Heritage Provider Network Health Prize

Round 2 Milestone Prize

How We Did It – Team 'Market Makers'

Phil Brierley
Tiberius Data Mining
philb@tiberius.biz

David Vogel
Voloridge Investment Management
dvogel@voloridge.com

Randy Axelrod
Healthcare Consultant
rcaxelrod@cox.net

Overview

The Heritage Provider Network Health Prize is a competition to develop a predictive algorithm that, based on historical claims, will help identify those patients most likely to be admitted to hospital. The competition will run for two years, with milestone prizes awarded every six months. Team 'Market Makers' were second on the leaderboard on the 13th February 2012, the deadline date for the second milestone award, but first on the holdout data.

This document describes the methodology used in generating the *Market Makers* submission. It is an extension to our milestone 1 report, which should be read first to understand the full context. We will not replicate the information previously documented, but concentrate on the additional developments implemented.

In Brief

There were two areas which accounted for the progress;

1. *Building a wider variety of base models*
   In the milestone 1 submission, our base models were derived from four main algorithms and concentrated on our data set 1. For milestone 2, we utilised two additional algorithms and extended our previous algorithms to data set 2.

2. *Improved Ensembling*
   We implemented the blending algorithm as referenced by Willem Mestrom in section 5 of his milestone 1 report. Slight tweaks were made that further protected against overfitting.

The final model was primarily based on a linear combination of 79 base models. This compares with the 20 base models utilised in the milestone 1 blend.

Model Building

The models in the milestone 1 submission are documented in Appendix B of the milestone 1 report. Of the 20 models used in the blend, only three were based on data set 2. The reason for this is that data set 1 gave the best performing individual models, so that is where the effort was focussed.

The median model, which was our best 'model' was the median prediction of nine models, of which four were from data set 2. Other median models were attempted with a bias in numbers towards data set 1 base models, but the leaderboard accuracy of the model with a high proportion of data set 2 models was never reached.

This pointed the us to the conclusion that more could be made of data set 2. This data set was revisited and more models built using the variety of algorithms and techniques we had employed on data set 1.

It is the variety of individual models that result in the most synergy. There were three ways that variety was introduced into our base models.

1. Using different algorithms
2. Using different data sets
3. Using different variable subsets

Adjusting parameter settings for a given algorithm (all other things being equal) is a fourth way variety can be introduced (and was used), but this will not be as effective in blending as using something totally different.

Two new algorithms were introduced that were not used in the milestone 1 solution;

1. Additive Groves [1]
2. Multivariate Adaptive Regression Splines

In our milestone 1 report we detailed the particular settings for each model built order to introduce this variety. The reality is that these specific settings are not prescriptive, they just need to be different. This was the approach also taken by Willem Mestrom -

'From the beginning of this contest I choose not to build a single very very good model but instead create different models each modeling the variation differently.'

This is the exact approach that we took. It is still necessary to build each model carefully ensuring it is the best you can do given the limitations you have set, but the secret is not to worry if it does not individually perform impressively.

1 We appreciate the guidance of Daria Sorokina

Ensembling

The term blend and ensemble are used interchangeably in this report.

Typically a blend is a linear combination of base models, with each model being assigned a weight. If each base model is a reasonable attempt then common sense implies that the weights should sum to 1.

The main concern when combining many models is overfitting - assigning weights that do well on the training data but do not generalise to new data. Overfitting is characterised by some models having extreme weights – large contributions to the blend. It is prudent to aim to share the weighting around as equally as possible, so as not to be too reliant on a single model (don't put all your eggs in one basket).

Linear regression is one way of determining the weights. Ridge regression is a modification of linear regression that helps reduce the extreme weights. The more variables there are in regression then the more likely it is that overfitting will occur – but also the more variables there are, then the more likely a better general solution exists somewhere, given the correct weightings are assigned.

