Sparse Regression Methodologies: An empirical analysis on credit card services dataset

Long Ngo

Table of Contents

1. Abstract
2. Exploratory Data Analysis
3. Data Preprocessing
4. Methods and Metrics
5. Result
6. Conclusion
7. Reference

Abstract

This project provides an empirical analysis on performance of the sparse regression methods, including the lasso, non-convex penalties and subset selection. As a consequence, we apply these sparse regression methods, attempting to solve a recent binary classification problem from Kaggle. The problem results from the fact that a bank experiences more and more customers leaving their credit card services. Therefore, it would be helpful if there is a model that can predict which customers are going to leave so that the bank can proactively provide them with better services and reduce their churn rate. Dataset provided by the bank consists of their 10,000 customers mentioning their age, salary, marital status, credit card limit, credit card category, etc. In order to obtain an evaluation of how the methods of sparse regression perform on real-world dataset, a logistic regression is employed together with the sparse methods as regularizations. Our computations are mainly conducted in R using the packets of *glmnet*, *ncvreg* and *L0Learn*. The methods are evaluated based on the metrics of accuracy (performance on validation test of dataset), computational time and the number of non-zero coefficients. Our findings show that non-convex penalty methods MCP and SCAD attain the best performance among these sparse methods while Lasso as expected, has the least accuracy. Subset selection method returns the highest time to termination, it, however, is not proved the best approach for this dataset.

Exploratory Data Analysis

Dataset contains 10,000 instances of the bank's customers, mentioning their related information. There are 20 attributes including 19 predictive attributes and a class attribute. The class attribute has two values, namely, Existing Customer and Attrited Customer which is considered the customers who left the bank's credit card services. The portions of these values in class attribute are 86.93% and 16.07%, respectively (Figure 1). Out of 19 features, there are 14 numerical attributes and 5 categorical attributes. There are no missing values in this dataset.

Figure 1: Attrition Flag

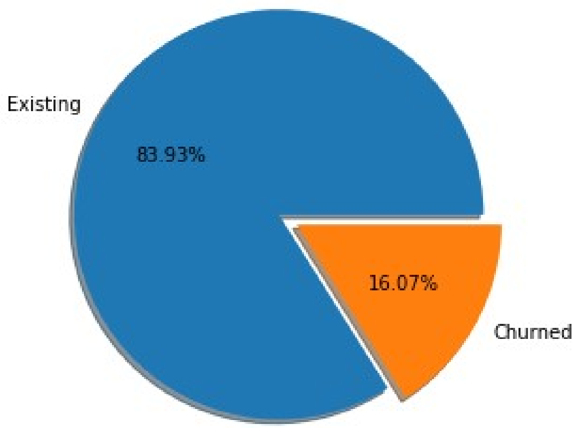
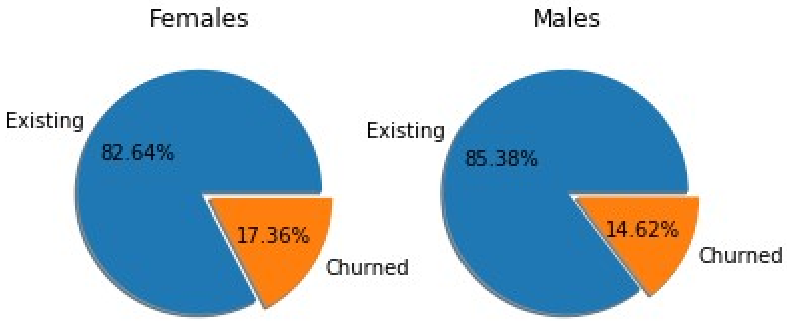
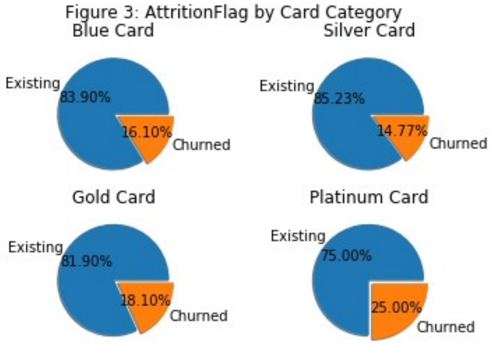


