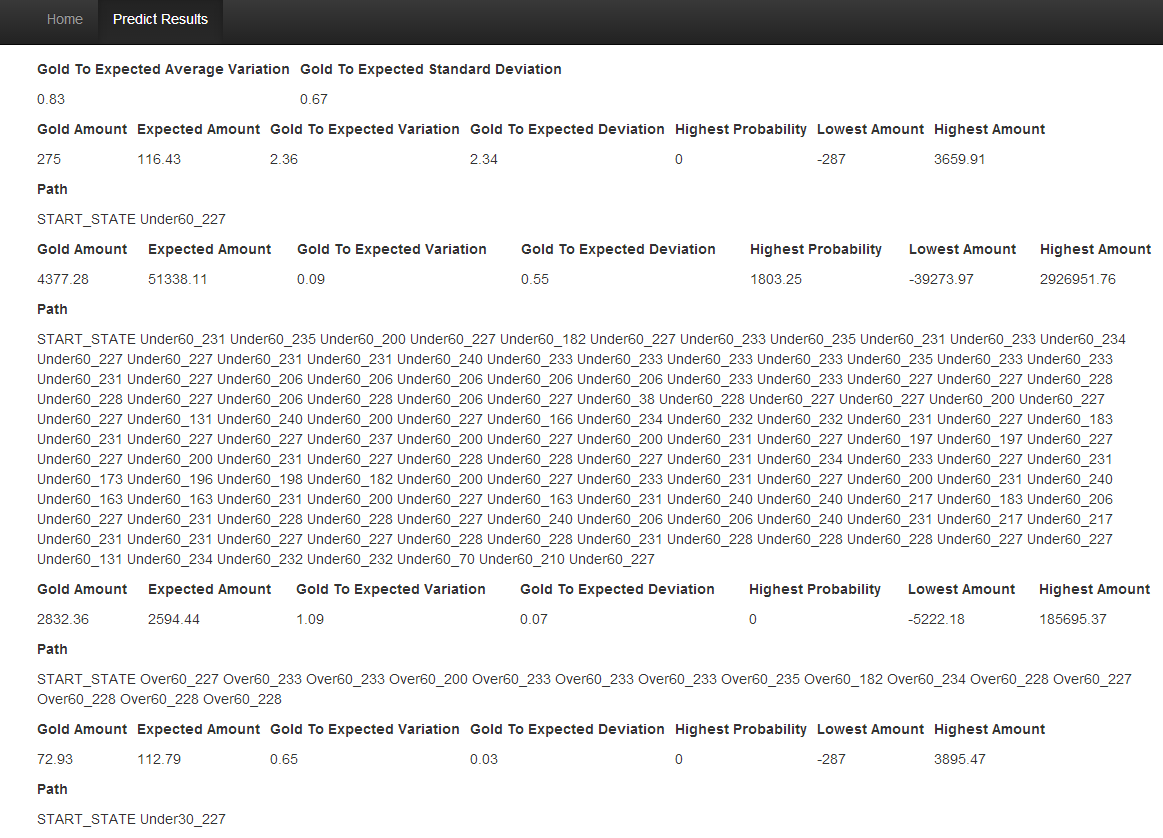
Emission – the emission dictionary looks at each bigram transition and records the amount and probabilities similar to the Transition dictionary. An example record looks like this:

**Under30\_231\_Under30\_240 -> 1425.22 : 0.0008103727714748784**

While the Transition and Emission dictionaries are being built, a small sample (3%) is omitted from the training dictionary and placed in a gold set. This gold set will be used to test the training set.

With the dictionaries built, the gold set is then tested. The results can be seen on the webpage:

<http://ec2-107-20-54-170.compute-1.amazonaws.com/HealthEquity/Home/PredictResults>



(Figure 1)

For each sequence in the gold set, the amount that’s recorded from each transition is compared against the expected amount (Expected Value) from the dictionaries.

In Figure 1:

**Overall Results:**

**Gold to Expected Average Variation** refers to the average value of Gold Amount / Expected Amount. This method predicted that, on average, Gold Amount was 83% of the Expected Amount.

**Gold to Expected Standard Deviation** refers to the standard deviation of the **Gold To Expected Average Variation** calculation. While the average was 83%, the standard deviation was quite large at .63. This meant that most of the results were within 20% to 146% of the expected amount.

