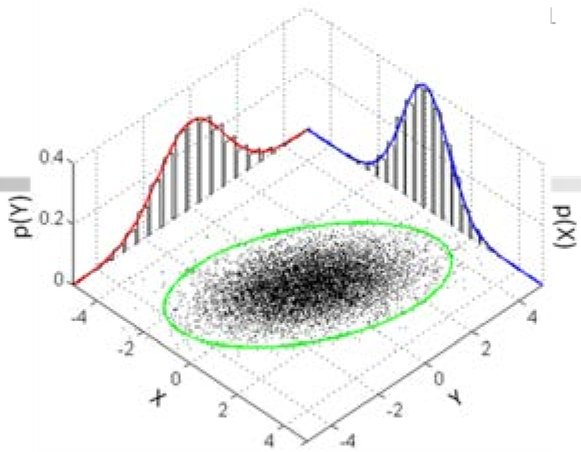
An Introduction to Binary Outcome Bayesian Classifiers

Stephen Denton

Some basic probability theory…



Joint probability – the joint probability of two (or more) events (X and Y).

$$P(X \cap Y) = P(X,Y)$$

Conditional probability – the probability of an event occurring given that another event is known to have occurred (Y given X).

$$P(Y|X)$$

How joint and conditional probabilities are related.

$$P(X,Y) = P(X) * P(Y \mid X) = P(Y) * P(X \mid Y)$$

Bayes Theorem

$$P(Y \mid X) = \frac{P(Y) * P(X \mid Y)}{P(X)}$$

Some more probability theory…

Bayes Theorem

$$P(Y \mid X) = \frac{P(Y) * P(X \mid Y)}{P(X)}$$

When there is more than one predictor variable:

$$P(Y \mid X_1, X_2, \ldots, X_p) = \frac{P(Y) * P\left(X_1, X_2, \ldots, X_p \middle| Y\right)}{P\left(X_1, X_2, \ldots, X_p\right)}$$

$$P(Y \mid \bar{X}) = \frac{P(Y) * P(\bar{X} \mid Y)}{P(\bar{X})}$$

Why do we care?

We are interested in predicting the likelihood that our client will perform some behavior:

- product purchase
- transaction
- churn
- default
- etc.

We will assume that Y is a binary outcome (the event either occurs, 1, or does not, 0).

$$P(Y = 1) = P(Sale) = P(S)$$

$$P(Y = 0) = \mathrm{P}(No\ Sale) = \mathrm{P}(N)$$

$$P(S \mid \bar{X}) = \frac{P(S) * P(\bar{X} \mid S)}{P(\bar{X})}$$

Motivating example - Foreign Exchange (FX) Model

Model Objective:

- Predict the likelihood of a client with a deposit account to perform an online foreign exchange (FX) currency request within the next 4 months.

Model Target:

- There were ~9k clients that requested a FX cash withdrawal in the 4 month period from Dec. 2015 to Mar. 2016.

Model Data:

- Used historical data from June 2015 through Nov. 2015
- Kept all targets and randomly selected 10% of 5.3M non-targets to generate sample dataset
- Sample dataset was split randomly into a development and validation dataset.

The full Bayesian model

Given there are only two outcomes we can write the **full Bayesian binary outcome model**:

$$P(S\,|\bar{X}) = \frac{P(S) * P(\bar{X} \mid S)}{P(\bar{X})}$$

$$P(S \mid \bar{X}) = \frac{P(S) * P(\bar{X} \mid S)}{P(S) * P(\bar{X} \mid S) + P(N) * P(\bar{X} \mid N)}$$

The full Bayesian model

Given there are only two outcomes we can write the **full Bayesian binary outcome model**:

$$P(S\,|\bar{X}) = \frac{P(S) * P(\bar{X} \mid S)}{P(\bar{X})}$$

$$P(S \mid \bar{X}) = \frac{P(S) * P(\bar{X} \mid S)}{P(S) * P(\bar{X} \mid S) + P(N) * P(\bar{X} \mid N)}$$

$$P(S \mid \bar{X}) = {}^{1}\Big/\left(1 + \frac{P(N) * P(\bar{X} \mid N)}{P(S) * P(\bar{X} \mid S)}\right)$$

$$P(S \mid \bar{X}) = {}^{1}\Bigg/\left(1 + exp\left(-ln\left(\frac{P(S) * P(\bar{X} \mid S)}{P(N) * P(\bar{X} \mid N)}\right)\right)\right)$$

$$P(S \mid \bar{X}) = logistic\left(ln\left(\frac{P(S)}{P(N)}\right) + ln\left(\frac{P(\bar{X} \mid S)}{P(\bar{X} \mid N)}\right)\right)$$

For any binary outcome classification problem, given a fixed set of predictor variables, $\bar{X}$, this is the optimal classifier.

