JMLR: Workshop and Conference Proceedings 7: 45-55 KDD cup 2009

Predicting customer behaviour: The University of Melbourne's KDD Cup report

Hugh Miller H.MILLER@MS.UNIMELB.EDU.EDU
Department of Mathematics and Statistics
The University of Melbourne
Parkville, Victoria, 3010, Australia

Sandy Clarke SJCLARKE@UNIMELB.EDU.AU
Department of Mathematics and Statistics
The University of Melbourne
Parkville, Victoria, 3010, Australia

Stephen Lane S.LANE@MS.UNIMELB.EDU.AU
Department of Mathematics and Statistics
The University of Melbourne
Parkville, Victoria, 3010, Australia

Andrew Lonie ALONIE@UNIMELB.EDU.AU
Department of Information Systems
The University of Melbourne
Parkville, Victoria, 3010, Australia

David Lazaridis D.LAZARIDIS@PGRAD.UNIMELB.EDU.AU
Department of Mathematics and Statistics
The University of Melbourne
Parkville, Victoria, 3010, Australia

Slave Petrovski SLAVEP@UNIMELB.EDU.AU
Department of Medicine
The Royal Melbourne Hospital
The University of Melbourne
Parkville, Victoria, 3010, Australia

Owen Jones ODJONES@UNIMELB.EDU.AU
Department of Mathematics and Statistics
The University of Melbourne
Parkville, Victoria, 3010, Australia

Editor: Gideon Dror, Marc Boullé, Isabelle Guyon, Vincent Lemaire, David Vogel

Abstract

We discuss the challenges of the 2009 KDD Cup along with our ideas and methodologies for modelling the problem. The main stages included aggressive nonparametric feature selection, careful treatment of categorical variables and tuning a gradient boosting machine under Bernoulli loss with trees.

Keywords: KDD Cup 2009, nonparametric feature selection, generalised boosting machine, decision trees

1. Introduction

The KDD Cup 2009[1] was organised by Orange Labs, France. The data consisted of information about telecommunication customers, with 15,000 predictor variables. The competition involved producing binary classifiers for three types of consumer behaviour:

- churn, which is whether someone ceases to be a customer,
- appetency, being the propensity to buy a service or product, and
- upselling, where a more profitable or additional service is sold to a customer.

Competitors were provided with a training set of 50,000 observations, with an additional 50,000 in the test set, which was used by the organisers for model evaluation. The measure for predictive accuracy was the area under the ROC curve (AUC), which integrates sensitivity over all possible specificities of the model. The average of the AUC for the three different classification tasks was used to rank competitors. A reduced dataset of 230 variables was also available, which our team did not make use of for our primary entry.

The challenge had a fast component, with predictions for the test data due within 5 days of the full data being released, and a slow component, where predictions had to be submitted within 5 weeks. IBM Research produced the best model for both components, but as the competition rules stated that no team could win both parts, The University of Melbourne team won first prize in the slow component, having the second best model. Table 1 shows the final results for both IBM research and The University of Melbourne. Our model was based entirely on the large dataset, making no use of the other smaller dataset provided to competitors.

	Model			
Team	Churn	Appetency	Upselling	Average
IBM Research	0.7651	0.8819	0.9092	0.85206
Univ. Melbourne	0.7570	0.8836	0.9048	0.84847

Table 1: Final model performance for IBM research and The University of Melbourne.

The dataset provided for the KDD Cup 2009 is typical of many contemporary data-mining problems. There are a large number of observations, which enables many signals to be resolved through the noise, allowing complex models to be fit. There are also a large number of predictors, which is common since companies and other organisations are able to collect a large amount of information regarding customers. However many of these predictors will contain little or no useful information, so the ability to exclude redundant variables from a final analysis is important. Many of the predictors have missing values, some are continuous and some are categorical. Of the categorical predictors, some have a large number of levels with small exposure; that is, a small number of observations at that level. For the continuous variables, the distribution among the observations can have extreme values, or may take a small number of unique values. Further, there is potential for significant interaction between different predictors. Finally, the responses are often highly unbalanced; for instance only 7% of the upselling observations were labelled "1". All these

1. www.kddcup-orange.com

factors need to be considered in order to produce a satisfactory model. Sections 2 to 4 detail the stages of our modelling for the KDD Cup, while Section 5 makes some comment on the computational resources used.

