Assignment 3: Adding Classifiers

CAPP 30254: Machine Learning for Public Policy

Christine Chung
Spring 2016

1. Introduction

Historical data was provided on 250,000 borrowers in order to generate a delinquency score. Decision makers are interested in predicting the probability that a person would experience 90 days past due delinquency or worse on a credit payment.

The goal was to evaluate and compare multiple classifiers in order to determine the best performing classifier for this task.

2. Content

I compared four models over six metrics. For classifiers, I focused on Decision Trees, K-Nearest Neighbors, Naive Bayes, and Logistic Regression. Since the model will be serving decision makers, I focused on the most easily interpretable models.

The results and comparisons the four classifiers are show in Figures 1 and 3. Graphed comparison of Decision Trees, K-NN, and Naive Bayes are shows in Figure 3.

Model	Training Time (s)	Accuracy	F1 Score
Decision Trees	49.807	0.9356	0.1706
K-Nearest Neighbors	1482.363	0.9341	0.0054
Naive Bayes	0.492	0.9332	0.0470
Logistic Regression	114.608	0.9325	.0709

Figure 1. Comparison of Training Times (seconds), Accuracy, and F1 Scores

Model	Precision	Recall	AUC
Decision Trees	0.5644	0.1006	0.5475
K-Nearest Neighbors	0.5000	0.0027	0.5013
Naive Bayes	0.3978	0.0249	0.5111
Logistic Regression	0.5225	0.0381	0.5178

Figure 2. Comparison of Precision, Recall, and Area Under the Curve

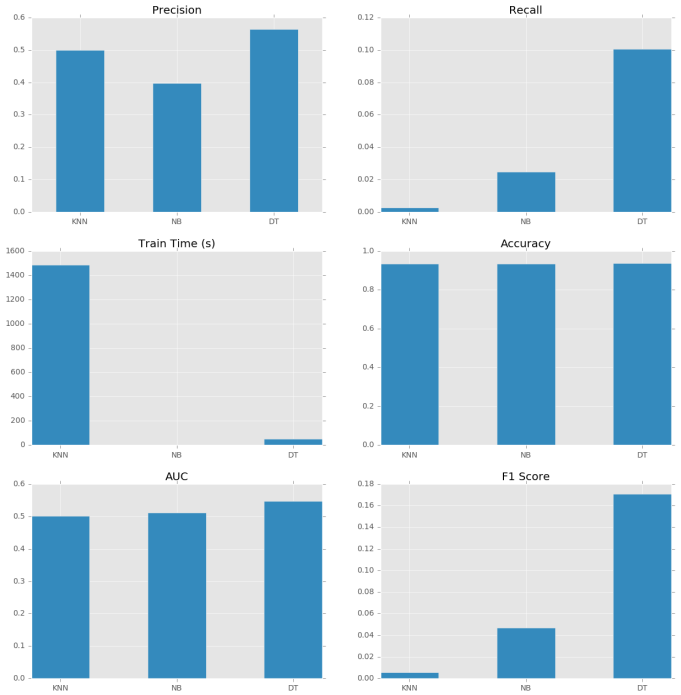

Figure 3. Graphed comparisons of Decision Trees, K-Nearest Neighbors, and Naive Bayes across six metrics.

3. Analysis

Decision trees performed best overall over all metrics. Logistic Regression was a distant second, and took more than twice as along to complete.

We care about different metrics depending on the intervention. If a credit card company wanted to reduce risk, for example, it would care most about denying all people who were truly at risk of delinquency, even if it occasionally denied people who are not actually likely to be delinquent. In this case, you would care most about Recall, because you are most interested in identifying all delinquent borrowers. On this metric, Decision Trees out-performed the next best metric by three times.

If, however, you wanted to minimize the number of false positives, you would instead look at precision. On this metric, Decision Trees also performs the best. However, Logistic Regression and K-Nearest Neighbors also performed decently.

In terms of time, Naive Bayes was the clear winner in this category. Training took less than half a second. By comparison, Decision Trees performed reasonable well, taking a little under a minute to complete training. K-Nearest Neighbors was by far the slowest, taking a whopping 25 minutes to complete.

4. Conclusion

The final recommendation is to use the Decision Trees classifier to predict potentially delinquent borrowers. It performs the best across all metrics, and so is suitable for any application. Additionally, while not the fastest, it completes in a reasonable amount of time. The significant gains in precision and recall make it more attractive than using Naive Bayes.