MA678 Midterm project

Yuxi Wang

2020/11/09

Abstract

Santander Group wants to identify the value of transactions for each potential customer. They use some indicators to describe the value of transactions for each potential customer. I want to use such data is because I want to engage in related work in the field of financial data analysis. Moreover, I think it is necessary for every company to learn and try to judge the value of its customers or clients. For the dataset, there are 4993 features and 4459 observations in it. Through EDA and basic dimensionality reduction methods, like PCA and clustering, I sorted out the entire messy data set. Then I used a multilevel regression model for modeling the data.

1. Introduction

In this project, Santander invites people to help them identify which customers will make a specific transaction in the future. The data provided for this competition has the same structure as the real data they have available to solve this problem.

2.Exploratory Data Analysis

2.1 Data Summary

After watching the first ten rows of two datasets, we can conclude: 1. As the project had presay, the names of every columns are anonymized, so we do not know what these variables' meaning. 2. There are many zero values present in the data.

```
## [1] "The number of NAs in the training set is: 0"
```

```
## [1] "The number of NAs in the test set is: 0"
```

3. Neither the test set nor the training set has NAs.

However, since the data size is a little huge, we cannot easily use summary() or str() to have a simple understing of the data-set. So, I just count the rows and columns of training set and test set.

```
## [1] "The number of records in the training set: 4459"
```

```
## [1] "The number of predictors in the training set: 4993"
```

```
## [1] "The number of records in the test set: 49342"
```

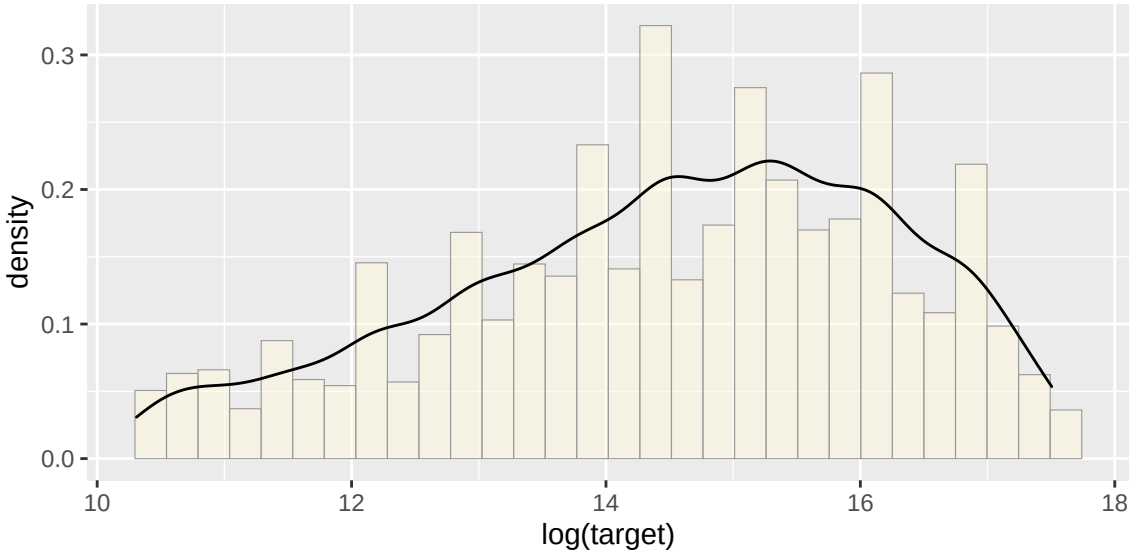
```
## [1] "The number of predictors in the test set: 4992"
```

As we can see, features in test set and training set are different. But this is because the tset set does not have target, and that is what I will do in test set if I take part in the competition. So, it does not metter.

2.2 Target variable

Because the dataset is very massive, so I have to rank the value to see the distribution plot of it. If we analysis the plot we can find that the distribution with majority of the data points having low value, and a huge amount of the data is 0.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



So, now we know the desity of the target variable. Also, we know that if we use target as dependent variable to fit the model, it is better to use log(target) instead of target.

