




Annuity Purchaser Case Study



Background: A financial institution wants to predict who will purchase variable annuity product.

- Applying predictive modelling on customer data: comparing logistic regression, decision tree, and random forest algorithm.
 - Feature selection: stepwise selection using AIC criteria.
 - Data set split: training (70%) and holdout (30%).
 - Decision Tree: grow and prune the tree.
 - RF: select optimal number of variables in each split.
 - Model evaluation: AUC, ROC on testing data set.

Variable Description:

- Ins: purchased variable annuity (1=yes, 0=no).
- DDA: checking account (1=yes, 0=no).
- DDABal: checking account balance.
- Depamt: amounts deposited.
- CashBk: number of times customer received cashback.
- Checks: number of checks.
- Phone: number of times customers use telephone banking.
- Sav: saving account (1=yes, 0=no).
- Savbal: saving account balance.
- IRA: has retirement account (1=yes, 0=no).
- Mtg: has mortgage account (1=yes, 0=no).

Approach and tools: using R glm().

- R is used to build model, evaluate model and plot charts.
 - Libraries: MASS, tree, stats, ROCR, caret, randomForest
 - Data cleansing: imputation of missing values, remove outliers.
 - Split data to training (70%) and holdout (30%).
 - Stepwise regression: select model has lowest AIC.
Var selected: dda, ddabal, depamt, cashbk, checks, phone, sav, savbal, ira, mtg.
 - *Model: `logi_all <- glm(ins~dda+ddabal+depamt+cashbk+...)`*
 - *Prune tree: `pruned_model <- prune.tree(tree_model,best=5)`*
 - *Random Forest: `randomForest(as.factor(ins)~. , data=...)`*

Result discussion: logistic regression

- Stepwise logistic regression

- Check overall significance by chi-square test:

1- pchisq(n1-n2, df1-df2)

```
> 1-pchisq(logi_all2$null.deviance-logi_all2$deviance,  
+          logi_all2$df.null-logi_all2$df.residual)  
[1] 0
```

- This “0” indicates the overall model is significant.

Result discussion: logistic regression (continued)

■ Stepwise logistic regression

- Promote to people have saving account or IRA account.
- Having a checking account, or frequent user of phone banking or cash back have negative impact on purchasing. (lower odds)
- Table at right-hand is standardized coefficients. Key factors are: checking acct (-), saving acct balance (+), checking acct balance (+), saving acct balance (+), IRA acct (+).

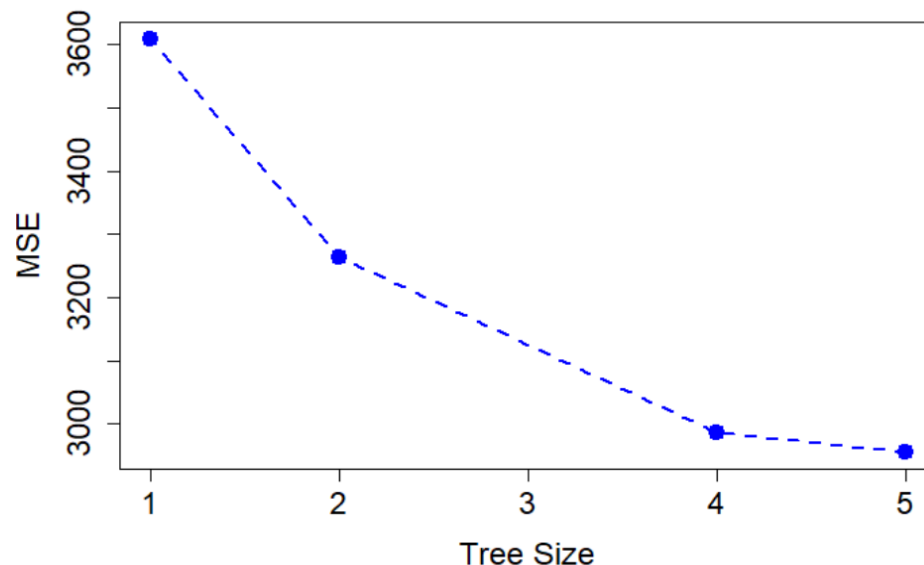
Coefficients:

	Estimate	Std. Error	z value	Pr(> z)		Overall
(Intercept)	-2.059e-01	4.200e-02	-4.904	9.41e-07 ***	dda1	21.553904
dda1	-1.086e+00	5.040e-02	-21.554	< 2e-16 ***	ddabal	14.025995
ddabal	6.923e-05	4.936e-06	14.026	< 2e-16 ***	depamt	5.367315
depamt	2.750e-05	5.123e-06	5.367	7.99e-08 ***	cashbk	2.466321
cashbk	-3.813e-01	1.546e-01	-2.466	0.0137 *	checks	2.387524
checks	-1.027e-02	4.300e-03	-2.388	0.0170 *	phone	6.641631
phone	-1.364e-01	2.053e-02	-6.642	3.10e-11 ***	sav1	12.103647
sav1	4.813e-01	3.977e-02	12.104	< 2e-16 ***	savbal	15.426338
savbal	5.229e-05	3.390e-06	15.426	< 2e-16 ***	ira1	9.432574
ira1	7.620e-01	8.078e-02	9.433	< 2e-16 ***	mtg1	1.531410
mtg1	-1.328e-01	8.674e-02	-1.531	0.1257		

