Big Data Competition

Health Equity Data Set

Mike Roylance, David Crockett

Introduction (Health Equity): 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.

Solution Provided: Predictions for whether a member should add money to their account, and how much, can be seen here:

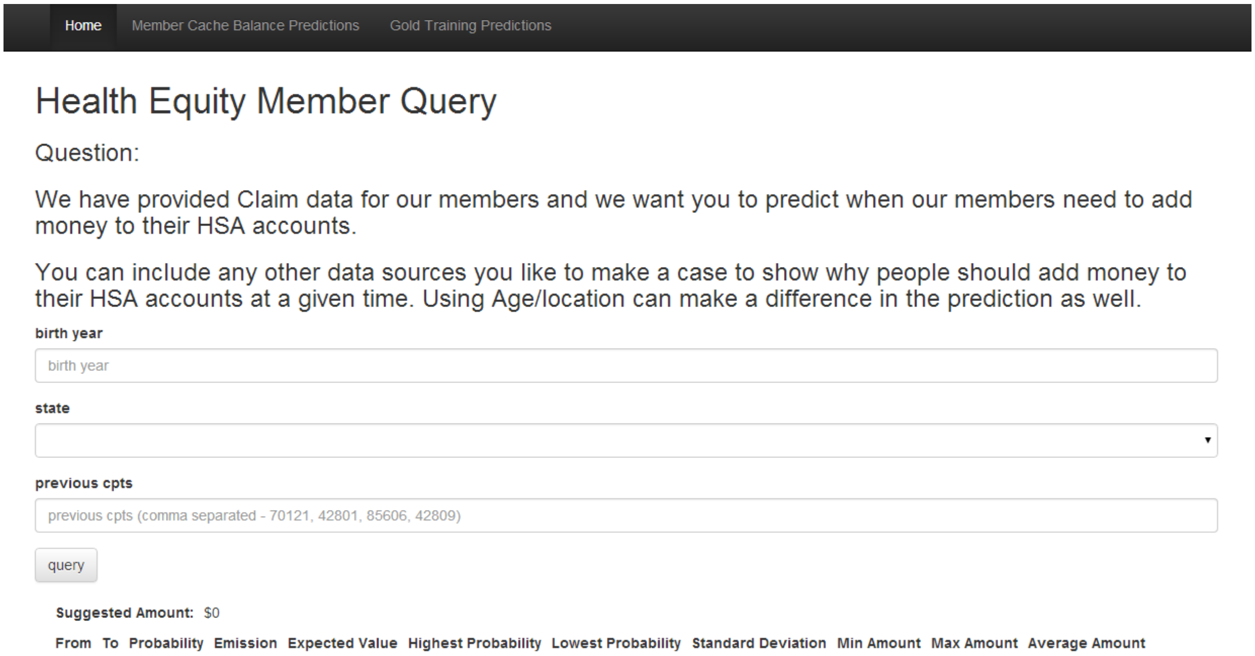
http://ec2-107-20-54-170.compute-1.amazonaws.com/HealthEquity/Home/NewMemberPredictions