**Gold Amount** refers to the actual amount recorded in the gold set.

**Expected Amount** refers to each amount recorded multiplied by its probability of occurring.

**Gold to Expected Variation** the result of Gold / Expected.

**Gold to Expected Deviation** the result of Gold / Expected – Average.

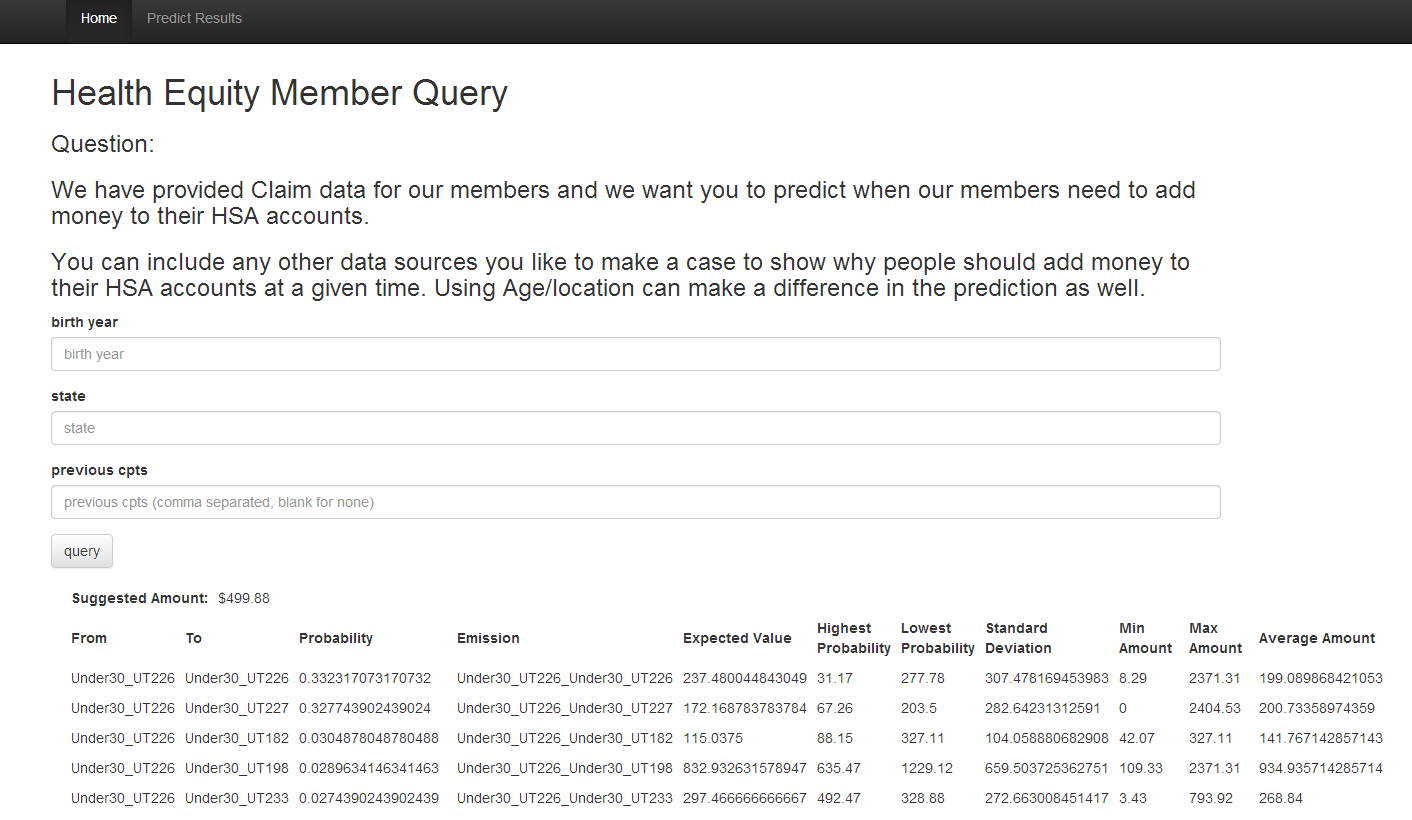
**Highest Probability** refers to the amount that has the highest probability of occurring.

**Lowest Amount** refers to the lowest amount found.

**Highest Amount** refers to the highest amount found.

We then built a page that allows a user to query the specific results of a person given an age, location and preceding CPT codes:

<http://ec2-107-20-54-170.compute-1.amazonaws.com/HealthEquity>



(Figure 2)