Making the Most of Your Data: KDD Cup 2007 "How Many Ratings" Winner's Report

Saharon Rosset, Claudia Perlich,
Yan Liu
IBM T.J. Watson Research Center
P. O. Box 218
Yorktown Heights, NY 10598
{srosset, perlich, liuya}@us.ibm.com

ABSTRACT

We describe the ideas and methodologies that we developed in addressing the KDD Cup 2007 *How Many Ratings* task, and discuss how they contributed to our success.

1. INTRODUCTION — WHAT MAKES A MODEL SUCCESSFUL?

At the Data Analytics Research group at IBM Research we aim to combine theoretical rigor with practical usefulness in our research and the projects we develop for IBM groups and external customers. Our projects often include aspects of data analysis, algorithm development, application development and product delivery [4, 3, 1]. Based on our experience there are several important components to success in modeling data — whether it be in a competition or in real-life modeling problems. One possible characterization of these components divides them into three general categories:

1. **Data and domain understanding.** The focus here is on understanding how the data and modeling problem were generated; how the data can best be used to address the problem at hand; what transformations or preprocessing are required to make the data most appropriate; and how this knowledge can be put together in a coherent manner.

2. **Statistical insights.** This aspect of the modeling process is concentrated on making sure that we make the best and most correct use of the data — based on our data and domain understanding — to get models and insights that are statistically sound and optimal. While the statistical analysis may be closely inter-twined with the data and domain understanding, as the two components feed each other, it is distinct in relying on probabilistic and statistical insights rather than knowledge about the data generating processes.

3. **Modeling or learning approach.** This is the step that typically generates the most "scientific" interest in the data mining and machine learning communities, of choosing and/or developing and/or implementing the best algorithmic approach to solve the modeling problem at hand.

From our experience, the ordering of these categories above is consistent with their typical importance in maximizing the success of practical modeling projects. The ability to understand the data and the domain well, figure out their correct interpretation and appropriate use is by far the most useful way to gain "an edge" and improve models above and beyond what any statistical insights or modeling approaches can. Correctly formulating a statistical or probabilistic framework is also of critical importance, when possible. Finally, in our opinion, the learning approach, while highly influential on bottom-line performance in many cases, cannot be counted on as a way to circumvent the need to understand the data and the statistical setup properly.

In this short paper we use this classification of the elements of modeling success to describe our approach. After describing the general setup of the challenge in Section 2, we detail the data insights we used in Section 3. We discuss out key statistical insights in Section 4, and show how everything comes together to guide our modeling approach in Section 5. Finally, we briefly analyze the competition results and how the different components of our approach affected our performance in Section 6.

2. PROBLEM SETUP

The second task in the KKD-CUP was to predict the total number of reviews that a movie received during 2006 from the whole set of users in the Netflix competition training set. This task can generally be viewed as a regression problem where the number of ratings that a movies receives in a given period of time depends on a number of relevant factors that contribute to the popularity and in turn the number of ratings of a movie. Such factors include its age, arrival in the Netflix database, genre, rating and importantly also the characteristics and history of the roughly 480,000 reviewers.

These factors naturally do not only impact the ratings in the current time frame but also the number of rating that the movie received in previous periods. This suggests a temporal dynamic in the rating with different periods in the movie

KDDCup.07, August 12, 2007, San Jose, California, USA.

life-cycle: prior to Box-office release, prior to DVD release, prior to availability in Netflix and finally the slow decrease of interest as the movie ages. So the historical reviewing behavior of a movie is another vital piece of information to capture the dynamic life-cycle of a movie. Such lagged rating counts can be extracted from the Netflix competition dataset with time-stamped ratings from 1998 through 2005.

One way of formalizing the supervised modeling problem of The *How Many Ratings* task is to build a time-series prediction model that estimates the number of rating in the next time period as a function of past ratings and movie-specific features on historical data and roll the model over to the next time period.

