Team WAAT

Wojciech Maliszewski
Adithya Harsha
Ahmed Safdar
Talha Ahmed
DSC- 540, DePaul IL

Adult Census Income Dataset

Predict if individual income is >50k or <50k

- 30,000+ observations
- 14 predictor variables
- Binary response variable
- Source: https://www.kaggle.com/datasets/uciml/adult-census-income (1994)

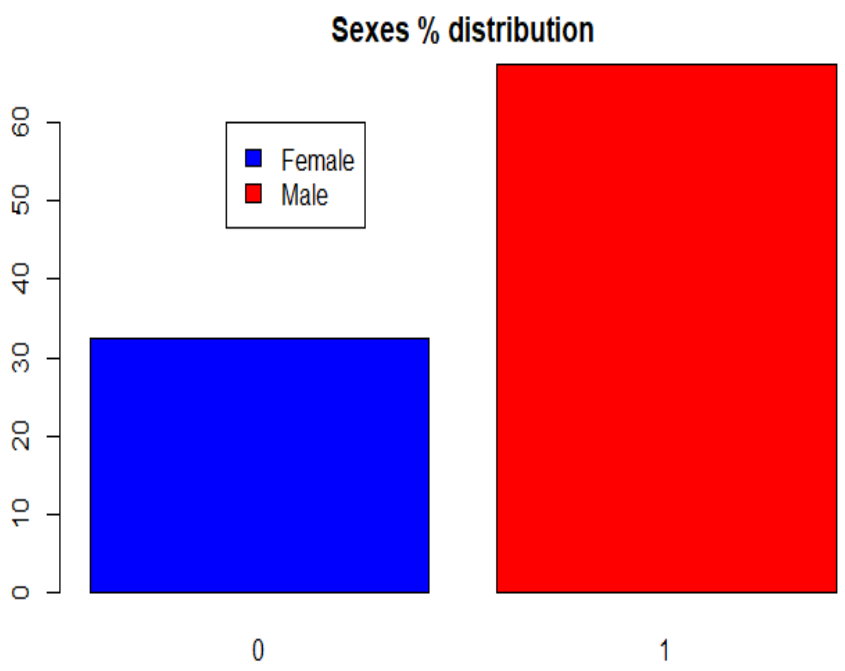
Snapshot of dataset

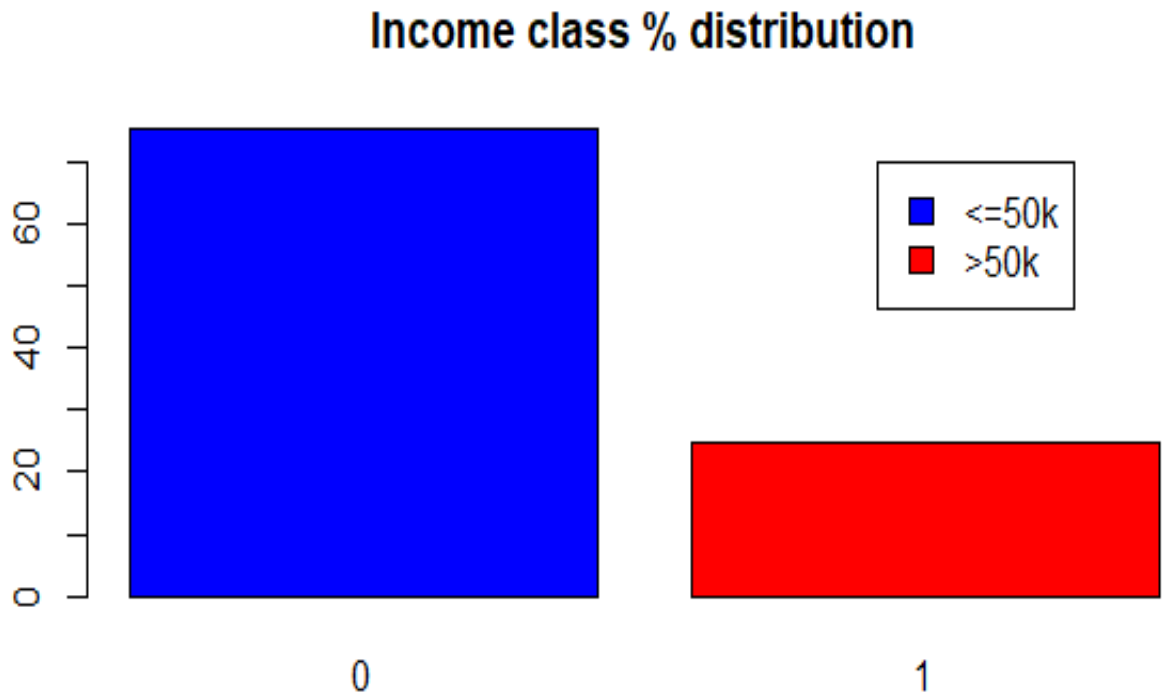
age	workclass	fnlwgt	education.num	marital.status	occupation	relationship	race	sex	capital.gain	capital.loss	hours.per.week	native.country	income
82	Private	132870	9	Widowed	Exec-managerial	Not-in-family	White	Female	0	4356	18	United-States	<=50K
54	Private	140359	4	Divorced	Machine-op-inspct	Unmarried	White	Female	0	3900	40	United-States	<=50K