2.3 Other predictors

There are near 5000 columns in the training set, which means I need to do the feature selection, in order to better fit a model. We know that there are a lot of 0s in the dataset, so we first remove them, and then, we can calculate the condition number to see if there is multicollinearity.

```
## [1] "The number of predictors that are not all 0s are: 4737"
```

So we know that there are 256 variables that all equals 0. Also, since they are all 0, they are meaningless in the model.

However, there are still too many predictros.

```
## [1] 6.187769e+19
```

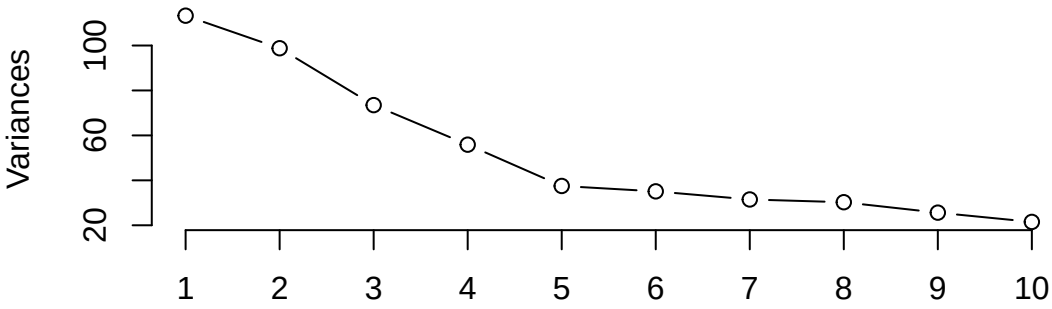
Since $6.187769e + 19 >> 1000$, it shows that there is serious multicollinearity between the independent variables.

We can use the Characteristic root judgment or step regression which may deal with this method. But they are both too slow to use step by step. In the meantime, the result I get may not meet my ideas. So, I add a PCA model in the modeling part in order to deal with the multicollinearity.

3. Modeling

3.1 PCA

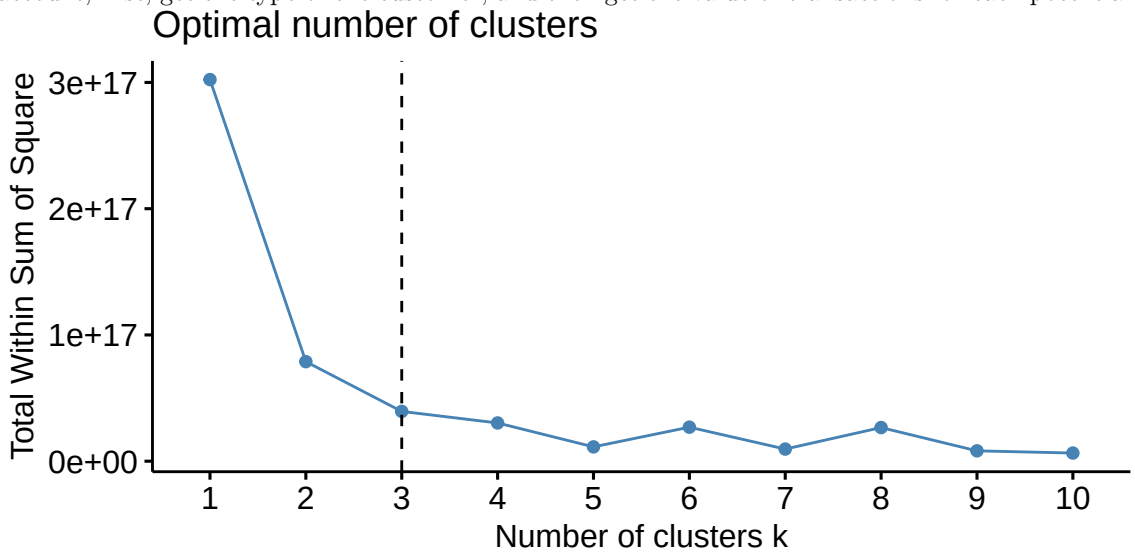
Since the predictor I used is encrypted, so I don't need to estimate the actual meaning of each new variable temporarily by using the principal component. Although I acknowledge that in the specific analysis process, it is very necessary to understand the actual meaning of each variable as much as possible. Now, I am going to do PCA. Here, since there are more features than observation, the function I use is "prcomp".