So why not use the full Bayesian binary outcome model?

The full Bayesian binary outcome model:

$$P(S\,|\bar{X}) = \frac{P(S) * P(\bar{X}\mid S)}{P(S) * P(\bar{X}\mid S) + P(N) * P(\bar{X}\mid N)}$$

$$P(\bar{X}\mid S) = P\left(X_1, X_2, X_3 \ldots, X_p \mid S\right)$$

$$P(\bar{X}\mid S) = P(X_1|S) * P(X_2|S, X_1) * P(X_3|S, X_1, X_2) * \cdots * P\left(X_p\middle|S, X_1, X_2, \ldots, X_{p-1}\right)$$

Assume (unrealistically) that each X is discrete with only 2 possible values, you need to estimate 2^p unique probabilities…

A Naïve Assumption

Given the computational burden of computing the full Bayesian solution, you can make a simplifying assumption that all predictors, $X_1,\ldots X_p$, are conditionally independent given the outcome, Y.

$$P(\overline{X}\mid S) = P(X_1|S) * P(X_2|S,\cancel{X_1}) * P(X_3|S,\cancel{X_1},\cancel{X_2}) * \cdots * P\big(X_p\big|S,\cancel{X_1},\cancel{X_2},\ldots,\cancel{X_{p-1}}\big)$$

$$P(\overline{X}\mid S) \coloneqq P(X_1|S) * P(X_2|S) * P(X_3|S) * \cdots * P\big(X_p\big|S\big)$$

$$P(\overline{X}\mid S) \coloneqq \prod_{j=1}^{p} P(X_j|S)$$

A Naïve Assumption

Given the computational burden of computing the full Bayesian solution, you can make a simplifying assumption that all predictors, $X_1,\ldots X_p$, are conditionally independent given the outcome, *Y*.

$$P(\bar{X} \mid S) = P(X_1|S) * P(X_2|S,\cancel{X_1}) * P(X_3|S,\cancel{X_1},\cancel{X_2}) * \cdots * P\left(X_p\middle|S,\cancel{X_1},\cancel{X_2},\ldots,\cancel{X_{p-1}}\right)$$

$$P(\bar{X} \mid S) \coloneqq P(X_1|S) * P(X_2|S) * P(X_3|S) * \cdots * P\left(X_p\middle|S\right)$$

$$P(\bar{X} \mid S) \coloneqq \prod_{j=1}^{p} P(X_j|S)$$

Conditional Independence Definition:

We say *X* is conditionally independent of *Y* given *Z*, if and only if the probability distribution governing *X* is independent of the value of *Y* given Z (i.e., $P(X|Y,Z) = P(X|Z)$)

As an example, consider the current weather described as 3 Boolean random variables: *Thunder* is conditionally independent of *Rain* given *Lightning*.

$$P(Thunder \mid Lightning, Rain) = P(Thunder \mid Lightning)$$

Once we know the value of *Lightning*, no additional information is provided by knowing the value of *Rain*.

The Naïve Bayesian Classifier

If we substitute the product of the marginal probability densities for the full joint probability distribution, we get the naïve Bayesian classifier

$$P(S\ |\bar{X}) = logistic\left(ln\left(\frac{P(S)}{P(N)}\right) + ln\left(\frac{P(\bar{X}\mid S)}{P(\bar{X}\mid N)}\right)\right) \qquad P(\bar{X}\mid S) \rightarrow \prod_{j=1}^{p} P(X_j|S)$$