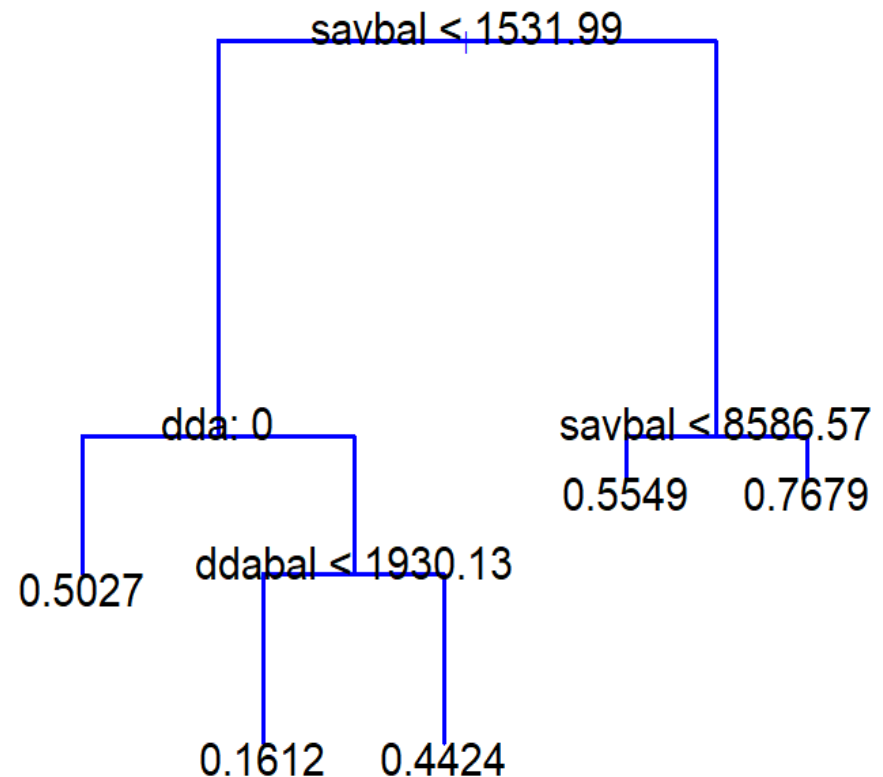
By Simon Liu

Result discussion: decision tree

■ Decision tree (after pruning)



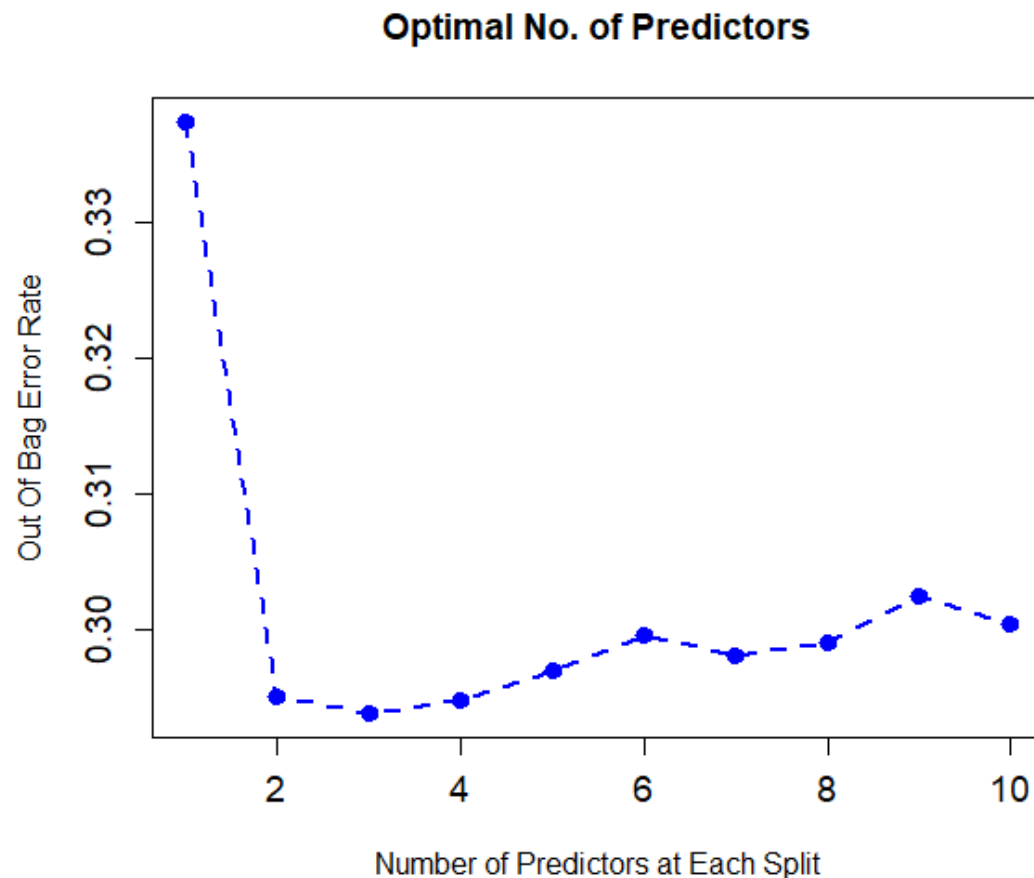
- Most important factor is saving account balances.
- The more savings deposited, the more likely you will purchase annuity product.



