Tree-based Models for Variable Annuity Valuation: Parameter Tuning and Empirical Analysis

Insurance Data Science Conference

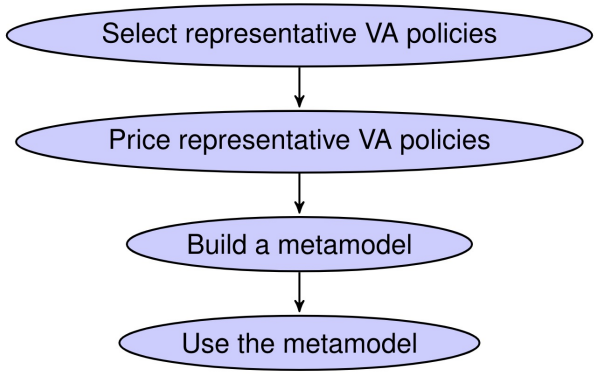
Zhiyu Quan, Assistant Professor
University of Illinois at Urbana-Champaign

Variable annuity valuation dynamic hedging and its drawbacks

- Dynamic hedging is a popular approach to mitigate the financial risk, but
  - It requires calculating the dollar Deltas of a portfolio of variable annuity policies within a short time interval.
  - The value of the guarantees cannot be determined by closed-form formula.
  - The Monte Carlo simulation model is time-consuming.

Metamodeling

- Metamodeling has been applied to address the computational problems arising from valuation of variable annuity portfolios.
    - Select a small number of representative VA policies.
    - Use Monte Carlo simulation to calculate the fair market values of the representative policies.
    - Build a predictive model, called a metamodel, based on the representative policies and their fair market values.
    - Use the predictive model to estimate the fair market value for every VA policy in the portfolio.



Tree-based models and its advantages

- **Tree-based models** can be broadly described by:
    - repeatedly partitioning the space of the explanatory variables and thereby creating a tree structure for predicting the response variable.
- Advantages:
    - Nonparametric approach - distribution free
    - Partially solve multicollinearity
    - Detect non-linear effects and interactions among the explanatory variables
    - Interpret by visualizing the tree structure
    - Variable selection by assessing the relative importance
    - Robust to the outliers and handle missing data in a natural way
    - Less data preparation

VA data - continuous

Response variables	Description	Min.	1st Q	Mean	Median	3rd Q	Max.
fmv	Fair market value	-68.37	-5.55	64.63	11.7	64.84	1210.32
Continuous variables							
gmwbBalance	GMWB balance	0	0	27.8	0	0	422.26
gbAmt	Guaranteed benefit amount	51.88	183.98	323.29	306.89	437.36	920.62
FundValue1	Account value of the 1th fund	0	0	32.02	12.62	46.76	629.89
FundValue2	Account value of the 2nd fund	0	0	36.54	16.08	56.31	571.59
FundValue3	Account value of the 3rd fund	0	0	26.78	11.81	36.64	458.78
FundValue4	Account value of the 4th fund	0	0	25.8	10.48	38.29	539.36
FundValue5	Account value of the 5th fund	0	0	22.29	10.54	34.71	425.92
FundValue6	Account value of the 6th fund	0	0	37.15	19.64	53.96	654.64
FundValue7	Account value of the 7th fund	0	0	28.78	12.88	42.56	546.89
FundValue8	Account value of the 8th fund	0	0	31.27	15.59	46.24	529.57
FundValue9	Account value of the 9th fund	0	0	31.93	13.9	45.17	599.44
FundValue10	Account value of the 10th fund	0	0	32.6	13.86	45.09	510.43
age	Age of the policyholder	34.52	42.86	50.29	51.36	57.21	64.46
ttm	Time to maturity in years	0.75	10.09	14.61	14.6	19.12	27.52

VA data - categorical

Categorical variables	Description	Proportions
gender.M	Male policy holder	64.71 %
gender.F	Female policy holder	35.29 %
productType.ABRP	Indicate type GMAB with return of premium	8.82 %
productType.ABRU	Indicate type GMAB with annual roll-up	4.26 %
productType.ABSU	Indicate type GMAB with annual ratchet	6.03 %
productType.DBAB	Indicate type GMDB + GMAB with annual ratchet	5.00 %
productType.DBIB	Indicate type GMDB + GMIB with annual ratchet	5.88 %
productType.DBMB	Indicate type GMDB + GMMB with annual ratchet	5.74 %
productType.DBRP	Indicate type GMDB with return of premium	4.85 %
productType.DBRU	Indicate type GMDB with annual roll-up	6.62 %
productType.DBSU	Indicate type GMDB with annual ratchet	4.41 %
productType.DBWB	Indicate type GMDB + GMWB with annual ratchet	4.41 %
productType.IBRP	Indicate type GMIB with return of premium	5.74 %
productType.IBRU	Indicate type GMIB with annual roll-up	4.71 %
productType.IBSU	Indicate type GMIB with annual ratchet	4.85 %
productType.MBRP	Indicate type GMMB with return of premium	4.56 %
productType.MBRU	Indicate type GMMB with annual roll-up	5.29 %
productType.MBSU	Indicate type GMMB with annual ratchet	5.29 %
productType.WBRP	Indicate type GMWB with return of premium	4.12 %
productType.WBRU	Indicate type GMWB with annual roll-up	3.97 %
productType.WBSU	Indicate type GMWB with annual ratchet	5.44 %