The Naïve Bayesian Classifier

If we substitute the product of the marginal probability densities for the full joint probability distribution, we get the naïve Bayesian classifier

$$P(S\,|\bar{X}) = logistic\left(ln\left(\frac{P(S)}{P(N)}\right) + ln\left(\frac{P(\bar{X}\mid S)}{P(\bar{X}\mid N)}\right)\right) \qquad P(\bar{X}\mid S) \rightarrow \prod_{j=1}^{p} P(X_j|S)$$

$$P(S\,|\bar{X}) = logistic\left(ln\left(\frac{P(S)}{P(N)}\right) + ln\left(\prod_{j=1}^{p}\frac{P(X_j\mid S)}{P(X_j\mid N)}\right)\right)$$

$$P(S\,|\bar{X}) = logistic\left(ln\left(\frac{P(S)}{P(N)}\right) + \sum_{j=1}^{p} ln\left(\frac{P(X_j\mid S)}{P(X_j\mid N)}\right)\right)$$

$$P(S\,|\bar{X}) = logistic\left(\alpha + \sum_{j=1}^{p} g_j(x_j)\right)$$

$$\alpha = ln\left(\frac{P(S)}{P(N)}\right) \qquad g_j(x_j) = ln\left(\frac{P(X_j\mid S)}{P(X_j\mid N)}\right)$$

Alpha is the prior log odds (of a sale) – estimated from historical data.

The marginal effect of a predictor variable

The term

$$g_j(x_j) = ln\left(\frac{P(X_j \mid S)}{P(X_j \mid N)}\right)$$

is the log odds of the predictor likelihoods or the marginal effect of X_j - also known as the weight-of-evidence (WOE) for X_j or the naïve effect.

For discrete predictor variables simple histogram estimators can be used to estimate the probabilities.

For continuous predictor variables, the marginal effects can be estimated using parametric or nonparametric density estimation.

Each can be estimated from historical data.

Example: Estimating Weight-of-Evidence

$$g_j(x_j) = ln\left(\frac{P(X_j \mid S)}{P(X_j \mid N)}\right)$$

For discrete predictor (e.g., gender) with 3 levels (F, M, Missing)

Level	*N:Y=0*	*S:Y=1*	*P(X|N)*	*P(X|S)*	*g(x)*
Female	133743	2297	0.508	0.516	0.0162
Male	123635	2100	0.469	0.472	0.0054
Missing	6081	54	0.023	0.012	-0.6432
Total	263459	4451	1	1	

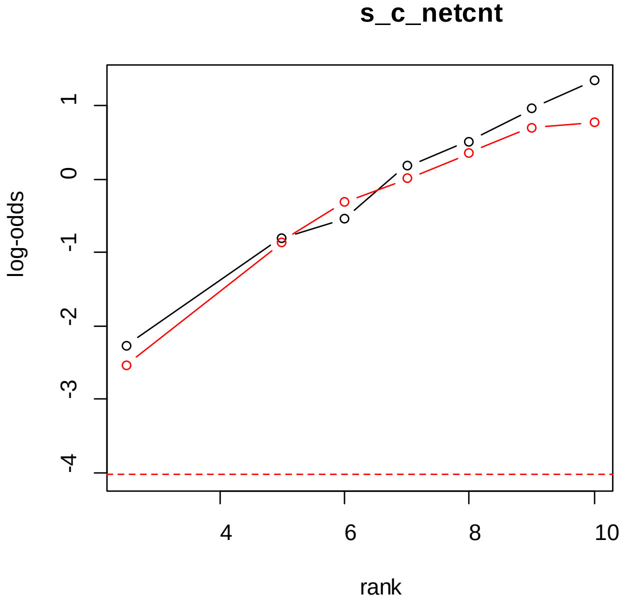
A $g_j(x_j)$ close to zero indicates average odds or no change in the log odds of a sale due to the predictor variable relative to baseline.

Example: Estimating Weight-of-Evidence

A continuous predictor: # of internet banking transactions

Level	*N:Y=0*	*S:Y=1*	*P(X|N)*	*P(X|S)*	*g(x)*
Missing	6610	1	0.025	0.000	-4.716
0	115863	202	0.440	0.045	-2.271
1-3	15998	121	0.061	0.027	-0.804
4-11	24466	240	0.093	0.054	-0.544
12-24	27264	554	0.103	0.124	0.185
25-40	24621	695	0.093	0.156	0.513
41-67	24327	1075	0.092	0.242	0.962
68+	24310	1563	0.092	0.351	1.336
Total	263459	4451	1	1	

$$g_j(x_j) = ln\left(\frac{P(X_j \mid S)}{P(X_j \mid N)}\right)$$



For continuous predictors, marginal conditional densities can be estimated using K nearest neighbors density estimation with missing values treated as their own separate neighborhood.

