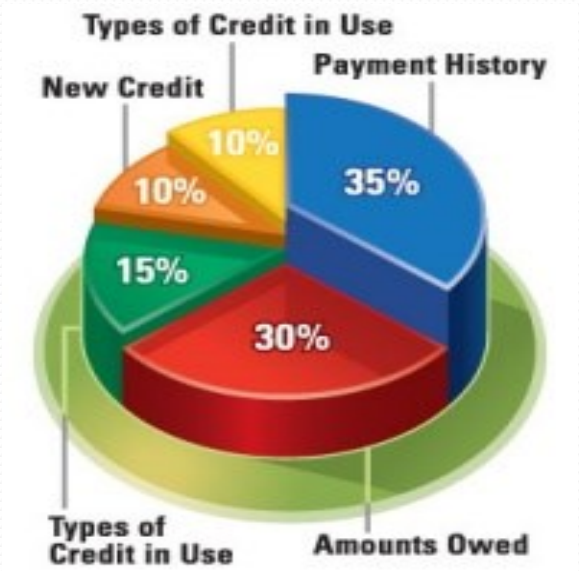
Improving Credit Scoring with Random Forests

Better credit models benefit us all



Agenda

- Credit Scoring - Overview
- Random Forest - Overview
- Random Forest outperform logistic regression for credit scoring out of the box
- Interaction term hypothesis for random forest advantage
- I* algorithm and Synthetic experiment
- Conclusion

Credit Scoring - Overview

- One of the most successful operations research/data science techniques
- Over-arching purpose is to predict risk of serious delinquency default
- Unique characteristics of credit scoring models include:
  - Highly correlated variables (multicollinearity)
  - P-values are relatively unreliable
  - Alternative specifications of models using completely different but correlated variables give similar results
- High visibility (it impacts us all very personally)
  - U.S. law mandates that components of the score must be analyzed to ensure fair and equal treatment
  - Borrowers must be given top 3 denial reasons for decision most important variables affecting their score
- Guide to credit scoring in r ref: http://*cran.r-project.org/doc/contrib/Sharma-CreditScoring.pdf*

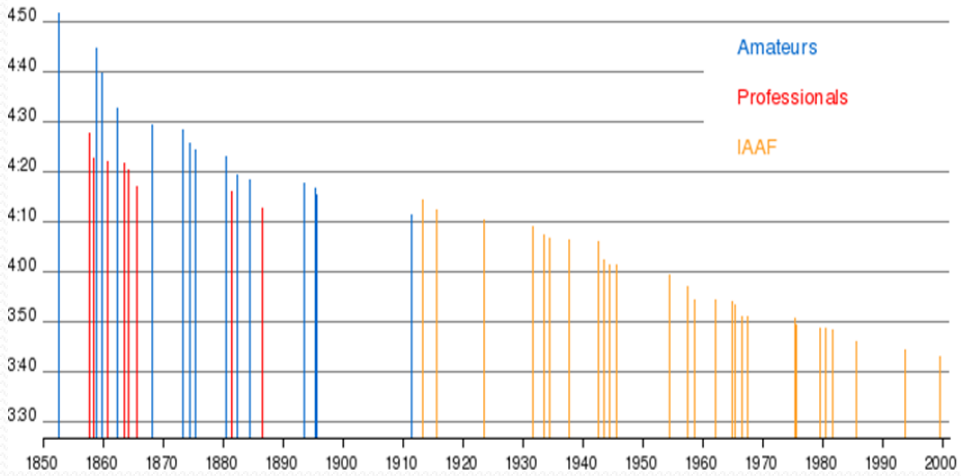
Roger Bannister and credit scoring…

Credit scoring is plagued by Flat max or 4 minute mile problem

- Scholars have argued that model performance cannot be enhanced further and has approached its limits:
- 'The predictive ability of linear models is insensitive to large variations in the size of regression weights and to the number of predictors.' (Lovie and Lovie, 1986; Overstreet etal 1992;)
  - More recently Classifier Technology and Illusion of progress (Hand, D; 2006)
    - Hand's point is 80-90% of model lift with simple model;
    - Still true but 5-7% relative lift is still stellar for modelers and data scientists;

"There was a mystique, a belief that it couldn't be done, but I think it was more of a psychological barrier than a physical barrier."

