CLASSIFYING CREDIT CARD CUSTOMERS BASED ON SPEND USING MULTI-CLASS LINEAR SVM CLASSIFIER

Gopesh Dwivedi, Saurabh Semwal, Tanmoy Chakraborty

Institute of System Science, National University of Singapore

ABSTRACT

The report reviews an approach to classify credit card customers according to their spending behavior. Segmentation is an important aspect in understanding the customer and running effective and profitable marketing campaigns. This report takes an alternative approach to the task using Support Vector Machines, unlike traditional logistic regression or discriminant analysis. Customers' demographic data, employment details and lifestyle play vital role in how they spend. There are hidden factors as well like likeness towards shopping, buy now pay later attitude which cannot be measured directly. Support Vector machine can be used for both regression and classification problems. In this research, we used multi-class linear SVM as our proposed model for classifying the category variable. Though, we have also evaluated other kernels including radial and polynomial for a range of hyperparameters. We have also taken into account the biasness in the data and treated it before proceeding to model development.

Keywords: SVM, credit card customer segmentation, spend analytics, classification

1. INTRODUCTION

Identifying and ability to classify customers based on the wallet size has been always an area of interest for banking institutions and Credit Card companies. It is an important aspect in customer relationship management and helps in increasing the revenue from existing customers. Various attempts have been made in this regard. Credit card issuers have traditionally targeted consumers by using information about their behaviors and demographics. By analyzing and segmenting customers, based on their buying and payment behaviors, organizations can develop a deeper understanding of their habits, need, and spend behavior. Understanding the debit card customers through their past spending behavior and level of engagement, is the key to unlock the strategy of effective marketing.

Support vector machines (SVMs) have been applied successfully in many classification problems such as text categorization, image recognition and gene expression analysis. Experiments using SVM for classification of credit card customers are relatively new, however. Several papers have recently been published assessing the performance of SVM for credit card customer segmentation. We would be experimenting various SVM kernels (linear, radial, polynomial) for our classification problem at hand. These implementations would be graded based on the overall accuracy as well as the sensitivities of individual classes since overall accuracy might be misleading in such cases. We have also used 10-fold validation to avoid over-fitting of our models and therefore, the obtained accuracy can be a good estimator of the model performance and fit.

2. BASELINE APPROACH

The dataset in consideration is from a reputed credit card issuing company. It has 5K records of customers and their spending behavior attributes. It has 117 variables which include unique customer ID, demographic indicators like age, gender etch and various attributes which describes the lifestyle and employment of the customer. The data has lots of missing values which may be because of two main reasons 1) The data was not recorded for those instances or 2) The respondent did not supply the information to the company. In both cases, we assume the data to be missing at random. This was treated with the use of PMM i.e. Predictive mean matching which is an attractive way to do multiple imputation for missing data, especially for imputing quantitative variables that are not normally distributed. This was done with the help of (mice) package in R. Compared with standard methods based on linear regression and the normal distribution, PMM produces imputed values that are much more like real values. If the original variable is skewed, the imputed values will also be skewed.

Also, for the validation and testing, we have split our data in 80:20 ratio. Thus, 80 % (4K) records are used for training the model and 20% (1K) are used for testing the performance of the model on unseen or test data.

After exploring the dataset, classes were identified as Low Spending customer, Medium Spending Customer and High Spending Customer. The data was used as input to a support

vector machine model with default parameters and settings implemented in R (using e1071 machine learning package). After running frequency analysis on the response variable i.e. the customer category, we noticed the data to be highly biased as there were very few instances of High spending customers (only 2%) which can be expected in such scenarios. The largest cases are of medium spending customers (68.4%) followed by Low spending customers (~30%). Since, our problem does not give preference to any class, this biasness should be removed before finalizing the solution. A dataset is imbalanced if the classification categories are not approximately equally represented. There are various ways to remove the bias, like under-sampling the majority class or over-sampling the minority class which is replicating the records of minority class. However, we have used Synthetic Minority Oversampling. Rather than replicating the minority observations (e.g., defaulters, fraudsters, churners) in our case the High Spending customers, Synthetic Minority Oversampling (SMOTE) works by creating synthetic observations based upon the existing minority observations. When the SVM classifier was evaluated on the untreated data, the overall accuracy and sensitivity of individual classes were not up to acceptable standards and far less than statistical classification methods. Hence, there is a need to make adjustment in the process and refine our approach. These refinements are discussed in section 3. Note that the default kernel for R package e1071 SVM is radial with cost function of 1 and gamma at 0.00105042. Below are the performance indices of our base model.

