

Model Understanding: The objective of this report is to build a model to predict whether or not a loan was approved or denied. This approval or denial will serve as the binary response variable (1 = denial and 2 = approval) that the model aims to predict, and several metrics will be evaluated to gauge the capabilities of different models, including ROC curves, lift curves, and the confusion matrix. This report will also look at the topic of ensemble modeling, which is done to combine the outputs of several very distinct models into one modeling average for better prediction accuracy.

Data Understanding: There were over 166,000 total records in the data set. Certain variables were excluded from the creation of the model due to missing values when continuous and only having one variable level when nominal. For example, the *state* and *year* variable only had values of 2016 and LA (for Louisiana), respectively. *STATE_FIPS* was also excluded for the same reason, as it is simply a two-digit code representing Louisiana. *RateSpread_num* was also excluded from analysis because it contained over 150,000 missing values, too many to impute. Several continuous variables were candidates for imputation because they had a small number of missing values. *HOEPADescription* had only two different level values (non-HOEPA and HOEPA loans), and 99.97% of all loans in the data set were non-HOEPA. Therefore, this variable was also excluded. Several variables like *race* and *loan type* were binned. For example, VA and FHA loans had very similar denied = 1 rates. See Figure 5 for the differences in the response variable based on the type of loan being requested. This led to the creation of three bins for that variable in particular. The entire data set was then split for validation in the following proportions: 60% training, 30% validation and 10% test.

Analysis: After cleaning the data set and generating the validation column, different model types were then created to compare for prediction: logistic regression, boosted tree, k-nearest neighbor (k-NN), and neural network.

Confusion Matrix: All confusion matrix figures for each model are shown in Figure 1. The false positive rate represents the percentage of incorrect Denied = 1 predictions, and the false negative rate represents the number of incorrect Denied = 0 predictions. Sensitivity and specificity indicate the model's ability to correctly predict true positives (Denied = 1) and true negatives (Denied = 0), respectively. In other words, the sensitivity shows the percentage of predicted denied = 1 classes that were truly denied, and specificity shows the percentage of predicted denied = 0 classes that were truly not denied. The overall error is calculated as the sum of all incorrect predictions divided by the total number of cases in the data set.

ROC Curve: The ROC curve is a metric that allows easy comparison between models. The classification models discussed in the **Confusion Matrix** section all predict the probability that an observation belongs the target class of Denied = 1 (with the exception of k-NN). A cutoff value (set to 0.5 in this case) is then chosen to decide if the probability (which is between 0 and 1) is high enough to generate a Denied = 1 prediction. For example, if one prediction model generates a 0.51 value for a data point, that case would be predicted as Denied = 1 because it is higher than the 0.5 cutoff value. A comparison of all ROC curves is shown in Figure 2. The straight line shows a default prediction method that assigns classes at

random. The area between the curve of each model and the straight line is called the AUC (area under the curve), and the higher this value is, the better the model is at predicting Denied = 1 when compared to random classifications. Having an ROC curve closer to the top left of the graph is more desirable. Before the ensemble model was created, the model with the highest AUC is the boosted neural network. The model with the lowest AUC is the logistic regression, meaning it is the worst out of all models when compared to random predictions at a cutoff value of 0.5. Figure 6 shows the cumulative gains curve, which indicates the percentage of the number of Denied = 1 cases "gained" by targeting a percentage of the total number of cases. For each of the models, including the ensemble, around 85% of all loans in the data are covered or gained in the first 50% of data.

Lift Curve: The lift curve is another method of model comparison. Like the ROC curve, it also compares the prediction capabilities of multiple models to a random classification. That random classification is flipping a coin in this case. A "good" lift curve is one that begins high above a value of 1 on the left side of the graph and then falls steeply to 1 towards the right side. Figure 3 shows the lift curves for the prediction models. Each model exhibits the qualities of a good lift curve. One note about the logistic regression lift curve shown in red: it starts high above 1 but not as high as the other three models. The other models appear to begin at values above 3, whereas the regression curve begins around a value of 2.

Conclusion: Figure 4 shows the measures of fit for all validation levels for each model developed for the report, besides the k-NN as JMP does not produce a statistic for it. This is because the k-NN model does not generate a probability nor does it really on a probability formula to create its predictions. Instead, this type of model is made by comparing every data point to every other data point, hence the consideration surrounding long computation times. The logistic regression model had the highest sensitivity of all, but also the highest false positivity rate at 42%. With this type of model, there is a concern of overfitting. Overfitting is also an issue when it comes to decision tree models, but the boosted tree platform has penalty settings to protect against that effect. The boosted tree model had the lowest sensitivity at 25%. Therefore, it is the worst at correctly predicting the true classes of Denied = 1. Neural networks have an advantage over regression models because they are able to capture complex relationships between predictors. With all of that being said, the logic behind ensemble modeling is simple: there is usually a benefit of combining multiple different models for prediction instead of just one. Because each model has its own aspects that pose limitations on its capabilities, averaging models together helps to increase the accuracy of prediction. Based on the metrics of the ensemble model produced for this report, it would be the best candidate for prediction purposes. Although it does not have the highest AUC or sensitivity, it has the second highest in each of those, and the lowest false positive rate. It also has the same false negative rate and overall error rate as the k-NN and boosted neural network. It offers a good balance of all evaluation metrics, instead of having only metric it is superior in.