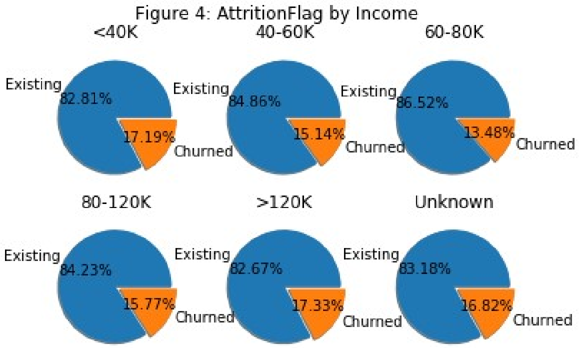
Figure 2 portrays the portions of existing customers and churned customers by gender, indicating that female customers are more prone to leave the bank's service than male customers. Only 14.62% of male are churned while this number from females is 17.36%

Figure 2: Attrition Flag by Gender



In Figure 3 and 4, the chart from Card Category describes the fact that customers who own Platinum card are not too happy with the bank's service and more likely to leave. There are up to a quarter of them are churned while only around 17% of customers of the other cards experience the same. Attrition Flag by Income does not show any pattern. The figures from different income range are similar.

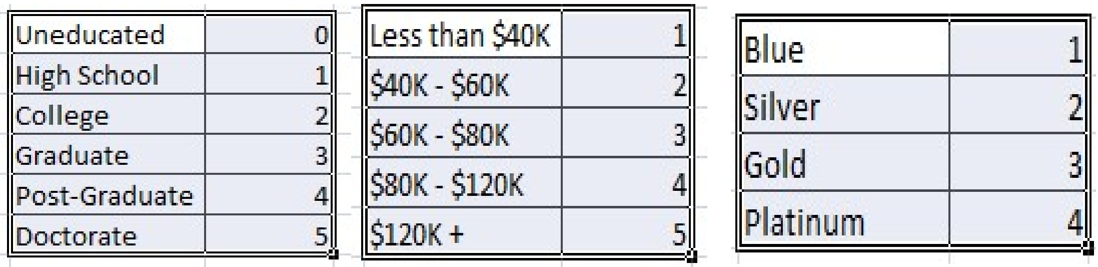




Data Processing

The dataset is fairly clean, which has no missing data. Hence, it only requires a few processes to make it ready for modeling. We first convert all categorical features into numerical values. Figure 4 shows mapping of how ordinal features, i.e. Education, Card Category and Income Category, are converted to numerical.

Gender and Marital Status are nominal data types. Hence, one-hot technique is used to encode these features to numerical. There are some instances whose income and education values are "unknown". As we can see from the table 1 that while income distribution is skewed right, education is more normally distributed. Therefore, we replace "unknown" income by median and "unknown" education by median. Lastly, dataset is split to train set and test set for validation purpose.



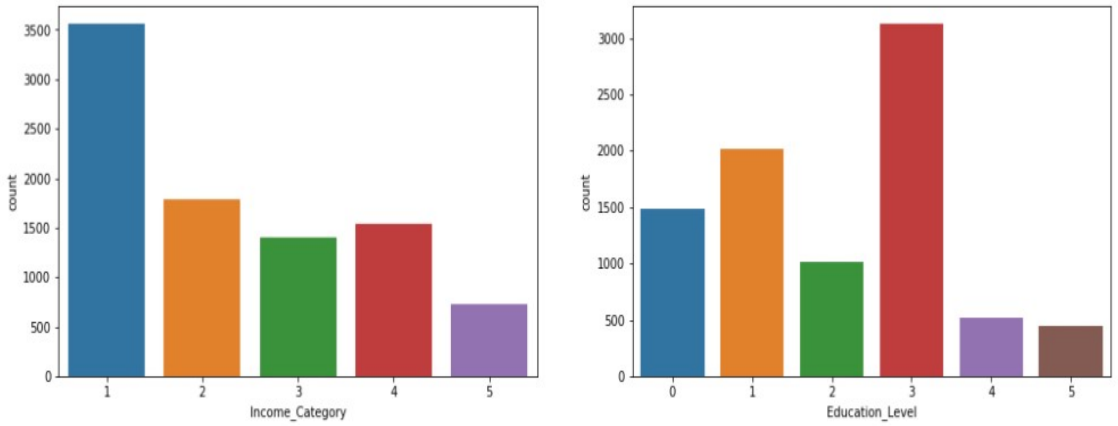
Uneducated	0
High School	1
College	2
Graduate	3
Post-Graduate	4
Doctorate	5

