University of Malta
M.Sc. in Artificial Intelligence

ICS5110:
Applied Machine Learning

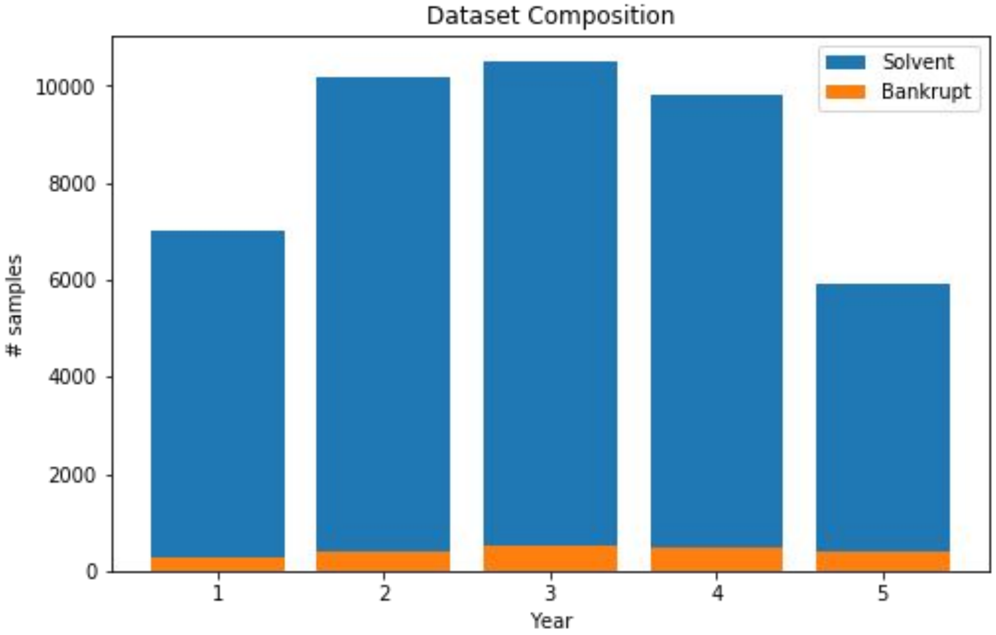
Bankruptcy Prediction using Logistic Regression & Bagged Decision Trees

Albert Bezzina, Daniel Farrugia & Ivan Salomone

22nd January 2019

Overview

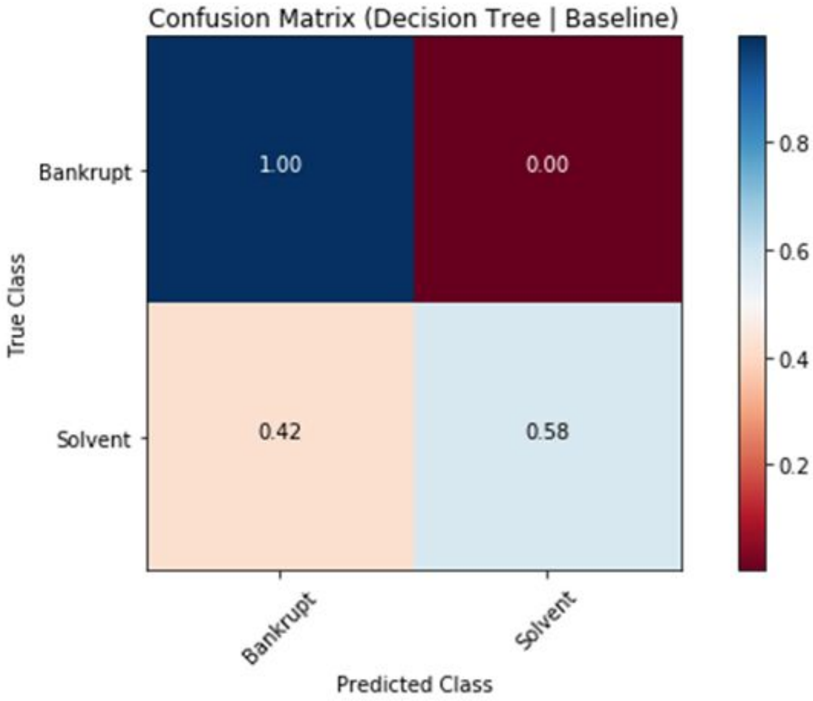
- Objective: to classify companies as bankrupt/solvent
- Dataset consists of 5 years of financial ratios of Polish companies
- 64 features
- Dataset characteristics
  - Class imbalance
  - Missing values
  - Outliers
  - Correlated attributes



