551 Project

Chen Xie, Xinye Jiang, Xun Wang

2019/4/13

1 Introduction

Logistic regression is a very famous and widely applied technique in supervised learning. It is usually used to perform predictive analysis when the response variable is binary.

Logistic regression can also be approached by Bayesian modeling. In general, Bayesian analysis is more flexible, and it is proved to be superior for small samples. For Bayesian modeling, it can incorporate prior information. For example, if we want to know exactly which factors are most effective on the response, we can use shrinkage prior to implement variable selection.

In practice, predicting the exit status (binary: 0 or 1) of customers of a bank can be an application of the standard logistic regression as well as Bayesian approach. In this report, we delve into a data set about account status of a bank in three European countries. The objective is to predict if the clients will leave the bank based on part of their information, such as geography, gender and account balances, etc. In this process, we will also explore the different effects of specific predictors to response variable in different models.

2 Data Exploration

2.1 Data Set

The data set we use for this report is from Kaggle website. It has 10000 observations and 11 variables. The Exit Status (Exited) is our response variable. The Exit States is 1 if the customer closed account with bank and 0 if the customer is retained. The other variables can be treated as predictors, and they are Credit Score, Geography, Gender, Age, Tenure, Balance, Number Of Products (NumOfProducts), Has Credit Card (HasCrCard), Is Active Member, Estimated Salary. This data set has mixture of types of independent variables, where 6 of them are numeric and 4 are categorical. And the Table 1 below shows more details of all the variables in this data, including names, types and brief descriptions.

Table 1: Brief Descriptions of Predictors

Variables	Type	Description
Exited Status	Binary	1 if the customer closed account and 0 if the customer is retained
Credit Score	Continuous	Credit Score of the customer
Geography	Categorical	The country from which the customer belongs
Gender	Binary	Male or Female
Age	Continuous	Age of the customer
Tenure	Continuous	Number of years for which the customer has been with the bank
Balance	Continuous	Bank balance of the customer
Number of Products	Discrete	Number of bank products the customer is utilizing
Has Credit Card	Binary	Whether the customer holds a credit card with the bank or not
Is Active Member	Binary	Whether the customer is an active member with the bank or not
Estimated Salary	Continuous	Estimated salary of the customer in Dollars

Table 2: Summary of Numeric Variables

	CreditScore	Age	Tenure	Balance	NumOfProducts	EstimatedSalary
Min.	350.0000	18.0000	0.0000	0.00	1.0000	11.58
1st Qu.	584.0000	32.0000	3.0000	0.00	1.0000	51002.11
Median	652.0000	37.0000	5.0000	97198.54	1.0000	100193.91
Mean	650.5288	38.9218	5.0128	76485.89	1.5302	100090.24
3rd Qu.	718.0000	44.0000	7.0000	127644.24	2.0000	149388.25
Max.	850.0000	92.0000	10.0000	250898.09	4.0000	199992.48

Table 3: Summary of Categorical Variables

Geography	Gender	HasCrCard	IsActiveMember	Exited
France: 5014	Female: 4543	0: 2945	0: 4849	0: 7963
Germany: 2509	Male: 5457	1: 7055	1: 5151	1: 2037
Spain: 2477				

2.2 Data Exploration

The Table 2 and Table 3 are summary statistics of all the variables. To be more precise, Figure 1 shows barcharts of categorical variables. In the Figure 1, the left figure shows the overall distribution of Exited statu, where we could find that a large proportion of clients keep their bank accounts. The number of Exited Status=0 is almost four times of number of clients who left the bank. The data is very imbalanced, which may be problems for further inference or prediction. The right part in Figure 1 shows how categorical predictors are distributed by Exited Status. It implies that whether the customer has a credit card extremely affects the Exited Status.

Next, we also want to explore the relationship between predictors and the response variable. The left part in Figure 2 shows the boxplots of continuous factors by Exited Status. It provides some evidence that some predictors have strong impact on Exited Status. For instance, Age and Balance are inluential factors for Exited Status. The right part in Figure 2 is the pairwise plot of numeric predictors, including scatterplots, correlations bewteen each two of them, as well as histograms. We could gain a general knowledge of distributions for every numeric variable. According to the right plot in Figure 2, the numeric predicors are not highly correlated with each other.

