Results

An application to Insurance data

Introduction

- **Objectives:**
  - Comparison of 4 different model computing an Insurance Pure Premium
    - GLM,
    - Single Tree,
    - Random Forest,
    - GBM

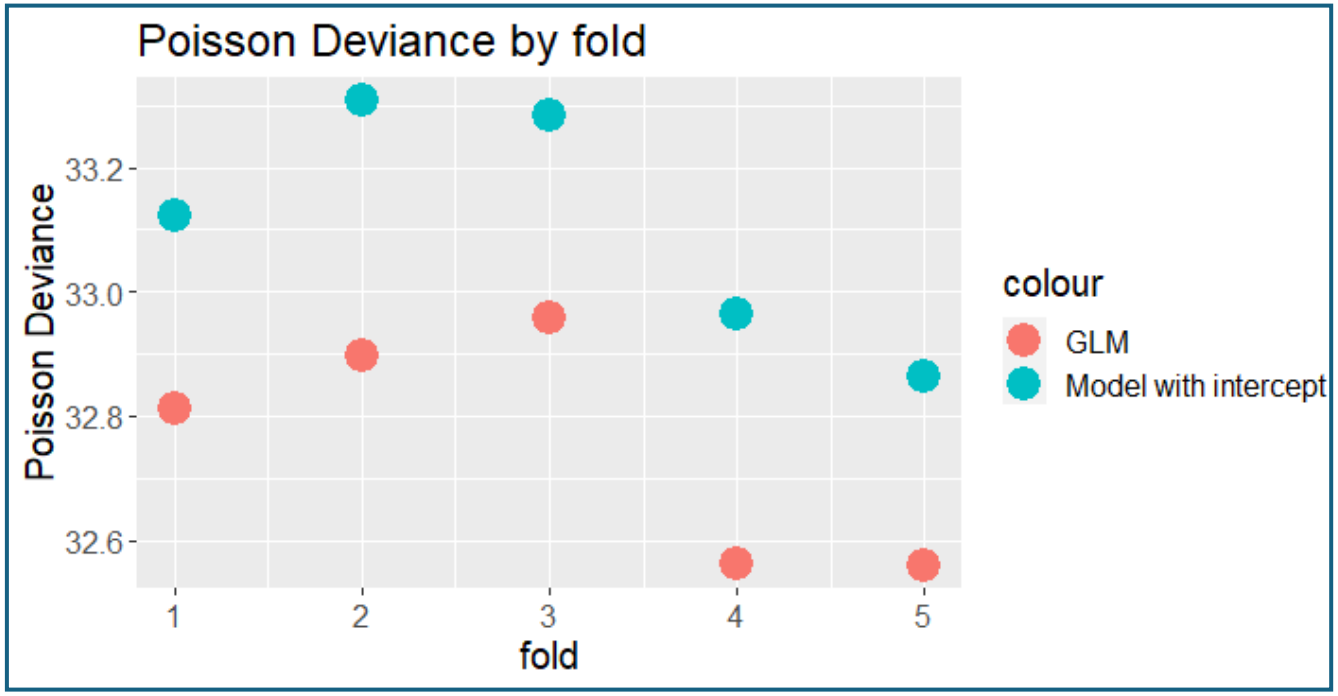
Data in use

```
# Define column class for dataset
colCls <- c("integer",         # row id
            "character",       # analysis year
            "numeric",         # exposure
            "character",       # new business / renewal business
            "numeric",         # driver age (continuous)
            "character",       # driver age (categorical)
            "character",       # driver gender
            "character",       # marital status
            "numeric",         # years licensed (continuous)
            "character",       # years licensed (categorical)
            "character",       # ncd level
            "character",       # region
            "character",       # body code
            "numeric",         # vehicle age (continuous)
            "character",       # vehicle age (categorical)
            "numeric",         # vehicle value
            "character",       # seats
            rep("numeric", 6), # ccm, hp, weight, length, width, height (all continuous)
            "character",       # fuel type
            rep("numeric", 3)  # prior claims, claim count, claim incurred (all continuous)
)
```

```
## 'data.frame':    40760 obs. of  27 variables:
##  $ row.id        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ year          : chr  "2010" "2010" "2010" "2010" ...
##  $ exposure      : num  1 1 1 0.08 1 0.08 1 1 0.08 1 ...
##  $ nb.rb         : chr  "RB" "NB" "RB" "RB" ...
##  $ driver.age    : num  63 33 68 68 68 68 53 68 68 65 ...
##  $ drv.age       : chr  "63" "33" "68" "68" ...
##  $ driver.gender : chr  "Male" "Male" "Male" "Male" ...
##  $ marital.status: chr  "Married" "Married" "Married" "Married" ...
##  $ yrs.licensed  : num  5 1 2 2 2 2 5 2 2 2 ...
##  $ yrs.lic       : chr  "5" "1" "2" "2" ...
##  $ ncd.level     : chr  "6" "5" "4" "4" ...
##  $ region        : chr  "3" "38" "33" "33" ...
##  $ body.code     : chr  "A" "B" "C" "C" ...
##  $ vehicle.age   : num  3 3 2 2 1 1 3 1 1 5 ...
##  $ veh.age       : chr  "3" "3" "2" "2" ...
##  $ vehicle.value : num  21.4 17.1 17.3 17.3 25 ...
##  $ seats         : chr  "5" "3" "5" "5" ...
##  $ ccm           : num  1248 2476 1948 1948 1461 ...
##  $ hp            : num  70 94 90 90 85 85 70 85 85 65 ...
##  $ weight        : num  1285 1670 1760 1760 1130 ...
##  $ length        : num  4.32 4.79 4.91 4.91 4.04 ...
##  $ width         : num  1.68 1.74 1.81 1.81 1.67 ...
##  $ height        : num  1.8 1.97 1.75 1.75 1.82 ...
##  $ fuel.type     : chr  "Diesel" "Diesel" "Diesel" "Diesel" ...
##  $ prior.claims  : num  0 0 0 0 0 0 4 0 0 0 ...
##  $ clm.count     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ clm.incurred  : num  0 0 0 0 0 0 0 0 0 0 ...
```

