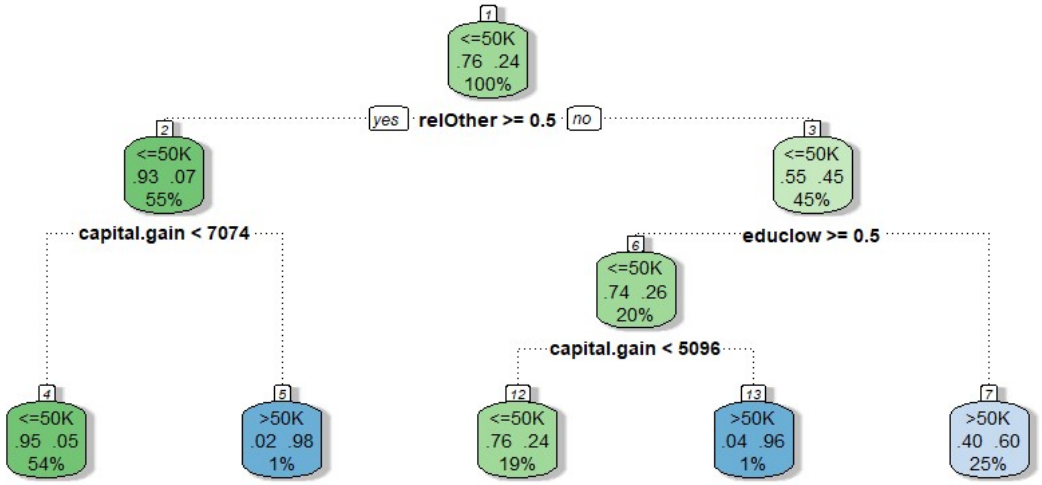
This project seeks to identify high income customers through information attained in previous analysis and through the use modeling to produce predictions of a customer's income. These predictions are based on information about them such as their level of education, capital gains, and marital status. The data consists of 32561 individuals of which 7841 or 24.01% of them are high-income.

- Base line model performance was established using naïve models. An all-positive model and an all-negative model were produced. The all-negative model was chosen as the baseline to beat due to its higher accuracy of 75.92% compared to the all-positive's accuracy of 24.10%
- 4 Metrics were used to analyze the quality of models produced, accuracy, error rate, sensitivity and specificity. The model trained with the data referred to as CART was able to outperform the baseline models in both accuracy and error rate, with an accuracy of 82.04% and error rate of 17.95%. The CART model's sensitivity and specificity were 69.56% and 86% respectively.
- The decision tree produced by the CART model terminated in 5 leaf nodes which allowed the creation of 5 different decision rules for the classification of customer income. 3 of these rules, rules 2,4, and 5 use customer information to predict high income. While rule 1 and 3 predict low income
- Using all the information found it was possible to create a profile for high and low-income customers. Key features that identify a high-income customers include, being married those who were married were 7 times as likely to also be high income compared to those with other relationship status. A customer who has received some college education is also more likely to be high-income. They also more frequently have some form of capital gains, and the amount of capital gains can be used to make predictions of a customer's income.

Now that a model has been made that outperforms baseline model it can be deployed to help predict future customers incomes monitoring and adjusting for new real-world data. Decision rules and customer profiles can assist in targeting and marketing to the right people to find future customers who are more likely to be high income such as those with college degrees.



Rattle 2022-Apr-18 15:02:21 alexa

The root node of the tree splits on the relationship value which indicates whether the person is married or some other status. The tree has 5 leaf nodes 3 of which predict high income and 2 predicting low income. The leaf node which has the most support is leaf node 1 which predicts low-income with 54% support

	Accuracy	Kappa	Resample
1	**0.8089681**	0.5143105	Fold02
2	0.8323096	0.5608556	Fold01
3	0.8273956	0.5509825	Fold03
4	0.8323096	0.5715740	Fold06
5	**0.8402948**	0.5757527	Fold05
6	0.8206388	0.5337915	Fold04
7	0.8144963	0.5105329	Fold07
8	0.8138821	0.5117837	Fold10
9	0.8230958	0.5178281	Fold09
10	0.8238183	0.5474277	Fold08

The cross folds show no serious evidence of overfitting. The variance of accuracy between the lowest fold at 80.89% accuracy and the highest accuracy fold at 84.03%, is only 3.14%.