Possible drawbacks in CART

- The CART algorithm employs recursive binary partition.
    - Overfitting
    - Bias in variable selection especially when the explanatory variables present many possible splits or missing values.

Conditional Inference Trees

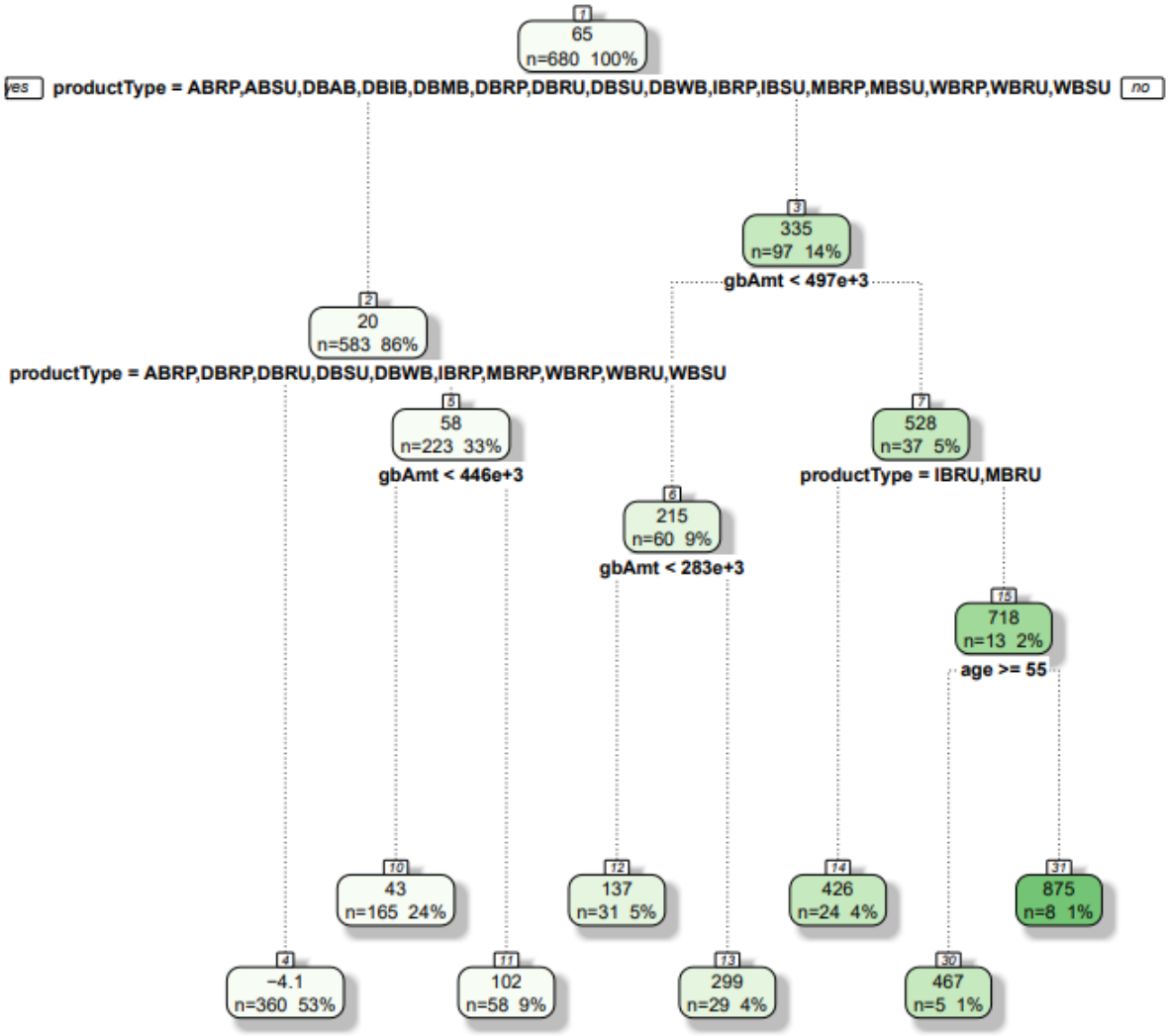
- Unbiased recursive binary splitting
    - Apply statistical permutation test to determine if there is any dependency between the response variable and the explanatory variables.
    - Find the most significant (strongest association) explanatory variable to perform the split.