2. Feature selection

As mentioned in the introduction, many of the predictors were irrelevant for predictive purposes and thus needed to be excluded. In fact, some variables were absolutely redundant, having the same entry in all cells. Over 3,000, about 20%, of the variables had either entirely identical observations, or had fewer than 10 observations different to the baseline, so these were obvious candidates for removal.

For those features remaining, we assessed the individual predictive power with respect to the three responses (churn, appetency and upselling). To do this we split the data into two halves, one to make predictions and the other to measure the resulting AUC, so that the measure of predictor performance was directly related to the measure used in the competition. For categorical values, the proportion of positive responses for each level was used as the prediction that was applied to the second half of the data. For continuous variables we separated the observations into bins based on 1% quantiles and used the proportion of positive responses for each quantile bin as the prediction. In both cases missing values were treated as a separate level. An AUC score could then be calculated for each variable using the second half of the training data and the process was repeated to increase reliability

The above feature selection technique is very simple; it involves taking the mean of the responses for each level, and so amounts to a least squares fit on a single categorical variable against a 0-1 response, with the categories in the continuous case defined by quantiles. Despite its simplicity, it had a number of advantages:

- **Speed:** Computing means and quantiles is direct and efficient
- **Stability with respect to scale:** Extreme values for continuous variables do not skew predictions as they would in many models, especially linear models, and the results are invariant under any monotone transformation of the continuous variables. Therefore this is robust to unusual distribution patterns.
- **Comparability between continuous and categorical variables:** Predictive performance of the two types of variables is measured in a similar way and so they are directly comparable.
- **Accommodation of nonlinearities:** Since a mean is estimated for every quantile in the continuous case, nonlinear dependencies are just as likely to be detected as a linear pattern.

Naturally there were some drawbacks to this approach as well. For instance, by underemphasising linear patterns, any genuine linear or nearly linear patterns were less likely to be detected. Also, the choice of 1% for the quantiles was somewhat arbitrary, but judged to maintain a reasonable balance between shape flexibility and reliability. Figure 1 shows the quantile fit for the most important variable in the churn model, as recorded in Table 4, against the response. Although the fit does exhibit substantial noise when compared to

the smoothed overlay, created using a local kernel regression fit, there remains a strong detectable signal and the noise is mitigated by testing on a separate portion of the data. It is also noteworthy that this variable exhibits significant nonlinearity.

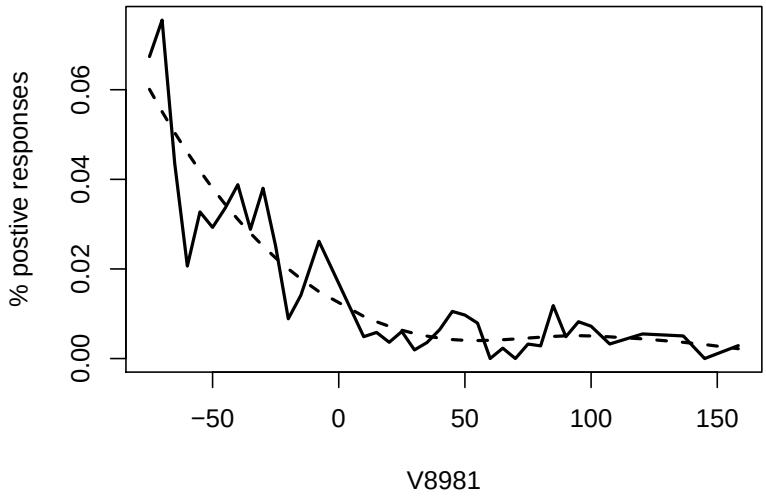

Figure 1: Quantile and smooth fits for variable V8981 against the churn response.

This method of feature selection can be considered as a special case of generalised correlation as in Hall and Miller (2009). There the generalised correlation of the jth variable is the maximum correlation of the response with a nonlinear transformation of that variable:

$$\rho_j = \sup_{h \in \mathcal{H}} \text{cor}\{h(X_j), Y\}\,,$$

where $\mathcal{H}$ is the allowable set of nonlinear transformations. When this set has a finite basis then the choice of h is equivalent to the least squares fit under the basis of $\mathcal{H}$. In our case the finite basis was the collection of quantile based indicator functions (in the continuous case), or indicator functions for each category (for categorical variables). Thus the feature selection may be thought of as maximising the nonlinear correlation between each variable and the response, making use of a large number of degrees of freedom, as permitted by the relatively large number of observations.