Weight-of-evidence values are smoothed for the final model.

The Naïve Bayesian Classifier

$$P(S\,|\bar{X}) = logistic\left(\alpha + \sum_{j=1}^{p} g_j(x_j)\right)$$

$$\alpha = ln\left(\frac{P(S)}{P(N)}\right) \qquad g_j(x_j) = ln\left(\frac{P(X_j\,|\,S)}{P(X_j\,|\,N)}\right)$$

The NBC is simply the sum of the marginal effect/WOE functions.

If it is the case that each X_j is conditionally independent of all other X_j's given the outcome Y, then the NBC is the optimal classifier.

However, in practice, the conditional independence assumption is rarely true.

Summing the marginal effects of each X_j produces a highly biased classifier that generally overshoots the true probabilities.

Although estimated probabilities are heavily biased, they have low variance and are generally rank-ordered correctly.

Thus, despite its simplistic and unrealistic assumptions, the NBC is rarely systematically outperformed when compared to more sophisticated classification techniques.

Logistic Regression

There are obvious connections between the NBC and logistic regression

$$P(S\,|\bar{X}) = logistic\left(\alpha + \sum_{j=1}^{p} g_j(x_j)\right)$$

The logistic regression model

$$P(S\,|\bar{X}) = logistic\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j\right)$$

Logistic regression is a parametric algorithm that uses training data to directly estimate $P(S\,|\bar{X})$.

Bayesian classifiers that relax the independence assumption

If you create a hybrid of the NBC and logistic regression you have:

The semi-naïve Bayesian classifier (SNBC)

$$P(S\,|\bar{X}) = logistic\left(\alpha + \sum_{j=1}^{p} \beta_j g_j(x_j)\right)$$

Here a set of parameters, $\beta_j' s$, are multiplied by the marginal effects. The resulting adjusted effects, $\beta_j g_j(x_j)$, have the same shape as the marginal effects but have been linearly scaled to better approximate $ln\left(\frac{P(\bar{X}|S)}{P(\bar{X}|N)}\right)$.

The SNBC can be estimated quickly and efficiently using the same methods used to estimate parameters in logistic regression, and hence can be very useful for variable selection.

More Bayesian classifiers that relax the independence assumption

If you create a hybrid of the NBC and generalized additive models (GAMs) you have:

The generalized naïve Bayesian classifier (GNBC)

$$P(S\ |\bar{X}) = logistic\left(\alpha + \sum_{j=1}^{p}\left(g_j(x_j) + b_j(x_j)\right)\right)$$

Where $\sum_{j=1}^{p}\left(g_j(x_j) + b_j(x_j)\right)$ is an additive approximation of $ln\left(\frac{P(\bar{X}|S)}{P(\bar{X}|N)}\right)$.

