R data mining for insurance retention modeling

R in Insurance conference
CASS Business School
London 14th 2014

Giorgio Alfredo Spedicato
Ph.D C.Stat ACAS

UnipolSai Assicurazioni Reserach And Development

14th July 2014

Cass Business School
CITY UNIVERSITY LONDON

Table of contents

Outline

Retention modeling

- Modeling policyholders' retention consists in predicting whether an existing policyholder will renew its contract in order to take underwriting decisions accordingly.
- It means enriching a classical actuarial loss cost modeling with marketing and competitive business considerations in order to propose an "optimized premium" upon renew.
- Similarly, "conversion" analysis incorporates in prospective policyholders' premium behavior considerations.

Retention modeling

- Logistic regression is the most widely used approach in actuarial practice, but little attention has been given to other machine learning techniques.
- An application that showes the use of Random Forest for life insurance policy retention can be found in [Milhaud et al., 2011].
- Good introductions on the business side can be found in [Arquette and Guven, 2013] and [Anderson and McPhail, 2013].

Retention modeling

- A statistician would consider retention modeling as a classification tasks. In fact, policyholders can be classified in two groups (Renew/Not Renew) according to features that lie within: demographics (age, gender, ...), insurance policy history (years since customer, claim history, coverage dept, ...), marketing and business environment (year to year price change and relative competitiveness, ...).
- An adequately tuned retention model can be integrated within an underwriting algorithm to define the final renewal price.
- Redefining renewal price will need to recalculate price variation and competitiveness variables.

The business problem

- A 50K policies dataset has been available containing demographics, insurance and business variables as well as recorded Renew/Lapse decision.
- Various retention models will be fit to model the probability that the policyholder will renew upon termination date, p_i.
- The aim is to get a function $p_i = f(P_i)$, depending renewal probability by the proposed renewal premium, P_i.

The business problem

- The final aim consists in maximizing the existing business gross premium $P^T = \sum_{i=1...N} (P_i * p_i)$.
- The renewal premium can be modified by $\pm 1\%$ with respect to initial proposed value in our renewal premium optimization scenario.
- This exercise will not consider the loss cost component in the analysis, thus the portfolio premium volume will be optimized and not the global underwriting margin.

The approach

- A standard logistic regression will be fit and compared with various machine learning techniques in Section 2.
- The models found in Section 2 will be used to calculate an optimized premium for each policyholder in Section 3.
- Brief considerations will be drawn also.

Outline

Data and Methods

- The dataset was bundled into an actuarial pricing software package. It can be deemed representative of a common personal lines retention problem.
- Most statistical fitting process will be based on the caret [from Jed Wing et al., 2014] package framework to fit predictive models.
- The reference model will be a logistic regression where continuous predictors have been handled using restricted cubic splines from **rms** packagem [Jr, 2014]. package.

Data and Methods

- Logistic regression key advantages are: relatively fast estimation time, easy model interpretation and no need of parameters' tuning.
- Most machine learning techniques have at least one or more parameter that cannot be estimated from data and need to be tuned (tuning parameters).
- The caret package, [from Jed Wing et al., 2014], provides a very valuable infrastructure that standardizes the process for fitting, validating and predicting models that have been implemented within different packages.