41	Private	264663	10	Separated	Prof-specialty	Own-child	White	Female	0	3900	40	United-States	<=50K
34	Private	216864	9	Divorced	Other-service	Unmarried	White	Female	0	3770	45	United-States	<=50K
38	Private	150601	6	Separated	Adm-clerical	Unmarried	White	Male	0	3770	40	United-States	<=50K
74	State-gov	88638	16	Never-married	Prof-specialty	Other-relative	White	Female	0	3683	20	United-States	>50K
68	Federal-gov	422013	9	Divorced	Prof-specialty	Not-in-family	White	Female	0	3683	40	United-States	<=50K
45	Private	172274	16	Divorced	Prof-specialty	Unmarried	Black	Female	0	3004	35	United-States	>50K

Cleaning and Encoding

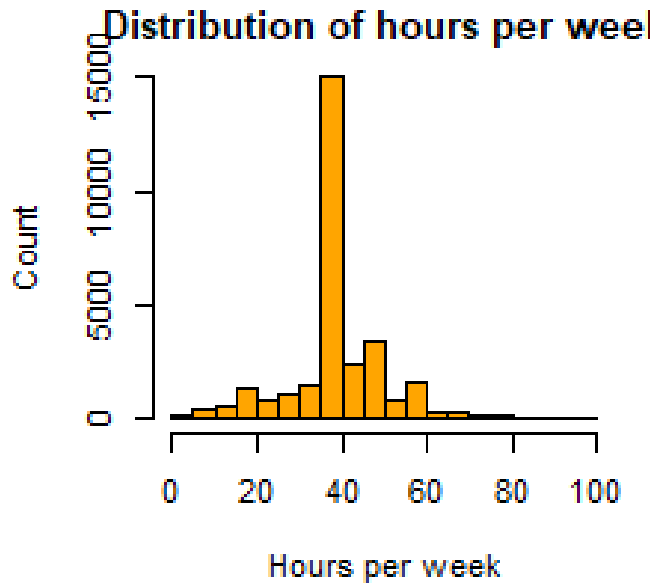
- No. of rows deleted due to missing values: 4262
- Dimensions: 30162 x 14
- Sex ----> binary
- Income variable ----> 0 for <50k, 1 for >50k
- Creation of dummy variables