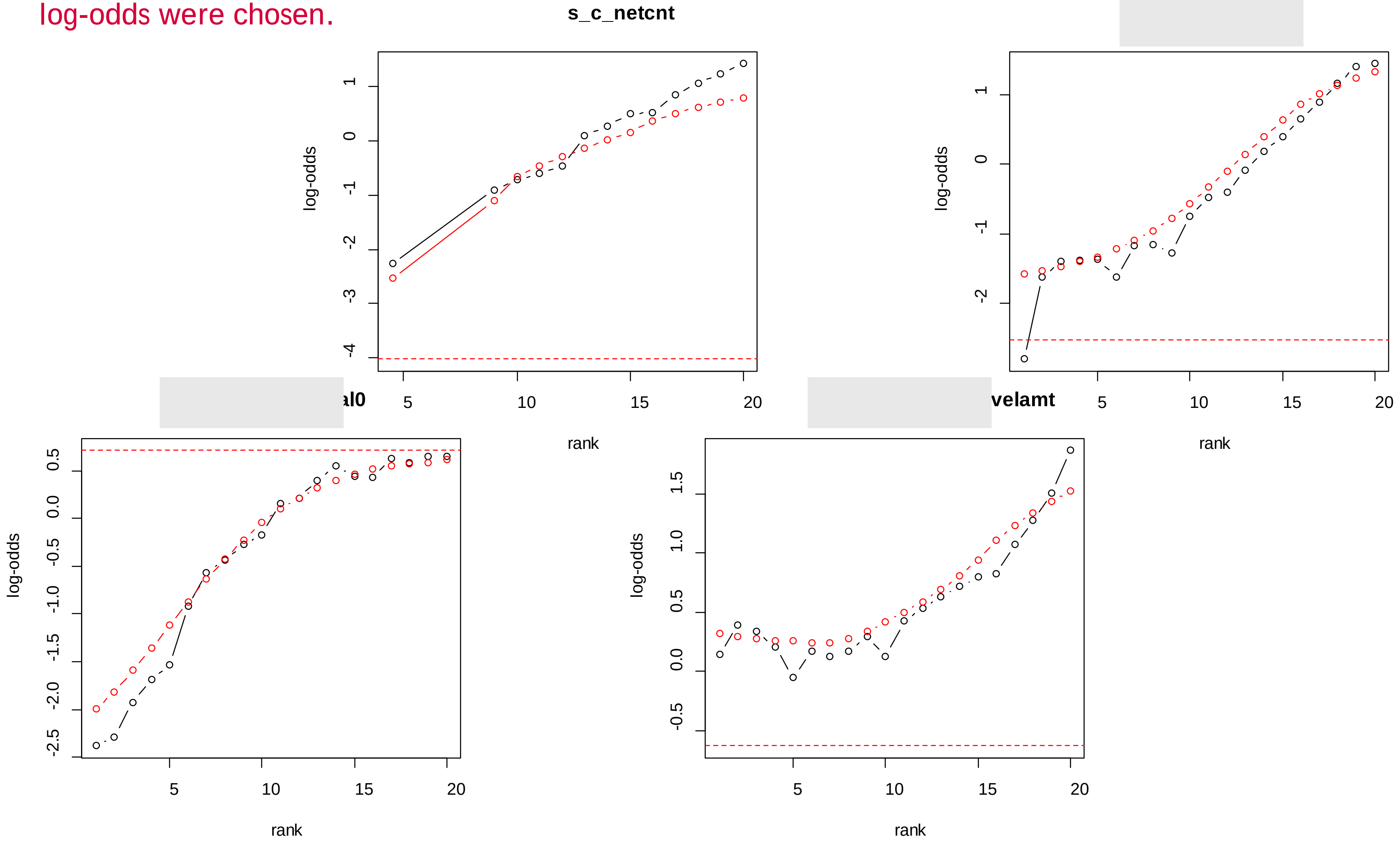
Each adjusted effect, $g_j(x_j) + b_j(x_j)$, can be interpreted as the marginal effect of X_j adjusted for all other predictor variables because the adjustment function $b_j(x_j)$ accounts for the marginal bias attributable to X_j from all other predictor variables.

For discrete predictors $b_j(x_j)$ is a step-function, and for continuous predictors, it is a smooth function.

Given $b_j(x_j)$ is a function, GNBC can adjust the shape of the marginal effects, hence GNBC is more flexible than SNBC (but not as computationally friendly).

Back to the FX Model

For the final model, 4 predictor variables that have nice linear relationship with the predicted variable in log-odds were chosen.



Main Features of FX currency requesters – Key Drivers

Clients who are more likely to initiate a foreign exchange (FX) currency request are more likely to have the following characteristics:

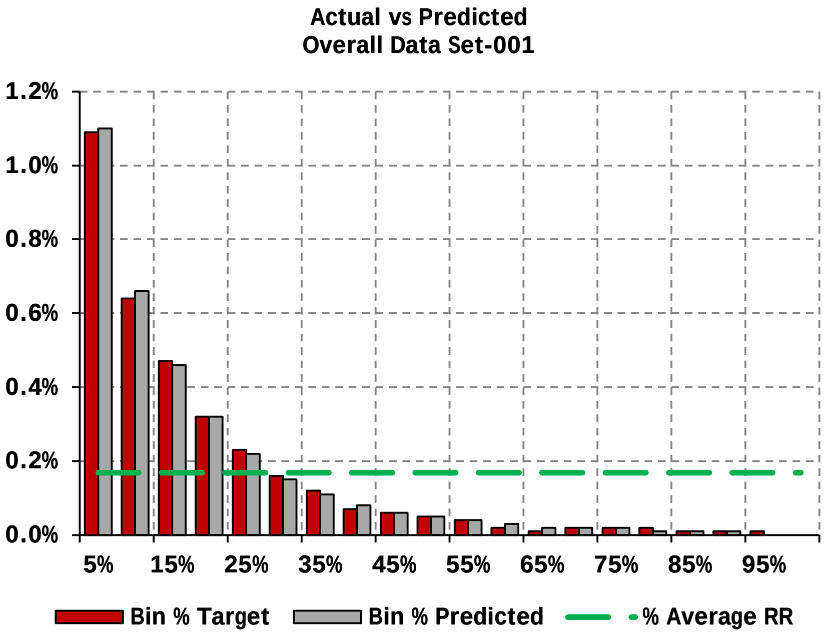
- Perform more internet banking transactions using their deposit account.
- …

Decile	# Clients	# Acquirers	# Predicted	% Actual Rate	% Predicted Rate	PCA internet banking transactions (6 month sum)
1	530,031	4,587	4,661	0.87%	0.88%	81.9
2	530,031	2,079	2,055	0.39%	0.39%	51.0
3	530,031	1,052	972	0.20%	0.18%	38.0
4	530,032	506	508	0.10%	0.10%	21.0
5	530,031	295	293	0.06%	0.06%	13.8
6	530,031	160	165	0.03%	0.03%	6.5
7	530,032	87	114	0.02%	0.02%	0.7
8	530,031	83	83	0.02%	0.02%	0.6
9	530,031	46	43	0.01%	0.01%	0.1
10	530,032	26	20	0.00%	0.00%	0.0
Total	**5,300,313**	**8,921**	**8,915**	**0.17%**	**0.17%**	21.9

FX Model Performance: Overall Model

5.3M personal clients who had a deposit account at the end of November 2015:

- 0.17% requested foreign currency in the 4 months between Dec 1, 2015 and March 31, 2016.
- Out of the top 5% of clients (265k), 2.9k of them (1.1%) performed a FX in the next 4 months, which is 6.5 times better than average (0.17%).
- Top 30% of the model captures 87% of the requesters.


Actual vs Predicted
Overall Data Set-001

Depth	# of Clients	Bin % Target	Bin % Predicted	Cum Actual Lift	Cum % Target
5%	265,016	1.09%	1.10%	6.49	32.5%
10%	265,016	0.64%	0.66%	5.14	51.4%
15%	265,015	0.47%	0.46%	4.36	65.3%
20%	265,016	0.32%	0.32%	3.74	74.7%
25%	265,016	0.23%	0.22%	3.27	81.6%
30%	265,015	0.16%	0.15%	2.88	86.5%
35%	265,016	0.12%	0.11%	2.57	90.0%
40%	265,016	0.07%	0.08%	2.3	92.2%
45%	265,015	0.06%	0.06%	2.09	94.1%
50%	265,016	0.05%	0.05%	1.91	95.5%
55%	265,016	0.04%	0.04%	1.76	96.6%
60%	265,015	0.02%	0.03%	1.62	97.3%
65%	265,016	0.01%	0.02%	1.5	97.7%
70%	265,016	0.02%	0.02%	1.4	98.3%
75%	265,015	0.02%	0.02%	1.32	98.7%
80%	265,016	0.02%	0.01%	1.24	99.2%
85%	265,016	0.01%	0.01%	1.17	99.5%
90%	265,015	0.01%	0.01%	1.11	99.7%
95%	265,016	0.01%	0.00%	1.05	99.9%
100%	265,015	0.00%	0.00%	1	100.0%
Total	**5,300,313**	**0.17%**			

Summary and questions

- Probability theory and Bayes theorem dictate how to formulate an optimal classifier for any classification problem.
- This a optimal Bayesian model that is generally unattainable computationally.
- The best we can do is approximate the Bayesian ideal with ever increasing precision given additional data and computational power.

Further reading:

http://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf

http://www.kdd.org/exploration_files/11-Larsen.pdf