- **Data:** *Predictive Modelling Applications in Actuarial Science, Vol.2* (E. Frees & al.): https://instruction.bus.wisc.edu/jfrees/jfreesbooks/PredictiveModelingVol1/glm/v2-chapter-1.html
- Data already explored in a previous study (cf. EDA for Insurance) stored in another repository where a description of the fields is also available.

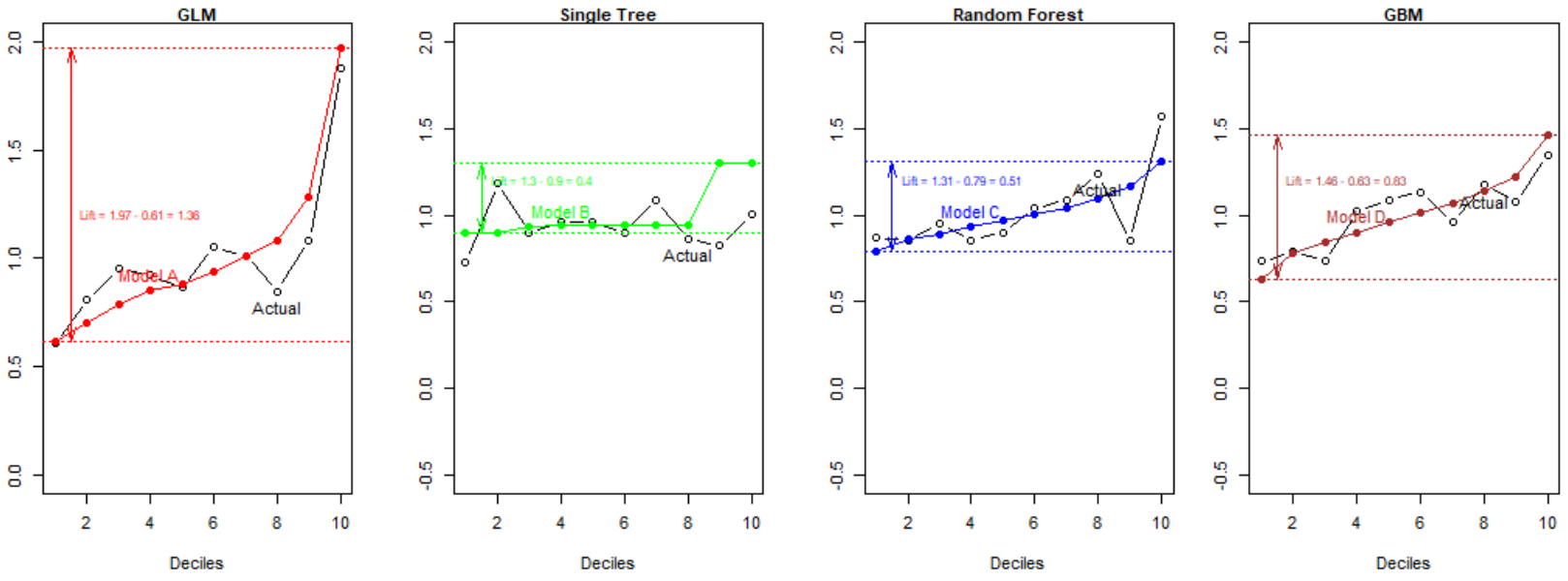
Cross-Validation



- For each fold, we can compute a Poisson Deviance and compare between models.
- Here, a model with predictors is better than a Null model.