	Reference or Actual		
Predicted	High Spend	Low Spend	Med Spend
High Spend	0	0	0
Low Spend	0	172	30
Med Spend	33	140	625
Total	33	312	655

Table 1: Confusion Matrix for Base Model on original Data

As it can be inferred from Table 1 and Table 4 that the base model on original data i.e. without correcting the bias for minority classes gives superficial better accuracy than on SMOTED data. However, if we look at the minority class i.e. High spending customer, none of which are being predicted by our model on test data. Thus, sensitivity of minority class is very poor. Going forward, we only train the models on the SMOTED dataset to improve individual class prediction accuracy rather than overall accuracy. The next section will enlist the steps taken to improve the model performance and our proposed approach to the given problem.

3. PROPOSED APPROACH

Above section revolved around the base model. This section would list the various experiments and refinements to the base model to improve not only the overall accuracy but also sensitivity of individual classes. In the next sections, we discuss our final approach to the classification of credit card customers and evaluate its performance.

3.1 Implementation

As stated, the original dataset has been corrected for bias and the table 2 shows the distribution of classification classes or the response variable before and after applying Synthetic Minority Oversampling.

Classes	Before SMOTE	After SMOTE
High Spend	02.1 %	34.0 %
Medium Spend	68.4 %	45.8 %
Low Spend	29.5 %	20.2 %

Table 2: Class distribution before and after bias treatment

The SMOTE technique has worked well for this data as the representation of each class is quite balanced in the training data. Next, we applied the base model to the SMOTED training data and observed that now the same base model has improved sensitivity for High Spending customers and has predicted 13 out of 33 cases correctly. However, we can still improve the classification by changing hyperparameters and kernel of our SVM classifier.

	Reference or Actual		
Predicted	High Spend	Low Spend	Med Spend
High Spend	13	0	12
Low Spend	0	172	42
Med Spend	20	140	601
Total	33	312	655

Table 3: Confusion Matrix for Base Model on SMOTED Data

Parameter (in %)	Original Data	SMOTED
Overall Accuracy	79.5	78.6
Sensitivity (High Spend)	0	**40.4**
Sensitivity (Low Spend)	34.5	**55.1**
Sensitivity (Med Spend)	95.4	**91.8**
Specificity (High Spend)	100	98.7
Specificity (Low Spend	95.6	93.9
Specificity (Med Spend	49.2	53.5

Table 4: Performance of Base Model on Original and SMOTED Data

3.2. Training

The training was done using three different kernels of Support vector machines namely Liner, Radial and Polynomial. For each kernel, grid search along with 10-Fold validation as sampling method was performed.

Kernel	Cost	Gamma	Degree
Linear	2	-	-
Radial	2	0.01	-
Polynomial	2	-	2

Table 5: Best Performance Parameters

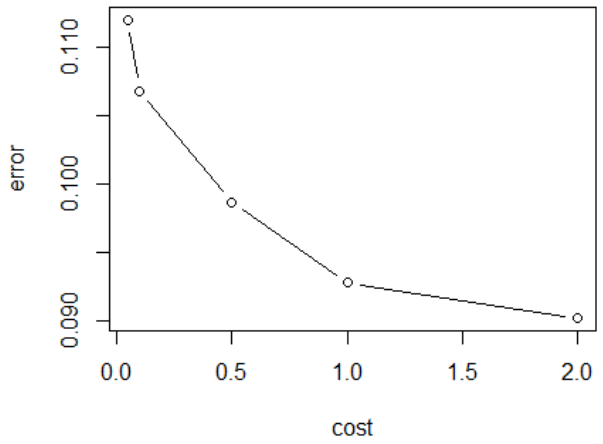

Figure 1: Performance of Linear SVM

3.3. Testing

Testing was done on the 20% data sample of the dataset. It was not corrected for the bias as we wanted to check the performance of the model on real world scenario.

