Home Work 6 (Case Study 3) – Final Project

Drew Morris

Introduction

Predicting whether or not a customer is likely to default on their credit card is important to the issuing institution as it can help manage risk. This project attempts to create an algorithm that classifies users as being likely to default or not. The classification of customers is done by several different algorithms but uses the same base data set.

Data Collection

The data was collected by a major Taiwanese bank over a six month period in 2005 from april to september. The response variable was the default status of a customer in the month of october in 2005. The data set has 30,000 unique data points and the data was split in a 70/30 distribution between training and testing data. The data set contains descriptive data about the customer like their age, sex, marriage status, and education level received along with some basic economic information like credit limit, payment status, history, and bill information. This data was used to then create additional synthetic features and had its categorical variables like marriage status and education level converted into a one-hot encoding scheme. The application allows the user to select which features are fed into the models but more complex relationships (like conditional information) is not included or used.

Motivation for the Project

Predicting whether or not a customer is likely to default in the next month was interesting to me for two reasons. To start, the way financial institutions manage risk and assess a customer's creditworthiness is a black box to me and working with this data might help me understand that area better. In addition, it would be interesting to discover the warning signs of an impending default and to later explore if there are intervention methods that could be employ to prevent a default from happening or to allow one to recover quickly from a default. Finally, this project was interesting because the problem formulation was relatively simple and is easy to convey while still being a challenging problem to solve.

Data Analysis

The data was first explored to understand what was contained in the data in an effort to help guide the generation of hypotheses towards the creation of new features. To start, the data was checked to ensure the all features were correct and did not have any missing values. While none of the features had missing values, some of the features did have unexpected values. For example, the payment status feature for each of the prior six months had records that contained a -2 which was not defined in the description of the data. Erroneous records were simply excluded from the dataset as it was the safest option in comparison to doing nothing or attempting to relabel the information. If greater access to the data and the data collectors was available an attempt to relabel the data would have been made but if no reason explanation was able to be found for why the data was mislabelled then the data would still be included.

Once the data was cleaned, a correlation matrix was generated to explore the relationships between the features and with the response. Of the raw features only two features had moderate correlationships with the response. The most recent month's (september) payment status and the limit balance had moderate correlations. This makes sense as the payment status represents the payment history of the customer. The

higher the number in the payment status for a month the greater the number of months a customer went without paying their bill. It also makes sense that the credit limit of a customer would be correlated –albeit negatively– with the target as the financial institution has some measure of creditworthiness when setting the limit balance and when a customer is less trustworthy then the institution is likely to give them a lower limit than a customer that is highly trustworthy. Strong correlations amongst the features were largely contained to features that were logically similar. For example, bill amounts for each month were strongly correlated which makes sense as customers are likely to maintain the same level of balance if they are using the card normally and are paying their bill off as expected. One correlation that did exist that was not immediately obvious was the strong negative correlation between age and the raw marital status feature. Further investigation explains the correlation as the older a person is, the more likely they are to be married. Once the correlations amongst the variables were explored it was time to generate new features.

The new features came from the hypotheses that were generated during the initial data exploration and problem formulation. The data revealed that certain variables were, in fact, categorical variables and would be better represented using a one-hot style of encoding as opposed to a single variable for each. The data also made it clear that payment status was a strong predictor of default and that perhaps more summary information around payment status would be helpful. The sum of all the months payment status was created as it could help stratify the customers based on their payment histories better than looking at each month individually. Similarly, the count of each month that a customer made no payments was created. The discovery that the credit limit of a customer was correlated with the response drove the generation of creating a ratio for each month's bill to the credit limit. The rationale is that the actual amount for a bill does not provide enough context to be useful as two customers with the same bill could have wildly different credit limits and would thus have different situations for paying it back. When created, these variables did have moderate correlations with the response. Other features were considered, like measuring the change in a customer's bill to limit ration over the past six months, but they were not included as they did not have a strong correlation with the response and ultimately did not improve the prediction accuracy.

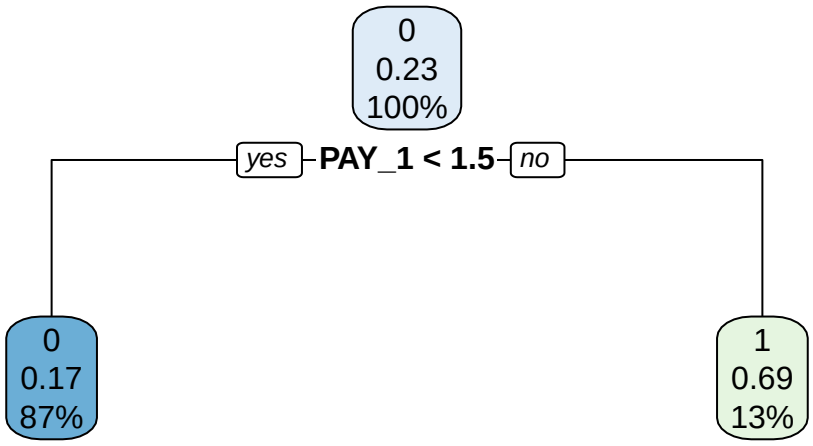
The three models were selected because of their varied approaches to prediction. The K Nearest Neighbors method was chosen due to its relative simplicity and the fact that it is an instance-based predictor. An instance-based predictor does not build an underlying model from the training data but rather directly uses the training data for each prediction it makes. The logistic regression was chosen because it is a member of the regression family and is quite good at calculating the probability of an input belonging to a class. It is also easy to interpret the resulting model and how each feature contributes to the output. Finally, it also allows for the direct modeling of more complex relationships. The classification tree was used as it is a non-linear model that can be built upon to create some very strong ensemble methods. Unlike the logistic regression which has high bias and low variance, the classifaction tree has low bias and high variance and it is great to explore how the two methods change as inputs are switched and the predictions are rerun.