Hyperparameter Optimization

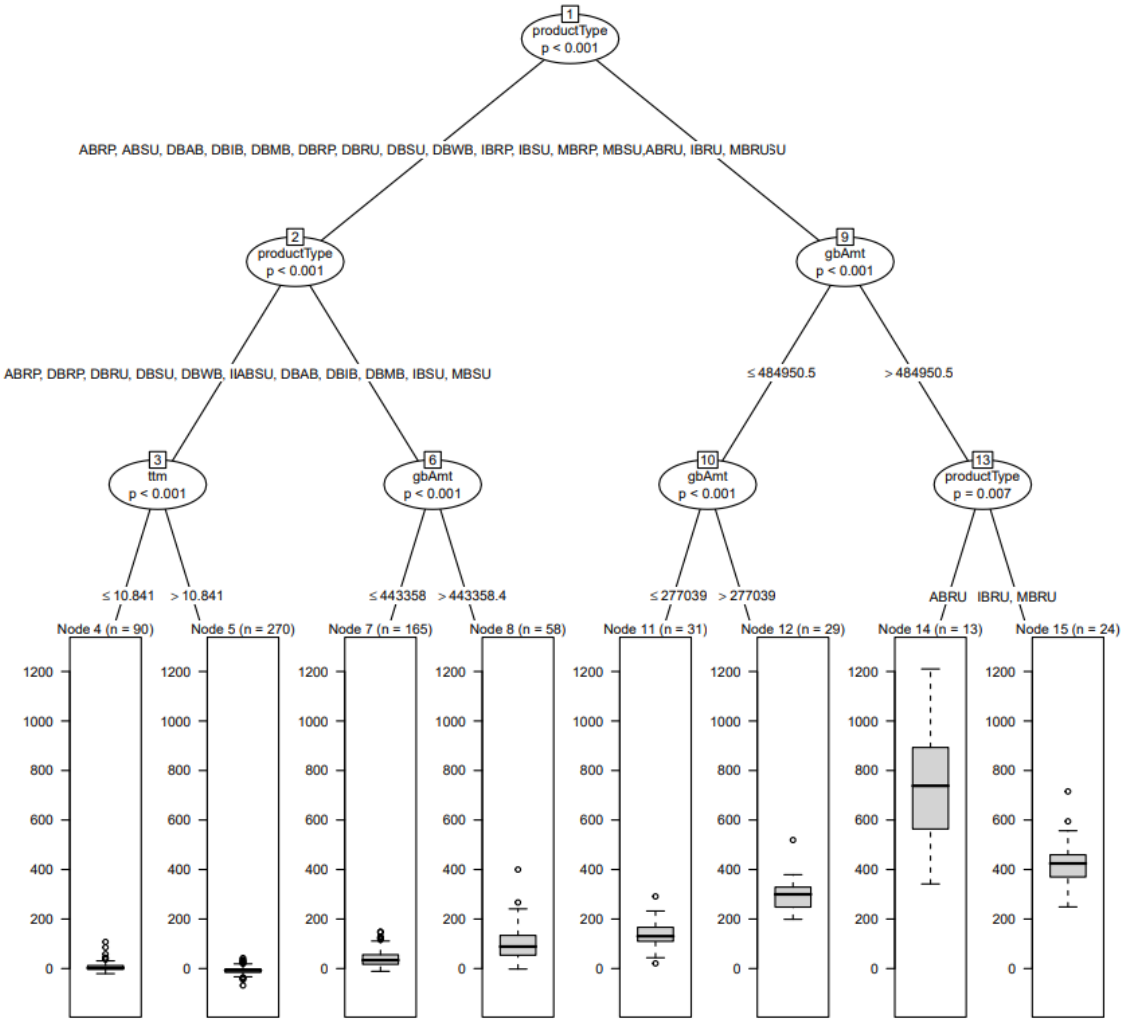
- Grid Search
- Random Search
- Automatic Hyperparameter Optimization
    - Bayesian optimization: models generalization performance as a sample from a Gaussian Process (GP) and creates a regression model to formalize the relationship between the model performance and the model hyperparameters.