The outcome is too huge, so I just show the scree plot of the PCA. At first glance, the results shown in this plot are not good. But this is mainly caused by the existence of a large number of variables with only one or a few values, and they are independent. After I understand the variable structure of our original data. When the principal component is 5, an obvious inflection point appears. After the five variables produced by the principal component can cover 60% of the original variable, the contribution rate of the principal component sharply decreased. Therefore, the method I use is to use the top five principal components, although they can only represent 60% of the original independent variables. As for the other predictor, without greatly expanding the number of principal components, it is difficult for me to use them without considering the specific application background because they are not universal behaviors to the bank's customers.

3.2 Clustering

Since the result we want is to determine whether the customer has value, we can get a structured data type after converting the original variable. First, we cluster all the customers that are owned by the observations, that is, we can divide customers into several categories according to the behavior of the account. At the same time, we can classify their value levels according to the 1/3 and 2/3 quantiles. Then, according to the behavior of their account, first, get the type of the customer, and then get the value of transactions for each potential customer.



3.3 Multilevel regression

Finally, do the multilevel linear regression on the processed data set.

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: log(y) ~ 1 + x1 + x2 + x3 + x4 + x5 + (1 | cluster) + valuable
##    Data: data
```

```
##
##      AIC      BIC   logLik deviance df.resid
##  10751.4  10809.1  -5366.7  10733.4     4450
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.0958 -0.4387  0.0547  0.5426  3.3190
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  cluster  (Intercept) 0.2210   0.4701
##  Residual             0.6475   0.8047
## Number of obs: 4459, groups:  cluster, 3
##
## Fixed effects:
##               Estimate Std. Error t value
## (Intercept) 11.5946281  0.2770679  41.848
## x1          -0.0030195  0.0011991  -2.518
## x2          -0.0002849  0.0012784  -0.223
## x3           0.0004816  0.0014107   0.341
## x4          -0.0014976  0.0016158  -0.927
## x5           0.0036888  0.0019738   1.869
## valuable     1.6586389  0.0203550  81.486
##
## Correlation of Fixed Effects:
##          (Intr) x1     x2     x3     x4     x5
## x1        0.008
## x2        0.019  0.102
## x3       -0.008 -0.014 -0.017
## x4       -0.011 -0.010 -0.013  0.005
## x5        0.003  0.021  0.018  0.000 -0.001
## valuable -0.186  0.068  0.000  0.030  0.046  0.008
```