Less than $40K	1
$40K - $60K	2
$60K - $80K	3
$80K - $120K	4
$120K +	5

Blue	1
Silver	2
Gold	3
Platinum	4

Table 1: Mapping of converting ordinal attributes

Figure 5: Distribution of income and education

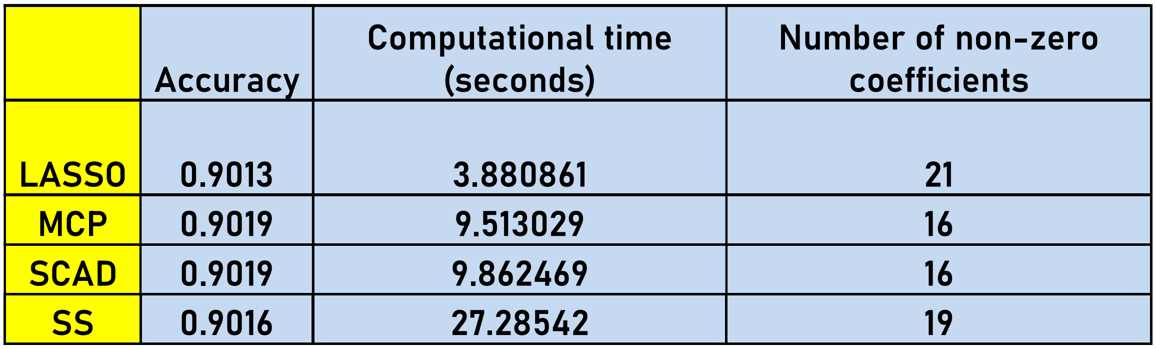


Methods and Metrics

In this binary classification problem, we apply logistic regression together with sparse penalties, i.e. lasso, non-convex penalties: MCP and SCAD, and subset selection, as regularization. Our computation is implemented mainly in R with the support of 3 packages: *glmnet* for lasso, *ncvreg* for non-convex penalties; and *L0Learn* for subset selection. In order to evaluate our regression methods, we use 3 metrics: accuracy, computational time and the number of non-zero coefficients. Our accuracy metric is computed by the accuracy that each method performs on test set. Processing time is measured by real time that computer requires for model fitting on training data. The number of non-zero coefficients represent for the ability to recover the true support of features. For all 3 models applied, we also employ cross validation to tune in the hyper parameter lambda of the penalties.

Result

As shown on table 2, all method attains a fine performance in terms of accuracy when their accuracies are above 90% correct in predicting whether one client will be churned or not. While the non-convex penalties MCP and SCAD achieve the best accuracy with 90.19%, Lasso is the one with lowest accuracy rate of 90.13%. Among all methods, Lasso also as expected requires shortest computational time for fitting process with approximately 3.88 seconds. However, in terms of ability to recover a true support of features, Lasso is the least accurate method when it includes up to 21 attributes into the model. MCP and SCAD, on the other hand, only needs 16 features but able to deliver a more robust performance in predicting. Subset selection, unsurprisingly, is the most computational expensive out of all 3. This method as theory is considered the performing method. However, with this real-world dataset, subset selection performs slightly better than Lasso, achieving an accuracy rate of 90.19% while only requires 19 features. All things considered, the MCP and SCAD penalization are the best approach for the dataset.



	Accuracy	Computational time (seconds)	Number of non-zero coefficients
LASSO	0.9013	3.880861	21
MCP	0.9019	9.513029	16
SCAD	0.9019	9.862469	16
SS	0.9016	27.28542	19

Table 2: Result

Conclusion

Our results show that among all methods, Lasso is the least computational expensive. However, this method has the lowest accuracy and highest False Detection Rate (based on number of non-zero coefficients). Subset selection, on the other hand, has a longest time for fitting process but only performs slightly better than Lasso. Non-convex penalty (MCP and SCAD) demonstrate they outperforms the other methods in terms of accuracy and False Detection Rate and definitely is the go-to method for this real-world credit card scenario.

Reference

Bertsimas, D., Pauphilet, J., & Van Parys, B. (2020, October 01). Institute of Mathematical Statistics: Statistical Science Future Papers. Retrieved November 01, 2020, from https://www.e-publications.org/ims/submission/STS/user/submissionFile/35175?confirm=f096a84a

sakshigoyal. (2020, December 2). 91% recall rate - churned customers. Retrieved December 8, 2020, from Kaggle.com website: https://www.kaggle.com/sakshigoyal7/91-recall-rate-churned-customers