Model	RdTrainMSE	BoTrainMSE	RdTestMSE	BoTestMSE	RdTime	BoTime
Regression tree (CART)	3846.0	3609.7	3717.8	3726.2	4.6 secs	1.7 mins
Bagged trees	1875.5	1992.3	1829.4	1860.2	120.7 secs	38.7 mins
Gradient boosting	2552.6	1487.3	3061.3	1918.2	251.1 secs	38.8 mins
Conditional inference trees	4090.0	3871.9	3778.5	3779.2	8.4 secs	6.4 mins
Conditional random forests	2792.1	1992.3	2421.1	1912.8	921.3 secs	29.4 mins

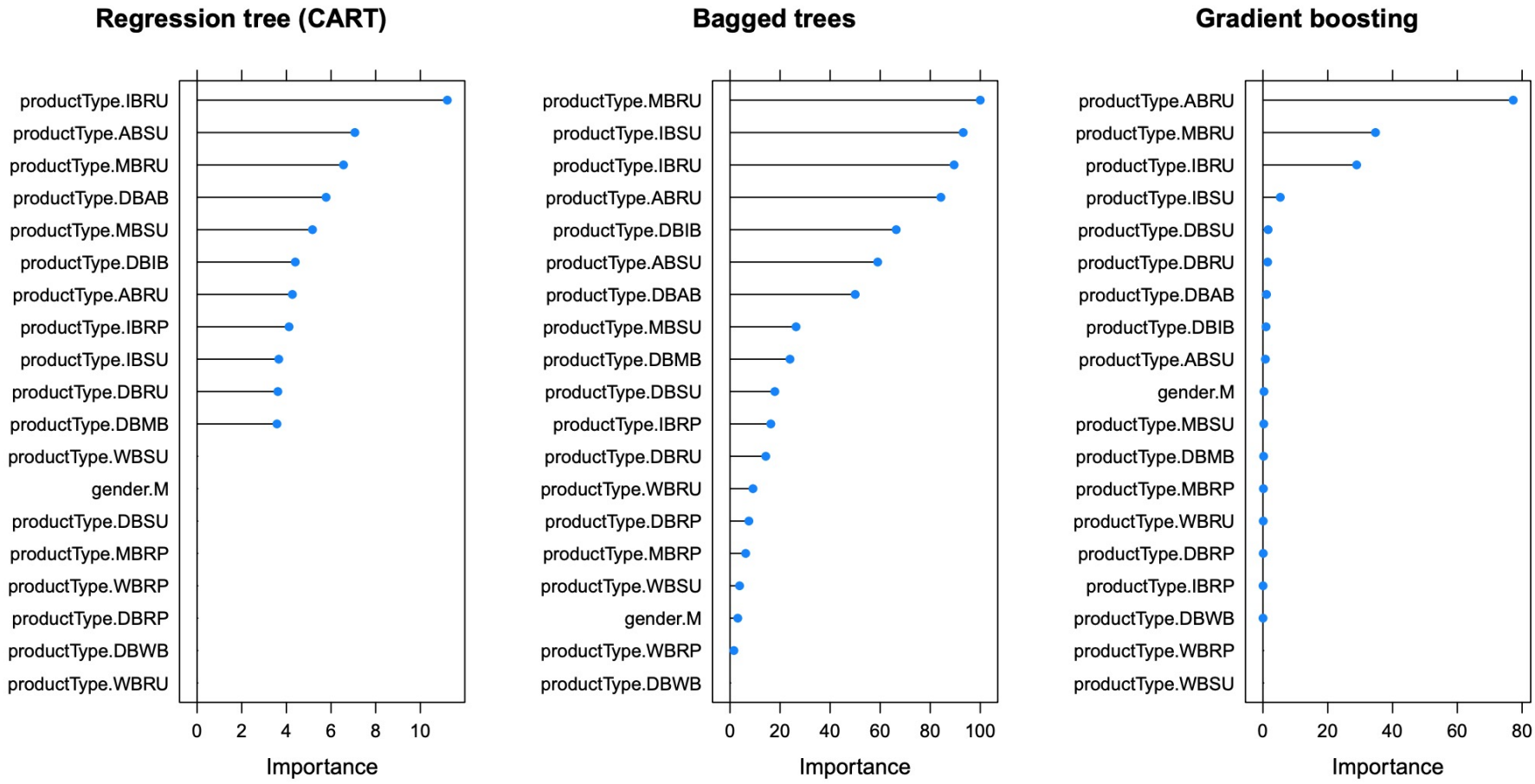
CART tree


1
65
n=680 100%
yes
productType = ABRP,ABSU,DBAB,DBIB,DBMB,DBRP,DBRU,DBSU,DBWB,IBRP,IBSU,MBRP,MBSU,WBRP,WBRU,WBSU
no
2
20
n=583 86%
productType = ABRP,DBRP,DBRU,DBSU,DBWB,IBRP,MBRP,WBRP,WBRU,WBSU
3
335
n=97 14%
gbAmt < 497e+3
5
58
n=223 33%
gbAmt < 446e+3
6
215
n=60 9%
gbAmt < 283e+3
7
528
n=37 5%
productType = IBRU,MBRU
15
718
n=13 2%
age >= 55
4
-4.1
n=360 53%
10
43
n=165 24%
11
102
n=58 9%
12
137
n=31 5%
13
299
n=29 4%
14
426
n=24 4%
30
467
n=5 1%
31
875
n=8 1%