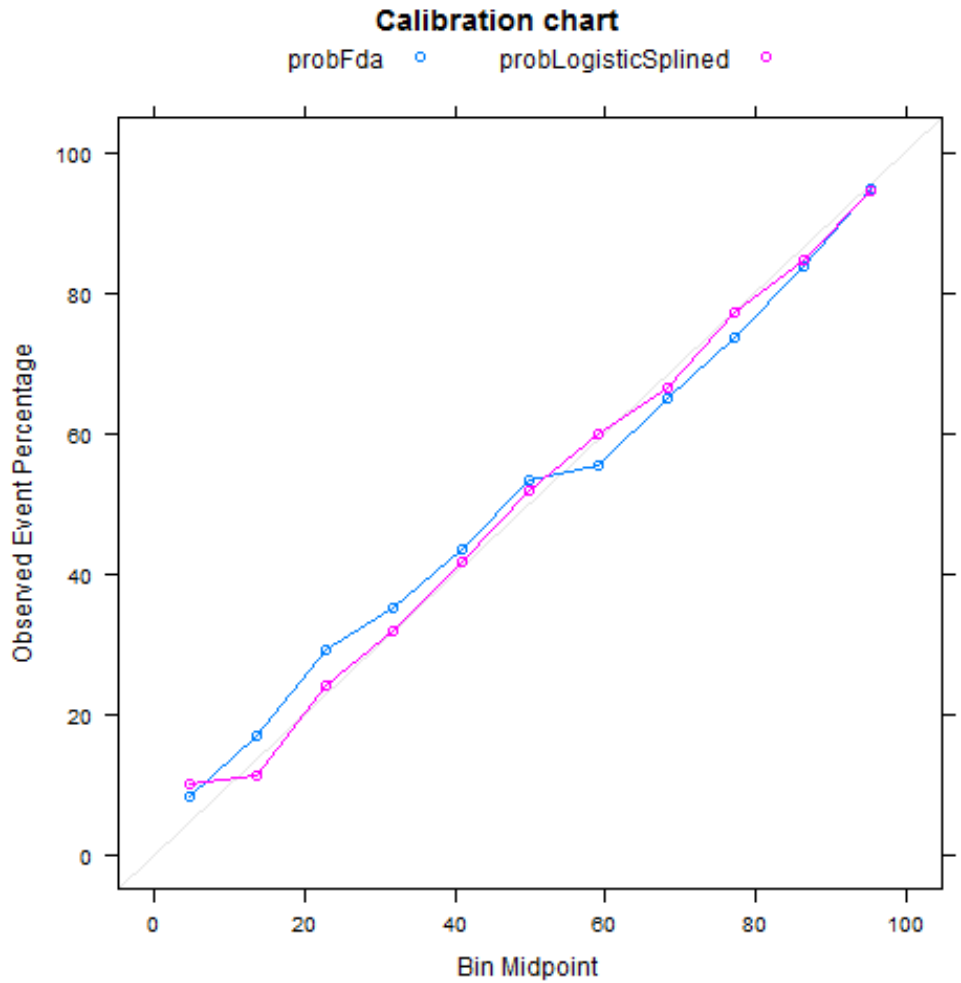
Approach

- A performance measure should be defined. For binary classification models, the area under the ROC curve Curve (AUC) is a very common choice.
- In addition, fitting the model over different samples (cross - validation) is a standard practice to better assess model performance in order to avoid overfitting.
- A final choice is to divide the dataset between a train set and a test set and to perform the final model evaluation under the test set.

Model fitting and assessment

- The following models were fitted on a train set split: logistic regression, C5.0 classification trees, Flexible Discriminant Analysis (FDA), Naive Bayes (NB), Support Vector Machines (SVM), Neural Networks (NN), Boosted Glms.
- Only FDA shows a slight increase in predictive performance in the provided dataset when compared with logistic regression (the reference model)
- Logistic regression and FDA appear very close in terms of predictive gain (Lift Curve, Figure 2) and precision in estimated probabilities (Calibration Chart, Figure 1) .