Record for mile run over time



Progress

- 4 min 1.3 seconds pre-Bannister
- 3 min 59 seconds Bannister
- 3 min 43 seconds Gourrouj
  - 17 sec/4 min ≈ 7% lift
    - Proportional to rf advantage on app scoring

Random Forests - Overview

- Ensemble method using random selection of variables and bootstrap samples (by Leo Breiman at berkeley, CA 1999)
  - Develop recursive partitioning trees using majority vote
    - 1) Each tree is grown on a bootstrap sample of the training set.
    - 2) A number m is specified much smaller than the total number of variables M. At each node, m variables are selected at random out of the M, and the split is the best split on these m variables.
      - (Breiman, 2002 Wald Lecture looking inside the black box)

- The most powerful aspect about random forests is variable importance ranking which estimates the predictive value of variables by scrambling the variable and seeing how much the model performance drops
- Unclear why ensembles work, research has shown a random forest is a form of adaptive nearest neighbor (Yin,2002)

Random Forests Outperform in Credit Scoring

- There are two credit modeling domains:
  - Origination score used to predict whether to give a loan to someone
  - Behavioral models used to predict customer history and behavior related variables to predict collection or line of credit performance later in the life of loan product
- As a credit risk manager I have found that random forests outperform logit models on numerous credit products (credit cards, mortgages, auto loans, home installments, and home equity line of credit)
- Random forests score an estimated 5-7% higher area under the curve ROC for origination scoring
- Random forests have been shown to improve behavioral models by as much as 15%-20%

Thrust of Research

- In practice, variable importance rankings can be used to test one way interaction effects and add them to logit models, resulting in performance which matches or sometimes outperforms random forests
- A hypothesis for the performance gap in the random forest and logit model is that ensemble methods ferret out predictive interactions effects between variables
- To test this hypothesis a synthetic data set was created where the Y variables is dependent on X,Z and X*Z interaction. Both the random forest and logit models were tested using the dataset and the performance gap or advantage of the random forest became clear

I* algorithm

- To tune logit to match or exceed random forest performance
- The i* algorithm. This algorithm takes the random forest rankings and searches for interaction terms and keeps adding terms in a step wise manner if adding them increases out of sample performance. Using the random forest ranking results in an ordered search with a smaller search space than all models and quickly converges to an effective model
  - Generate interaction terms with important variables n way one at a time vs. all interactions
    - using rf var imp to generate a manageable set of interactions one way and building step wise based on validation performance is effective
- Reason interactions not been explore more:
  - Unreliable p values; Modeling too many interactions at once doesn't work; overfit poor performance (Gayler, 2008)

Summary of improvements to logit vs. interaction terms

	Original glm logistic	I* glm w interactions	Original random forest
Credit card	0.60	0.64	0.62

Example 1: Credit card data

Results

30% out of sample results	
Algorithm	AUC
Rpart	0.56
Logit	0.5888
Random Forest	0.6075
Random Forest w Interaction	0.6357
Logit w Interaction	0.6266
Conditional Inf Tree with Interaction	0.614
Logit w more interactions	0.632

Common method	Logit	Rf	rf X logit	AVG of logit +rf	max of rf of logit prob.	square root of logit*rf
AUC	0.6159	0.6072	0.6196	0.6189	0.6073	0.6196