The above rankings were reasonably effective in capturing all the interesting behaviour for the churn and appetency models. However for the upselling model, spurious variables tended to appear high in the variable ranking. In this case, the list of top variables needed to be adjusted in the later, more sophisticated modelling stages to produce competitive results.

Figure 2 shows the sorted AUC scores for all the variables using the churn response. The plot is typical of the three different models, with the bulk of predictors having AUC close to 0.5, implying no relationship with the response. The dotted line represents our cutoff for admission into the boosted model of Section 4. The cutoff is reasonably aggressive, but there did not appear to be much gain in admitting more variables. Even if a more conservative cutoff was adopted, considering more than the top 2,000 variables for the final

model appears to be unnecessary, so a substantial dimensionality reduction is possible and preferred.

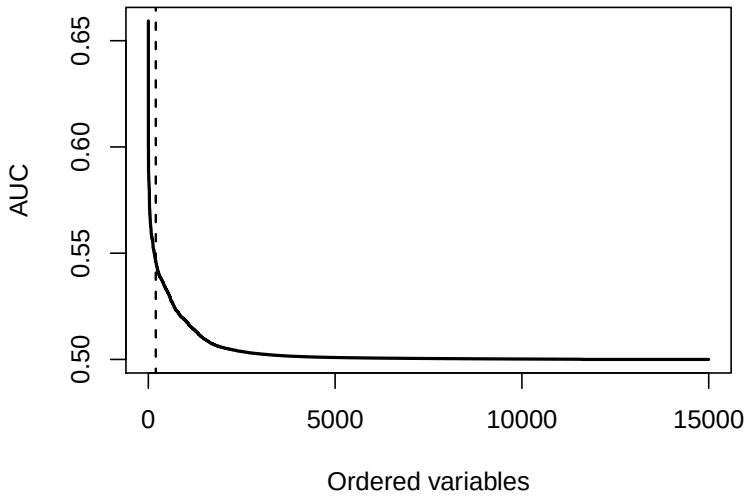

Figure 2: AUC scores for feature selection using churn response.

We also compared the results of this feature selection with a F-score based feature selection method as described in Chen and Lin (2006). In general, agreement was good, although this alternative method suggested a small number of new variables to also include at the modelling stage.

3. Treatment of categorical variables with a large number of levels

Many of the categorical variables had a large number of levels—some even having over 10,000—and some of these ranked high in our feature selection. In many modelling contexts such an abundance of levels is undesirable, since it encourages overfitting on the small exposure levels. This is particularly true of decision trees, which we used for building our final models. Another problem is that some levels appear in the training set but not the test set and vice versa. While some of the levels that had a large exposure were important, the other levels needed aggregation.

Our initial attempt to aggregate was to collapse the levels with less than 1,000 corresponding observations into 20 levels, with grouping based on the response. Thus levels with small exposure and a large proportion of positive responses were grouped together, while those with small exposure but lower proportion would be aggregated in a different level. This was the aggregation we used for the fast part of the challenge. Unfortunately this exacerbated the overfitting problem because we were artificially creating good predictors of the training set which depressed model performance on the test set, so an alternative means of aggregation was necessary.

To prevent this kind of overfitting, our second attempt at aggregation was completed independently of the response. If a categorical variable had more than 25 levels, we created a replacement variable by:

- keeping any levels that had at least 1000 observations worth of exposure,
- aggregating any levels with exposure between 500-999 into a new level,
- aggregating any levels with exposure between 250-499 into a new level, and
- aggregating any levels with exposure between 1-250 into a new level.

This removed the overfitting problem. It is not entirely clear whether the aggregating into three levels based on exposure did in fact provide any improvement compared to using a single level, although there is some supporting evidence. For instance some variables, such as V14788 and V14904, had only levels corresponding to the different exposures and were judged significant in some of our models. Also, Table 2 gives AUC scores on 5-fold cross-validated training set predictions for our final models using the exposure aggregation compared to a single aggregation. The churn and appetency models in particular seem to support the exposure based aggregation. While not conclusive, it is worth noting that if the differences in the table are representative then not including the exposure levels would have lowered the team's ranking.

Model	Exposure-based aggregation	Single level	Difference
Churn	0.7493	0.7478	0.0015
Appetency	0.8790	0.8784	0.0006
Upselling	0.9062	0.9063	-0.0001

Table 2: AUC scores comparing aggregation approaches for categorical variables

Another advantage of this approach compared to the initial attempt was that the processed categorical variables were the same across the three consumer behaviours.