Predicted category				
Actual Category		<=50k	>50k	Total
	<=50k	TN = 10630	FP = 1730	TAN = 12360
	>50k	FN = 1193	TP = 2727	TAP = 3920
	Total	TPN = 11823	TPP = 4457	GT = 16280

The model has an accuracy of 82.05% this level of accuracy is greater than the established baseline of 75.92% so, the model has outperformed the naïve baseline.

	All positive model	All negative model	CART model
TN	0	12360	10630
FP	12360	0	1730
FN	0	3921	1193
TP	3921	0	2727
Accuracy	24.10%	75.9%	82.04%
Error rate	75.9%	24.1%	17.95%
Sensitivity	100%	0%	69.56%
Specificity	0%	100%	86.00%

The all-positive model is the worst of the 3 models of the 4 metrics used to evaluate the models all positive model has the worst accuracy of 24.10%, error rate of 75.%, and specificity at 0%. The sensitivity of all positive model is the best of 3 models though along with it having the highest rate of truly identified positives but due to its naïve assumptions it also has the highest rate of false positives. The all-negative model has the opposite issue featuring the highest rate of true negatives and false negatives.
The all-negative model does better than all positive model in accuracy and error rate. With an accuracy of 75.9% and error rate of 24.1% but these values were outperformed. Meanwhile it has the high specificity and lowest sensitivity of the 3 models. This model was chosen as the baseline to beat for further modeling efforts due to its accuracy being higher than all positive model.

The CART model has the best accuracy 82.04% and error rate 17.95% of the 3 models. The relative decrease in error rate from the chosen baseline, the all-negative model to the CART model is 25.51%. Although it's sensitivity 86% and specificity 69.56% are not the best of the 3 models it has relatively good values for both metrics suggesting that the CART model can predict for both classes of income successfully, although its specificity is higher than its sensitivity suggesting that the CART model is better suited to identifying low-income customers.

Decision rules

1. If the customer is not married, and has capital gain less than 7074, then they are low income. With 54% support and 95% confidence.
2. If the customer is not married and has capital gain greater than 7074 then they are high income. With 1% support and 98% confidence
3. If the customer is married and has low education and has capital gain less than 5096 then they are low income. With 19% support and 76% confidence.
4. If the customer is married and has low education and has capital gain greater than 5096 then they are high income. With 1% support and 96% confidence.
5. If the customer is married and has high education, then they are high income. With 25% support and 60% confidence

If we are looking to find high income customers, then the decision rules that would best aide that task are 2, 4, and 5. These rules provide information on what a high-income customer's information may look like. I think that decision rule 5 would be most useful for identifying high income customers. Although it features less confidence than rules 2 and 4 it comes with much greater support of 25% compared to the other rules only 1% support each.

Profiles

High-income customer.

High income customers are more likely to be married. Analysis of the data showed that when a customer is married, they are 7 times more likely to be high income. This finding is further supported by the decision tree model which bases its root node split on the relationship variable. Following that branch of the decision tree we see that it next predicts High-income based on education. High-income customers are more likely to be highly educated, those who were highly educated were 2.55 times as likely to be high-income. Following the decision tree to the right of root node we see that a leaf node follows directly after a split made on the education variable. This leaf node predicts high income creating decision rule number 5 which has 25% support and 60% confidence.
High-income customers were also 3 times as likely to have some form of capital gains or losses. From the decision tree and rules derived from it we can see that, customers who had values of capital gain greater than 5096 or 7074 the number being dependent on other features of the customer's records were predicted to be high income.

Low-income customer.

Low-income records often indicated that their relationship status as something other than married. 93.65% of customers who indicated that they were not married were also low-income. Customers who had lower values for capital gains or losses or no capital gains or losses at all were more likely to be low income. Decision rule 1 supports these two findings, if a customer is not married and has a capital gain less than 7074 Then they are predicted to be low-income decision rule 1 has 54% support and 95% confidence in this prediction. A low-income customer has also usually received less education, 87% of customers who indicated that they had received no college education we're also low-income customers. Combing the information that low-income customers are not married and tend to have low capital gains and lastly have low education, leads us to decision rule 3 which predicts that customers who meet this criterion are low income with 19% support and 76% confidence.