Result discussion: Random Forest

- Determine the number of variables at each split.

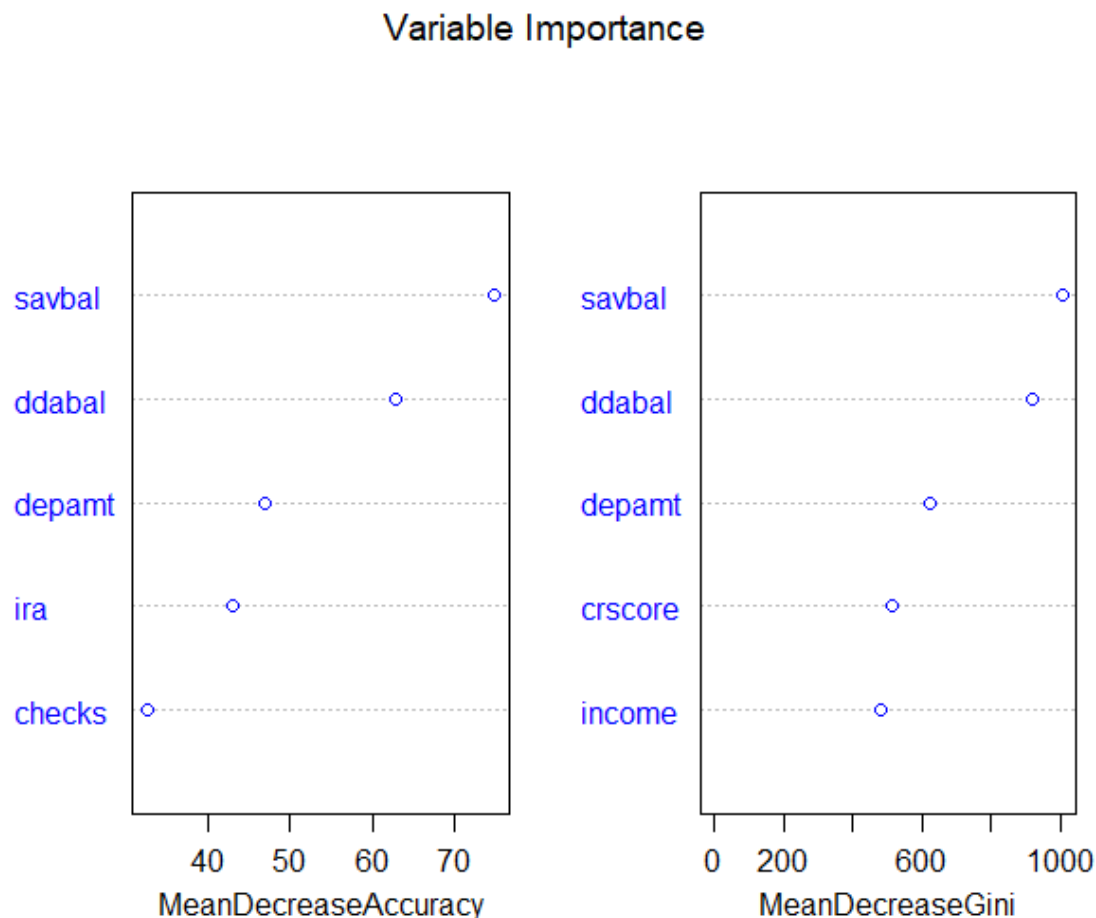
- Try different number of predictors in the model and see which one has min. OOB error rate.
- 3 is the optimal number of predictors as it has smallest OOB error rate.



Result discussion: Random Forest (continued)

- Chart shows variables' importance ranking in both decrease of Gini and decrease of accuracy.

- Ranking is not exactly the same with result of logistic regression. Still, saving acct balance and checking acct balance are important.



Model evaluation: AUC and ROC

ROC curve and AUC

- RandomForest has the best performance at almost every cutoff point. (AUC=0.77)
- Using RandomForest really increased model performance compared with simple trees.

