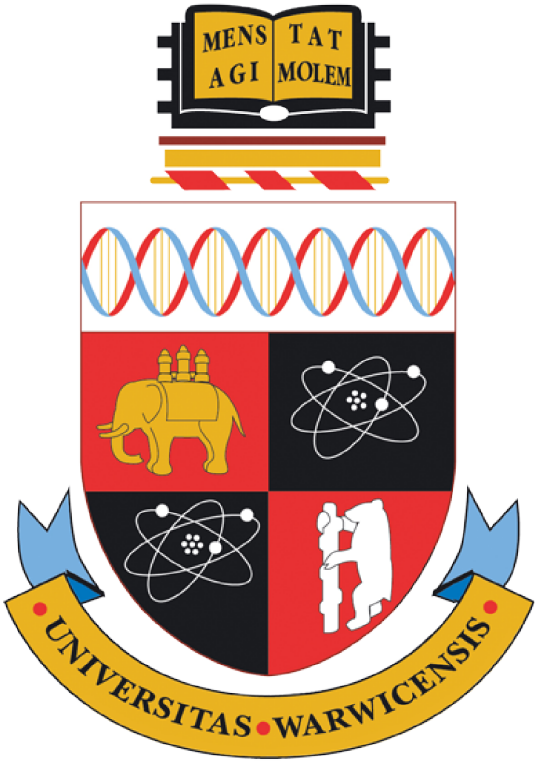
ST952:
Assessed Coursework 2

1620789, 1957259, 1966041



Department of Statistics
University of Warwick
United Kingdom
November 2019

Contents

1 Introduction

In the financial domain, with high volume of customers, automation of processes is of a high interest. Using the 1999 financial Czech data, we attempt to learn the best model describing whether an individual is issued with the Gold as opposed to Classic or Junior card. We employ deviance, step-wise and LASSO model reduction techniques to select the best model which could serve automatic Gold card issuance purposes. We find that card withdrawals, sex and age for those who have used their card are the most statistically relevant variables to answer this question. However, due to low quality of the data (small target distribution, unbalanced data), our model proved to be of low explanatory and predictive power.

2 Data description

The data in use is the extract of the data used prior to the conference of Principles and Practices of knowledge discovery in Prague in 1999. The data contains 5369 observations with 10 variables - 4 continuous and 6 discrete (including the response variable). The discrete variables are all transformed into factors prior the analysis. The resulting summary statistics are presented in the Tables 1 and 2.

Age	Card withdrawal	Cash credit	Cash withdrawal
Min. :17.1	Min. : 0	Min. : 0	Min. : 0
1st Qu.:31.4	1st Qu.: 0	1st Qu.: 0	1st Qu.: 2521
Median :45.6	Median : 0	Median : 7850	Median : 5442
Mean :46.3	Mean : 359	Mean :10362	Mean : 7229
3rd Qu.:59.1	3rd Qu.: 0	3rd Qu.:19736	3rd Qu.:11118
Max. :82.0	Max. :6160	Max. :31778	Max. :28481

Table 1: Summary statistics of the continuous variables

Gold card	Second card	Sex	Frequency of statement	Used card	Type
No: 5281	No: 3631	Female: 2640	After Transaction: 107	No: 4558	Owner: 4500
Yes: 88	Yes: 1738	Male: 2729	Monthly: 4989	Yes: 811	User: 869
			Weekly: 282		

Table 2: summary statistics of the categorical variables

2.1 Data Issues

Already from the table we can see that the data is skewed and have many zero values. The 1st quantile of credit card withdrawals and cash credit variables are zero. Furthermore, the median of the credit card withdrawals is zero too. Together with the fact that there are only 88 observations with Gold card (Table 2), this could seriously affect the analysis and even more affect the predictions.

Furthermore, duplicates in the data impose an additional potential problem. Perfect duplicated records are spotted if we were to ignore Type and Gold variables. In this situation, card owner and user use the same account, hence the spending information is combined together. As a result we do not know which characteristics made card user or owner to be issues with the Gold card or not. However, we lack information about the data and the relevant domain knowledge in credit card, thus do not know on what basis should we eliminate the duplicates. To avoid incorrectly change the information contained in the data, we will not drop the duplicates and will report the analysis of based on the full data.

3 Initial Explanatory Analysis

Figure 1 illustrates the distribution of the explanatory variables split by the category of the response variable Gold (1 or 0). In terms of the discrete variables, 4 out of 5 variables appear to be unbalanced. Alongside the Gold card variables (only 88 positive responses), this issue is present in the frequency of a statement variable as well as credit card user and owner vs user dummy. We can see that sex seems to be the only variables which distribution changes depending on the value of Gold - males seem to be more likely to to be issued with the Gold card than females.