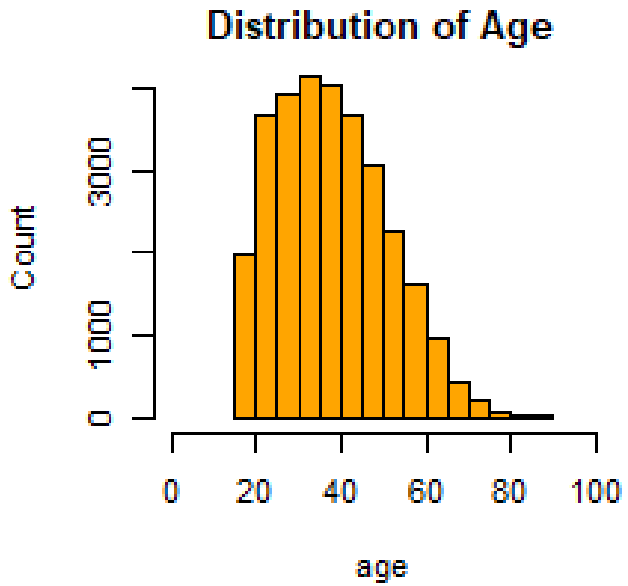
Exploratory Analysis

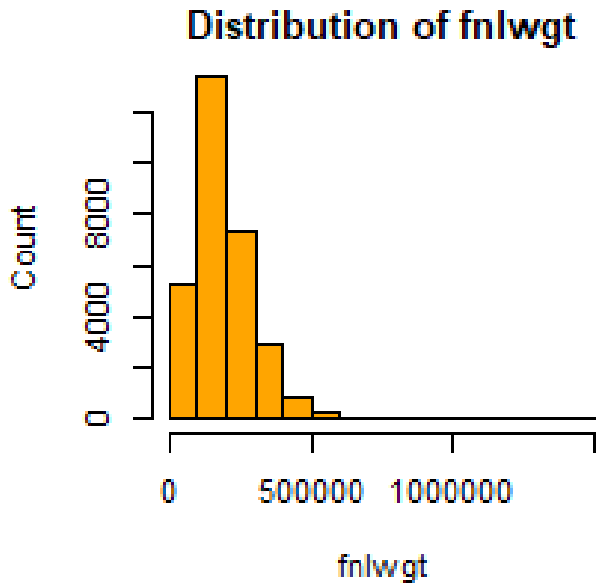




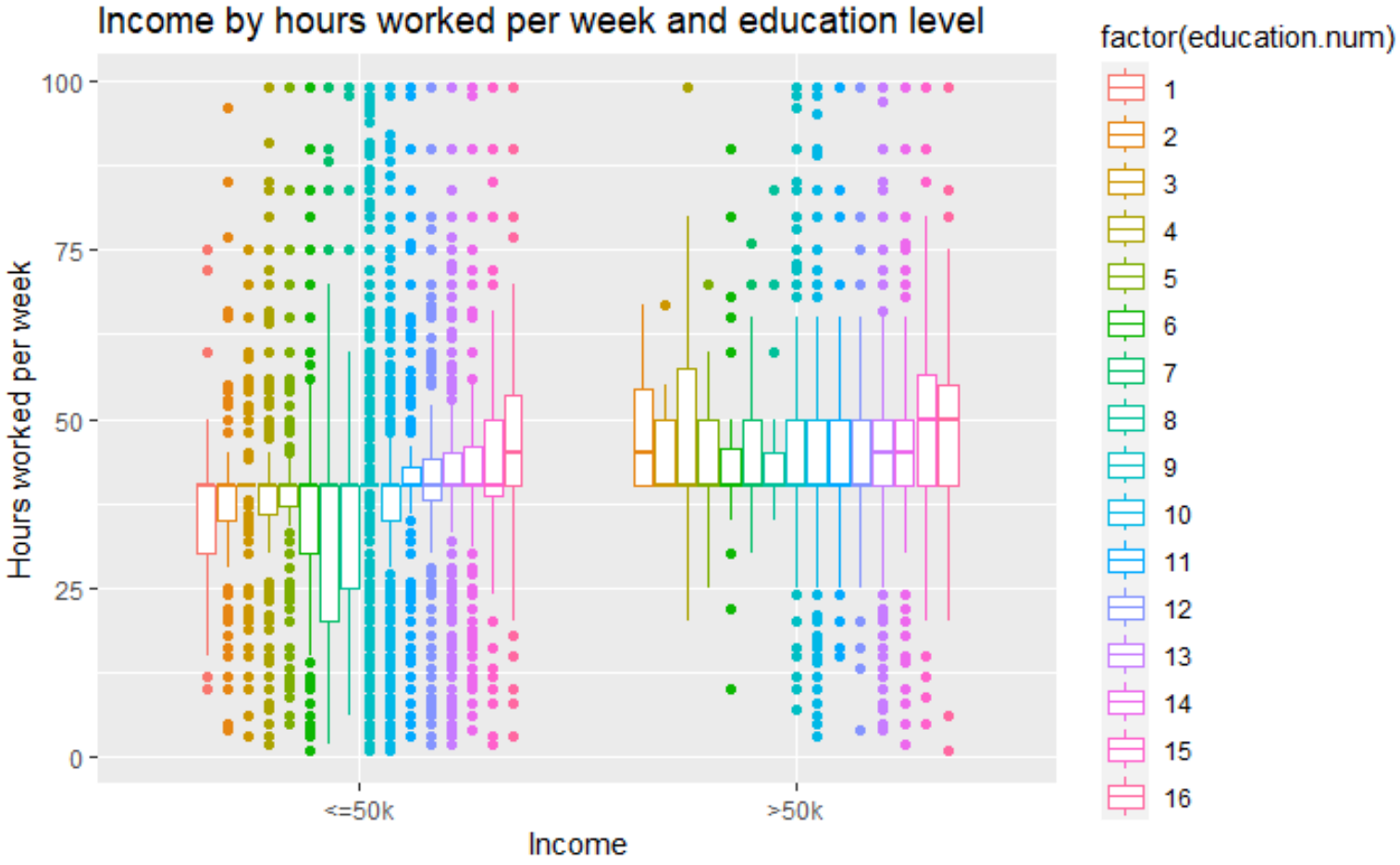
