**Final Assignment**

**Scenario:** Universal Bank has begun a program to encourage its existing customers to borrow via a consumer loan program. The bank has promoted the loan to 5000 customers, of whom 480 accepted the offer. The data are available in fle UniversalBank.csv. The bank now wants to develop a model to predict which customers have the greatest probability of accepting the loan, to reduce promotion costs and send the offer only to a subset of its customers.

**Data:** The Personal Loan Offer dataset . We will develop K-NN with k=3, Naive Bayes and classification tree, then combine them in an ensemble. Finally we do the bagging and boosting and compare the results.

**Data preparation:** Load the data and remove unnecessary columns (ID, ZIP Code). Split the data into training (60%) and validation (40%) sets (use random\_state=1).

**PART A :** We develop K-NN with k=3, Naive Bayes (after binning the continuous predictors) and classification tree, then combine them in an ensemble.

1. The models are fit to the data for (1) k-nearest neighbors with k = 3, (2) Naive Bayes and (3) classification trees. Personal Loan is used as the outcome variable. The validation confusion matrix for each of the three models is reported.

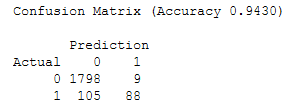
**Refer to notebook file**

1. A data frame is created with the actual outcome, predicted outcome, and probability of being a "1" for each of the three models. The first 10 rows of this data frame is reported

**Refer to notebook file**

1. Two columns are added to this data frame for (1) a majority vote of predicted outcomes, and (2) the average of the predicted probabilities. By using the classifications generated by these two methods a confusion matrix is derived for each method and the overall accuracy is reported.

The combined confusion matrix and overall accuracy is as follows :



**Refer to notebook file**

1. The error rates are compared for the three individual methods and the two ensemble methods.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Methods** | **On it’s own** | | **Ensemble** | | | |
| **Bagging** | | **Boosting** | |
| **Accuracy** | **Error Rate** | **Accuracy** | **Error Rate** | **Accuracy** | **Error Rate** |
| K-nearest neighbors with k = 3 | 0.9545 | 1-0.9545=0.0455 | 0.9585 | 0.0415 | Since KNeighborsClassifier doesn't support sample\_weight, boosting cant be performed on it | |
| Naive Bayes | 0.8795 | 1-0.8795=0.1205 | 0.8800 | 0.12 | 0.9125 | 0.0875 |
| Classification trees | 0.8975 | 1-0.8975 =0.1025 | 0.9015 | 0.0985 | 0.9030 | 0.097 |

**PART B :** Using Bagging and Boosted Trees to compare their performance with all the methodologies used in PART A.

**Refer to notebook file and table above**

Bagging and boosting definitely improved the performance of all 3 methods. Infact Boosting values were higher than Bagging

Predicting the loan acceptance behavior for three new customers, with the following profiles. Each of the developed methods were used to predict these three new customers.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Customer | Age | Experience | Income | Family | CCAvg | Education | Mortgage | Securities Account | CD Account | Online | CreditCard |
| #1 | 40 | 10 | 84 | 2 | 2 | 2 | 0 | 0 | 0 | 1 | 1 |
| #2 | 25 | 6 | 50 | 1 | 1.8 | 1 | 1 | 0 | 0 | 1 | 1 |
| #3 | 59 | 30 | 120 | 3 | 1.9 | 3 | 0 | 0 | 1 | 1 | 0 |

**Refer to notebook file**