Figure 1 also plots the histograms of the continuous variables, with and without the zero values of card withdrawal and cash credit variables. Age seems to follow a rather uniform distribution over the working age people domain, with lower representation occurring for people above 60 years old. The credit card withdrawal and cash withdrawal variables both exhibits the extremely positive skew, while cash credit variable is though positive skewed but shows less skewness. Furthermore, the low presence of Gold card holders in the data results in sparser distributions among all the continuous variables.

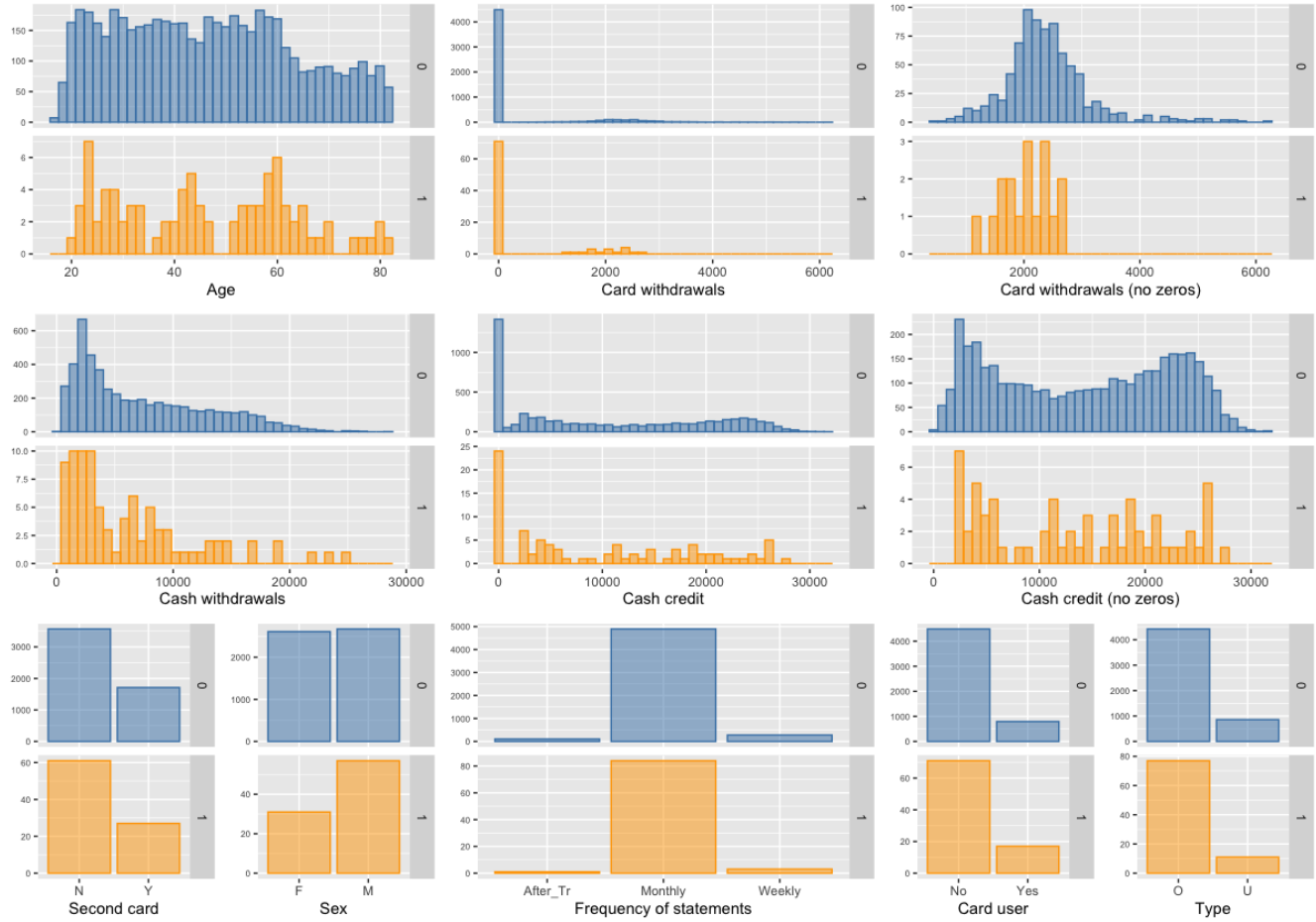


Figure 1: Historgrams

3.1 Initial variable relationship analysis

Since we use logistic regression model to describe the relationship between independent variables and response variables, we employ empirical logit to analyse the potential relationships. In short, empirical logit is the logit transformation of empirical odds. It works by dividing data into equally sized bins, calculating odds in each bin, transforming it by link function $f(x) = \log(x)$ and plotting on a 2D plane.