Exploratory Analysis





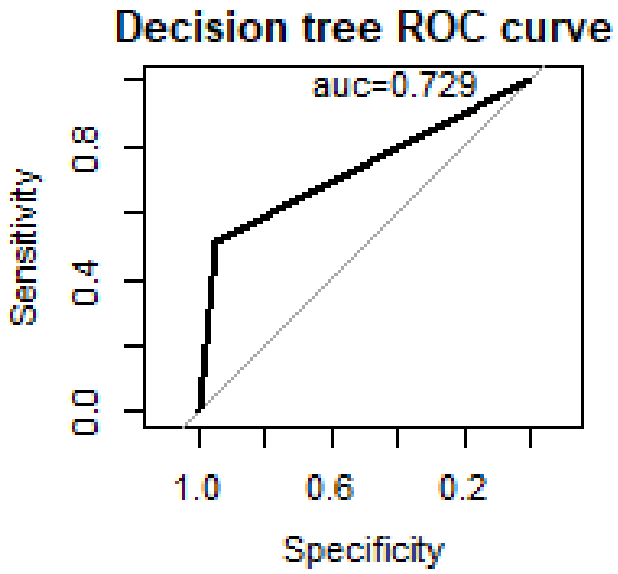


Exploratory Analysis

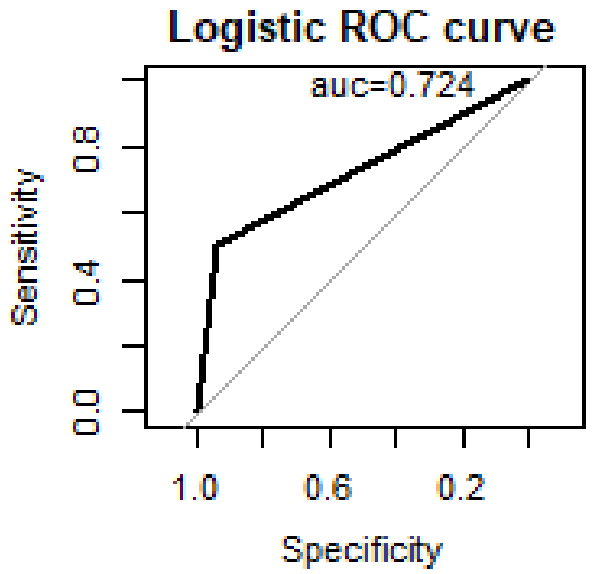


Initial Models (80:20 split)