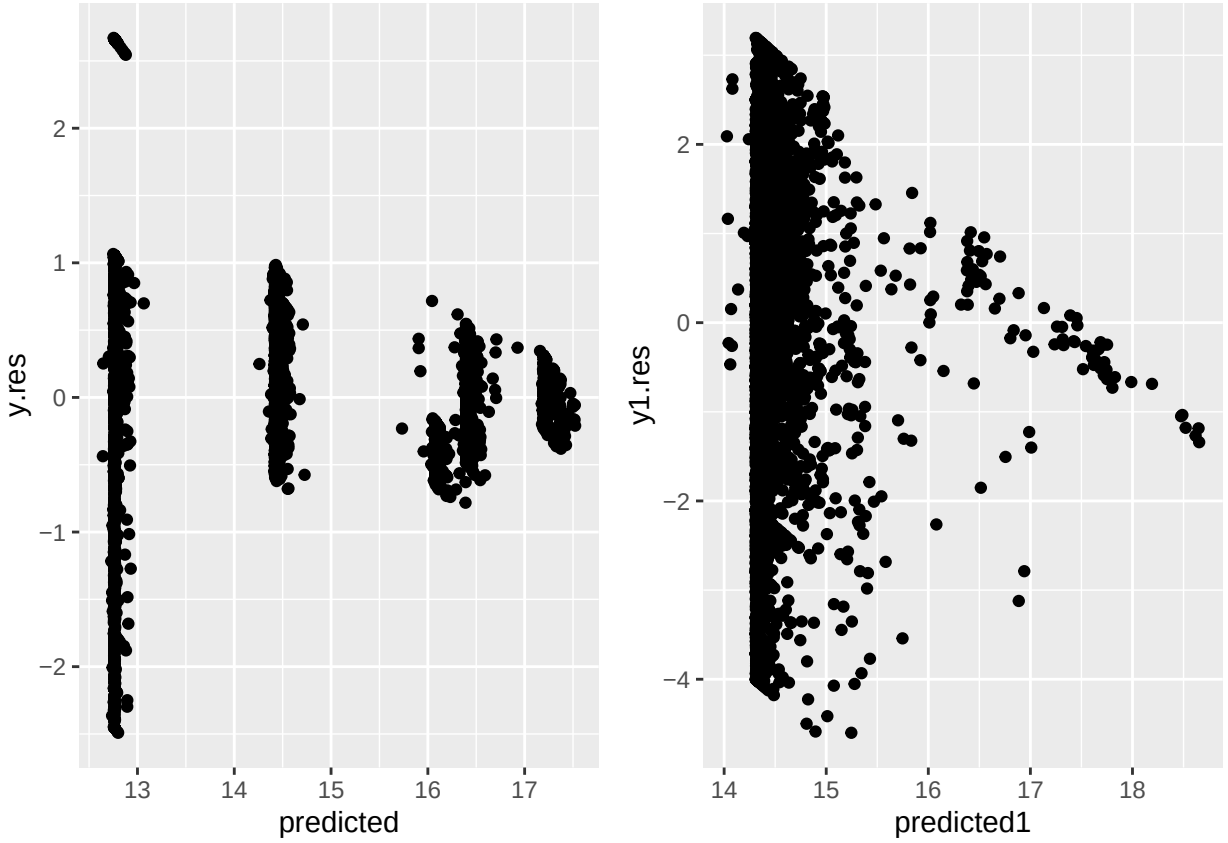
4.Results

So far, I have completed the modeling part. So, it is time to see the results of the model. First of all, I need to admit that compared with Fixed effects, Random effects have a very large influence in predicting the target. From the output results, the fixed effects: intercept and valuable have the greatest impact, followed by X1, X4, and X5.

It is common to calculate the confidence interval. Computing profile confidence intervals ...

	2.5 %	97.5 %
.sig01	0.2461	1.395
.sigma	0.7883	0.8217
(Intercept)	10.83	12.36
x1	-0.005372	-0.0006668
x2	-0.002794	0.002224
x3	-0.002284	0.003247
x4	-0.004665	0.00167
x5	-0.0001807	0.007558
valuable	1.619	1.699

From the confidence interval, we can see that all the coefficients of fixed effects are pretty small than random effects. After calculating the confidence interval I make a residual plot to see the residence of the model. It can be seen from the left residual plot that points are not evenly distributed randomly throughout the area, and there are certain groups of points. In the right plot, through spending the residual plot of a simple linear model, we can see that the result, in this case, maybe better than the multilevel results.



5.Discussion

I have to admit that my work has some flaws, and I will pay special attention to future projects. 1. First of all, I understand that a more critical step is missing, which is to bring the coefficients of all the principal components used in the final model into the initial variables. The reason why I did not do this is that, even if I include the coefficients into the original independent variables, I cannot explain the specific economics since they are anonymous variables. 2. The second point is that the model I used is not completely suitable. I understand that PCA is not a suitable method now for specific industry problems. When I find that the results of PCA are not good enough for modeling, I tried to learn some machine-learning descending dimension algorithm, but in the process of running the program I often break down, and I don't have a deep understanding of those models, so I finally used PCA's outcome as results. 3. Because I want to combine the data with the knowledge learned this semester, I define some variables myself in order to achieve the goal. I know this is not a good habit, especially when I structure them without an in-depth understanding of the specific business.