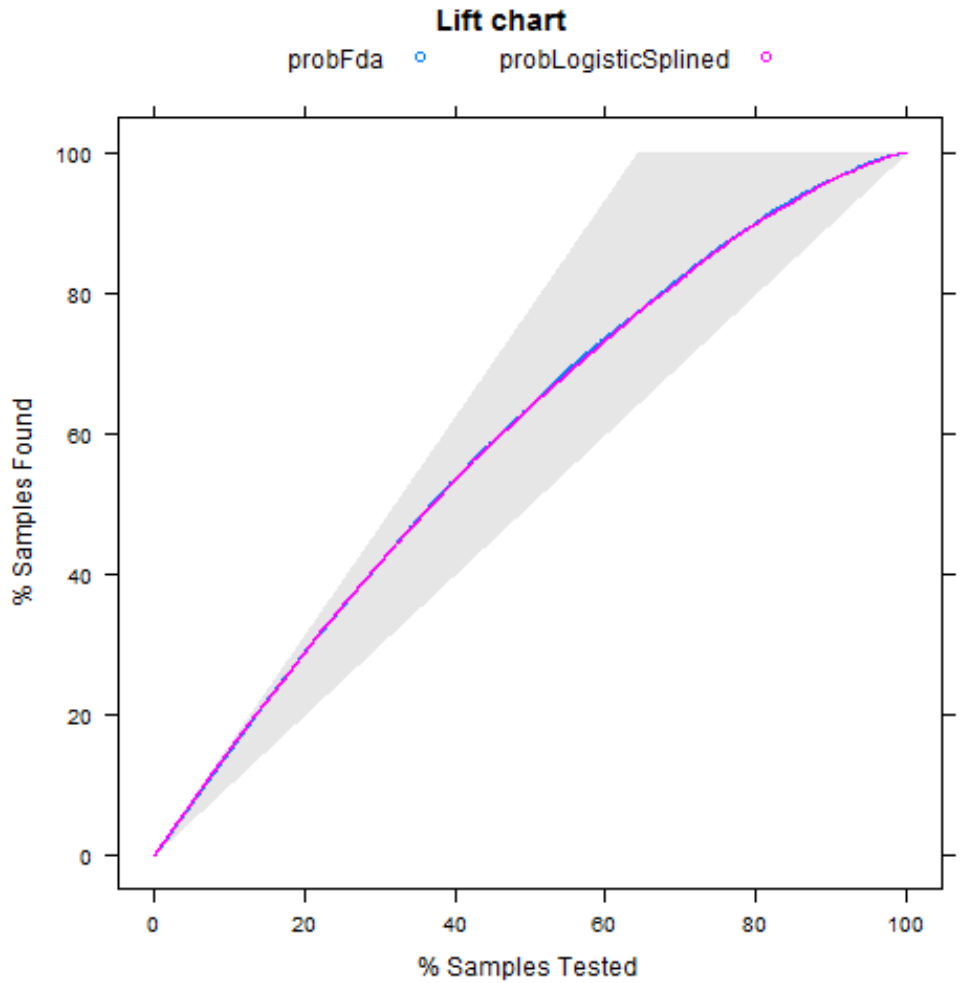
Model fitting and assessment



models.png

Figure: Calibration Plot

Lift chart



models.png

Figure: Lift Chart

Outline

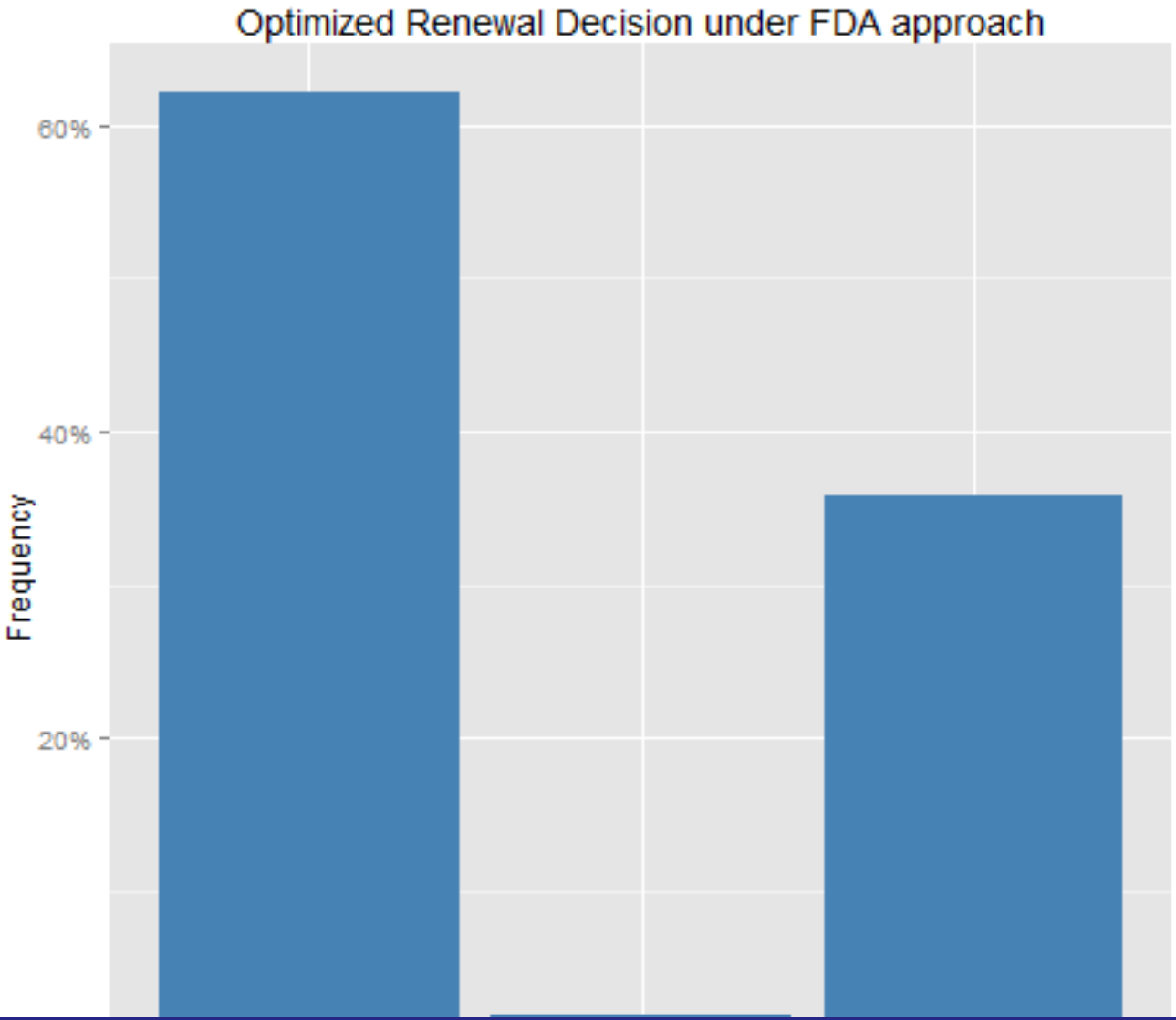
Optimizing

- The retention model has been used to optimize the renewal premium, P_i. The proposed P_i was varied in different scenarios and the expected renewal premium inflow, $p_i * P_I(p_i)$, was calculated for each individual policyholder.
- Three premium variation scenarios were proposed: no change, +1% increase, -1% decrease.
- The highest expected renewal premium scenario for each customer has been selected.

Results

- Each policyholder would be proposed the highest expected renewal premium according to the three scenarios.
- The final renewal decision for each policyholder is shown in Figure 3.
- In the available data set, the overall variation in expected renewal premium is very close using either FDA or Logistic approach

Renewal decision



Considerations

- R appears to be a valid environment, not only to perform standard loss cost modeling but also to estimate renewal probabilities.
- Logistic regression requires less fitting effort than other machine learning techniques (that not always offer improved predictive performance).
- The caret package [from Jed Wing et al., 2014] provides a good framework to unify the process of building and assess competing machine learning techniques.

Outline

Infrastructure

- Various machine learning techniques have been compared in the analysis: glms (linear logistic regression and its boosted version); tree based methods (C5.0, Random Forest); non - linear methods (neural network, Naive Bayes Classifiers, Support Vector Machines and Flexible Discriminant Analysis).