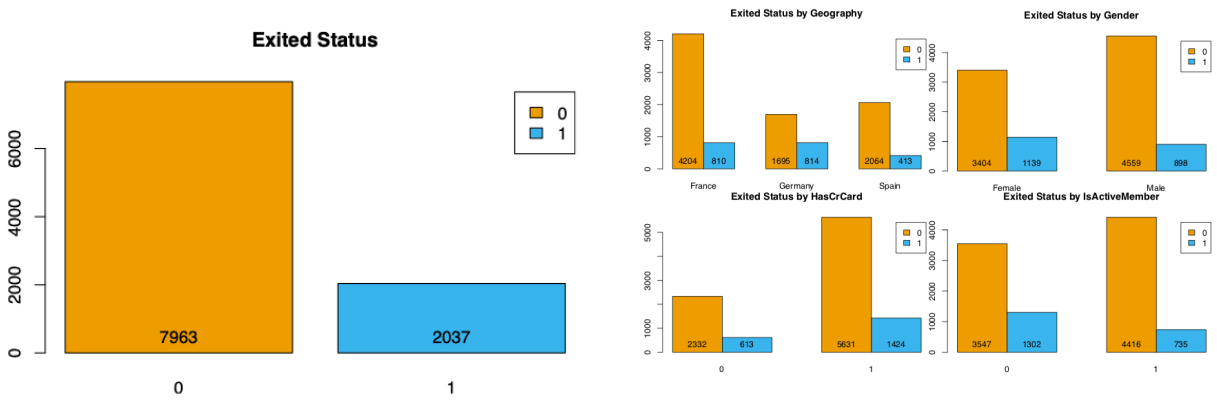


Figure 1: Barcharts

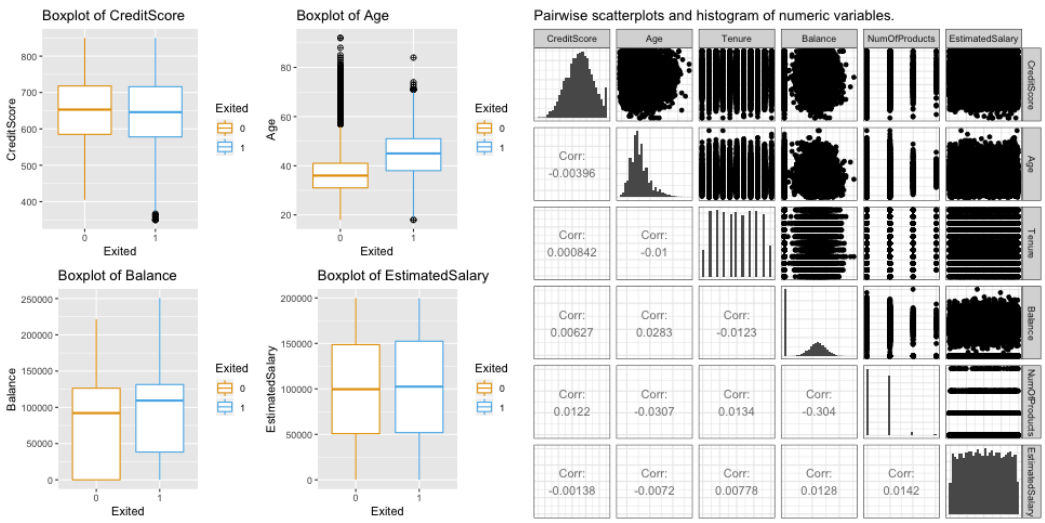


Figure 2: Boxplots and Scatter Plots

However, although we can observe some information of the relationship between predictors and the response variable, we need more credible proofs.

2.3 Problems of the Data

Imbalanced data unclear source biased data collection It may need more . . . we can still do some learning from this data.

3 Regression Models

In this part, we will briefly introduce three logistic regression models, including the standard one and two Bayesian models.

3.1 Logistic Regression

In addition to the most significant reason that our response variable is binary, there are several advantages of logistic regression against other methods. First, it can be interpreted more easily comapred with some complex models, while holding the same level of prediction precision. It helps to explain the relationship between the predictors and response variable. Next, the logistic regression can also handle mixed types of exploratory variables. Finally, logistic regression is an informative method that it provides both the size and the direction of the effects of its predictors.