However, there is another less obvious approach that can be taken to formalize The *How Many Ratings* task as a supervised learning problem that takes advantage of the *Who reviewed what* test set as discussed in the next section.

3. OBSERVATIONS ON THE DATA AND THE DOMAIN

We will discuss two important observations about the generation of the test set for the KDD-CUP and training data that strongly affect the design of our modeling approach.

3.1 Using *Who Reviewed What* test set to model *How Many Ratings*

The two tasks for the KDD-CUP were constructed based on the 2006 reviews of the 17770 movies in the Netflix competition dataset. The organizers randomly assigned 8863 movies to The *How Many Ratings* task and the remaining movies were used to construct the test set for *Who Reviewed What*. Let us take a more detailed look at the construction of the *Who Reviewed What* test set and how it can be utilized to build a model for The *How Many Ratings* task.

In order to achieve a reasonably high base rate for the classification task *Who Reviewed What*, the sampling probability for a movie,user pair was based on the product of the marginal rating distributions by movie and user in 2006. The marginal is directly proportional to the number of ratings a movie received in 2006. So the number of times a movie appears in a Task 1 pair is again proportional to the total number of reviews the movie received in 2006.

This suggests a very interesting modeling approach. We can use the number of ratings in the *Who Reviewed What* test set as the dependent variable to estimate a model that predicts the number of 2006 ratings. We can then apply this model to the movies in the *How Many Ratings* test set.

This idea has two major advantages:

1. we capture directly the dynamics of the 2006 rating behavior; and
2. we can make optimal use of all recent data up to end of 2005 to construct independent variables

However, there are two issues to consider. We are still missing a scaling parameter. The counts in the test set of *Who Reviewed What* are relative to a sample of 100,000 movie-rating pairs. What remains unknown and of critical importance is the total number of reviews in 2006 to use as a scaling factor for our prediction. This modeling problem is described in detail in Section 5.

Another important observation is the fact, that the counts in the *Who Reviewed What* test set are not really proportional to the marginal because the organizers had to remove pairs that had received ratings prior to 2006. The probability of rejection is a function of the marginal distribution and affects highly rated movies much more than less popular movies. We resolve this problem by correcting the counts as outlined in Section 4.2.

3.2 Dynamics of the total rating counts

One of the missing piece of information is the scaling parameter that is needed to predict the total number of ratings in 2006. We first analyzed the number of total ratings over time. One initially observation is the clear drop in the rating counts in the last quarter of 2005 - even after including the qualifying dataset. The missing 3 million ratings (or about 20% compared to the third quarter of 2005) need explanation for two reasons. First, the counts in the last quarter are the obvious choice as independent variables for predicting 2006. But also, they are the clear choice of dependent variable for time-series models that will be rolled forward.

So what happened in the last quarter of 2005? Are indeed the dynamics such that all of a sudden the Netflix users decrease their rating activity significantly?

The first explanation can be found in the Netflix description of the training data: the selection criterion for users to be included in the dataset was to have at least 20 ratings as of the end of 2005. This procedure affects a high number of new users that joined only recently and have no accumulated 20 ratings yet. This is easily verified looking at the arrival rates of new users - they decrease clearly in the last quarter of 2005. However, this does not fully explain the phenomenon. The arrival rate of new movies also decreases in the last period. In addition, when we simulate the effect of the user cutoff on earlier dates (say 3rd quarter of 2005), we cannot reproduce a rating decrease comparable to the decline we see in the original data.

We were not able to resolve this issue completely. However, we tried to assess whether this decline was caused by missing movies (which is no problem since the task is limited to included movies) or by missing ratings for included movies. We applied the user cutoff on an earlier quarter (3rd quarter 2005) and trained a model on the resulting 3rd quarter ratings for each movies based on the rating counts in previous quarters. When we role this model forward on the original data to predict the last quarter of 2005 for the existing movies, we find a very similar total number of ratings as in the original. We therefore conclude that whatever causes the overall drop in ratings, is likely to be caused by missing movies, not to missing ratings within movies and we therefore feel comfortable using the quarterly data from 2005 to build time-series models.

