Customer Default Records Prediction

USING SUPPORT VECTOR MACHINES TO PREDICT CUSTOMER DEFAULT RECORDS FOR A CREDIT CARD COMPANY

NAITIK SHUKLA – (A0163426H)

NAVAL KUMAR NANJAPPA – (A0163256E)

NAVNEET GOSWAMI – (A0163221W)

SHOBHIT JAIPURKAR – (A0163331R)

VIGNESH SRINIVASAN – (A0163246H)

TEAM-22

M.TECH, KNOWLEDGE ENGINEERING | INSTITUTE OF SYSTEM SCIENCES, NUS

Introduction

In this project, we aim to fit a Support Vector Machine (SVM) model to predict ***default*** of customers as “No” or “Yes”, using the other variables as predictors. The entire project was carried out in R studio, and R was used to implement the SVM with the default data set. We made use of some additional packages in R Studio to further improve the quality of our prediction. These packages are ***E1071***, ***ISLR*** and ***SMOTEFAMILY***.

Data Description

We made use of the Default data set (data size = 10,000; No = 9667; Yes = 333), which are the customer records for a credit card company. This data set was found to be imbalanced with respect to the target variable.

```
> table(defaulters$default)

  No  Yes
9667  333
```

It was observed that 97% of the data contributed to the class “No” while only 3% towards “Yes”. To account for this irregularity, we implemented a method to synthetically input data using SMOTE. This method makes use of the k-Nearest Neighbour framework to generate minority class observations. With synthetic data generation, we create an artificial data based on the feature space from the minority sample. It generates random samples of minority class observations to make the data more proportionately equivalent and well-rounded.

The subsequent data was classified into training and test sets in the ratio of 80:20.

Procedure:

- Respective datasets were loaded with Data with Synthetic Imputation, and without.
- Training and test data is in the ratio 80:20
- The Kernels used by the SVM to map the input features into a higher dimension were- Radial and Sigmoid.
- Using above kernel, we looked for tuning the SVM for the best Gamma and Cost variable to obtain the Support Vectors. The accuracy of the obtained Support Vectors will define the overall prediction accuracy of the model.

Results:

We experimented with different combinations of predictor variables in the data and computed the classification results.

For imbalanced classification data

Predictor Variables	Best Gamma (Radial, Sigmoid)	Best Cost(Radial, Sigmoid)	Accuracy (Radial, Sigmoid)
Student, Income, Balance	0.5, 4	10, 0.1	97.8%, 95.5%
Student, Balance	2, 2	1, 0.1	97.7%, 97.3%
Student, Income	0.5, 0.5	0.1, 0.1	97.1%, 97%
Balance, Income	0.5, 0.5	10, 0.1	97.8%, 95.4%

For balanced classification data (After Synthetic Data Generation)

Predictor Variables	Gamma (Radial, Sigmoid)	Cost (Radial, Sigmoid)	Accuracy(Radial, Sigmoid)
Student, Income, Balance	4, 0.5	10, 0.1	90.5%, 80%
Student, Balance	0.5, 0.5	0.1, 0.1	90.4%, 83.6%
Student, Income	4, 0.5	10, 0.1	57.7%, 52.08%
Balance, Income	4, 0.1	10, 0.1	90.21%, 82.3%

When the test data is used, the results generated as follows (For Oversampled Data)

With all the features

Kernel – Radial (**Best** Gamma: 4; Cost: 10)

```
       truth
predict    0    1
      0 1727  127
      1  239 1772
```

Accuracy: 90.5%

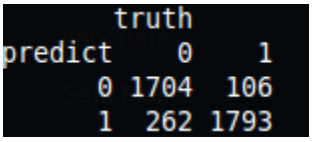
Kernel – Sigmoid (**Best** Gamma: 0.5; Cost: 0.1)

```
       truth
predict    0    1
      0 1566  373
      1  400 1526
```

Accuracy - 80%

With Student and Balance feature

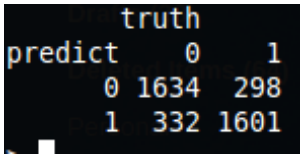
Kernel – Radial (**Best** Gamma: 0.5; Cost: 0.1)



predict	truth 0	1
0	1704	106
1	262	1793

Accuracy - 90.4%

Kernel – Sigmoid (**Best** Gamma: 0.5; Cost: 0.1)

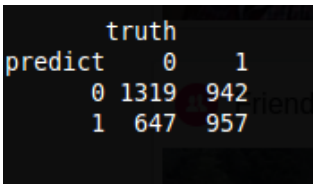


predict	truth 0	1
0	1634	298
1	332	1601

Accuracy - 83.6%

With Student and Income as feature

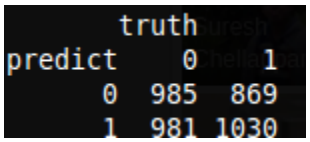
Kernel –Radial (**Best** Gamma: 4; Cost: 10)



predict	truth 0	1
0	1319	942
1	647	957

Accuracy - 57.7%

Kernel – Sigmoid (**Best** Gamma: 0.5; Cost: 0.1)

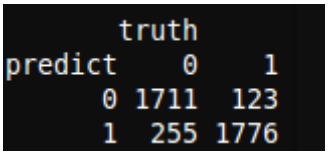


predict	truth 0	1
0	985	869
1	981	1030

Accuracy - 52.13%

With Balance and Income as feature

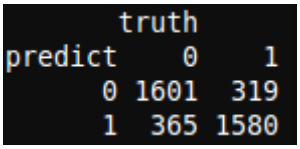
Kernel – Radial (**Best** Gamma: 4; Cost: 10)



predict	truth 0	1
0	1711	123
1	255	1776

Accuracy - 90.21%

Kernel – Sigmoid (**Best** Gamma: 0.1; Cost: 0.1)



predict	truth 0	1
0	1601	319
1	365	1580

Accuracy - 82.3%

Conclusion:

- SVM with radial kernel gives the better performance than rest, with Minimum Least Square Error.
- Default Data set was biased towards one class (Yes).
- Gamma and Cost is lower for operations using the Radial Kernel function, as compared to that of the Sigmoid Kernel function.
- Support vectors generated are large for the balanced dataset generated.