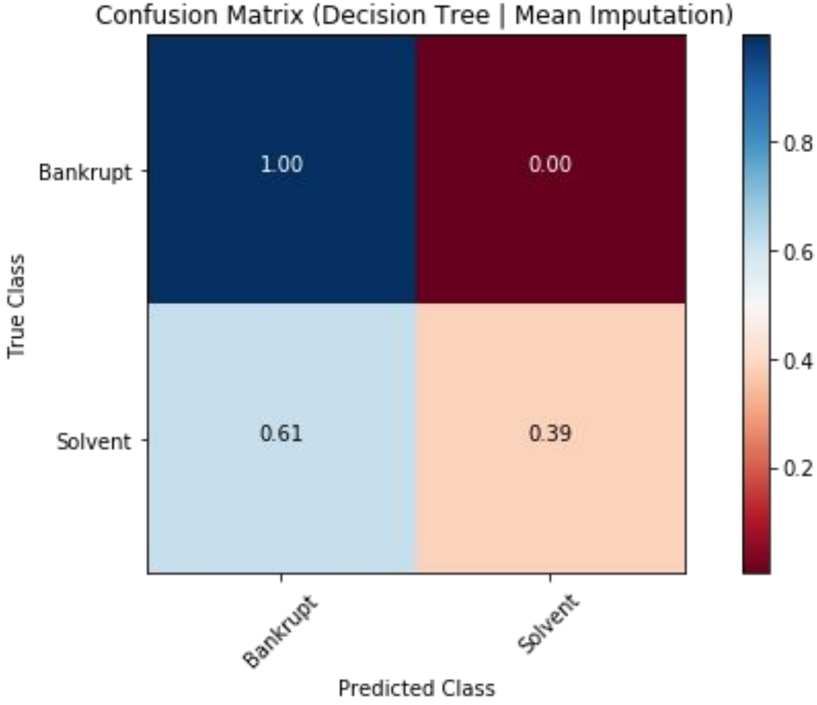
Experiments Setup

- Established Baseline models for comparability
  - Random search Cross Validation used to find optimal hyper parameters
  - Stratified k-Folds
  - Hyper parameters chosen based on best AUC
- Hyper parameters kept constant when comparing against library models
- 10-fold Cross-Validation used throughout
- Evaluation metrics AUC, sensitivity, specificity on out-of-sample examples (20% of dataset)
- Experiments: Class Imbalance, Normalisation, Feature Selection / Dimensionality Reduction, Imputations