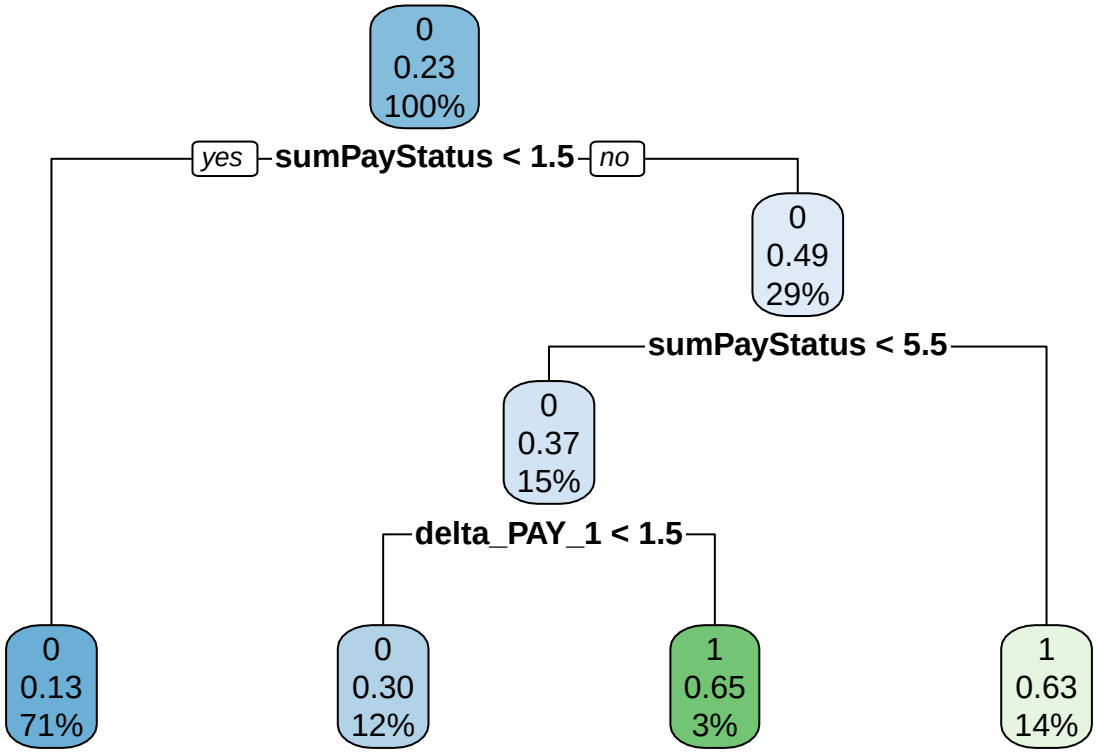
The final model chosen was an ensemble of the three base models to allow for the exploration of how basic ensembling can be used to increase the accuracy of an algorithm and attempt to leverage the strengths of each model. The rudimentary weighted average ensemble method is not ideal because it is attempting to combine three already strong learners it does a good job of conveying the point of what ensembling does and thus it was chosen. More popular ensemble methods like boosted trees perform better but are more complex than is reasonable to convey in the application when the primary purpose is to demonstrate and explore various methods.

Findings of the algorithms

One of the most interesting findings of the whole process was that when allowed to choose freely amongst all data points while also considering complexity the classication tree only had one level which made the prediction solely off whether or not the customer made any payment in the month of September. This also was frequently the most accurate model of the three base algorithms. This result was unexpected but is not unreasonable due to the temporal aspect of the data's structure. It would make sense that if a customer is going to default in october they are unlikely to be able to make a payment in september. When this feature is removed from the data set, the classifcation tree does then employ some of the engineered features but leverages August's payment status with less accuracy but not by a large amount. The structure of this resulting tree also makese

sense that leveraging the customers most recently available payment information would have predictive power.





The intended use case of the algorithm should be considered when selecting the right algorithm to use and which features should be used and the results contained in the confusion matrix can help guide that decision. For example, if the intended use case is to use the output to decide whether or not to employ some intervention strategy to prevent a default then using the classification tree that only uses the previous month's payment information could be ideal as it is responsive but would likely trigger many false positives. However, because the use case is an attempt to help and positively affect the outcome the high number of false positives is not harmful. However, if the intended use case is to determine whether or not to cancel or limit a customer's line of credit then using the classification tree that does not simply predict based on last month's payment status would be more ideal as minimizing the number of false positives is better.

While all of the algorithms performed well, the classification tree routinely performed the best, but the logistic regression was not significantly worse. The logistic regression might be preferrable to use if the interpretability of the algorithm and assessing how each feature contributes to the output is important. K nearest neighbors routinely performed the worst of the three options and this is likely due to the large number of unnecessary inputs and its ability to confuse the output. Again, the logitics regression is also a worthy option as it allows the user to tune the model to its responsiveness by adjusting the threshold for classifcation up or down to shift the false positive and negative rates to better suit the intended use case. In addition, the ensemble method was not able to generate an output that had greater accuracy than the classification tree on its own when the classification tree was allowed to simply use the previous month's payment status.

Overall the logistic regression and classification tree were able to increase the accuracy of the model 5% above the no information rate of the model. While this number is not large it is statistically significant, when one considers that the no information rate was already at 77%, it is a strong gain in accuracy. The approach in this project was largely focused on generating some predictive new features and employing classification methods in a shiny app. Exploring the use of outlier detection methods could have strong predictive power due to the low rate of default in the data. It is likely that manipulating the data to search for abnormalities within a given dataset could improve the accuracy particularly for methods like k nearest neighbors where

the current dataset is not likely grouping the data on predictive values.