As the transformation from probability of being issued with gold card to log odds is a monotonic, we can use this plot to reveal the relationship between probability of being issued with gold card and variables.

To analyse the relationship between discrete variables and target, we introduce the ratio plots by plotting target ratio of each level for a certain factor.

$$Ratio\, A = \frac{\#\, Target}{\#\, A}$$

where A is a level in a factor and Target is the cases in A with Y = 1

Ratio plots show the ability of variables to distinguish target to some extent. Its advantage is that it disregards absolute count, which is extremely useful in our case due to the unbalance in factor values.

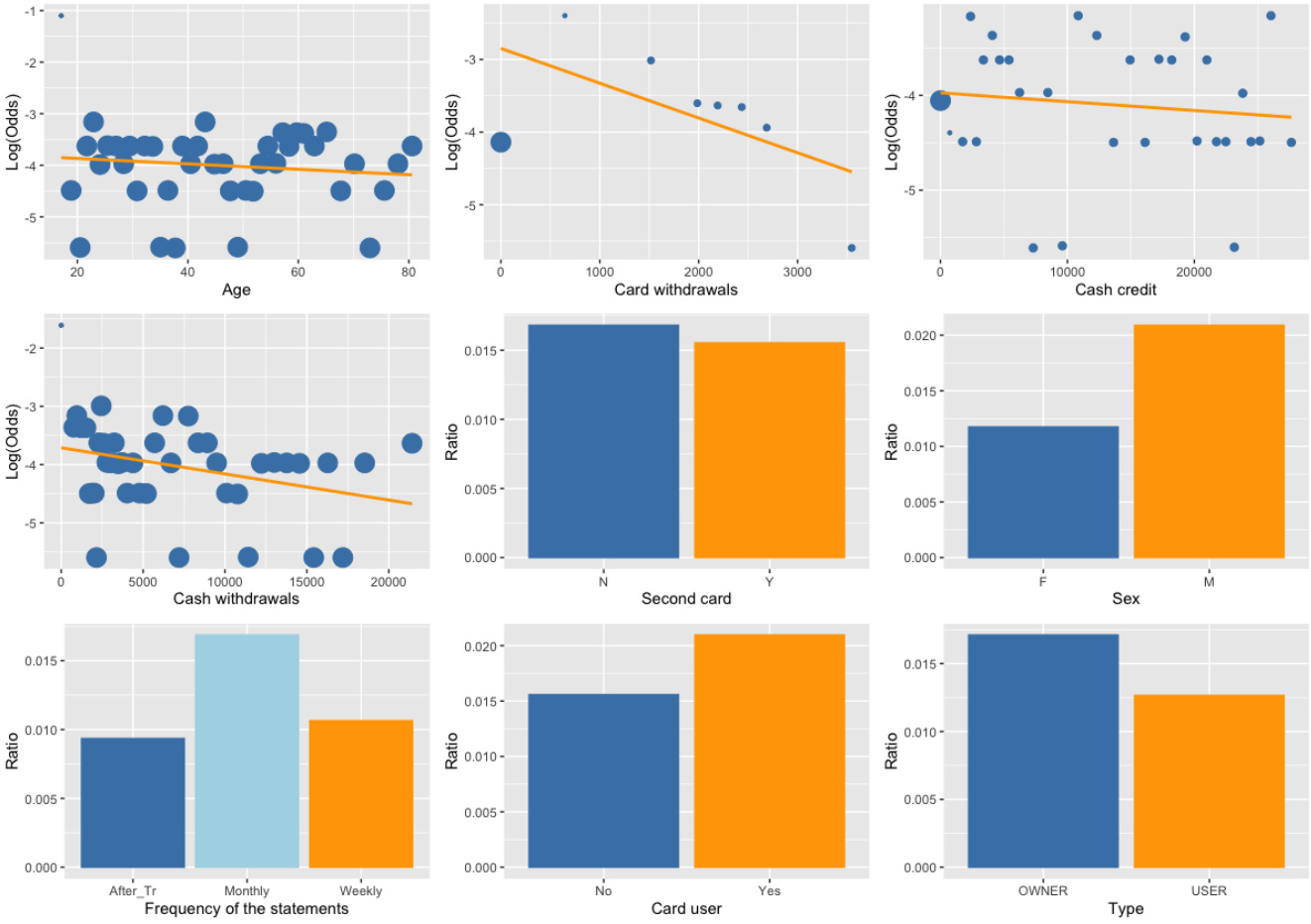


Figure 2: Empirical logit and Ratio bars

Figure 2 shows the empirical logit plots for the continuous variables and ratio plots for the discrete variables. Beginning with the continuous variables, we observe that credit card withdrawals and cash withdrawals seem to both follow a strong negative relationship with log odds, thus a negative relationship with probability of being issued with gold card. The relationship between log odds and the cash credit variable is also negative, yet much weaker and even negligible than that of the credit card withdrawals. Age, seems to show insignificant relationship with log odds by only minor negative slope. Thus, only the cash and card withdrawal variables seems to exhibit potential to be good predictors.