Imputations for Missing Data



Class-partitioned Mean

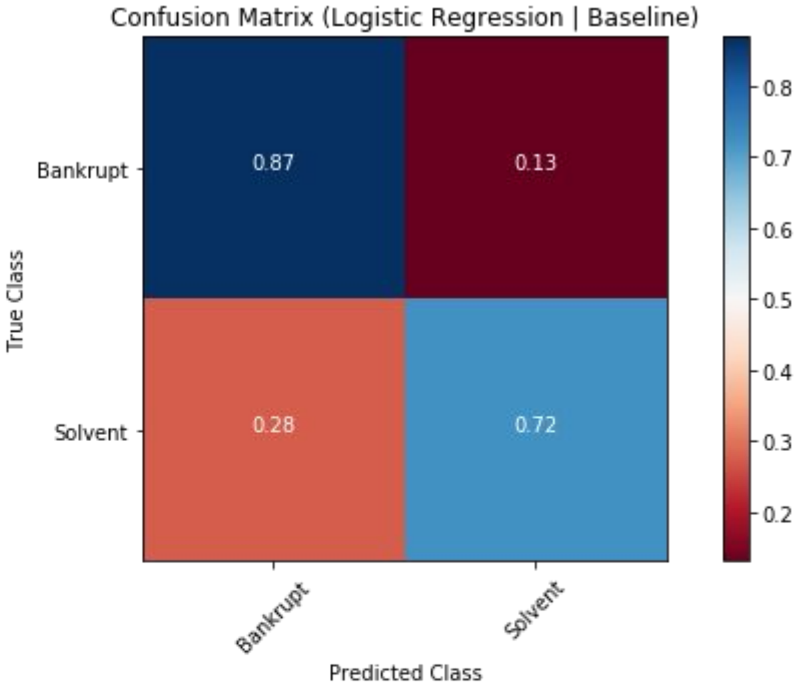


Mean

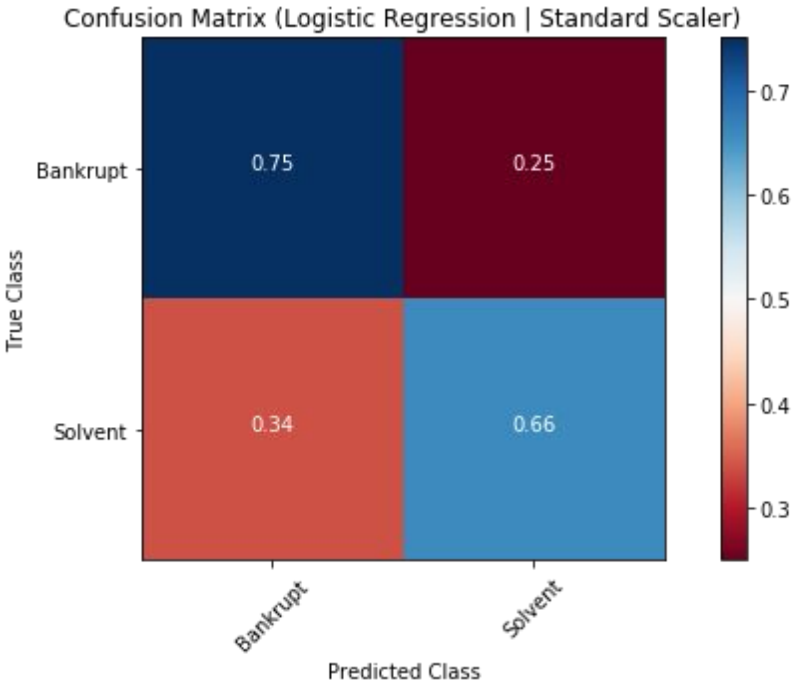
Class Imbalance

	Logistic Regression			**Bagged Decision Trees**		
	Class Weights	*SMOTE*	*Under-sampling*	*Class Weights*	*SMOTE*	*Under-sampling*
AUC	0.875	0.879	0.867	0.909	0.918	0.915
Sensitivity	0.725	0.775	0.758	0.581	0.612	0.801
Specificity	0.869	0.845	0.826	0.997	0.991	0.858

Normalisation



Quantile Normalisation



Z-Score Normalisation

Feature Selection

	Logistic Regression				**Bagged Decision Trees**	
	Baseline	*PCA*	*L2*	*k-Best*	*Baseline*	*PCA*
AUC	0.875	0.877	0.874	0.865	0.909	0.875
Sensitivity	0.725	0.775	0.725	0.706	0.581	0.352
Specificity	0.869	0.839	0.870	0.866	0.997	0.996

Logistic Regression

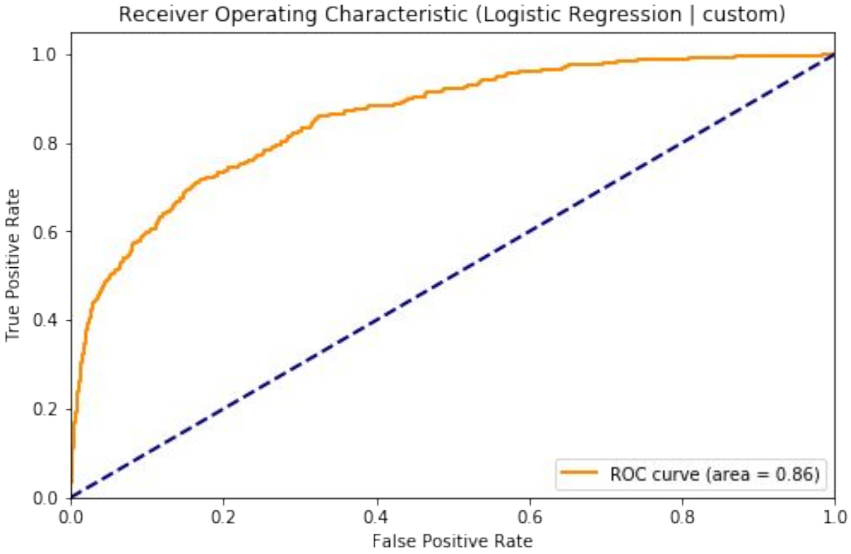
- Vectorised implementation
- Early stopping (halt if no progress after 5 iterations)
- Comparable to Scikit-learn's `SGDClassifier()` with *log* loss function
- In-built class imbalance handling using weights