Home | Member Cache Balance Predictions | Gold Training Predictions

MemberID	DependentID	Birth Year	State	Last CPT Code	Cached Balance	Recommended Balance	Sufficient Amount
11592	0	1962	OR	99213	$1848.440	$4452.99	No
25126	0	1953	OR	84153	$13758.200	$4452.99	Yes
25126	1	1951	OR	84153	$13758.200	$4452.99	Yes
11699	0	1954	UT	99396	$7693.890	$3905.92	Yes
12387	0	1976	UT	99213	$1059.210	$3905.92	No
11784	0	1962	WA	J7030	$173.550	$1989.03	No
11784	1	1965	WA	J7030	$173.550	$1989.03	No
11784	2	1994	WA	J7030	$173.550	$1989.03	No
11784	3	1998	WA	J7030	$173.550	$1989.03	No
12423	0	1959	WA	99213	$534.000	$1707.06	No
12423	1	1963	WA	99213	$534.000	$1707.06	No
12423	2	1993	WA	99213	$534.000	$1707.06	No
12423	3	1995	WA	99213	$534.000	$1707.06	No
12423	4	1997	WA	99213	$534.000	$1707.06	No
12850	0	1971	WA	99212	$406.260	$1707.06	No
12850	1	1974	WA	99212	$406.260	$1707.06	No
12155	0	1973	MD	84702	$56.460	$1520.16	No
11322	0	1967	GA	99213	$1688.880	$1151.15	Yes
11322	1	2001	GA	99213	$1688.880	$1151.15	Yes
11322	2	1956	GA	99213	$1688.880	$1151.15	Yes
12096	0	1964	GA	99203	$5999.870	$1151.15	Yes
12096	1	1959	GA	99203	$5999.870	$1151.15	Yes
12096	2	1992	GA	99203	$5999.870	$1151.15	Yes
12096	3	1995	GA	99203	$5999.870	$1151.15	Yes
12096	4	1990	GA	99203	$5999.870	$1151.15	Yes
1069	0	1969	WA	99213	$1400.710	$1104.87	Yes
1069	1	1970	WA	99213	$1400.710	$1104.87	Yes

We have also uploaded a query system:

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



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 determine when a person should add money into their account, we wanted to predict what the most likely rendered next service and costs are.

First, we knew we needed to track the CPT codes. However, CPT codes are very granular in their service description. This would create a model that is too tightly fit. We decided to group the CPT codes together using the same groupings as described here:

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

The set is 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 built 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

Home | Predict Results

Gold To Expected Average Variation	Gold To Expected Standard Deviation
0.83	0.67

Gold Amount	Expected Amount	Gold To Expected Variation	Gold To Expected Deviation	Highest Probability	Lowest Amount	Highest Amount
275	116.43	2.36	2.34	0	-287	3659.91

Path

START_STATE Under60_227

Gold Amount	Expected Amount	Gold To Expected Variation	Gold To Expected Deviation	Highest Probability	Lowest Amount	Highest Amount
4377.28	51338.11	0.09	0.55	1803.25	-39273.97	2926951.76

Path

START_STATE Under60_231 Under60_235 Under60_200 Under60_227 Under60_182 Under60_227 Under60_233 Under60_235 Under60_231 Under60_233 Under60_234 Under60_227 Under60_227 Under60_231 Under60_231 Under60_240 Under60_233 Under60_233 Under60_233 Under60_233 Under60_235 Under60_233 Under60_233 Under60_231 Under60_227 Under60_206 Under60_206 Under60_206 Under60_206 Under60_206 Under60_233 Under60_233 Under60_227 Under60_227 Under60_228 Under60_228 Under60_227 Under60_206 Under60_228 Under60_206 Under60_227 Under60_38 Under60_228 Under60_227 Under60_227 Under60_200 Under60_227 Under60_227 Under60_131 Under60_240 Under60_200 Under60_227 Under60_166 Under60_234 Under60_232 Under60_232 Under60_231 Under60_227 Under60_183 Under60_231 Under60_227 Under60_227 Under60_237 Under60_200 Under60_227 Under60_200 Under60_231 Under60_227 Under60_197 Under60_197 Under60_227 Under60_227 Under60_200 Under60_231 Under60_227 Under60_228 Under60_228 Under60_227 Under60_231 Under60_234 Under60_233 Under60_227 Under60_231 Under60_173 Under60_196 Under60_198 Under60_182 Under60_200 Under60_227 Under60_233 Under60_231 Under60_227 Under60_200 Under60_231 Under60_240 Under60_163 Under60_163 Under60_231 Under60_200 Under60_227 Under60_163 Under60_231 Under60_240 Under60_240 Under60_217 Under60_183 Under60_206 Under60_227 Under60_231 Under60_228 Under60_228 Under60_227 Under60_240 Under60_206 Under60_206 Under60_240 Under60_231 Under60_217 Under60_217 Under60_231 Under60_231 Under60_227 Under60_227 Under60_228 Under60_228 Under60_231 Under60_228 Under60_228 Under60_228 Under60_227 Under60_227 Under60_131 Under60_234 Under60_232 Under60_232 Under60_70 Under60_210 Under60_227