4. Modelling with decision trees and boosting

The basic approach for constructing the final models involved the collection of shallow decision trees with boosting and shrinkage as in gradient boosting machines. Friedman (2001) serves as a primary reference for this approach; other literature on boosting includes Freund and Schapire (1997), Ridgeway (1999), Friedman et al. (2000) and Friedman (2002). Decision trees have been studied for many years, and include the work of Morgan and Sonquist (1963), Breiman et al. (1984) and Quinlan (1993). The basic principle is to fit a relatively simple tree-based model many times, each time focusing on the observations that are hardest to classify correctly by means of a weighting scheme. Bernoulli loss was used to compute the deviance, and the class weights were chosen so that the two classes had roughly equal weight. For example the churn model used a weight of 12 for the positive class, to better balance the trees.

Decision trees have a number of advantages which suited this year's KDD data, in particular. These are well-known, but worth restating here:

- Predictions are possible even when an important predictor has a missing value, through the use of surrogate variables.

- They are not affected by extreme values or strange distributions in continuous variables. In fact, they are invariant under monotone transformations of the predictors.
- They can easily handle both continuous and categorical variables.
- They can effectively model interactions between predictors.
- They allow for nonlinear dependencies.

Model validity was tested both by cross-validation and using the online feedback on the 10% test sample provided by the organisers. We aimed to build a model using about 200 predictors, partly for computational reasons and partly because adding extra predictors to our final subsets did not appear to noticeably improve performance. These variables were chosen on the basis of the feature selection ranking. However an important part of tuning the models involved discarding variables that did not appear useful in the model and adding some lower down the feature selection ranking. Here usefulness refers to the relative amount of deviance (Bernoulli loss) reduction each variable contributes to the model. Details of the variables used in each of the models are given in Appendix A. Model parameters for each of the fits are presented in Table 3. These were selected to maximise the AUC performance, using the test set feedback and cross-validation.

	Model		
	Churn	Appetency	Upselling
Number of variables	198	196	201
Class weight	12	20	12
Shrinkage parameter	0.01	0.01	0.01
Number of trees	1300	1300	3000
Tree depth	5	3	5

Table 3: Model parameters for boosted tree models

The final models suggest that there are some significant interactions between predictors in the models, most strikingly between continuous and categorical variables. Figure 3 shows one example of this, plotting the partial dependence between the two most important variables in the appetency model, V9045 and two levels of V14990. Note that this is not the change in the response excluding the effect of all the other variables, but rather integrating over them. The different behaviour in the continuous variable for the different levels is visible.

5. Computational details

The analysis and modelling work was performed almost entirely in the free open source program R.[2] We say "almost", because the original data chunks were too large to be read into R with our limited hardware, so it was first read into SAS[3] and exported in batches of 200 variables, each of which could then be read into and then deleted from R.

All modelling was conducted on individual desktop and laptop computers; the computer that did the most modelling was a mass-market laptop running Windows XP with 2Gb of

2. http://cran.r-project.org/
3. http://www.sas.com/

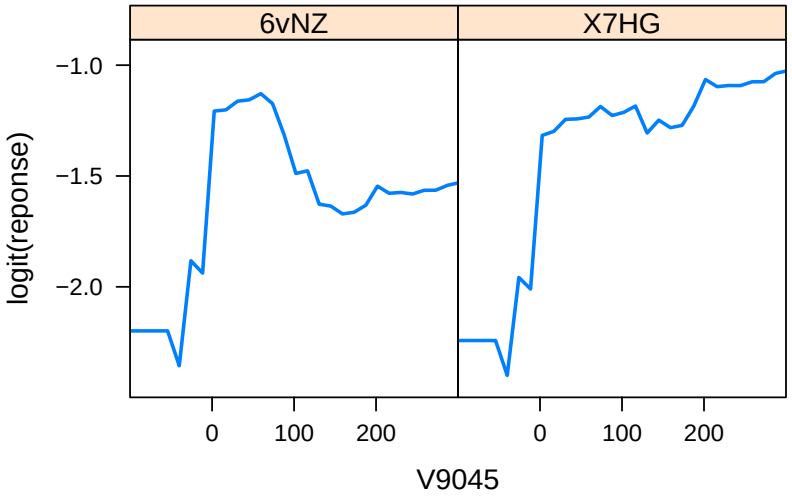

Figure 3: Partial dependence plots in the appetency model for variables V9045 and two levels of V14990. The different shapes, particularly for higher values of V9045, suggest interactions are present.