The basic assumption of standard logistic regression model is that the observations $y_1, ..., y_n$ are independent and following binomial distribution (1, p_i), where p_i is the probability of response $y_i = 1$. We could estimate the parameters β by maximum likelihood method, that is, maximizing $P(y_1, ..., y_n; \beta)$. The estimate is also called ML estimate. The statistical description is as below:

$$y_1, ..., y_n, \text{ are independent}$$

$$y_i \sim \text{ Binomial } (1, p_i)$$

$$\log(\frac{p_i}{1-p_i}) = x_i^T \beta$$

$$p(y_i; \beta) = (p_i)^{y_i}(1-p_i)^{1-y_i}$$

$$p(y_1, ..., y_n; \beta) = \prod_{i=1}^{n}(p_i)^{y_i}(1-p_i)^{1-y_i}$$

Among them, x_i is the i-th row of observation and β is the parameters.

3.2 Bayesian Logistic Regression with Normal prior

In the standard logistic regression above, we treat β as a column of unknown but fixed parameters. Actually, β could be seen as a vector of random variables from the prospective of Bayesian analysis. That is the preliminary thought of Bayesian analysis. In this case, we can also get a point estimate for the parameters by maximizing the posterior distribution, that is, maximizing $P(\beta|y_1, ..., y_n) \propto P(y_1, ..., y_n; \beta)P(\beta)$. The maximum of posterior distribution, that is, the mode of posterior, is called MAP estimation. We can see theat the ML estimate of standard logistic regression is the MAP estimate when the prior of β is uniform.

But sometimes, it is hard to get analytic form of mode of the posterior distribution. Instead, we can estimate the parameters by mean or median of posterior samples, which is ususally generated by Markov chain Monte Carlo (MCMC) sampling methods. For this paper, we also use this idea for inference and prediction.

In the Bayesian logistic regression, basiclly it assumes that β follows a multivariate normal prior distribution. For simplicity, we also assume β_i are indenpendent with each other. and μ, σ. It has some shrinkage effects on β, but not strong.

The assumptions are:

$$y_1, ..., y_n, \text{ are independent}$$

$$\text{Likelihood: } y_i \sim \text{ Binomial } (1, p_i)$$

$$\text{Parameters: } \log(\frac{p_i}{1-p_i}) = x_i^T \beta$$

$$\text{Prior: } \beta \sim N(\mu, \mathbf{\Sigma})$$

$$\text{Also Assume: } \beta_1, ..., \beta_p \text{ are independent}$$

$$\beta_i \sim N(0, 0.01^{-1})$$

where μ is the mean vector of β and Σ is a covariance matrix.

3.3 Bayesian Logistic Regression with NE prior

To figure out the most important predictors to the exited status response and to improve the prediction precision, we could use the LASSO Interpretation, cross validation methods,...

$$y_1, ..., y_n, \text{ are independent}$$

$$\text{Likelihood: } y_i \sim \text{ Binomial } (1, p_i)$$

$$\text{Parameters: } \log(\frac{p_i}{1-p_i}) = x_i^T \beta$$

$$\text{prior: } \beta_i \sim N(0, \sigma_i^2)$$

$$\text{Assume: } \beta_1, ..., \beta_p \text{ are independent}$$

$$\text{Hyper Prior: } P(\sigma_i^2) \sim \text{Exponential } (\lambda)$$

4 Inference and Prediction

4.1 Logistic Regression

Standard logistic regression could be easily fitted by using a very common function `glm` in `stats` package. Printing out the summary table, we get a model like:

By normal approaximation, we can get Estimate, standard errors, z statistics, and p-values of every β, as Table 4 shows. We can find ...

Table 4: Summary of Logistic Regression Fit

	Estimate	Std.Error	z.value	p.value
(Intercept)	-3.4001716	0.2818085	-12.0655372	0.0000000
CreditScore	-0.0007415	0.0003240	-2.2887402	0.0220945
GeographyGermany	0.8303893	0.0774359	10.7235652	0.0000000
GeographySpain	0.0634949	0.0818061	0.7761631	0.4376527
GenderMale	-0.5157407	0.0626479	-8.2323745	0.0000000
Age	0.0723626	0.0029592	24.4533208	0.0000000
Tenure	-0.0093102	0.0107122	-0.8691283	0.3847770
Balance	0.0000026	0.0000006	4.3894024	0.0000114
NumOfProducts	-0.0534129	0.0539250	-0.9905047	0.3219275
HasCrCard1	-0.0757810	0.0685853	-1.1049159	0.2691960
IsActiveMember1	-1.0558242	0.0662013	-15.9486858	0.0000000
EstimatedSalary	0.0000000	0.0000005	0.0344362	0.9725293

Table 5 is prediction.