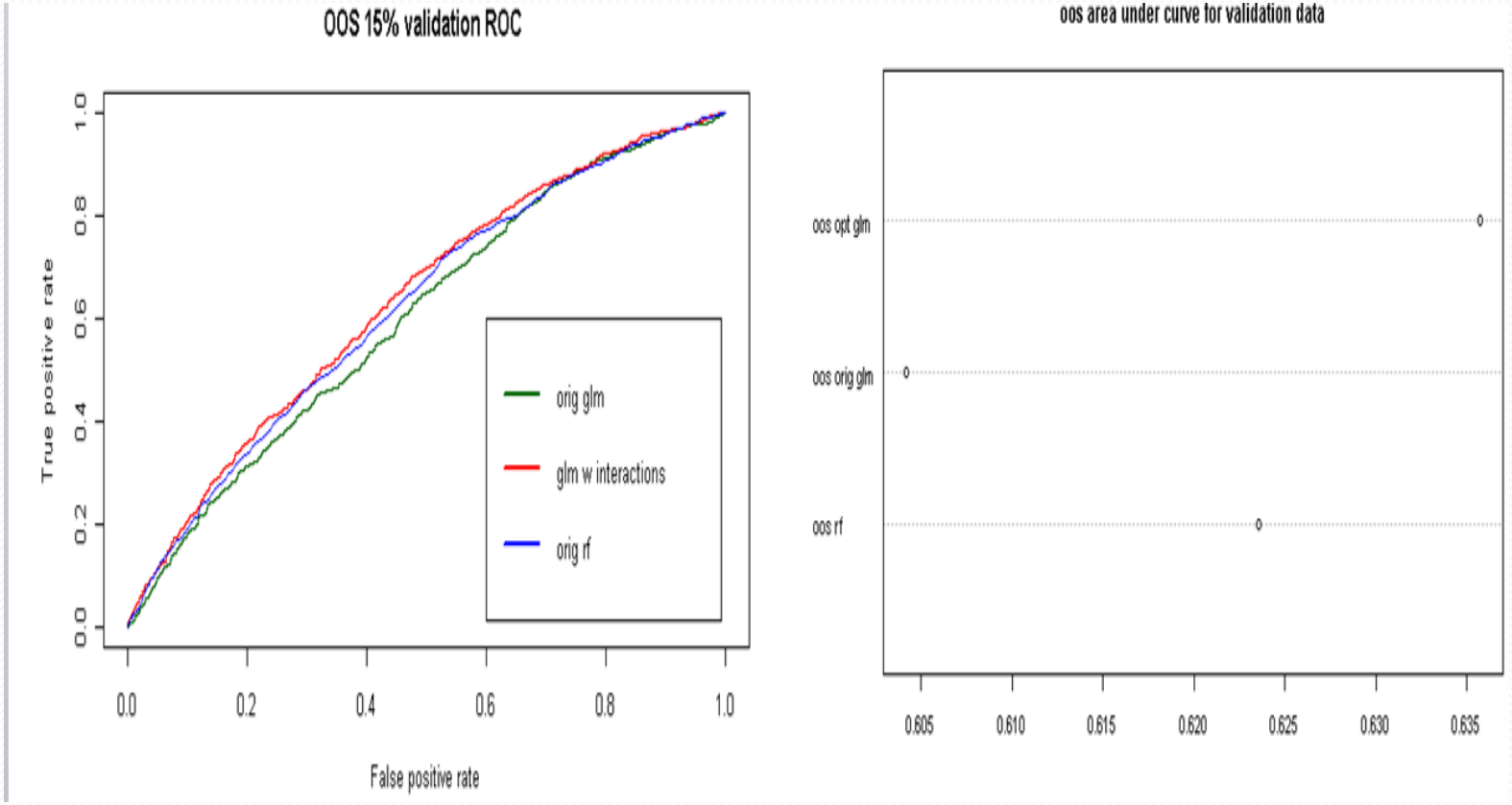
Pacific Asian KDD 2009 source:
Brazilian credit card data set of 50k credit cards