4. SOME STATISTICAL OBSERVATIONS

There are two statistical aspects to this data modeling problem that captured our attention.

4.1 Is Poisson the right likelihood?

Consider a set of m objects (in this case, movies) with counts $n_1, ...n_m$ (in this case, number of reviews per movie in a given period of length t). Our first observation is that under mild and reasonable assumptions about the arrival process of new reviews for each movie, these counts have a marginal Poisson distribution:

$$m_i \sim \text{Pois}(\lambda_i \cdot t)$$

Consequently, if we decide to use a linear model (or a kernel-based non-linear model) to describe the dependence of the observed movie counts on a set of features $\mathbf{x}_1, ..., \mathbf{x}_p$ (these are vectors of length n), a good candidate modeling approach would be a Poisson regression, a generalized linear model [2], with the natural (log) link function:

$$log(\lambda_i) = \sum_j \beta_j x_{ij}$$

And a corresponding maximum likelihood modeling problem:

$$\begin{aligned} \hat{\beta} &= \arg\min_\beta -l(\mathbf{n}|\lambda) = \\ &= \arg\min_\beta \sum_i [\lambda_i \cdot t - n_i \cdot \log(\lambda_i \cdot t)] \end{aligned} \quad (1)$$

where l denotes log-likelihood, and t a (known) time-period length for the data collection.

A more interesting situation is when we use the set of counts $\tilde{n}_1, ..., \tilde{n}_m$ from *Who Reviewed What* test set as our modeling target. This test set was sampled proportional to the true counts $n_1, ...n_m$ (subject to the rejection sampling correction we discuss next), and is constrained to sum to a fixed number (say, 100000). It is easy to show that:

$$\tilde{n}_1, ..., \tilde{n}_m | \sum_i \tilde{n}_i = 100000 \sim \text{Multinomial}(100000, p_1, ..., p_m)$$

where $p_i = \lambda_i / \sum_k \lambda_k$ is the relative rate of movie i.

Now, if we look at each of the $\tilde{n}_i$'s, their "marginal" conditional distribution is Binomial$(100000, p_i)$ and since this is a *large n, small p* situation, the distribution of $\tilde{n}_i$ is well approximated by Pois$(100000, p_i)$ distribution. Although we have not proven it rigorously, it seems reasonable that the joint distribution can then be approximated as a product of independent Poissons, since the covariance between $\tilde{n}_i, \tilde{n}_k, i \neq k$ is small compared to their variance:

$$\text{Cov}(\tilde{n}_i, \tilde{n}_k) = 100000 p_i p_k << 100000 p_i(1 - p_i) = \text{Var}(\tilde{n}_i)$$

Thus we propose to use a similar formulation to Eq. (1):

$$\hat{\beta} = \arg\min_\beta \sum_i [\lambda_i - \tilde{n}_i \cdot \log(\lambda_i)]$$

where we have eliminated the known time period t, and we will have to estimate a *scaling factor* as discussed in the previous section to scale the estimate λ_i's to use them for prediction.

4.2 Rejection sampling correction

Our discussion in the previous section assumed that the $\tilde{n}_i$'s were sampled *proportionally* from the original n_i's. As we discussed in the previous section, this is not exactly true, because after this proportional sampling, some of the sampled movies were rejected, based on previously having been ranked (prior to 2006). To obtain $\tilde{n}_i$s that are indeed proportionally sampled this rejection would have to be inverted. Here we describe our algorithm for this inversion.