RAM, a 2.66GHz Intel Core 2 Duo processor and a 120Gb hard drive. The feature selection and categorical collapsing was programmed ourselves, while the boosted decision tree used the "gbm" package, also freely downloadable[2].

The feature selection stage took a few hours of computing time for each response, while the boosted decision tree models typically took just over an hour to fit, depending on the number of trees and variables involved. This demonstrates that a linux cluster is not necessary to produce strong predictive results, although the authors suspect it would help; in our case, it would have enabled more comprehensive investigation of the effect of choices in category collapsing and feature selection. Interested readers are encouraged to contact the first author regarding any questions of coding or computation, or with any suggestions and comments.

References

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and regression trees*. Wadsworth, Belmont, 1984.

Y. W. Chen and C. J. Lin. Combining svms with various feature selection strategies. In *Feature extraction, foundations and applications*. Springer-Verlag, 2006.

Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

J. H. Friedman. Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38(4):367–378, 2002.

J. H. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *Annals of Statistics*, 28(2):337–374, 2000.

P. Hall and H. Miller. Using generalised correlation to effect variable selection in very high dimensional problems. *Journal of Computation and Graphical Statistics (to appear)*, 2009.

J. N. Morgan and J. A. Sonquist. Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, 58:415–434, 1963.

R. Quinlan. *C4.5: Programs for machine learning*. Morgan Kaufmann, San Mateo, 1993.

G. Ridgeway. The state of boosting. *Computing Science and Statistics*, 31:172–181, 1999.

Appendix A. Tables relating to final models

Rank	Churn		Appetency		Upselling	
	Name	Rel. Inf.	Name	Rel. Inf.	Name	Rel. Inf.
1	V8981	20.13	V9045	23.78	V9045	45.52
2	V14990	10.25	V8032	13.56	V14990	7.86
3	V10533	4.65	V14995	10.79	V8981	5.32
4	V14970	4.60	V14990	6.07	V12507	4.96
5	V5331	2.36	V5826	3.72	V6808	4.65
6	V14995	2.19	V8981	3.23	V1194	2.58
7	V14822	2.10	V10256	3.03	V14970	2.16
8	V9045	2.00	V12641	2.72	V14871	1.33
9	V2570	2.00	V14772	1.72	V1782	1.15
10	V14923	1.88	V14939	1.68	V10256	1.05
11	V14765	1.19	V14867	1.62	V5026	0.96
12	V14904	1.14	V14970	1.42	V8032	0.91
13	V5702	1.13	V11781	1.14	V14786	0.81
14	V11047	1.12	V14871	0.89	V7476	0.62
15	V14778	0.97	V14788	0.86	V11781	0.59
16	V14795	0.90	V13379	0.81	V14795	0.57
17	V990	0.90	V5216	0.71	V6255	0.57
18	V12580	0.86	V14795	0.70	V5216	0.50
19	V9075	0.86	V11315	0.66	V2591	0.50
20	V647	0.85	V12702	0.62	V12641	0.46