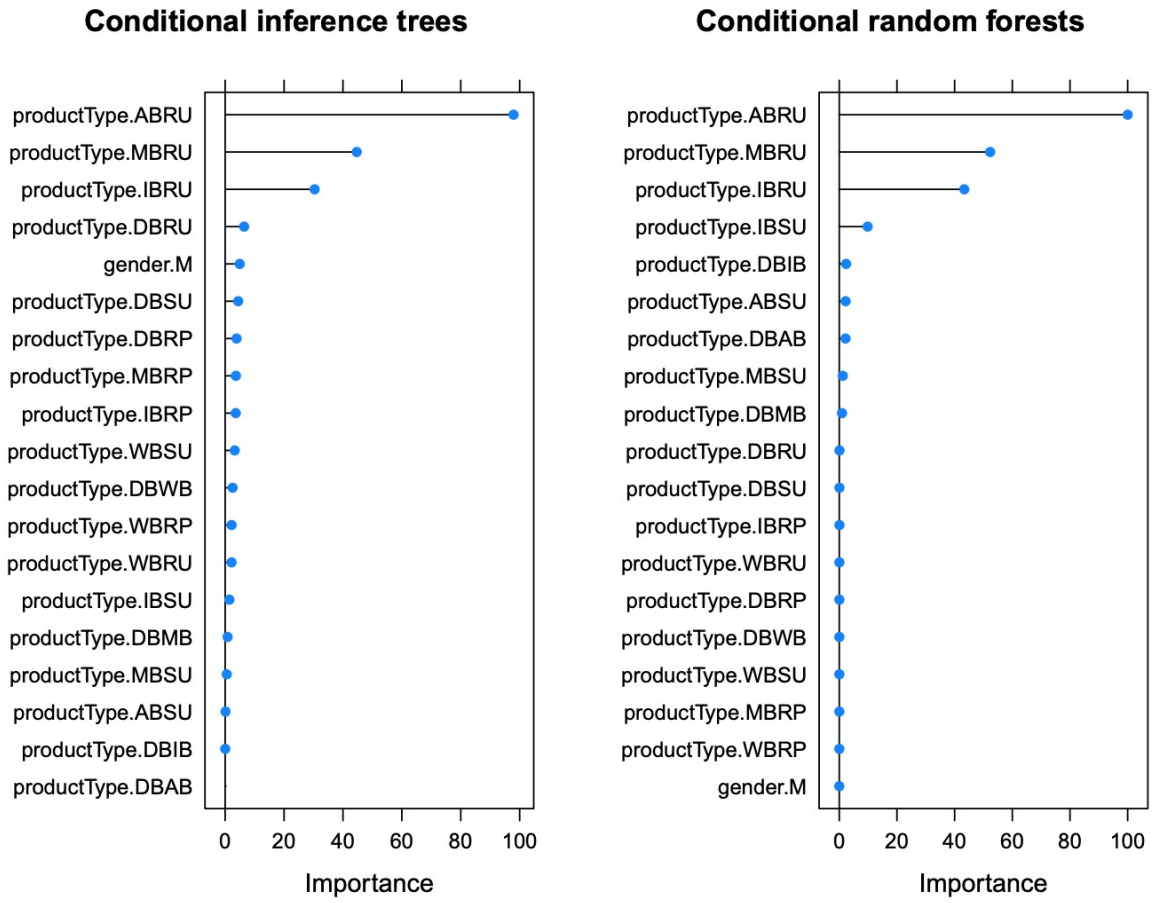
Conditional inference trees



Variable importance - recursive binary splitting



Variable importance - unbiased recursive binary splitting

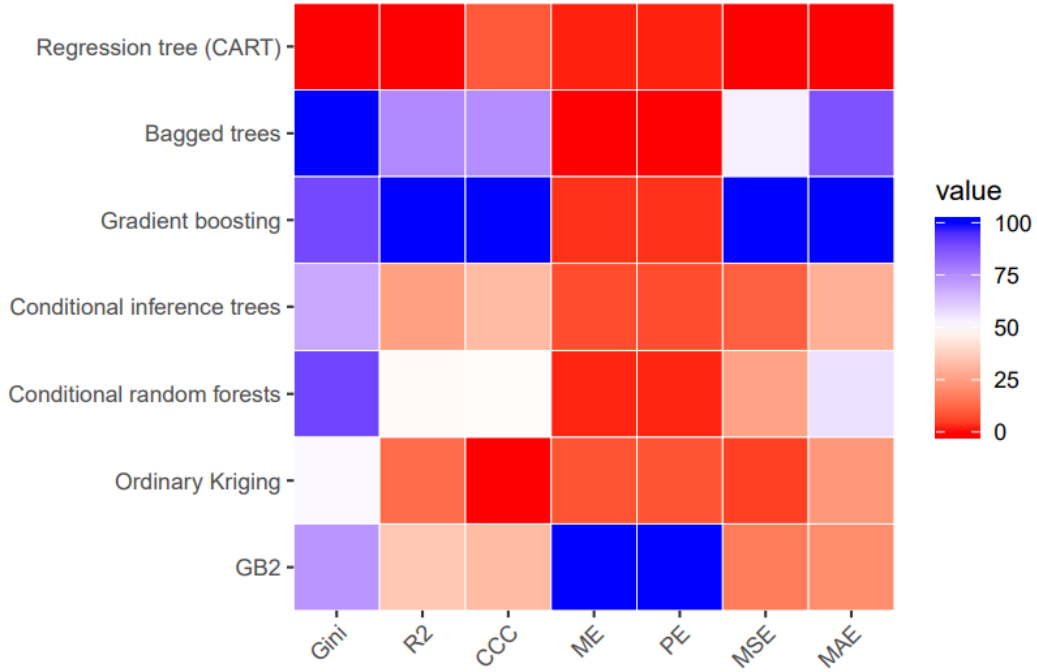


Computational expenses - tree-based models are efficient