Table 5: Prediction Result of Logistic Regression

	Predicted: No	Predicted: Yes
Actual: No	1942	71
Actual: Yes	389	98

Table 7: Prediction Result of Bayesian Logistic with Normal Prior

	Predicted: No	Predicted: Yes
Actual: No	1942	71
Actual: Yes	389	98

4.2 Bayesian Logistic Regression with Normal prior

Intepretation of parameters will be showed in next part together with other two models.

Monte Carlo Markov Chain is an extremely effective way to sample from posterior distribution in Bayesian analysis system. In this problem, we then use MCMC method to obtain the posterior distribution of β parameters from normal prior. Here `MCMClogit` function in `MCMCpack` package could be a good choice for us.

This function runs 11000 default iterations, starting with mean 0 and precision matrix of $0.01\mathbf{I}$ under a multivariate normal prior. Discarding first 1000 iterarions, we save the big c(10000,12) posterior matrix into an object `res_mcmc`.

normal approximation: z statistic, p-values Table 6, se

Table 6: Bayesian logistic with Normal prior

Table 6: Summary of Bayesian Logistic with Normal Prior

	Estimate	Std.Error	z.value	p.value
(Intercept)	-3.3805762	0.2726979	-12.3967823	0.0000000
CreditScore	-0.0007657	0.0003175	-2.4115229	0.0158861
GeographyGermany	0.8391580	0.0785673	10.6807606	0.0000000
GeographySpain	0.0761547	0.0824086	0.9241113	0.3554284
GenderMale	-0.5218670	0.0648600	-8.0460581	0.0000000
Age	0.0724382	0.0028751	25.1951166	0.0000000
Tenure	-0.0103392	0.0101812	-1.0155201	0.3098580
Balance	0.0000026	0.0000006	4.1315712	0.0000360
NumOfProducts	-0.0580465	0.0511432	-1.1349800	0.2563837
HasCrCard1	-0.0705661	0.0670334	-1.0527013	0.2924779
IsActiveMember1	-1.0571533	0.0672026	-15.7308313	0.0000000
EstimatedSalary	0.0000000	0.0000006	0.0002936	0.9997658

Table 7: Prediction same result

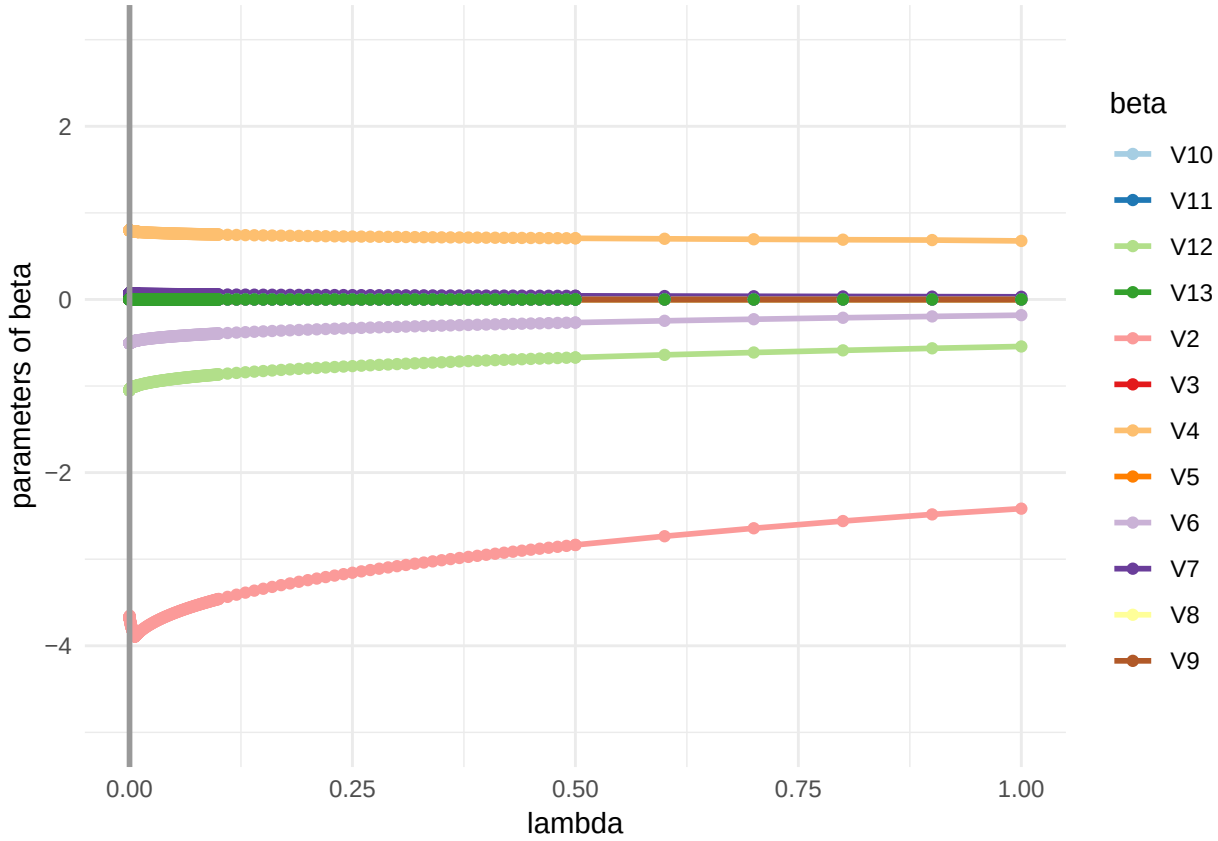
4.3 Bayesian Logistic Regression with NE prior