Table 4: Relative influence of top 20 variables in final models

Churn			Appetency			Upselling		
V47	V5216	V10447	V28	V5723	V11315	V28	V5216	V10136
V173	V5245	V10513	V83	V5808	V11322	V169	V5405	V10256
V384	V5277	V10533	V134	V5826	V11392	V173	V5462	V10402
V559	V5331	V10557	V182	V5873	V11396	V182	V5521	V10443
V621	V5360	V10589	V193	V5899	V11642	V213	V5576	V10521
V635	V5365	V10687	V282	V6003	V11777	V542	V5632	V10538
V647	V5559	V10808	V647	V6016	V11781	V559	V5723	V10687
V698	V5613	V10985	V698	V6238	V11916	V749	V5815	V11051
V706	V5666	V11047	V855	V6310	V12058	V941	V5826	V11083
V724	V5702	V11068	V941	V6424	V12102	V959	V5840	V11092
V749	V5723	V11172	V959	V6468	V12147	V975	V5985	V11115
V843	V5808	V11247	V1026	V6503	V12252	V1004	V6032	V11135
V941	V5820	V11315	V1075	V6565	V12264	V1045	V6228	V11160
V953	V5833	V11322	V1204	V6620	V12321	V1194	V6246	V11196
V990	V5895	V11392	V1275	V6659	V12483	V1362	V6255	V11277
V1036	V5982	V11480	V1476	V6735	V12507	V1376	V6503	V11315
V1095	V6016	V11671	V1514	V6751	V12517	V1596	V6514	V11369
V1227	V6049	V11731	V1543	V6812	V12548	V1623	V6565	V11566
V1254	V6255	V11985	V1596	V6825	V12638	V1782	V6637	V11781
V1392	V6310	V12199	V1969	V7004	V12641	V1853	V6735	V11832
V1428	V6468	V12200	V2120	V7055	V12670	V1925	V6778	V11859
V1501	V6534	V12264	V2157	V7180	V12702	V2095	V6808	V12011
V1565	V6551	V12370	V2284	V7212	V12747	V2120	V6837	V12058
V1604	V6636	V12381	V2334	V7335	V12840	V2157	V6892	V12147
V1996	V6653	V12580	V2352	V7356	V12884	V2249	V6894	V12199
V2284	V6722	V12702	V2413	V7575	V13084	V2321	V7004	V12221
V2315	V7071	V12840	V2418	V7579	V13104	V2434	V7014	V12264
V2370	V7146	V12993	V2453	V7651	V13362	V2453	V7029	V12507
V2450	V7212	V13008	V2531	V7653	V13379	V2531	V7055	V12539
V2453	V7229	V13038	V2544	V7904	V13492	V2591	V7230	V12548
V2456	V7425	V13053	V2591	V7950	V13653	V2849	V7308	V12641
V2570	V7500	V13153	V2715	V7960	V13871	V2852	V7476	V12702
V2773	V7511	V13210	V2822	V8003	V13952	V2890	V7485	V12884
V2822	V7670	V13350	V2849	V8032	V14221	V2892	V7521	V12952
V2852	V7706	V13571	V2852	V8343	V14246	V2985	V7522	V13038
V2961	V7758	V13572	V2966	V8458	V14334	V3128	V7575	V13135
V3080	V7817	V13573	V3000	V8591	V14344	V3219	V7579	V13153
V3104	V7964	V13644	V3128	V8619	V14362	V3305	V7631	V13162
V3264	V8032	V13663	V3130	V8787	V14374	V3487	V7737	V13287
V3305	V8181	V13714	V3199	V8936	V14377	V3558	V7874	V13362
V3339	V8375	V13849	V3202	V8981	V14517	V3568	V7987	V13379
V3439	V8484	V14087	V3219	V9001	V14643	V3711	V8032	V13467
V3508	V8605	V14187	V3249	V9045	V14696	V3962	V8070	V13469
V3515	V8621	V14226	V3305	V9248	V14721	V3999	V8122	V13592
V3624	V8709	V14274	V3339	V9311	V14732	V4048	V8181	V13653
V3719	V8717	V14334	V3704	V9408	V14772	V4075	V8338	V13705
V3759	V8854	V14359	V3719	V9409	V14786	V4221	V8458	V13727
V3766	V8863	V14429	V3759	V9655	V14788	V4316	V8505	V13952
V3886	V8981	V14487	V3863	V9671	V14795	V4566	V8561	V14015
V3905	V9001	V14502	V4186	V9704	V14834	V4585	V8591	V14138
V3972	V9037	V14765	V4248	V10032	V14846	V4614	V8619	V14157
V4028	V9045	V14778	V4340	V10130	V14867	V4659	V8833	V14170
V4088	V9075	V14788	V4347	V10212	V14871	V4665	V8981	V14362
V4098	V9342	V14791	V4585	V10256	V14878	V4686	V9045	V14721
V4218	V9375	V14795	V4590	V10333	V14923	V4735	V9051	V14773
V4389	V9408	V14822	V4614	V10343	V14928	V4802	V9069	V14778
V4393	V9498	V14846	V4665	V10405	V14939	V4856	V9230	V14786
V4563	V9536	V14871	V4902	V10415	V14970	V4996	V9294	V14795
V4669	V9608	V14904	V4957	V10443	V14974	V5021	V9311	V14862
V4735	V9616	V14906	V5026	V10450	V14980	V5026	V9386	V14871
V4856	V9686	V14923	V5065	V10521	V14990	V5053	V9409	V14890
V4986	V9704	V14970	V5185	V10589	V14995	V5065	V9431	V14928
V5025	V9711	V14990	V5213	V10594		V5097	V9574	V14946
V5026	V9799	V14995	V5216	V10739		V5138	V9658	V14965
V5031	V10073		V5405	V10843		V5144	V9708	V14970
V5166	V10183		V5462	V11196		V5182	V9797	V14990
V5170	V10256		V5554	V11247		V5213	V10097	V14995

Table 5: Variables used in final models