Let $p_i = \lambda_i / \sum_k \lambda_k$ be the *true* sampling rate for movie i, and $q_j = \eta_j / \sum_l \eta_l$ be the corresponding sampling rate for user j. A naive approach to get an estimate of p_i and q_j is to assume the reviewers in the test set were chosen uniformly randomly. Then we could correct this sampling effect easily as follows:

$$\widetilde{q}_j = n_j^{<2006} / \sum_k n_k^{<2006}$$

$$\widetilde{n}_i = n_i/(1 - \widetilde{q}_i),\ \widetilde{N} = \sum_i \widetilde{n}_i$$

$$\widetilde{p}_i = \widetilde{n}_i / \widetilde{N}$$

where $n_j^{<2006}$ is number of reviews by user j before 2006.

A more sophisticated approach to estimate p and q can be achieved by a simultaneous estimation approach. Suppose the sample size 100,000 is large enough for us to estimate p_i and q_j. We have a hidden variable, that is, the number of samples rejected because they have appeared before 2006, which is denoted as N. We have the following constraints so that p_i, q_j, and N have to satisfy:

$$\sum_i p_i = 1,\ \sum_j q_j = 1$$

In addition, we observe n_i appearances of movie i in the final sample set, which satisfies:

$$E[n_i|N] = p_i(100,000 + N)(1 - \sum_{t \in U_t} q_t), \quad (2)$$

where U_t is the set of users that has reviewed the movie t before 2006. On the right hand side of eq(2), the first product corresponds to the total number of samples with movie i (before rejection), and the last term is the proportion of pairs that are been eliminated because they appear before 2006. Similarly, we observe m_j appearances of user j in the final sample set, which satisfies:

$$E[m_j|N] = q_j(100,000 + N)(1 - \sum_{k \in M_k} p_k), \quad (3)$$

where M_k is the set of movies that has been reviewed by user j before 2006. We implemented an ad-hoc iterative procedure for solving the equations (2,3), by alternating between fixing the q_j's and solving (2), and fixing the p_k's and solving (3). This gives us a more accurate estimate of p_k, q_j and N (our interest is, of course focused on the p_k's). This correction can be thought of as increasing the marginal for movies that are likely to have been rejected a lot, because they have been heavily reviewed before 2006, while also taking into account which reviewers reviewed them.

5. MODELING APPROACH

The culmination of all the discussion in the previous sections led us to the learning approach we took, which we describe here briefly:

1. Extract a list of features for each movie:
   - log(Number of reviews by month for the most recent quarter+1) (three features)
   - log(Number of reviews by quarter for the most recent year+1) (four features)
   - log(Number of reviews by year for the last four years+1) (four features)
   - Movie's age in the Netflix database (days since first review), capped at two years, and also transformed into log and square scale. (three features)
   - Some characteristics of the movie's ratings (% of 5s, average rating, etc.) (typically one-two features)
   - Movie's Genre (categorical feature, usually taking only the most common genres and binning all the rest into "other")
2. Use the test set of *Who Reviewed What* as a *response* for training a model:
   - Apply the rejection sampling correction discussed in Section 4.2.
   - Build a Poisson regression model describing log(Poisson rate for movie i in *Who Reviewed What* 2006 test set) as a function of the features extracted from the full Netflix dataset (i.e., all reviews until 12/2005, including those in the Netflix qualifying set)
3. Go through a separate modeling exercise to estimate the *scaling factor*, i.e., the total number of reviews that were given to all movies in 2006:
   - Create *lagged* datasets, which are "anchored" in previous quarters. For example, for a lagged dataset for Q205 would only contain movies and reviewers which appeared in the Netflix data before end of June 2005. This is consistent with our modeling approach in some of our business modeling projects [4].
   - With these, build predictive models which use subsequent quarters as response. For example, for the Q205 lagged dataset, we may build a model which uses Q405 review numbers as response, *when limited to the set of movies and reviewers who were active by end of Q205*. These numbers would be much smaller than the actual numbers we see for Q405 in the complete dataset, which contains all movies and reviewers. This gives us a *two quarters ahead* prediction model.
   - These models can now be applied to our complete dataset to predict numbers of reviews in 2006. For example, applying the model built on the Q205 lagged dataset with Q405 as response, to the full Netflix dataset would comprise a prediction for Q206 (two quarters ahead of Q405 when the Netflix dataset terminates).
   - This prediction can be used either as an actual prediction for 2006 movie review counts, if it is better than the predictions generated by the models we built in step 2 above (see below on evaluation strategies for determining whether this is the case); or, if they are not better, as was the case in our evaluation, they can be used to determine the scaling factor between the *Who Reviewed What* test set and the total review numbers. This scaling factor can then be applied to the model's predictions on the *How Many Ratings* movies.

