Predicting Low Income Levels

Background and Aims

Governments interested in engaging in affirmative action to benefit lower income populations may find it useful to have a model that predicts whether people with certain characteristics would have earned less than 50 thousand dollars per year in 1994. This model would enable them to input the same characteristics for people today to see whether the income level they would likely have had in 1994 would have been less than 50 thousand dollars per year which could be used to evaluate and substantiate any findings on social welfare, say, improvement in living conditions or individual growth in paygrade. The aim of this project is to provide such a model that predicts whether income in the United States is less than fifty thousand dollars per year.

Summary of Results

The following table shows a list of the well-performing classifiers we evaluated with the best model in each method, or local optimal model, highlighted with a darker colour hue.

Summary of Results and Model Selection

Classifier	Resampling Approach	Kappa	Sensitivity	Specificity	AUROC
Logistic Regression, with Box Cox transformation	None	0.5629	0.9318	0.5947	0.906
	Over-sampling	0.5561	0.7999	0.8485	
	Under-sampling	0.5537	0.7979	0.8491	
	Both	0.5544	0.7998	0.8463	
Linear Discriminant Analysis, with Box Cox transformation	None	0.5269	0.9306	0.5559	0.892
	Over-sampling	0.5128	0.7592	0.8642	
	Under-sampling	0.5127	0.7569	0.8685	
	Both	0.5153	0.7596	0.8671	
Classification Trees	None	0.5629	0.9318	0.5947	0.875
	Over-sampling	0.5128	0.7592	0.8642	
	Under-sampling	0.5117	0.7562	0.8682	
	Both	0.5156	0.7601	0.8668	
C5.0	None	0.6099	0.9451	0.6224	0.917
	Over-sampling	0.5919	0.8442	0.8095	
	Under-sampling	0.5889	0.8102	0.8754	
	Both	0.5919	0.8442	0.8095	
Stacking (cart, lda, C5.0) with logistic	None	0.6129	0.9437	0.6289	0.920
Stacking (cart, lda, C5.0) with cart	None	0.6246	0.9103	0.7142	0.812

Figure 1: Model Summary Table

Conclusion and Takeaways

We obtained our best model through the use of the technique of Stacking, in which we combined three models (LDA, CART, and C5.0), attaining a sensitivity of 94.37%. This result means that given the set of feature values (age, education level, etc.) our model would correctly classify about 94% of people who make less than

fifty thousand dollars per year, although it would only correctly classify about 63% of people who make more than that.

For a welfare system that should benefit those in greatest need, even at the expense of wasting some subsidies on the less deprived, our model could be an effective tool for predicting which people have lower income based only on a set of survey characteristics.

There are some notable takeaways from this project. Having models trained on different resampling methods taught us that resampling a class imbalance dataset might not deliver results compatible with preliminary objectives. Although there is likely no algorithm that performs consistently well every time, understanding the strengths and weakness of each method still gives one an edge over randomly fitting a myriad of models and hoping for the best. Overall, we are pleased with our approach in optimizing predictive performance. Many other algorithms were attempted, but they were computationally slow to train and tune. Given excess time, we may look into implementing parallel processing in R to speed up some of the computationally expensive tasks, like tuning Support Vector Machines and Random Forest models. We could also perform a deeper analysis to justify the grouping of levels in categorical features, because the grouping was done intuitively without thorough justification.