Bayesian logistic regression with LASSO prior is a more complex hierarchical model compared with previous ones. We need to sample from the joint posterior formula of β, σ^2, and λ is a fixed but unknow paramter that we need to eatimate first. So we first use cross-validation to select the optimal λ which may minimize cross-validation error. By specifying λ, we fit a hierarchical Bayesian model to make references on test data. It is usually a little hard or tedious to realize the sampling process without cojugate priors. In terms of our knowledge, we could use another powerful tool Rstan to do Hamilton Monte Carlo sampling. See Appendix for Rstan script. we could also use a function contributed by other R users called `EBglmnet` to acheive the same objective, which is much concise. First, we use `cv.EBglmnet` function to select the optimal model. In last step using `EBglmnet` to sample from the established model. See references below to get more information about this fucntion.

Table 8: Prediction Result of Bayesian Logistic with Normal Prior

	Predicted: No	Predicted: Yes
Actual: No	1947	66
Actual: Yes	392	95

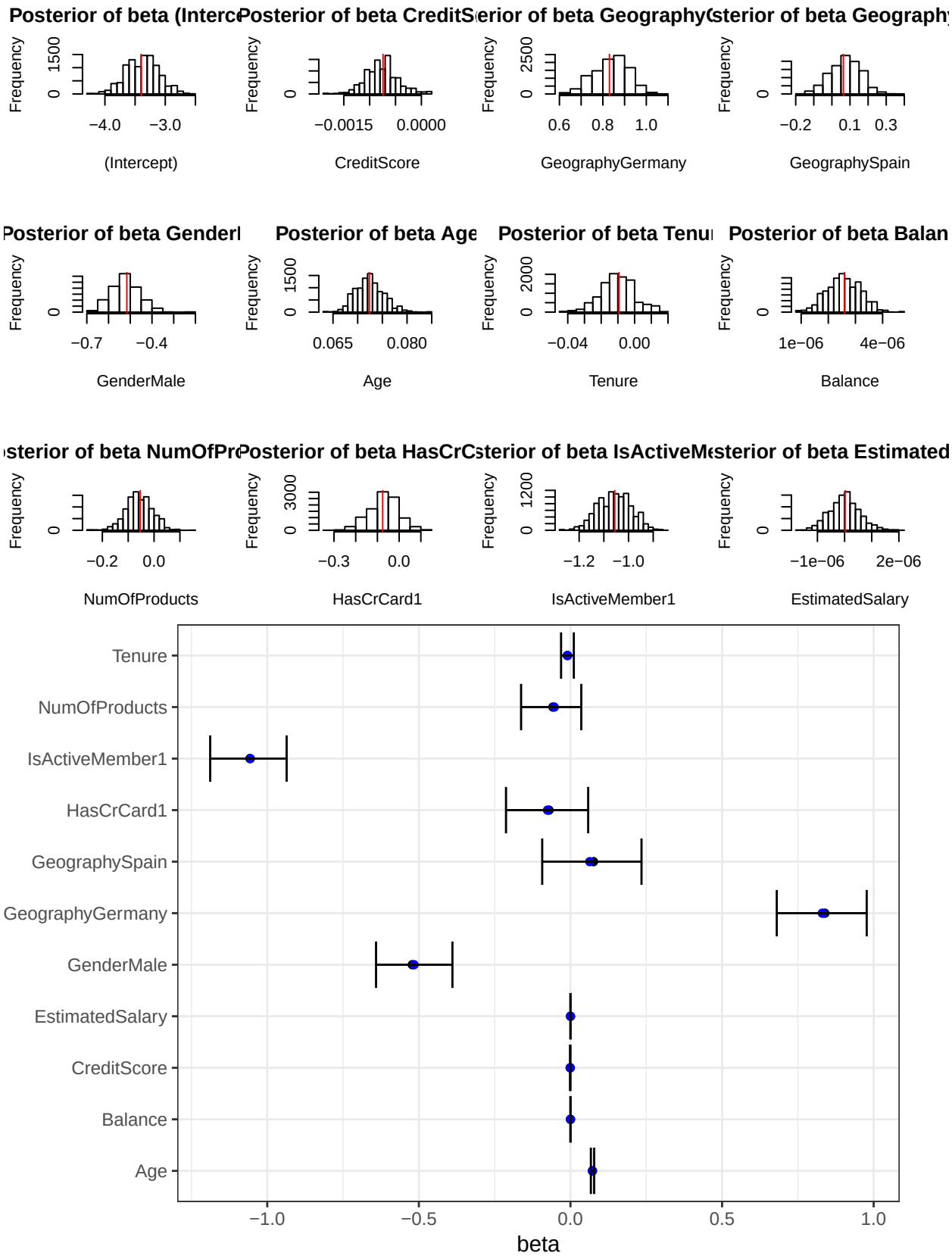
```
## EBLASSO Logistic Model, NE prior,Epis:  FALSE ; 5 fold cross-validation
```