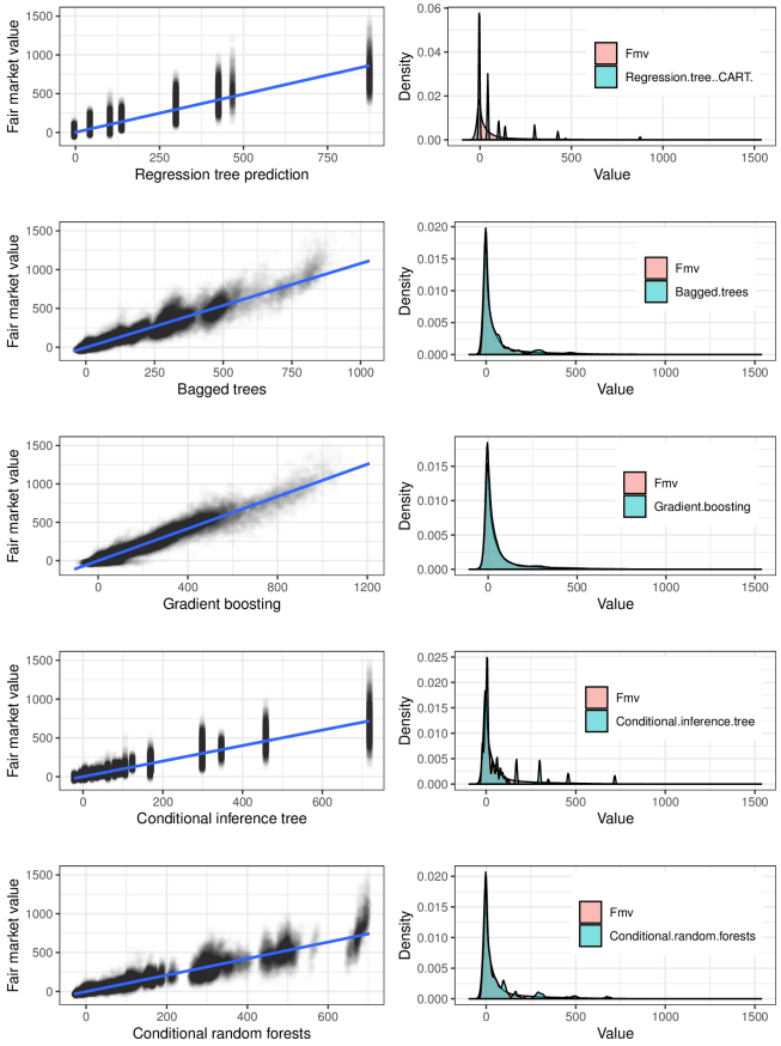
MODEL	COMPUTATIONTIME
Regression tree (CART)	0.13 secs
Bagged trees	2.70 secs
Gradient boosting	4.69 secs
Conditional inference trees	0.25 secs
Conditional random forests	1214.72 secs
Ordinary Kriging	277.49 secs
GB2	23.44 secs

Prediction accuracy - blue (good) vs red (bad)