Single Lift Plot (Simple quantile plot)

Single Lift Plots



Definition

The Lift measures the "Economic value of a model".
Used to compare the relative performance of two models following 3 criteria:

1. Predictive accuracy,
2. Monotonicity,
3. Vertical distance between first and last quantile.

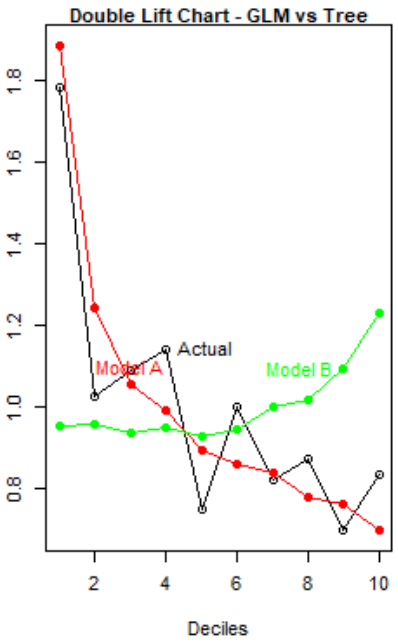
Models	Name	Lift
GLM	A	3.23
Single Tree	B	1.44
Random Forest	C	1.66
GBM	D	2.32

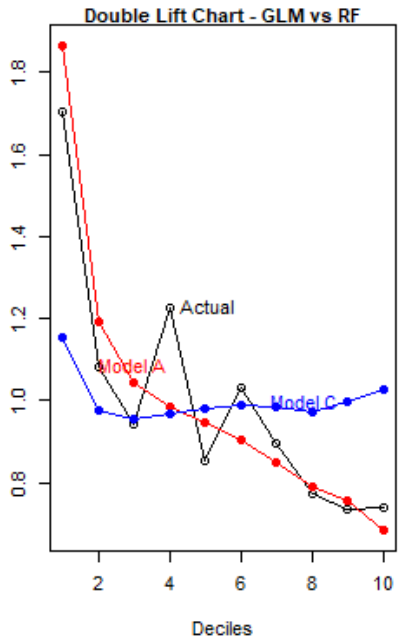
Results

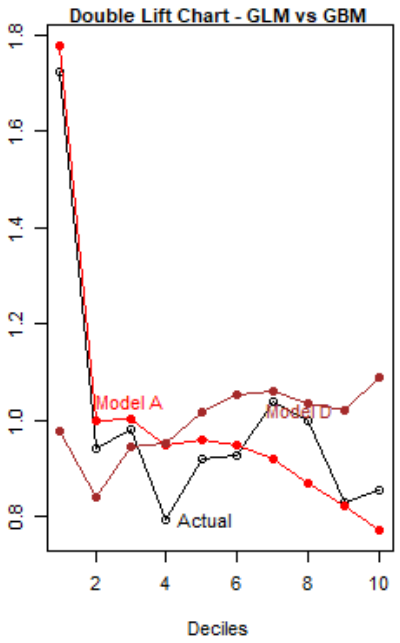
1. The plotted loss cost correspond more to the actual on the GBM plot. It seems to predict severity better than the other models.
2. No significant reversal. Pattern are stable.
3. The GLM's average predicted value for bin 1 is 0.61 and at the other extreme, in bin 10, the average predicted value is about 1.97 times the overall average predicted value: The lift that the GLM provides is 1.97-0.61=1.36, i.e. the distance between bin 1 and bin 10 is 36%, or a "lift of 1.97/0.61 = 3.23."

In Comparison, the single tree, random forest and GBM are more able to separate the risks than the GLM.