Example 2: Home Equity Data set from SAS

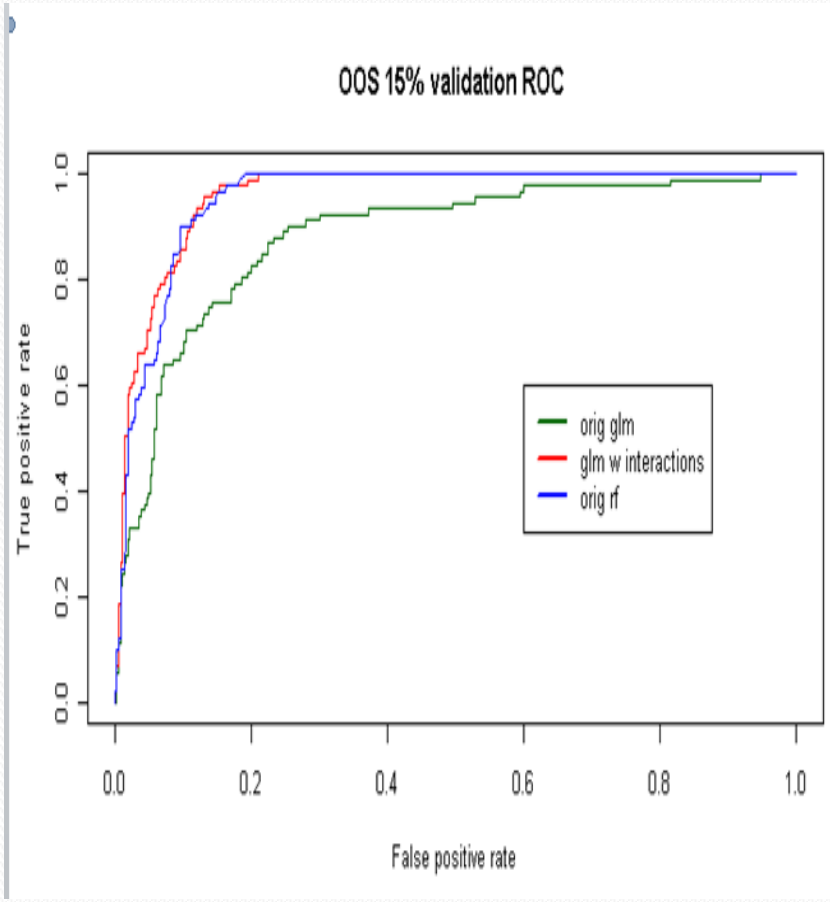
Home equity data set out of sample area under the curve:

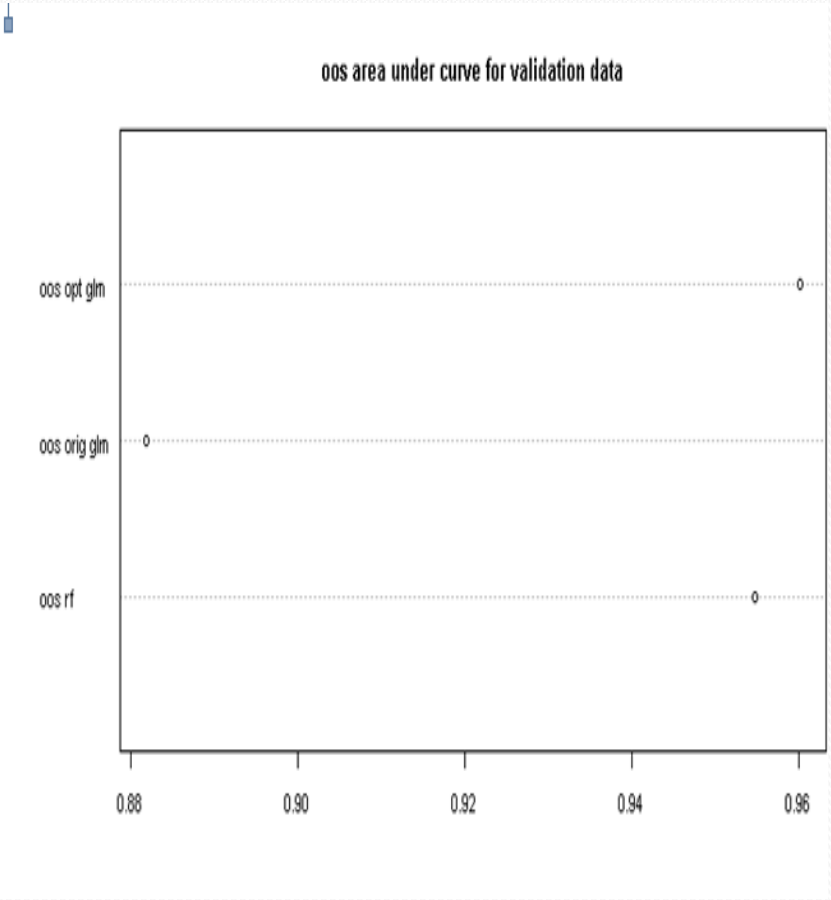
- Logistic .79
- Wallinga's neural net tuned logit link function .87
- Breiman's random forest .96 out of the box