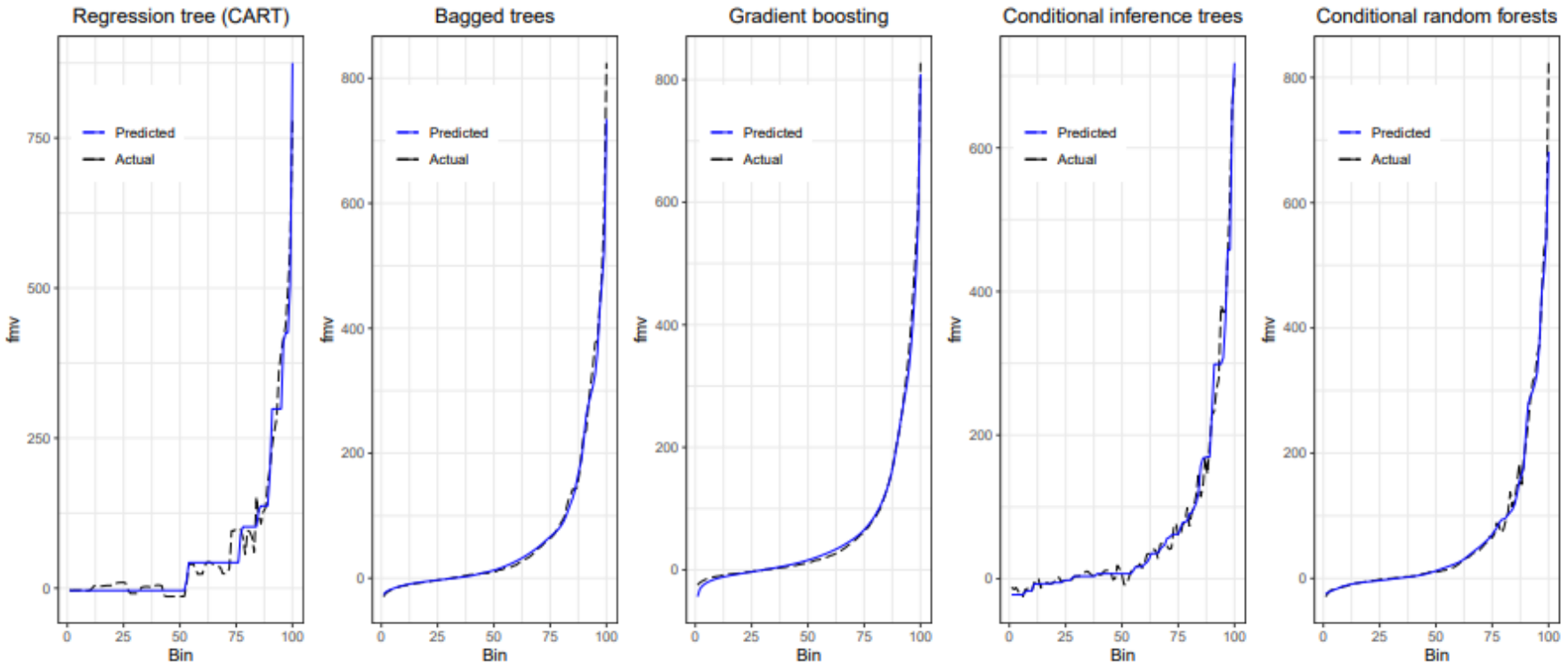
Model	*Gini*	R^2	*CCC*	*ME*	*PE*	*MSE*	*MAE*
Regression tree (CART)	0.786	0.845	0.917	1.678	-0.025	3278.578	31.421
Bagged trees	0.842	0.918	0.954	2.213	-0.033	1720.725	20.334
Gradient boosting	0.836	0.942	0.969	1.311	-0.019	1214.899	19.341
Conditional inference trees	0.824	0.869	0.930	0.905	-0.013	2754.853	26.536
Conditional random forests	0.836	0.892	0.940	1.596	-0.024	2273.385	23.219
Ordinary Kriging	0.815	0.857	0.912	-0.812	0.012	3006.192	27.429
GB2	0.827	0.879	0.930	0.106	-0.002	2554.246	27.772



Scatter plot - prediction vs actual



Lift curve - prediction performance used in insurance



Summary

- Tree-based models are generally efficient.
- Boosting performs best with respect to prediction accuracy.
- Variable importance for risk identification.

Reference

- Morgan, J. N. and Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. Journal of the American Statistical Association, 58(302):415–434.
- Loh, W. Y. (2014). Fifty years of classification and regression trees. International Statistical Review, 82(3), 329-348.
- Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). Classification and Regression Trees. Taylor & Francis Group, LLC: Boca Raton, FL.
- Breiman, L. (2001). Random forests. Machine Learning, 45(1):5–32.
- Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of Statistics, 29(5):1189–1232.
- Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical statistics, 15(3), 651-674.
- Gan, G., & Valdez, E. A. (2017). Valuation of large variable annuity portfolios: Monte Carlo simulation and synthetic datasets. Dependence Modeling, 5(1), 354-374.
- Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based recursive partitioning. Journal of Computational and Graphical Statistics, 17(2), 492-514.