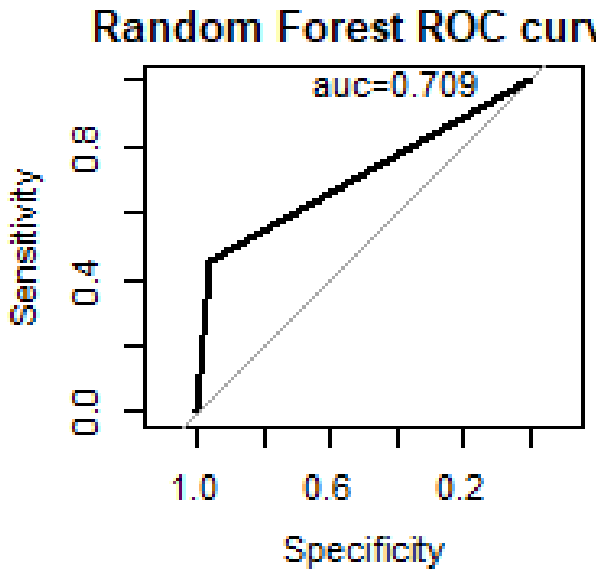
Decision Tree



Logistic Regression

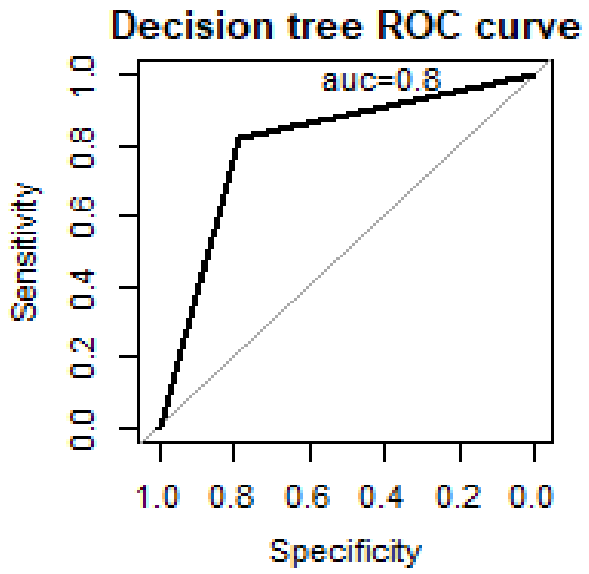
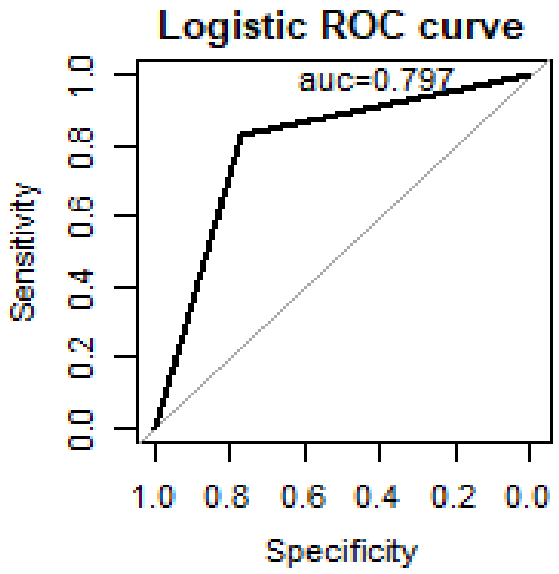
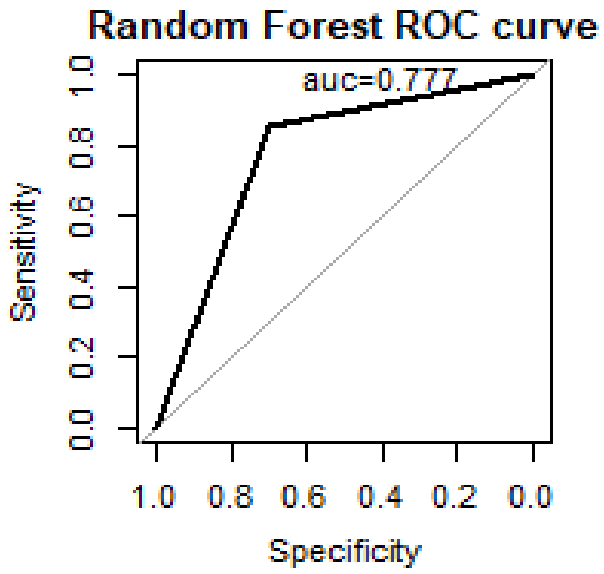