PA KDD CC Data: Tuned logit via i* vs. rf vs. logit



Home Eq: Tuned logit vs. rf vs. logit



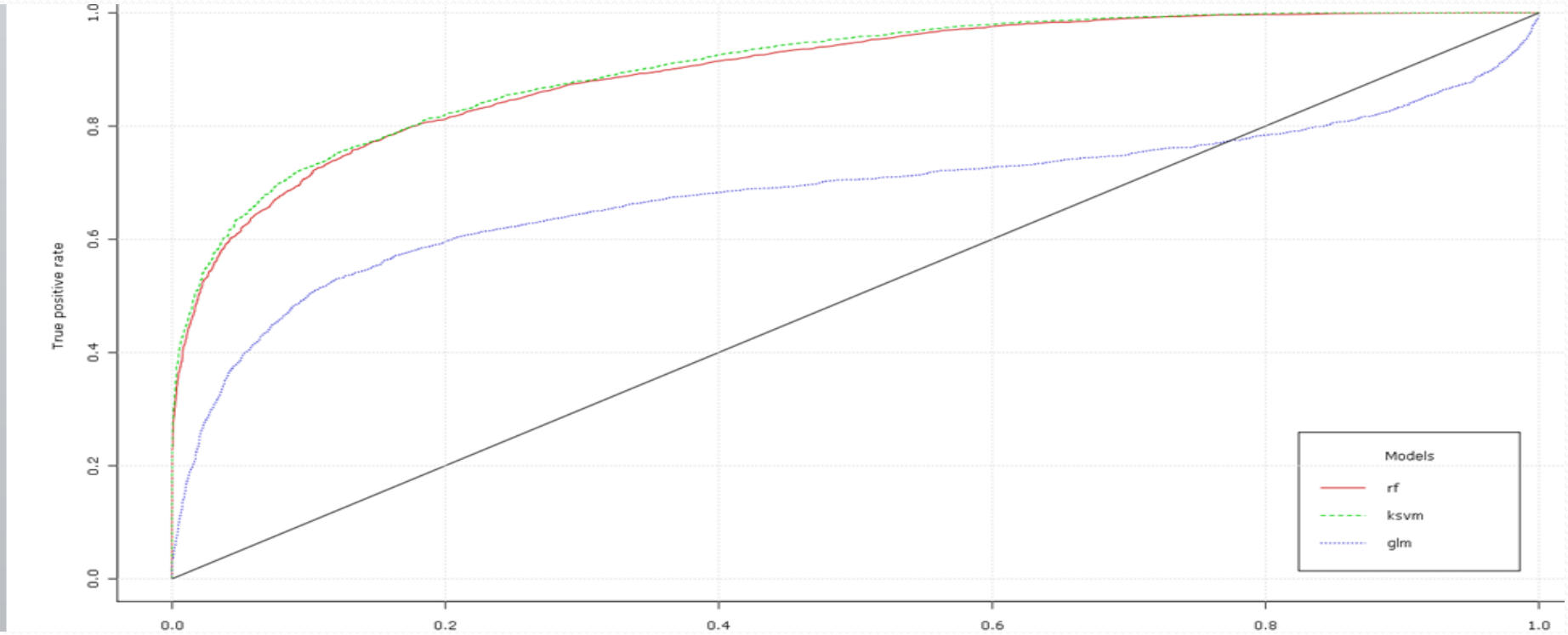


Synthetic data example

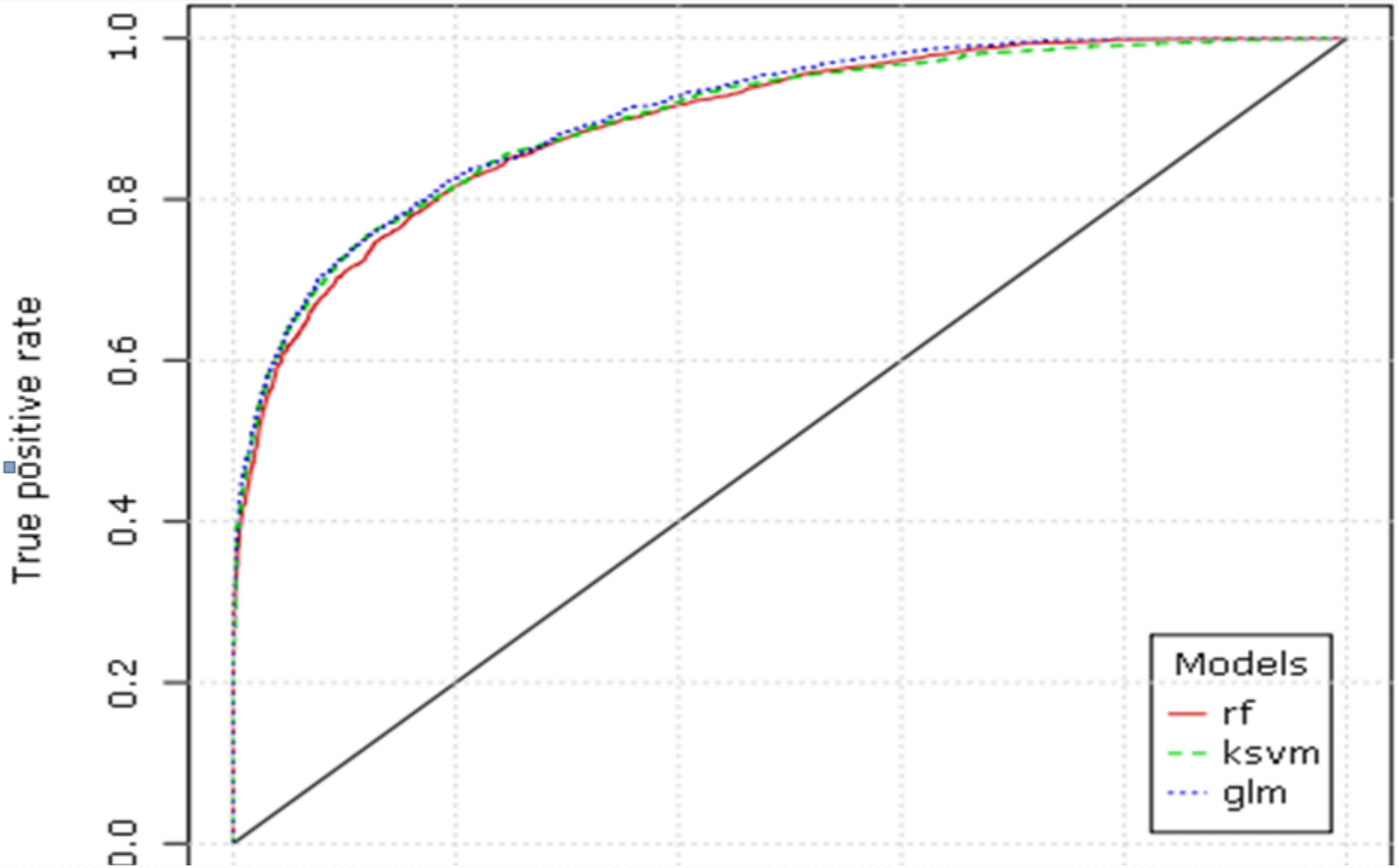
- Developed a synthetic data set where the Y= 1+ 0.5x1+ -1.5x2+2.0*x3
  - x3 is an interaction effect x3=x1*x2 where x1,x2, and x3 are normally distributed random numbers with a mean of 0 and a standard deviation of 1
- Three additional noise variables are created x4 (normal mean=1, stand dev=50) and x5 and x6 which are random numbers with mean 0 and st. dev =1
- X3 is not in the training set