References: Module 7 Resources

Appendix

Figure 1: Confusion Matrix Figures

	Sensitivity	Specificity	F. Neg.	F. Pos.	Error	AUC
K-Nearest Neighbor	27%	95%	20%	38%	24%	N/A
Boosted Neural Network	29%	95%	19%	38%	21%	0.8066
Logistic Reg.	32%	92%	19%	42%	23%	0.7920
Boosted Tree	25%	95%	20%	38%	22%	0.7983
Average Model	30%	94%	19%	36%	21%	0.8024

Figure 2: ROC Curves

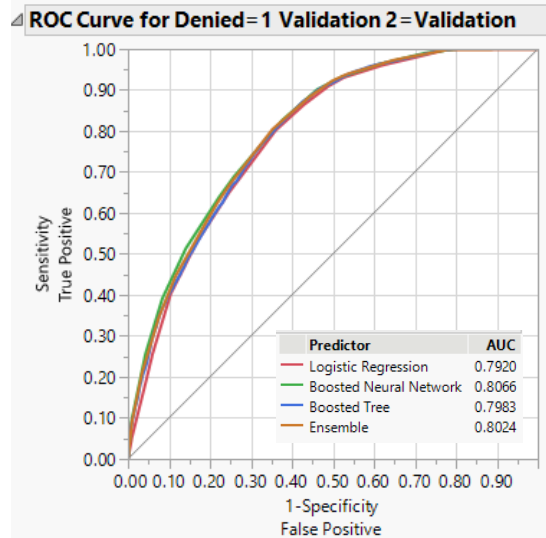


Figure 3: Lift Curves

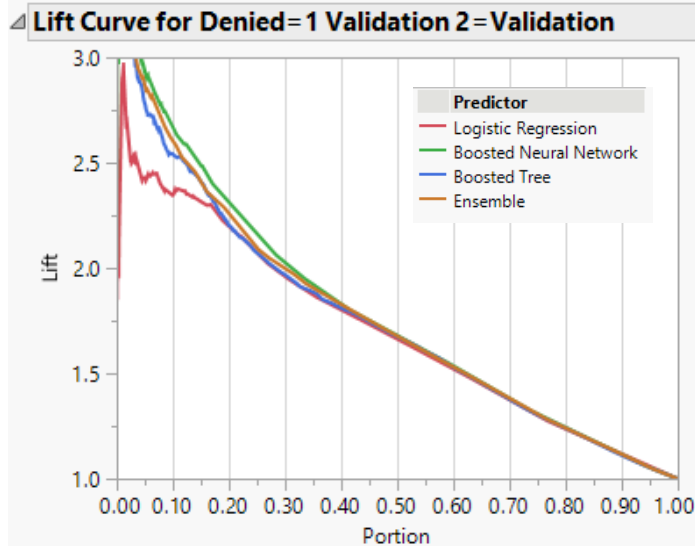


Figure 4: Measures of Fit/Model Comparison

Measures of Fit for Denied											
Validation 2	Creator		2.4.6.8	Entropy	Generalized	Mean -Log p	RMSE	Mean	Misclassification	N	AUC
				RSquare	RSquare			Abs Dev	Rate		
Training	Fit Nominal Logistic			0.2008	0.2995	0.4502	0.3885	0.3015	0.2282	99103	0.7935
Training	Neural			0.2295	0.3371	0.434	0.3800	0.2898	0.2158	99103	0.8113
Training	Boosted Tree			0.2091	0.3105	0.4455	0.3854	0.3052	0.2227	99103	0.8007
Training	Model Averaged			0.2189	0.3233	0.44	0.3830	0.2989	0.2192	99103	0.8055
Validation	Fit Nominal Logistic			0.1984	0.2952	0.4462	0.3866	0.2997	0.2251	49584	0.7920
Validation	Neural			0.2237	0.3283	0.4321	0.3790	0.2892	0.2133	49584	0.8066
Validation	Boosted Tree			0.2061	0.3053	0.4419	0.3835	0.3036	0.2203	49584	0.7983
Validation	Model Averaged			0.2152	0.3173	0.4368	0.3814	0.2975	0.2174	49584	0.8024
Test	Fit Nominal Logistic			0.1901	0.2837	0.4492	0.3878	0.3007	0.2249	16680	0.7851
Test	Neural			0.2158	0.3176	0.435	0.3800	0.2898	0.2142	16680	0.8015
Test	Boosted Tree			0.2018	0.2993	0.4428	0.3836	0.3039	0.2188	16680	0.7938
Test	Model Averaged			0.2085	0.3082	0.439	0.3822	0.2981	0.2171	16680	0.7967

Figure 5: Loan Types vs. Denied (before binning)

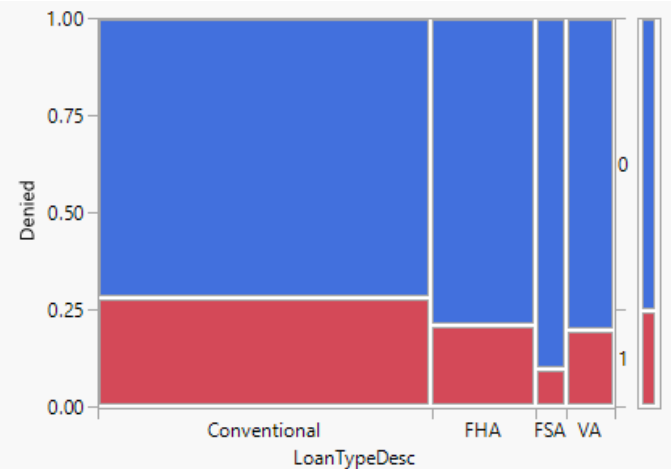


Figure 6: Cumulative Gains Curve