Random Forest

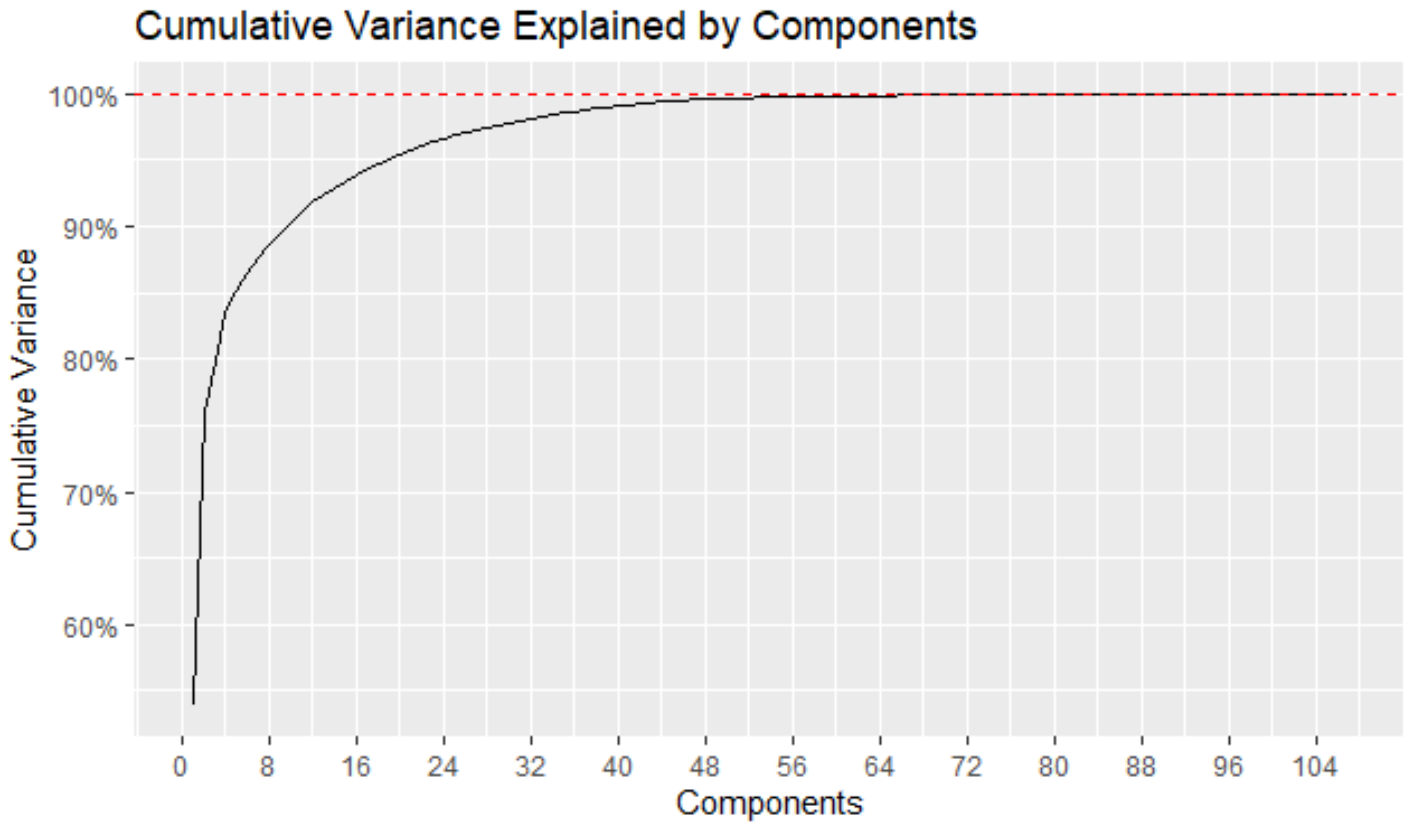


Balanced Accuracy		
0.7286	0.7236	0.7093

Initial Models (30:70 split to address class imbalance)

