Business Report: Assignment 3
Corinne Williams

Introduction

Working to predict whether firms in 2014 in the industry of 'Manufacturing computer, electronic and optical products', 6 different predictive classification models were used, each incorporating different amounts of the same predictor variables (some, including interaction terms) to compare each model's effectiveness in prediction. As such, the models used were (1) Logit Regression 1, (2) Logit Regression 2, (3) Logit Regression 3, (4) Logit Regression 4, (5) LASSO Logit Regression, (6) Random Forest Regression. After various data cleaning and feature engineering, in order to provide the best possible prediction of whether a firm defaulted or not, the following features were selected as predictor variables: amortization, current assets, current liabilities, extra expenditures, extra income, fixed assets, income before tax, intangible assets, inventories, liquid assets, material expenses, personnel expenses, profit loss year, share equity, subscribed capital, tangible assets, ceo count, foreign ceo percentage, percentage of female ceos, days a ceo was in-office, whether a firm was female-only, whether a firm was male-only, whether a firm was mixed-gender, whether the ceo is domestic, whether the ceo is foreign, whether the ceo(s) are of mixed origin (domestic and foreign), whether the headquarters are located in a capital city, whether the headquarters are located in a big city, whether the headquarters are located in another type of city, weather the headquarters are located in the Central United States, whether the headquarters are located in the East United States, and whether the headquarters are located in the West United States.

The holdout set was pre-defined as companies in the selected industry that are small or medium enterprises, in which the yearly sales in 2014 were between 1000 EUR and 10 million EUR. The training sample was then all other observations that did not include those in the holdout set. The holdout sample ended up including 1,037 firms.

Model 1: Logit Regression

The first model used all of the given predictor variables along with 7 interaction terms. The Logit model was cross-validated with 5 folds. The CV RMSE value was 0.164, meaning the fit is not awful, but also not the best. The AUC-ROC score was 0.61, meaning that the model has some discriminatory power to be able to 'discriminate' between firms that defaulted and firms that did not.

Model 2: Logit Regression

The second model used the first 9 predictor variables and no interaction terms. The Logit model was cross-validated with 5 folds. The CV RMSE value was 0.258, meaning the fit is alright. The AUC-ROC score was 0.70, meaning that the model has pretty acceptable discriminatory power (better than Model 1).

Model 3: Logit Regression

The third model used the first 18 predictor variables and no interaction terms. The Logit model was cross-validated with 5 folds. The CV RMSE value was 0.258, meaning the fit is alright. The AUC-ROC score was 0.67, meaning that the model has some discriminatory power (better than Model 1, worse than Model 2).

Model 4: Logit Regression

The fourth model used the all predictor variables and no interaction terms. The Logit model was cross-validated with 5 folds. The CV RMSE value was 0.165, meaning the fit was not bad, but also not great. The AUC-ROC score was 0.60, meaning that while the model has some discriminatory power, it has the weakest AUC-ROC values out of all the models so far.

Model 5: LASSO Logit Regression

The fifth model used the LASSO selected features (15 features) and was cross-validated with 5 folds using a penalization value of 0.1. The CV RMSE was 0.165, meaning the fit was not bad, but also not great. The AUC-ROC score was 0.63, meaning that while the model has some discriminatory power, it is still pretty weak.

Model 6: Random Forest Regression

The sixth model was a Random Forest using the features from Model 3 (due to Model 3 features providing the most favorable outcome in the Random Forest model). The model was cross-validated with 5 folds. The CV RMSE was 0.164, meaning the fit was not bad, but also not great. The AUC-ROC score was 0.84 meaning that the model has pretty excellent discriminatory power. Model 6 had the best AUC-ROC and relatively one of the lowest expected loss values.

Model Comparison:

Using predicted probabilities and then binary classifications, expected losses for each model were found. As expected loss is the predicted cost associated with misclassification, the lower the expected loss, the better.

Model 1 Expected Loss: 0.680
Model 2 Expected Loss: 0.345
Model 3 Expected Loss: 0.372
Model 4 Expected Loss: 0.375
Model 5 Expected Loss: 0.573
Model 6 Expected Loss: 0.352

Model 2 had the lowest expected loss, with Model 6 coming behind as a close 2nd.

Overall Model 6 (Random Forest) was found to be the best model (relatively low expected loss and very high AUC-ROC score in comparison to the other models).

	Models	# of Predictors	CV RMSE	CV AUC	Optimal Threshold	CV Expected Loss
0	M1	36	0.164350	0.602	0.026666	0.680350
1	M2	9	0.258435	0.696	0.449890	0.345080
2	M3	18	0.258443	0.679	0.404174	0.372925
3	M4	29	0.165728	0.599	0.066641	0.375868
4	LASSO	15	0.165272	0.626	0.025554	0.573800
5	RF	18	0.164587	0.844	0.020000	0.352320

Prediction with Random Forest on Holdout Sample:
Using model 6 to predict whether a firm defaulted or not in the holdout sample, 276 firms were predicted to default while 761 were predicted to stay alive in 2015.

Summary:
Despite the Random Forest model proving to be the best, its prediction was far off from what was expected. When given the assignment, it was stated that 56 firms should default; however, model 6 predicted 276. This large difference can also be seen in the change in metrics from the Random Forest training model to the Random Forest model used to predict the holdout sample. For example, in Model 6, the expected loss value was 0.352; however, when predicting the holdout sample, that value jumped to 0.758. Interestingly enough, the AUC-ROC score also increased to 0.89 despite the biased prediction. I suspect this is due to the large bias in the holdout sample classes. There is a large difference between the minority and majority classes. Therefore, I feel that perhaps the training sample is not a good representation of the holdout, leading to a pretty skewed prediction.