Synthetic performance without x3: RF dominates 30% OOS AUC

- AUC for random forest is .897, for glm (logistic) is .66



Results with interaction term x3 added to models



Synthetic experiments

- Random 50/50 assignment of binary outcome associated with an X variable generated with different probability distributions
  - RF dominates logit when X is 2-d gaussians with different var-covariance matrix
  - RF dominates when one of the X distributions is non-exponential (e.g. cauchy, sum of gaussians)
  - Logit dominates when distributions are 2-d gaussian but same variance-covariance structure, exponential distributions like gaussians

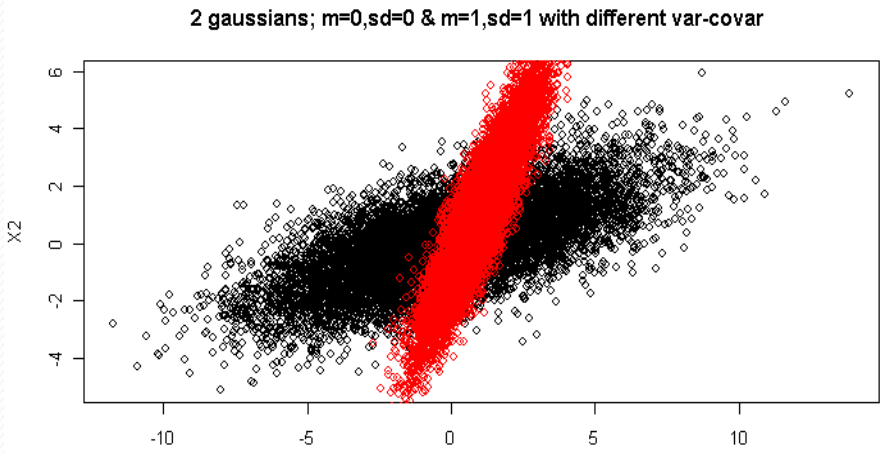
Results of synthetic experiments

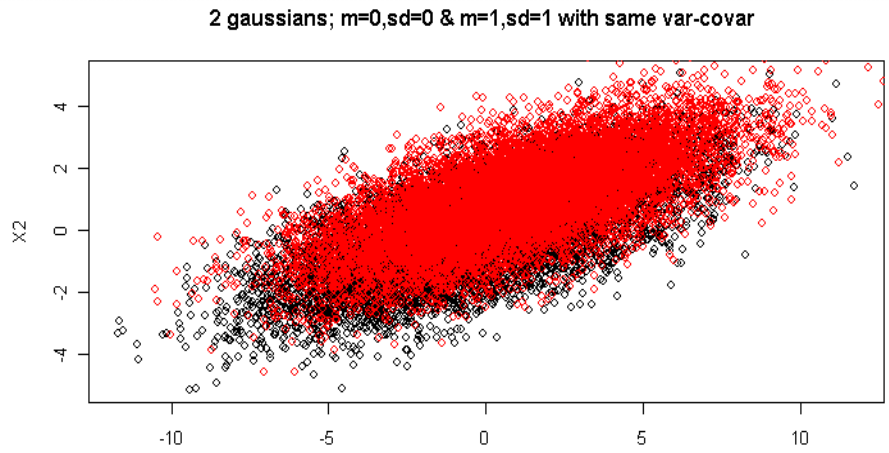
- The table below lists results from experiments on synthetic binary data
- Two classes were sampled randomly and X was sampled from two different distributions 10,000 draws