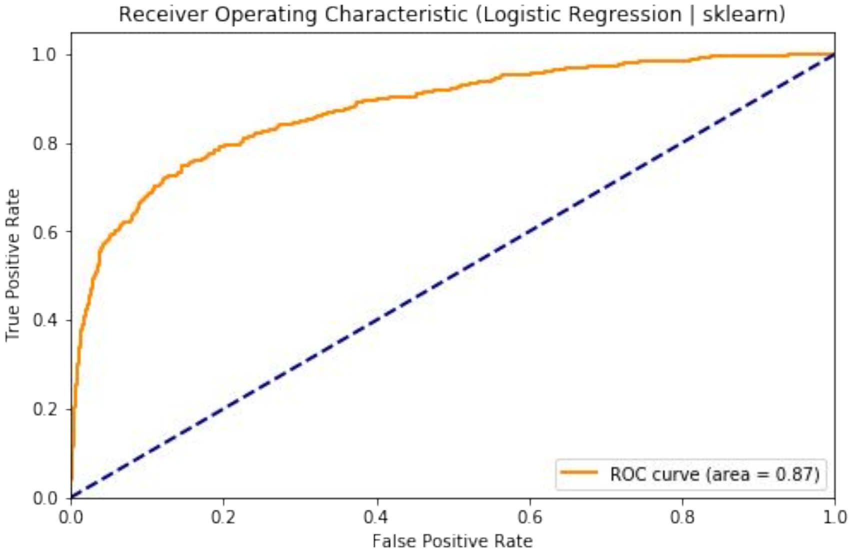
Bagged Decision Trees

- Bagging classifier that train a decision tree estimator
- Uses class weights to handle imbalance
- Decision tree uses information gain / entropy as criterion
- Needs to handle continuous feature set

Custom vs Scikit-learn (1)