- The caret package has been used in order to provide a common infrastructure to fit the diferent models, compare their performance and perform predictions.
- The data source was originally provided in SAS and imported in R thanks to the package **sas7bdat**, [Shotwell, 2014].

The caret package infrastructure

- The core caret function used is the **train** function that provides a common wrapper to fit more than one hundred machine learning techniques. A **predict** method is available for all **trained** objects
- Hyperparameter have been generally searched over a model pre - defined grid. 10 - fold cross validation has been used to better assess the predictive performance.
- Among the other package's functions, **lift** and **calibration** function have been used to project lift gain and calibration charts.

Logistic Regression and its extensions

- **glm** R base function was wrapped by **train** in order to fit logistic regression.
- **rms** package has been used to add restricted cubic splines to the prediction formula on the continuous predictors.
- A boosted version of the logistic regession has also be fit with the use of **mboost** [Buehlmann and Hothorn, 2007] package.

Tree based methods

- Tree based approaches perform classification by finding splits within available features that provides the most predictive performance within the subset. The splits sequence provides and effective way to visualize variables' dependency.
- The **C5.0** algorithm, one of the most widely used approach, has been used in this exercise via the **C50** package, [Kuhn et al., 2014].
- A boosted approach, Gradient Boosted Machines, **gbm**, has also be used thanks to package [with contributions from others, 2013].