This schematic description glosses over many details, like feature selection, interaction selection, exact form of the Poisson regression models, etc. We next discuss in a little more detail the critical elements of model evaluation and model selection.

Figure 1 gives a graphical representation of our modeling approach as described here.

5.1 Internal evaluation, validation and model selection

Our best asset for evaluation is the same as for modeling — the *Who Reviewed What* test set, after the rejection sampling correction. It can be used in a straight forward manner to evaluate the models built on it, through a cross validation approach or training-test splits.

For models built on the lagged datasets, the exercise is less trivial. To use *Who Reviewed What* test set for evaluation we need to invert the sampling scheme. To avoid various complications that stem from this, and to give the lagged models the best opportunity to surpass the *Who Reviewed What*-based models in terms of performance, we actually find the best possible scaling parameter to the lagged models predictions in terms of their performance on *Who Reviewed What* test set.

The end result of all these evaluations are model-performance scores for all models we consider for prediction.

The actual numbers we obtained for the models we had, after some work in optimizing the models within each class (*Who Reviewed What*-based vs. lagged data-based), were:

- *Who Reviewed What*-based model's prediction: hold-out RMSE on log scale about 0.24
- Lagged models' prediction: log-scale RMSE of about 0.31 on *Who Reviewed What* test set

We concluded that we should use the models we build on *Who Reviewed What* test set for prediction, and the models built on lagged datasets for scaling only.

6. ANALYSIS OF COMPETITION RESULTS

The log-scale MSE of our winning model on the *How Many Ratings* task 2006 reviews was 0.263. This error has two components:

