95-828: Machine Learning for Problem Solving
Case Study Phase 3: Predictive Modeling
Team Members: Ghazal Erfani (gerfani), Dan Lesser (dlesser), Joe Standerfer (josephst)

1.

i. We trained a Naïve Bayes, L1 and L2 Regularized Logistic Regression, Decision Tree, Random Forest, and Multi-Layer Perceptron to predict the probability that each loan defaults. Prior to training the models, we used a function that prepared the data by filtering down to the required date range, downsampled if required, prepared the training and test data sets, and scaled the variables. Next, to train these models, we used a function that loaded the data, fit the model, and evaluated the model. In training the model, it optimized sets of parameters using cross-validation.

ii. For Gaussian Naïve Bayes, we tuned var_smoothing. The smoothing parameter adjusts for variance in the features and ensures calculation stability. For L1 and L2 Regularized Logistic Regression, we tuned C, which is the inverse of the regularization strength. This controls how much features are penalized in the model. For Decision Tree, we tuned max_depth, the maximum depth of the tree, and min_impurity_decrease, the minimum impurity decrease required for a node to split. This helped ensure the tree does not have too many branches or leaves with few observations. For Random Forest, we tuned n_estimators, the number of trees in the forest, criterion, the function to measure the quality of the split, and max_depth and min_impurity_decrease. Lastly, for Multi-Layer Perceptron, we tuned hidden_layer_sizes, the number of neurons in each hidden layer, alpha, the L2 penalty parameter, and the activation function for the hidden layer.

iii. If the model did not predict probabilities, then accuracy was used for evaluation. Otherwise, the optimal threshold that maximized accuracy was first determined. We also reviewed the ROC curve with area under the curve, a sensitivity/specificity curve, and a calibration curve to evaluate performance. All of these models saw a similar accuracy of ~0.80 and F-1 scores of ~0.89. There was some greater variation in AUC ROC, with Naive Bayes having a score of 0.68, l1 logistic regression 0.69, l2 logistic regression 0.70, Decision tree 0.66, random forest 0.70 and multi-layer perceptron 0.71.

2. Random splitting of data is advantageous because it is straightforward to implement and is not biased towards some characteristics of the dataset. However, it may easily lead to overfitting due to an unfortunate split. Running each model through multiple train/test splits is one way to alleviate this concern.

Meanwhile, temporal splitting of the data can be advantageous because it enables us to use the most recent data as the test dataset. This may better reflect real-life performance of a model because we will be using historical data to predict future time periods. However, it may lead towards biased results as we are able to set the ‘cut-off’

point. It may also not take full advantage of the data we have been able to collect, especially since there are many more observations in more recent time periods.

3.

i. Features that are derived by LendingClub and others that reflect those include int_rate and all grade features (grade::B, grade::C, grade::D, grade::E, grade::F, grade::G, grade::nan).

ii. The predictive power of this L1 and L2 Regularized Logistic Regression models is similar to those trained in part 1. We used int_rate as we felt it would be reflective of whichever grade LendingClub assigned. This means LendingClub has done a good job using the available features to create a proxy feature that reflects the underlying data.

iii. The models refitted after removing the features identified in part (i) results in an accuracy of approximately 0.80, which is similar to models fitted with all features. However, the AUC scores have dropped by ~0.04 to ~0.05 each, meaning that the LendingClub features are helpful indicators.

4. To assess the extent to which our L1 Logistic Regression scores agree with the grades assigned by LendingClub, Kendall's tau, a measure of correspondence between two rankings, is calculated. Values closer to 1 indicate strong agreement between the rankings. We found that our L1 Logistic Regression's similarity to LendingClub's grade ranking was 0.375.

5. Our L1 Logistic Regression trained in 2010 performs worse in 2017 than Our L1 Logistic Regression trained on more recent data that includes 2016. This is because no defaults occurred in 2010 whereas many defaults occurred in 2016. This indicates that our model is not entirely stable. Moving forward, we may consider using only more recent data for training our model.

6. The performance of this model is significantly better than before. This is not surprising because we added in a few features that are incredibly indicative to whether or not the loan defaulted. Total_pymnt in particular is a feature that causes our results to be unnaturally high. This reiterates the importance of cleaning the data before running machine learning models.

7. Given that the R^2 values of these models is very low, these models are not performing well. This is surprising given the decent performance of the classifiers.

	Performance for each return calculation			
Model	**M1**	**M2**	**M3 (i=0.002)**	**M3 (i=0.005)**

L1 Regressor	-1.54e-9	-3.67e-5	-6.33e-6	-2.14e-6
L2 Regressor	0.0284	0.0202	0.0326	0.033
Neural Network Regressor	0.0096	0.0184	0.0358	0.0324
Random Forest Regressor	0.0553	0.0255	0.0619	0.0582

8.

i. See table below
ii. See table below

	Return Calculation			
Strategy	**M1**	**M2**	**M3 (i=0.002)**	**M3 (i=0.005)**
Rand	0.0036	0.0452	0.0102	0.056
Def	0.0067	0.0452	0.0107	0.0552
Ret	0.0096	0.0446	0.0111	0.0571
DefRet	0.017	0.0448	0.0114	0.0552
BEST	DefRet	Rand/Def	DefRet	Ret

iii. Based on the above table, Default & Return-based investment strategy performs best in most cases. Overall, the results are quite similar across the various models for any given return calculation.

9. As portfolio size increases, investment return decreases. The plot below shows that as the portfolio size increases there is a near-linear decrease in investment return. This may appear to run counter to traditional logic that diversifying is a good

thing. This is likely because with a larger portfolio size, there are more risky loans that need to be included in the portfolio and thus the overall average return is lower.