Our 'tweak' to the blending algorithm was aimed at reducing the magnitude of extreme weights while allowing us to use as many base models as possible. The process was to initialise all base model weights to zero, build many smaller sub blends on randomly selected base models, then just divide the accumulated weights by the number of sub blends built to determine the final ensemble weights.

Our candidate population contained 79 base models with each sub blend containing a randomly selected *n* base models. The process was repeated 1,000 times with a ridge parameter of 0.0001. We built models with various values of *n*, with generally increasing leaderboard performance as *n* increased, but also also with an increasing probability that the model has overfit to the leaderboard. The final choice of *n (*20) was a tactical choice that resulted in a final model slightly better on the leaderboard than the third placed team.

The model descriptions and weightings are listed in Appendix A.

Further Refinement

The result of the blending still has a chance of overfitting the leaderboard, as the leaderboard scores are used in the regression. To further protect against this, we then 'blended the blend' with another model blend, the weights of which were not derived using the leaderboard scores.

We reduced data set 2 to contain only those patients who had two consecutive years of history in the training set. This means we have the richest data

available at a patient level, but also means that we can only make predictions for those who also have two consecutive years of historical data.

An ensemble was built using this reduced data, with cross validation sets used to get the final weightings, which were;

Algorithm	Blend Weight
GBM	0.49
BT	0.19
RS	0.21
LM	0.12

See the table at the end of this report for definitions of these algorithm acronyms.

For those patients who have two years of history, we blended this model with the 79 model blend with the weights being 0.1 and 0.9 respectively. Those who did not have two years history just retained the 79 model blend predictions.

This technique resulted in a leaderboard score that was not significantly different to the 79 model blend, but gave us the comfort factor that a portion of this model was not going to be overfitting to the leaderboard.

A further adjustment was a calibration for those patients who had suppressed information (claims truncated, missing gender, missing age).

The 'models' at the bottom of the table in appendix A were just binary flags for certain features in the data, e.g. 'sex missing' was just a 1/0 indicating if the gender of the patient had been suppressed. Individually these models were the worst, but the blend weights were typically not insignificant.

We re-blended our incumbent model with three of these 'models' (claims truncated, missing gender, missing age). The resultant weightings told us that we were initially over estimating patients with these flags, and their scores needed to be reduced relative to the other patients.

The final submission was truncated at the lower boundary to 0.04.

The final leaderboard score of our selected submission was 0.455247.

Comments and Observations

Appendix A shows that our best individual models for this data were GBMs built on data set 1 (the table is ordered by leaderboard score). When the ensemble weights are ordered by absolute value to give the weight rank, we see the 2nd and 3rd ranked models are both GBMs but built on the two different data sets.

Although the 3rd ranked model in the blend is not individually special, it provides fantastic synergy. A top 10 model can be achieved by simply combining these two base models.

	Model A	Model B	(0.6 * A) + (0.4 * B)
Leaderboard Score	0.4599	0.4626	0.4578
Leaderboard Rank	18	106	8

We could not find any other two models that were built using the same data that when combined would give a higher leaderboard score. Hence it appears the largest benefit comes not in blending algorithms, it is the fact we are using different representations of the data.

A linear regression model was built resulting in a leaderboard score of 0.4618, which was particularly good for this algorithm. In order to achieve this, we ustilised data set 1 but also created a whole series of extra interaction variables. These interactions are simply the product of two of the existing variables, and are discovered by comparing the rmse of models built using each of the single variables with that of the rmse of a model built with the product of the variables. The interactions that result in the most synergy are retained.

When these interaction terms were also included in the GBM models, there was no gain in leaderboard performance. This demonstrates that some algorithms require variable pre-processing to improve, whereas others have the ability to achieve the same result internally.

The interaction terms were generally not used as they vastly increased the size of the modelling file.

Appendix A – the model weightings in the 79 model blend. The weights are applied to the log scale version of the predictions.