Neural Networks

- A neural network model consists in a set of adaptive weights (neurons) that can approximate non linear functions.
- Neural Networks represent a classical machine learning techniques, probably the one proposed first.
- Among the many R packages available, the **nnet** package, [Venables and Ripley, 2002], has been used in this exercise.

Support Vector Machines

- A support vector machine (SVM) fits an hyperplane that aims to separate categories within the dependent variable with least error. It is rated among the top classifiers.
- Non linear shapes can be used in order to define the hyperplane: polynomial, hyperbolic, radial basis functions.
- the **kernlab** package, [Karatzoglou et al., 2004] has been used in this exercise.

Flexible Discriminant Analysis

- The Flexible Discriminant Analysis (FDM) method extends the linear discriminant analysis by including non - linear hinge functions $h(x) = max(x - h, 0)$ in the calculation of the discriminant function.
- hinge function can be used within the discriminant function also in polynomial form.
- FDA models have been fit with package the **earth** package, [from mda:mars by Trevor Hastie and utilities with Thomas Lumley's

K Nearest Neighbours

- A KNN model classifies a sample with the most frequent item within its closest neighbours in the train set.
- It is a very simple approach, that can be used both in regression and classification.
- KNN models can be fit with the **caret** package, [from Jed Wing et al., 2014].

Naive Bayes

- A (Naive Bayes9 NB model classifies a sample using the conditional distribution of the predictors given the sample.
- Even if independency across predictors is assumed, the overall accuracy provided by the model is often competitive with respect to other techniques.
- NB models can be fit thanks to **KlaR** package, [Weihs et al., 2005].

Outline

Thanks

- The author is grateful to UnipolSai, his employeer, for the support.
- Nevertheless the opinion herewith stated are responsability of the author only. UnipolSai does not take any opinion within and shall not be deemed liable for any legal error found therein.
- A special thanks in addition is given to Dr. Andrea Dal Pozzolo for its suggestions.

Outline

Reference I


Anderson, D. and McPhail, M. (2013).
Price optimization for the u.s. market. techniques and implementation strategies.


Arquette, K. and Guven, S. (2013).
Intelligent use of competitive analysis.


Buehlmann, P. and Hothorn, T. (2007).
Boosting algorithms: Regularization, prediction and model fitting (with discussion).
Statistical Science, 22(4):477–505.


from Jed Wing, M. K. C., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., and the R Core Team (2014).
caret: Classification and Regression Training.
R package version 6.0-30.


from mda:mars by Trevor Hastie, S. M. D. and utilities with Thomas Lumley's leaps wrapper., R. T. U. A. M. F. (2014).
earth: Multivariate Adaptive Regression Spline Models.
R package version 3.2-7.


Jr, F. E. H. (2014).
rms: Regression Modeling Strategies.
R package version 4.2-0.

Reference II

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004).
kernlab – an S4 package for kernel methods in R.
Journal of Statistical Software, 11(9):1–20.

Kuhn, M., Weston, S., and code for C5.0 by R. Quinlan, N. C. C. (2014).
C50: C5.0 Decision Trees and Rule-Based Models.
R package version 0.1.0-19.

Milhaud, X., Loisel, S., and Maume-Deschamps, V. (2011).
Surrender trigger in life insurance: What main features affect the surrender behavious in an economic context.
Bullettin Francais d'Actuariat, 1(11).

Shotwell, M. (2014).
sas7bdat: SAS Database Reader (experimental).
R package version 0.4.

Venables, W. N. and Ripley, B. D. (2002).
Modern Applied Statistics with S.
Springer, New York, fourth edition.
ISBN 0-387-95457-0.

Weihs, C., Ligges, U., Luebke, K., and Raabe, N. (2005).
klar analyzing german business cycles.
In Baier, D., Decker, R., and Schmidt-Thieme, L., editors, *Data Analysis and Decision Support*, pages 335–343, Berlin. Springer-Verlag.

Reference III


with contributions from others, G. R. (2013).
gbm: Generalized Boosted Regression Models.
R package version 2.1.