Double Lift Plot



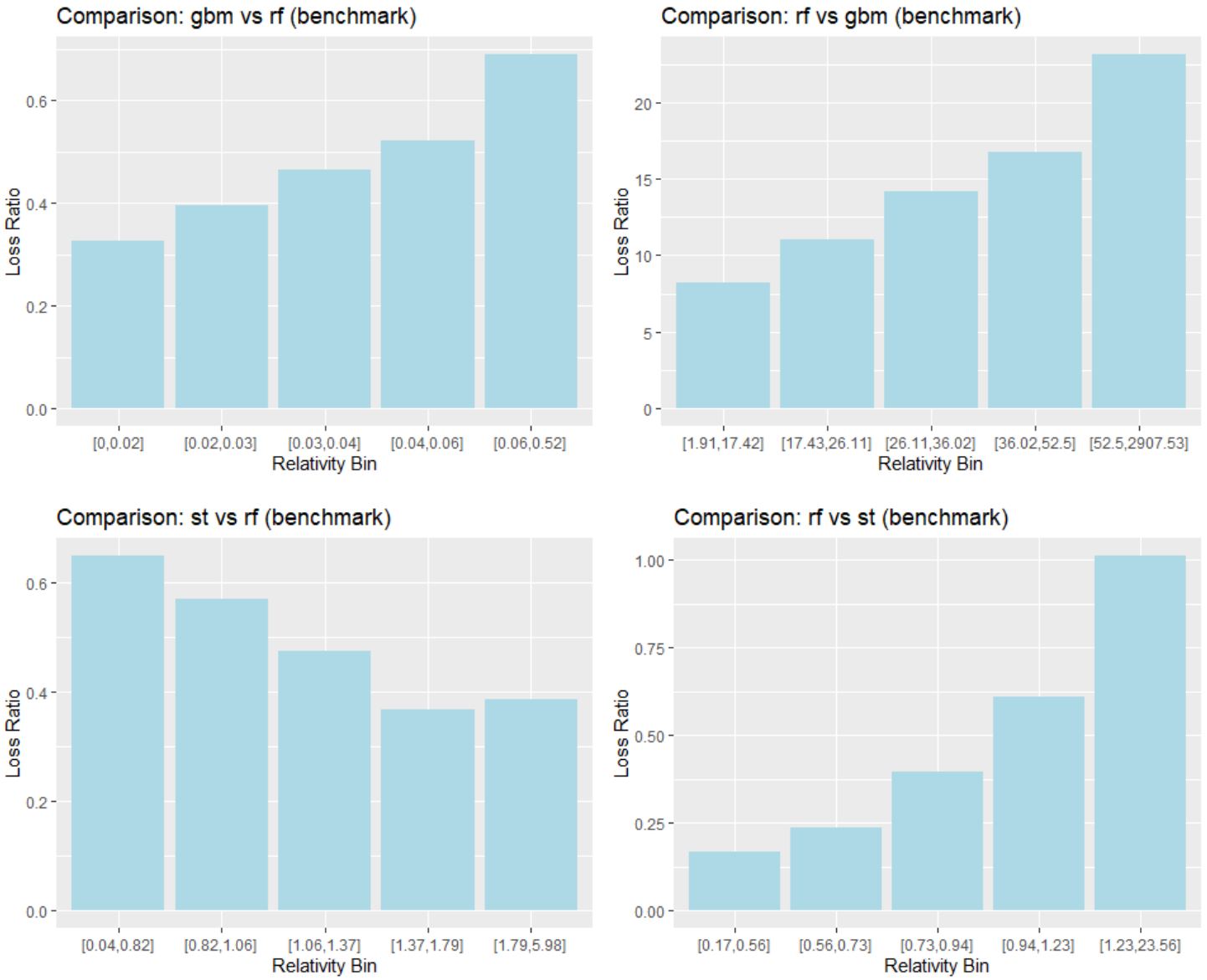




Similar to the simple quantile plot, but it directly compares two models.

The “winning’ model is the one that more closely matches the “Actual” model.

Loss Ratio Charts



Instead of plotting the Pure Premium for each bucket, the Loss Ratio is instead plotted.

Gini Indices

	glm_PP <dbl>	st_PP <dbl>	rf_PP <dbl>	gbm_PP <dbl>	max_gini <dbl>	bench <chr>
glm_PP	0.00000	32.70513	32.14586	16.79216	32.70513	glm
rf_PP	89.60435	-13.17079	0.00000	15.51544	89.60435	rf
gbm_PP	90.13804	12.41821	17.11281	0.00000	90.13804	gbm
st_PP	92.42702	0.00000	37.04613	35.74857	92.42702	st

4 rows

- Two-way comparison of Gini indices for all the methods tested and the GLM.
- We observe that the single Tree reaches the highest Gini index for all the benchmarks.

Sources

- *Predictive Modelling Applications in Actuarial Science, Vol.2* (E. Frees & al.)
- *Hands on ML with R (B. Broecke)* https://bradleyboehmke.github.io/HOML/