Leaderboard Score	Ensemble Weight	Weight Rank	Data Set	Algorithm
0.4590	0.02346	50		MEDIAN9
0.4593	0.08108	15	1	GBM
0.4599	0.22575	2	1	GBM
0.4599	0.06674	24	1	GBM
0.4600	0.14962	6	1	GBM
0.4602	0.01390	58	1	GBM
0.4603	0.00314	74	1	GBM
0.4603	0.00451	70	1	GBM
0.4603	0.05541	28		MEDIAN64
0.4607	0.03629	41	1	GBM
0.4607	0.05452	29	1	ENS
0.4608	0.03771	40	1	ENS
0.4608	0.00787	62	1	GBM
0.4609	0.05052	34	1	ENS
0.4610	0.00348	72	1	ENS
0.4610	0.07576	19	1	ENS
0.4612	0.00726	63	1	ENS
0.4613	0.13394	9	1	ENS
0.4613	-0.04907	35	1	ENS
0.4614	0.04150	39	1	ENS
0.4614	0.04185	38	2	GBM
0.4616	-0.01400	57	1	ENS
0.4617	0.10028	14	1	ENS
0.4618	0.08016	17	1	LM
0.4618	0.07228	21	1	ENS
0.4619	-0.00313	75	2	GBM
0.4619	0.05300	31	2	GBM
0.4621	0.06071	26	2	GBM
0.4622	0.05567	27	1	NN
0.4622	-0.00124	76	1	ENS
0.4625	-0.00603	65	1	BT
0.4626	0.20787	3	2	GBM
0.4628	0.00073	77	1	GBM
0.4629	0.01006	61	1	AG
0.4630	-0.07355	20	1	BT
0.4631	-0.00005	78	1	AG
0.4632	-0.00524	67	1	ENS
0.4633	-0.00391	71	2	ENS

0.4633	-0.01990	53	1	BT	
0.4633	-0.27159	1	1	BT	
0.4637	0.14179	8	1	NN	
0.4637	0.05384	30	1	ENS	
0.4637	-0.03024	46	1	BT	
0.4642	-0.02101	52	1	RS	
0.4644	0.03269	44	1	ENS	
0.4646	0.01487	56	1	ENS	
0.4650	0.00321	73	2	BT	
0.4658	-0.02632	49	2	NN	
0.4658	-0.07827	18	2	RS	
0.4662	0.12045	11	2	NN	
0.4664	0.11832	12	1	LM	
0.4664	-0.06722	23	1	NN	
0.4668	-0.06905	22	1	LM	
0.4668	-0.03198	45	2	LM	
0.4668	-0.02972	47	1	LM	
0.4671	0.02858	48	1	LM	
0.4673	-0.13284	10	1	LM	
0.4679	-0.19354	4	1	LM	
0.4679	-0.11737	13	1	RS	
0.4680	0.14790	7	1	LM	
0.4687	-0.08069	16	1	LM	
0.4690	0.03331	43	1	LM	
0.4692	-0.18733	5	1	LM	
0.4697	0.02145	51	1	LM	
0.4708	-0.00453	68	1	LM	
0.4710	0.01339	59	1	LM	
0.4762	0.04772	36	1	LM	
0.4796	0.03572	42	1	GBM	
0.4850	0.00575	66	1	LM	
0.4902	0.06470	25	1	LM	
0.4956	-0.01800	54	1	C	Constant
0.5096	-0.01563	55	1	C	Claims Truncated
0.5154	0.04285	37	1	ENS	
0.5243	-0.05272	32	1	C	Age Missing
0.5296	-0.01325	60	1	GBM	
0.5354	-0.05128	33	1	C	Sex Missing
0.6118	-0.00452	69	1	C	Sex Male
0.6180	-0.00680	64	1	C	Cohort 111
0.6238	0.00004	79	1	C	HasDrugs

GBM	Gradient Boosting Machine
BT	Bagged Trees (ie random forests)
LM	Linear Regression
NN	Neural Network
AG	Additive Groves
RS	Multivariate Adaptive Regression Splines
C	no model
ENS	ensemble of GBM,BT and LM