4. EXPERIMENTAL RESULTS

4.2. Evaluation

The main criterion for our evaluation is the confusion matrix, overall accuracy and the sensitivity and specificity of individual category classes. We are not only interested in the overall accuracy, that is, the number of cases our model classified correctly but also if it performs comparatively well in predicting each of the individual classes.

	Reference or Actual		
Predicted	High Spend	Low Spend	Med Spend
High Spend	20	0	21
Low Spend	0	220	71
Med Spend	3	92	563
Total	33	312	655

Table 6: Confusion Matrix using Linear Kernel and Parameter tuning

The linear Kernel with cost of 2 gave the best performance for a range of [0.05,0.1,0.5,1,2] with an accuracy of 80.3% which is the highest compared to others. The cost was limited at 2 so that the outliers does not affect the decision boundary creation extensively, and therefore avoiding overfitting.

Similarly, the data was fed to a SVM with Radial Kernel and grid search was performed for a wide range of cost and gamma values. However, again the cost was capped at 2 and the gamma was restricted to 0.001 to avoid overfitting. As we can see from the table 6, the performance degraded from the linear kernel. Hence, the radial kernel is not suitable for such problem and data. The accuracy also dropped which mean we should stick to linear kernel. However, we also evaluated Polynomial kernel with same approach.

	Reference or Actual		
Predicted	High Spend	Low Spend	Med Spend
High Spend	2	0	1
Low Spend	0	183	65
Med Spend	21	139	590
Total	33	312	655

Table 7: Confusion Matrix using Radial Kernel and Parameter tuning

The polynomial kernel also did not perform better than linear or provided any additional advantage. The cost and degree for best performance were found to be 2. It was also the most time consuming as compared to Linear and radial implementations. Polynomial performed the worst among the three implementations.

	Reference or Actual		
Predicted	High Spend	Low Spend	Med Spend
High Spend	0	0	0
Low Spend	0	8	2
Med Spend	33	304	653
Total	33	312	655

Table 8: Confusion Matrix using Polynomial Kernel and Parameter tuning

Here, we compare the overall accuracy, sensitivity and specificity of individual category classes. Clearly, linear kernel SVM outperforms the other approaches and should be used for such classification stated in the problem.

Parameter (in %)	Linear	Radial	Polynomial
Overall Accuracy	80.3	77.5	66.1
Sensitivity (High Spend)	60.6	6	0
Sensitivity (Low Spend)	70.5	58.6	2.5
Sensitivity (Med Spend)	85.9	90.1	99.6
Specificity (High Spend)	97.8	99.9	100
Specificity (Low Spend	89.7	90.7	99.7
Specificity (Med Spend	69.6	53.6	2.3

Table 9: Comparison of various SVM models

Balanced and improved sensitivity of all classes denotes that our classifier capable of identifying and classifying the customers into the spend category satisfactorily with good accuracy.

5. CONCLUSIONS

We already have acknowledged the need and power of customer segmentation in the credit cards and banking service sector. They help decision makes and marketers to make data driven decisions and have 360-degree view of the customer leading to better customer relation and effective campaigns. Our approach to classify customers based on their spending or card usage would be a step towards this direction.

The use of SVM or support vector machine is not new for classification. However, we experimented with various versions of it to come up the best performing algorithm or model to do such segmentation. The proposed approach not only has a good overall accuracy but also considers the correctness of classifying each category class. Thus, it does not favor one class and becomes blind to others. This is important for the stakeholders in credit card issuing companies.

We also noticed that out of all possible combinations, a multi-class linear SVM classifier performs the best in such time of scenarios.

The approach also has some limitations as to various other neural network approaches could have been implemented for feature engineering. Also, ensembles could have been useful in this case, however, we did not consider them.

REFERENCES

[1] Tony Bellotti and Jonathan Crook, “Support vector machines for credit scoring and discovery of significant features,” *Credit Research Centre*, 7 May 2017.

[2] Bernhard Schlkopf, Alexander J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*, The MIT Press, England

[3] http://www.dummies.com/programming/big-data/data-science/how-to-visualize-the-classifier-in-an-svm-supervised-learning-model/

[4] https://cran.r-project.org/web/packages/e1071/e1071.pdf