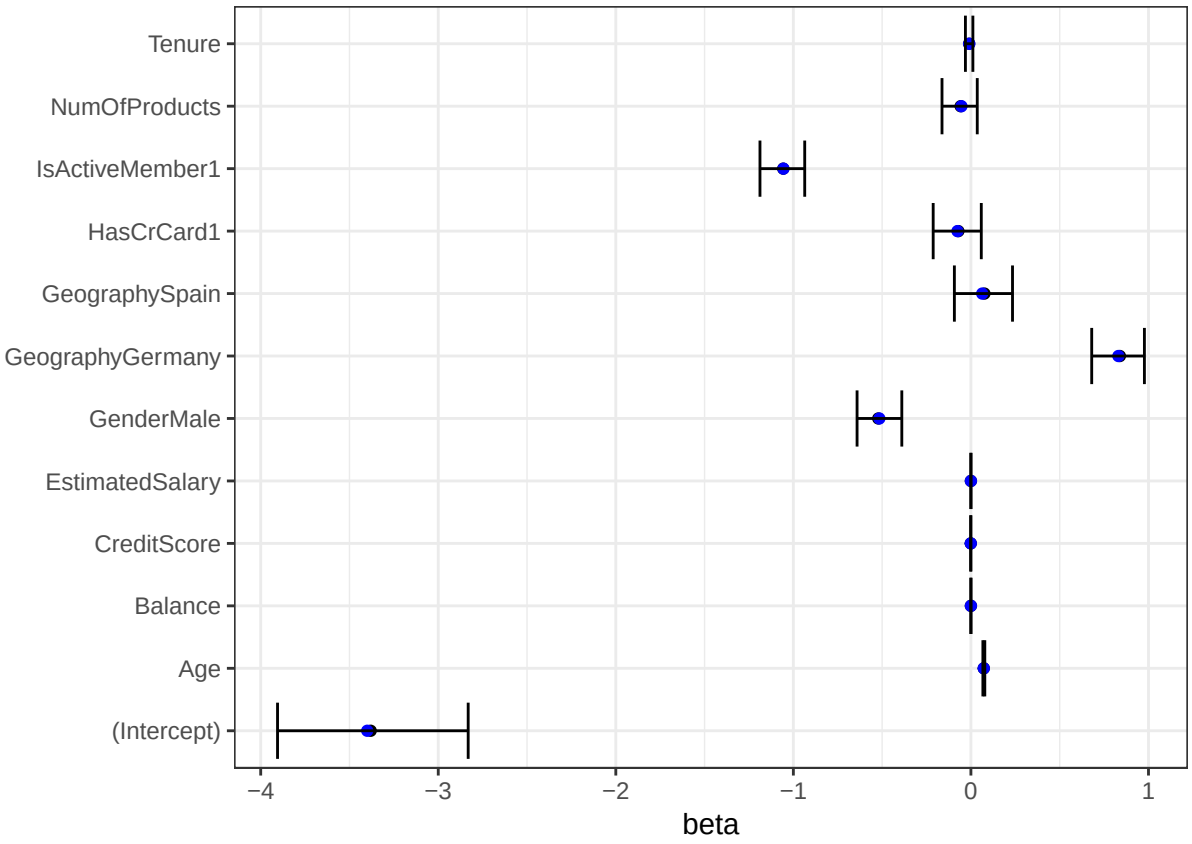


From the result of `cv.EBglmnet`, we could obtain an optimal λ, which is 8.201102e-05. Fit model again using training data set, we could get the hierachical model.

```
## [1] 0.8168
```

to be continue





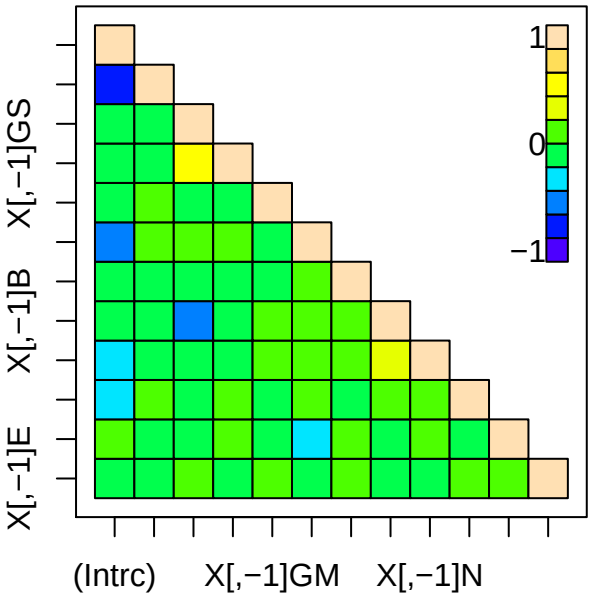


Table 9: Comparison

	p1	p2
(Intercept)	0.0000000	0.0000000
CreditScore	0.0220945	0.0158861
GeographyGermany	0.0000000	0.0000000
GeographySpain	0.4376527	0.3554284
GenderMale	0.0000000	0.0000000

Age	0.0000000	0.0000000
Tenure	0.3847770	0.3098580
Balance	0.0000114	0.0000360
NumOfProducts	0.3219275	0.2563837
HasCrCard1	0.2691960	0.2924779
IsActiveMember1	0.0000000	0.0000000
EstimatedSalary	0.9725293	0.9997658

5 Discussion

6 Conclusion

7 Future Work

8 References

Wei, R., and Ghosal, S. (2017). Contraction properties of shrinkage priors in logistic regression, Preprint at http://www4.stat.ncsu.edu/~ghoshal/papers.

Genkin, A., Lewis, D. and Madigan, D. (2007). Large-scale Bayesian logistic regression for text categorization, *Technometrics* **49**(3): 291–304.

Kapat, P., and Wang K. (2006). Classification Using Bayesian Logistic Regression: Diabetes in Pima Indian Women Example. Ohio State University, OH. https://www.asc.ohio-state.edu/goel.1/STAT825/PROJECTS/KapatWang_Team4Report.pdf

Anhui Huang and Dianting Liu (2016). EBglmnet: Empirical Bayesian Lasso and Elastic Net Methods for Generalized Linear Models. R package version 4.1. https://CRAN.R-project.org/package=EBglmnet

Andrew D. Martin, Kevin M. Quinn, Jong Hee Park (2011). MCMCpack: Markov Chain Monte Carlo in R. Journal of Statistical Software. 42(9): 1-21. URL http://www.jstatsoft.org/v42/i09/.

Li, L & Yao, W. (2017). Fully Bayesian logistic regression with hyper-LASSO priors for high-dimensional feature selection. Statistics 88, 1-25.