Gold Amount	Expected Amount	Gold To Expected Variation	Gold To Expected Deviation	Highest Probability	Lowest Amount	Highest Amount
2832.36	2594.44	1.09	0.07	0	-5222.18	185695.37

Path

START_STATE Over60_227 Over60_233 Over60_233 Over60_200 Over60_233 Over60_233 Over60_233 Over60_235 Over60_182 Over60_234 Over60_228 Over60_227 Over60_228 Over60_228 Over60_228

Gold Amount	Expected Amount	Gold To Expected Variation	Gold To Expected Deviation	Highest Probability	Lowest Amount	Highest Amount
72.93	112.79	0.65	0.03	0	-287	3895.47

Path

START_STATE Under30_227

(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.

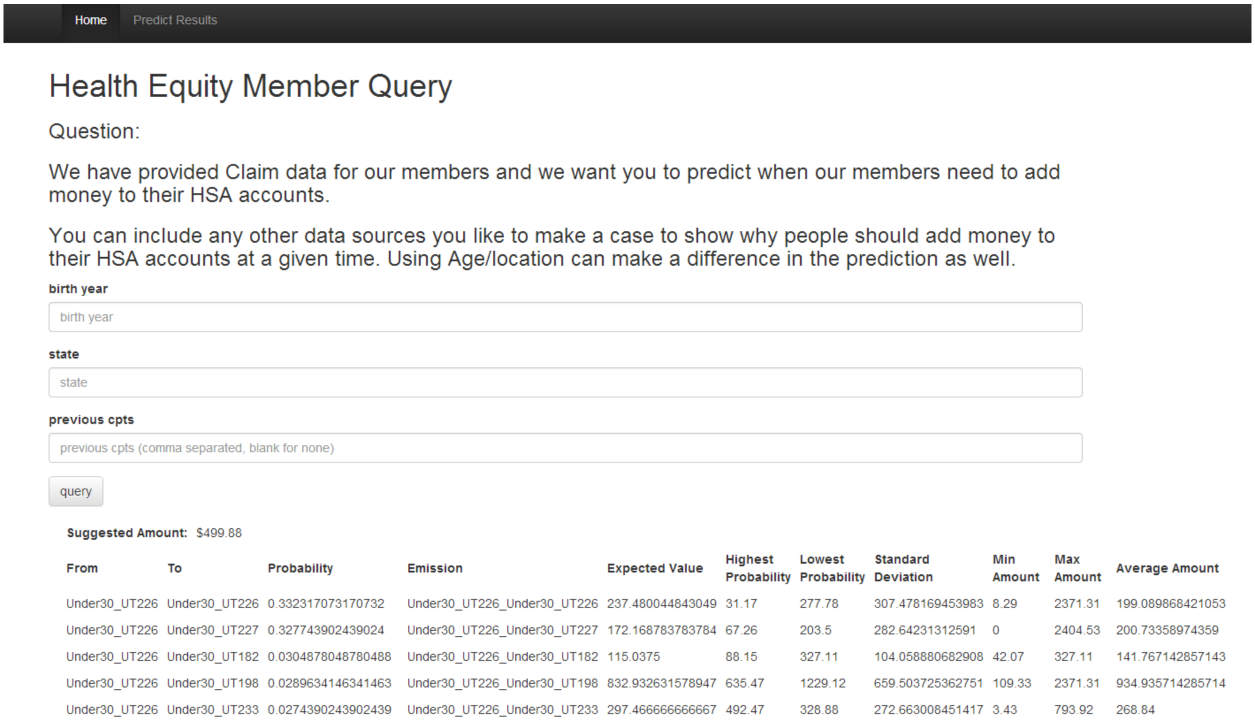
Highest Probability refers to the amount that has the highest probability of occurring.