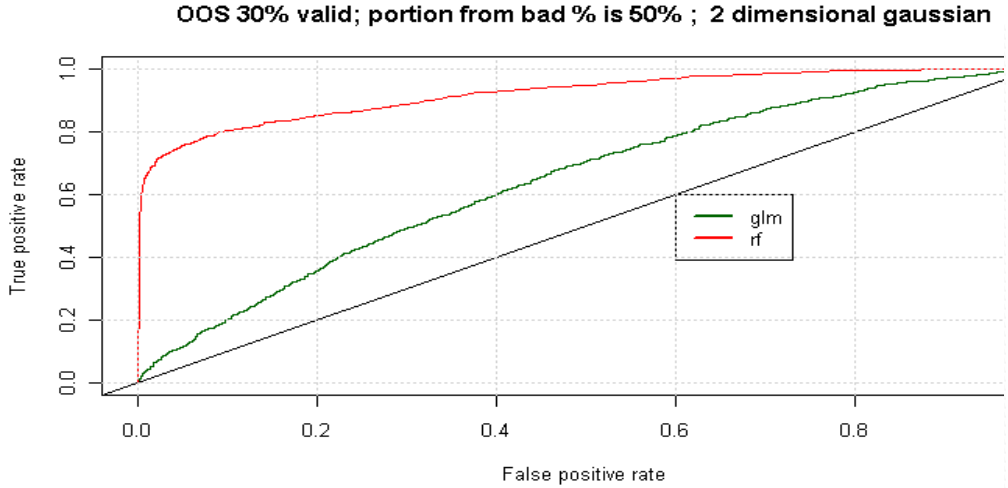
Gaussian (mean=0 sd=1) vs. Gaussian (mean=0, sd=1)	Logit
2d Gaussian with same variances/covariances	Logit
2d Gaussian with different variances/covariances	Random Forests
Data with interaction effects	Random Forests

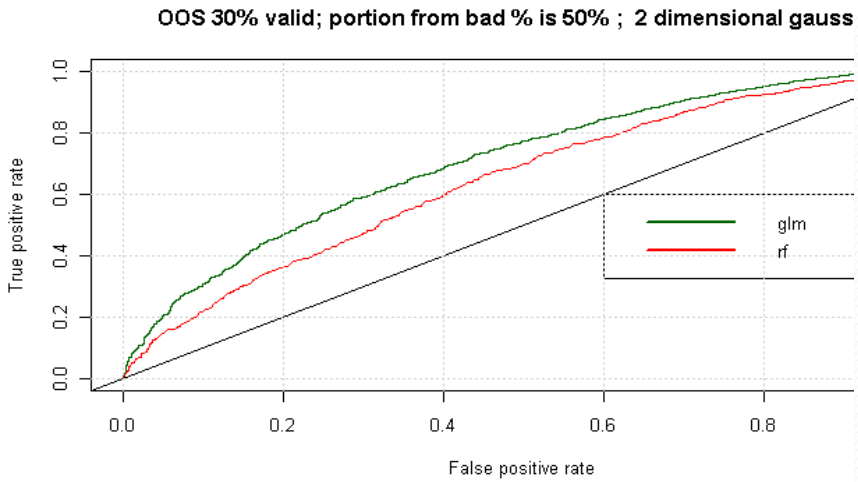
2-d Gaussians w/o same variance

- Random forest dominates when variance/co-variance is different but not when same









Conclusions

- Random forests outperform traditional logistic regression based score when there are interaction effects in data that are predictive
  - Resulting tuned logit can match or exceed rf classification
- The performance advantage of the random forest can be integrated with logit models via interaction terms identified through random forest variable rankings
  - In a sense using var imp one can algorithmically search for promising interaction effects
    - Which have always been in credit scoring like debt ratio (expenses * 1/income) or months reserves (assets/PITI etc)
    - Or even Hand's superscorecard (the product of 2 scores i.e. interaction effect)(2002)
- If a random forest model dominates logit it suggests that there are statistically significant interaction effects in the data which should be modeled
- Random forest variable importance is a powerful tool to identify interaction effects within a small search space

Details in papers

- *http://cran.r-project.org/doc/contrib/Sharma-CreditScoring.pdf*
- http://papers.ssrn.com/sol3/cf_dev/AbsByAuth.cfm?per_id=644975
  - Papers on i*, synthetic experiments, credit scoring, de-biased forest algorithm with r code appendices with detailed bibliographies
- http://en.wikipedia.org/wiki/Mile_run_world_record_progression
- http://www.stat.berkeley.edu/~breiman/wald2002-2.pdf
- http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1975801
- Data:
- Credit card
  - http://sede.neurotech.com.br/PAKDD2010/
- Home equity
  - www.*sasenterpriseminer*.com/*data*/HMEQ.*xls*