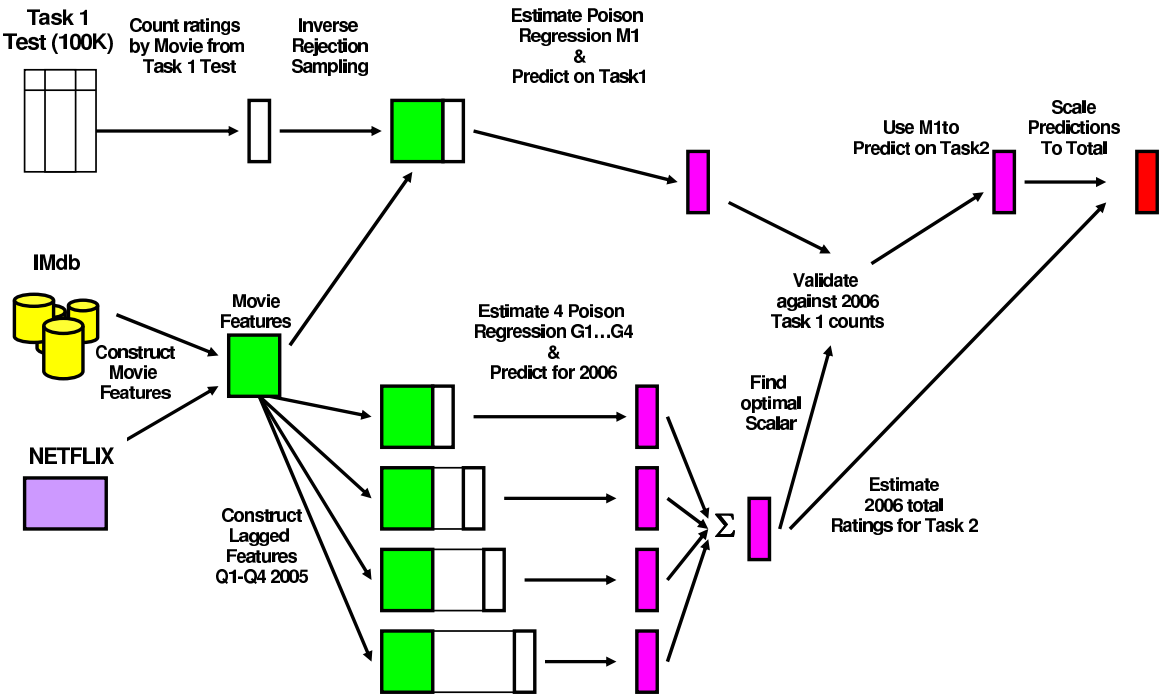

Figure 1: A schematic of our overall modeling approach

Table 1: Results for the top performing teams on the *How Many Ratings* task

Team	Score (RMSE)	MSE
IBM Research	0.513	0.263
NeoMetrics	0.523	0.273
Inductis	0.541	0.292
#4	0.607	0.368

- The error of the model for the scaled-down *Who Reviewed What* test set (which we estimated at about 0.24)
- The error from our incorrect scaling factor, i.e., the mismatch between the scaling factor we estimated from the lagged models and the true correct scaling factor.

In our case, the sum of our predictions was 9.35 million, and the sum of true responses was 8.7 million. Table 2 details the scores we would have attained if we had scaled our predictions differently. By correctly scaling to 8.7 million total, we would have attained MSE of about 0.234. Interestingly, by scaling down further, our best score improvement due to scaling only could have been as far as 0.208 MSE with a scaling factor of 0.8. This is possibly due to the quirkiness of the behavior of Poisson noise under the log transformation: the roughly symmetric noise (for large Poisson parameter) becomes long-left-tailed under the log transformation, and hence consistent under-prediction may lead to better performance.

7. CONCLUSION

Table 2: The effect of scaling on competition MSE. The first column is a hypothetical scaling factor applied to our submitted predictions, the second is the implied total 2006 reviews in millions, and the third the competition score.

Scaling	Total (mil.)	score	Comment
0.7	6.5	0.222	
0.8	7.5	**0.208**	Best performance
0.9	8.4	0.225	
0.93	**8.7**	0.234	Correct scale
1	9.35	**0.263**	Our actual score
1.1	10.285	0.316	

We can summarize our KDD Cup 2007 *How Many Ratings* experience in three short bullets:

- **We had fun** dealing with the data and understanding it, trying out different modeling approaches and speculating about outcomes.
- **We did well** and we believe that a combination of reasons drove this success, but possibly our "bootstrapping" of *Who Reviewed What* test set for training was the most important factor.
- **We encountered some interesting research problems**, most notably the *inverse rejection sampling* problem discussed in Section 4.2. While we applied the inversion here to correct an artifact of the competition sampling, we expect that this inversion problem may be encountered in various real-life problems. We

expect that that this question and others we have encountered will fuel our future research.

Acknowledgments

We thank Rick Lawrence and Zhenzhen Kou for help in useful discussions and in data formatting.

8. REFERENCES

[1] R. Lawrence, C. Perlich, S. Rosset, J. Arroyo, M. Callahan, M. Collins, A. Ershov, S. Feinzig, I. Khabibrakhmanov, S. Mahatma, M. Niemaszyk, and S. Weiss. Analytics-driven solutions for customer targeting and sales force allocation. *IBM Systems Journal*, 2007. To appear.

[2] P. McCullagh and J. Nelder. *Generalized Linear Models*. Chapman and Hall, London, 1989.

[3] C. Perlich, S. Rosset, R. Lawrence, and B. Zadrozny. High quantile moldeing for customer wallet estimation with other applications. In *KDD07*, 2007.

[4] S. Rosset and R. Lawrence. Data enhanced predictive moldeing for sales targeting. In *SIAM Data Mining*, 2006.