Decision Tree	Logistic Regression	Random Forest
		
Balanced Accuracy		
0.8003	0.7972	0.7775

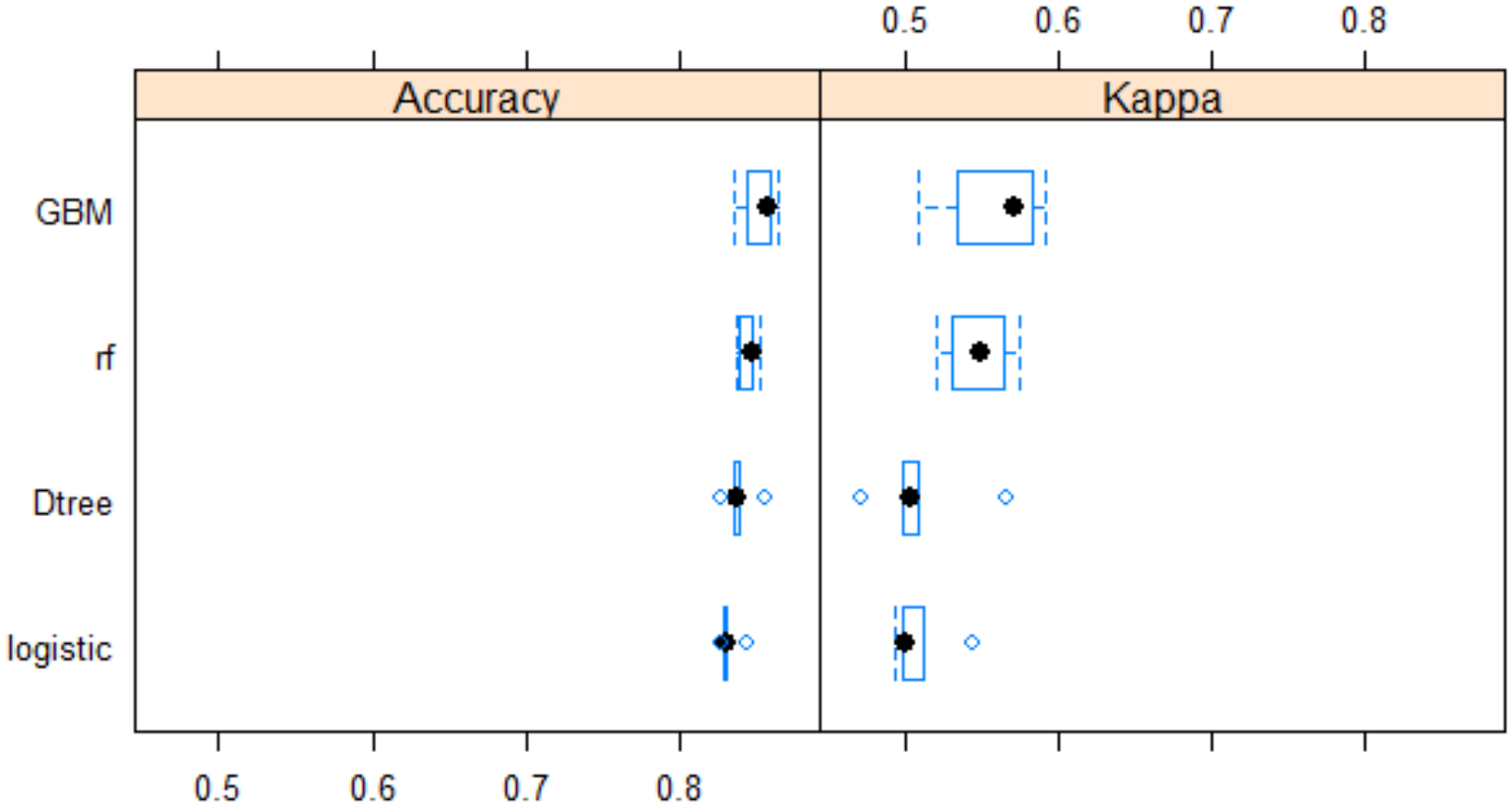
PCA



Dummy creation (107 features)
Log transformation
Principal loadings of 0.5

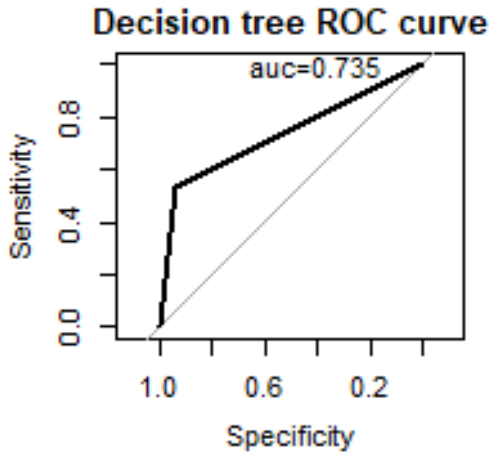
Cross-validated models on dimension-reduced data

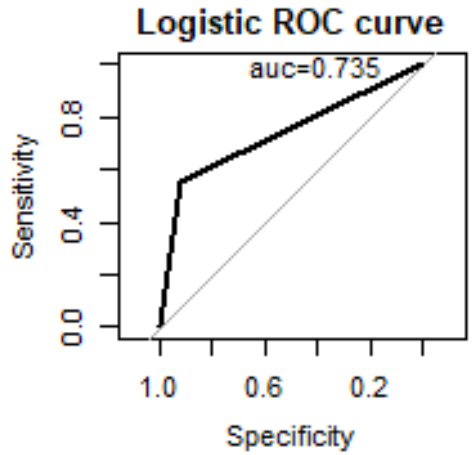
Accuracy on hold-out set (70:30 split)