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


Home Predict Results

Health Equity Member Query

Question:

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.

birth year

birth year

state

state

previous cpts

previous cpts (comma separated, blank for none)

query

Suggested Amount: $499.88

From	To	Probability	Emission	Expected Value	Highest Probability	Lowest Probability	Standard Deviation	Min Amount	Max Amount	Average Amount
Under30_UT226	Under30_UT226	0.332317073170732	Under30_UT226_Under30_UT226	237.480044843049	31.17	277.78	307.478169453983	8.29	2371.31	199.089868421053
Under30_UT226	Under30_UT227	0.327743902439024	Under30_UT226_Under30_UT227	172.168783783784	67.26	203.5	282.64231312591	0	2404.53	200.73358974359
Under30_UT226	Under30_UT182	0.0304878048780488	Under30_UT226_Under30_UT182	115.0375	88.15	327.11	104.058880682908	42.07	327.11	141.767142857143
Under30_UT226	Under30_UT198	0.0289634146341463	Under30_UT226_Under30_UT198	832.932631578947	635.47	1229.12	659.503725362751	109.33	2371.31	934.935714285714
Under30_UT226	Under30_UT233	0.0274390243902439	Under30_UT226_Under30_UT233	297.466666666667	492.47	328.88	272.663008451417	3.43	793.92	268.84

(Figure 2)

Finally, we then calculated the expected amount that each member should have (based on their last known CPT code) and indicated whether or not they should add money to their account.

Home | Member Cache Balance Predictions | Gold Training Predictions

MemberID	DependentID	Birth Year	State	Last CPT Code	Cached Balance	Recommended Balance	Sufficient Amount
11592	0	1962	OR	99213	$1848.440	$4452.99	No
25126	0	1953	OR	84153	$13758.200	$4452.99	Yes
25126	1	1951	OR	84153	$13758.200	$4452.99	Yes
11699	0	1954	UT	99396	$7693.890	$3905.92	Yes
12387	0	1976	UT	99213	$1059.210	$3905.92	No
11784	0	1962	WA	J7030	$173.550	$1989.03	No
11784	1	1965	WA	J7030	$173.550	$1989.03	No
11784	2	1994	WA	J7030	$173.550	$1989.03	No
11784	3	1998	WA	J7030	$173.550	$1989.03	No
12423	0	1959	WA	99213	$534.000	$1707.06	No
12423	1	1963	WA	99213	$534.000	$1707.06	No
12423	2	1993	WA	99213	$534.000	$1707.06	No
12423	3	1995	WA	99213	$534.000	$1707.06	No
12423	4	1997	WA	99213	$534.000	$1707.06	No
12850	0	1971	WA	99212	$406.260	$1707.06	No
12850	1	1974	WA	99212	$406.260	$1707.06	No
12155	0	1973	MD	84702	$56.460	$1520.16	No
11322	0	1967	GA	99213	$1688.880	$1151.15	Yes
11322	1	2001	GA	99213	$1688.880	$1151.15	Yes
11322	2	1956	GA	99213	$1688.880	$1151.15	Yes
12096	0	1964	GA	99203	$5999.870	$1151.15	Yes
12096	1	1959	GA	99203	$5999.870	$1151.15	Yes
12096	2	1992	GA	99203	$5999.870	$1151.15	Yes
12096	3	1995	GA	99203	$5999.870	$1151.15	Yes
12096	4	1990	GA	99203	$5999.870	$1151.15	Yes
1069	0	1969	WA	99213	$1400.710	$1104.87	Yes
1069	1	1970	WA	99213	$1400.710	$1104.87	Yes

(Figure 3)

Future Considerations:

The member calculations need to be completed, we only calculated a few thousand.

It would be good to experiment with different combinations of age / location grouping, as well as gender.

It would be good to experiment with better time amounts. Currently, we don’t include the exact time between transitions.