Acknowledgment

We would like to thank the Society of Actuaries for the funding support of this research project through our Centers of Actuarial Excellence (CAE) grant on data mining.

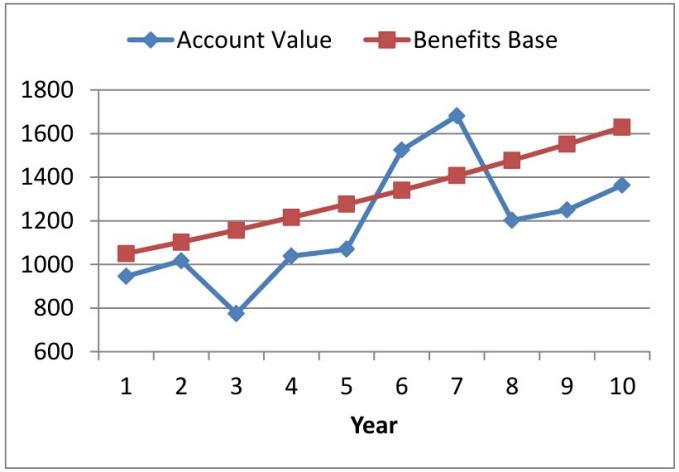
Q&A

Thank you for your attention!

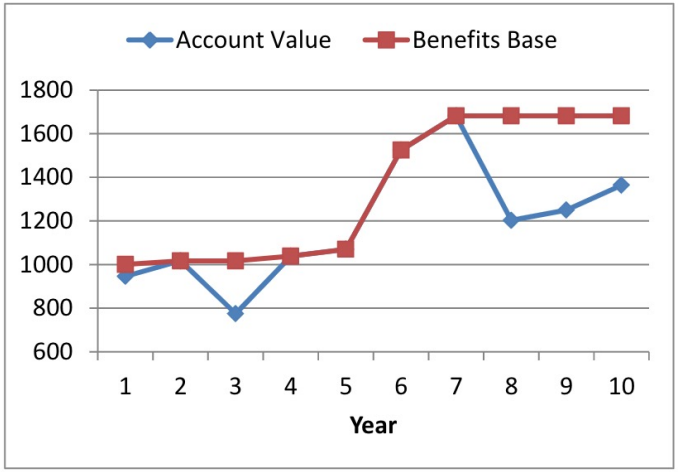
Appendix A: Prediction accuracy measures

Validation measure	Description	Interpretation
Gini Index	$Gini = 1 - \frac{2}{N-1}\left(N - \frac{\sum_{i=1}^{N} i\tilde{y}_i}{\sum_{i=1}^{N}\tilde{y}_i}\right)$ where $\tilde{y}$ is the corresponding to y after ranking the corresponding predicted values $\widehat{y}$.	Higher Gini is better.
Coefficient of Determination	$R^2 = 1 - \frac{\sum_{i=1}^{N}(\widehat{y}_i - y_i)^2}{\sum_{i=1}^{N}\left(y_i - \frac{1}{n}\sum_{i=1}^{n} y_i\right)^2}$ where $\widehat{y}$ is predicted values.	Higher R^2 is better.
Concordance Correlation Coefficient	$CCC = \frac{2\rho\sigma_{\widehat{y}_i}\sigma_{y_i}}{\sigma^2_{\widehat{y}_i}+\sigma^2_{y_i}+(\mu_{\widehat{y}_i}-\mu_{y_i})^2}$ where $\mu_{\widehat{y}_i}$ and μ_{y_i} are the means $\sigma^2_{\widehat{y}_i}$ and $\sigma^2_{y_i}$ are the variances ρ is the correlation coefficient	Higher CCC is better.
Mean Error	$ME = \frac{1}{N}\sum_{i=1}^{N}(\widehat{y}_i - y_i)$	Lower $\lvert ME\rvert$ is better.
Percentage Error	$PE = \frac{\sum_{i=1}^{N}\widehat{y}_i - \sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} y_i}$	Lower $\lvert PE\rvert$ is better.
Mean Squared Error	$MSE = \frac{1}{N}\sum_{i=1}^{N}(\widehat{y}_i - y_i)^2$	Lower MSE is better
Mean Absolute Error	$MAE = \frac{1}{N}\sum_{i=1}^{N}\lvert\widehat{y}_i - y_i\rvert$	Lower MAE is better.

Appendix B: Variable annuities provide guaranteed appreciation of the benefits base



(Roll-up)



(Ratchet)