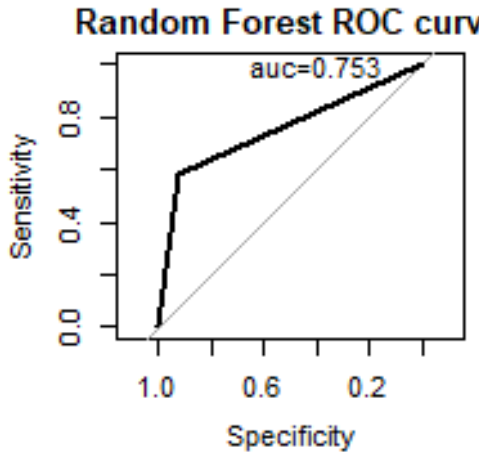


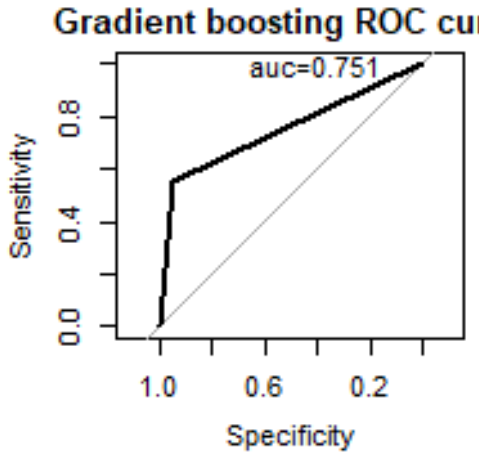
Cross-validated models of dimension-reduced data

Performance on test data









Conclusion

- Initial models ----> Same models with stratified sampling ----> Models after PCA
- GBM outperformed others in most cases
- Different sampling ratios and experimentation with loadings threshold are future ideas for exploration