- ROC: 0.86 vs 0.87



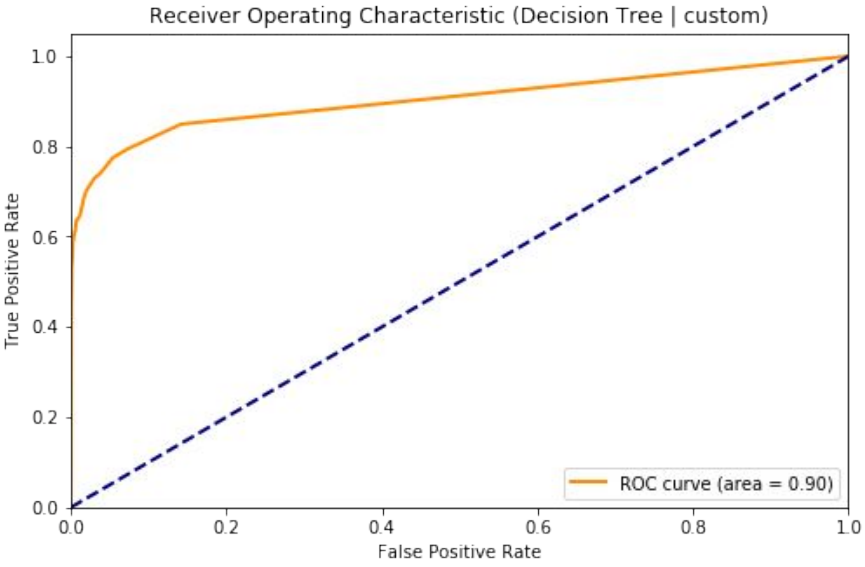


Custom logistic regression model

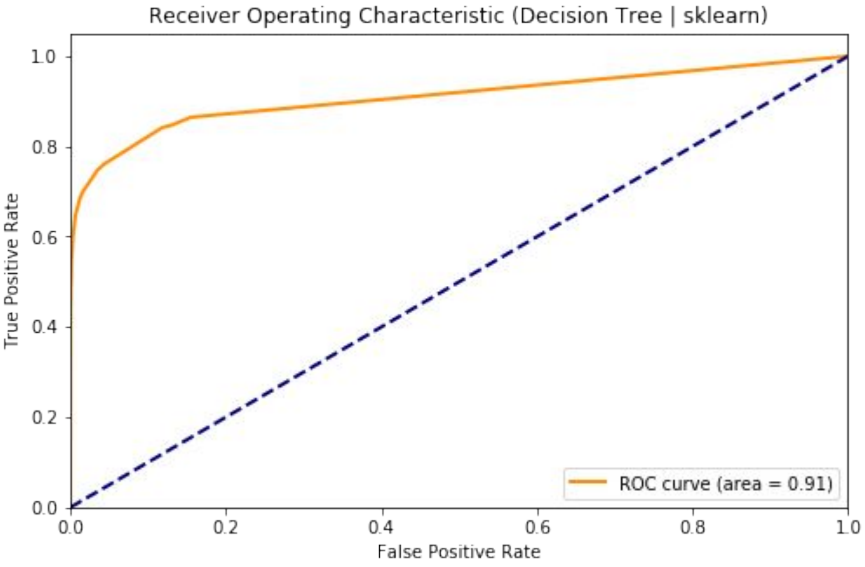
Scikit-learn logistic regression model

Custom vs Scikit-learn (2)

- ROC: 0.90 vs 0.91



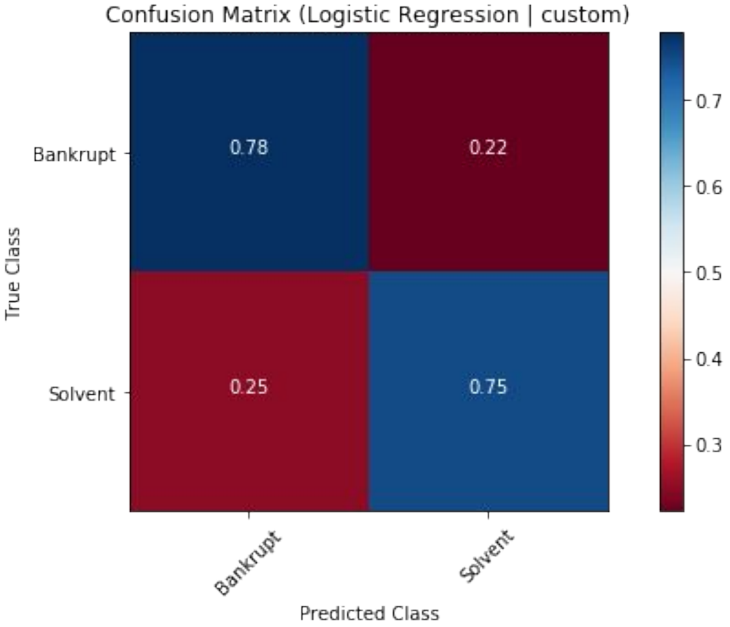
Custom bagged decision trees model



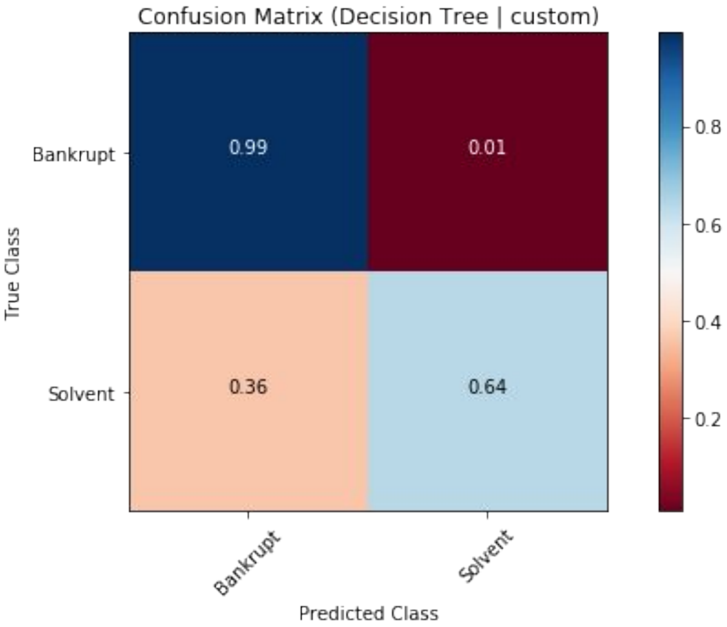
Scikit-learn bagged decision tree model

Conclusion

- Class-partitioned mean imputation helped for both models
- For Logistic regression - quantile normalisation improved results
- For Bagged decision tree - sensitivity score is below average



Logistic regression model



Bagged decision tree model