Moving to the discrete variables, it seems that males, owners of the account, those who have used their cards and those who gets statements monthly are more likely to be issued with gold card. Holding

a second card yields almost identical ratio bars suggesting that it might not be an important variable in our analysis.

4 Model Selection Methods

4.1 Data Preparation

The following four models will be trained on the transformed data, where dummy variables and interaction term are created before training steps. The transformation ensured that all variables have only one degree of freedom (hence frequency and interaction variables were affected). This is to easily add penalty to step-wise model so that we can compare the coefficients of four models in the same significance level. In addition, Lasso model requires its input a matrix with transformed variables. Considering above, we decide to use transformed data for all models to ensure consistency.

4.2 Full Model

The full model follows the following form:

$$log\left(\frac{\boldsymbol{p}}{\mathbf{1}-\boldsymbol{p}}\right) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{p} = (p_1, p_2, \ldots, p_n)$, $p_i = P(Y_i = 1)$ is an associated probability of being issued with a Gold card, $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is the matrix of covariates including the intercept and the interaction term, $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of coefficients and $\boldsymbol{\epsilon} \in \mathbb{R}^n$ is a vector of error. The resulting summary of the model fitted is presented in Table 3.

	Estimate	Std. Error	z value	Pr($>\|z\|$)
(Intercept)	-4.1297	1.1233	-3.68	0.0002
age	-0.0078	0.0071	-1.10	0.2700
Card withdrawals	-0.0007	0.0004	-1.69	0.0916
Cash credit	0.0000	0.0000	0.20	0.8399
Cash withdrawals	-0.0000	0.0000	-1.24	0.2150
Second card	0.1307	0.2875	0.45	0.6495
Male	0.5926	0.2260	2.62	0.0087
Monthly statements	0.2629	1.0317	0.25	0.7988
Weekly statements	0.1590	1.1627	0.14	0.8912
Card user	0.3830	1.2499	0.31	0.7593
User	-0.3823	0.3955	-0.97	0.3338
Interaction	0.0365	0.0170	2.14	0.0322
Log Likelihood	−437.900	Null Deviance		898.10
AIC	899.900	Model Deviance		875.85

Table 3: Summary table for full model

The model reduced the deviance by 22.25 points with null deviance being 898.10. This means that the model does explain some amount of deviance in Gold variable. Akaike Information criterion (AIC) yielded a statistic of 899.8. We will take the full model as a benchmark and try to find a model with less complexity but relatively small deviance and AIC .

	LR Chisq	Df	Pr(>Chisq)
Card withdrawals	6.71	1	0.0096
Male	6.71	1	0.0096
Interaction	9.50	1	0.0021
Model Deviance	881.44	AIC	889.44

Table 4: Anova II for the final deviance model

4.3 Deviance analysis using backward elimination method

Starting with the full model, we implement manual model reduction technique by comparing deviance of each model. The starting model acts as an initial benchmark and the less complex model will be accepted if an elimination of one variables does not add deviance significantly. Each time the variable is dropped if it leads to least decrease of deviance when it enters the model after all other variables have already entered the model. Continue the above step until dropping any variable will change the deviance significantly. Since our interaction is transformed, we conduct anova type 2 analysis. The eliminating order is as follows: weekly statements, cash credit, monthly statements, card user, second card holder, user, age, cash withdrawals. The final deviance model is presented in the Table 4 and its summary table is in Table 5

	Estimate	Std. Error	z value	Pr($>\|z\|$)
Intercept	-4.4665	0.1865	-23.95	$< 2e-16$
Card withdrawals	-0.0007	0.0003	-2.43	0.0148
Male	0.5715	0.2253	2.54	0.0118
Interaction	0.0386	0.0116	3.33	0.0009
Log Likelihood	-440.720		Null Deviance	898.10
AIC	889.44		Model Deviance	881.44

Table 5: Summary table for deviance model

4.4 Automated step-wise reduction

	Estimate	Std. Error	z value	Pr($>\|z\|$)
Intercept	-4.4665	0.1865	-23.95	$< 2e-16$
Card withdrawals	-0.0007	0.0003	-2.43	0.0148
Male	0.5715	0.2253	2.54	0.0118
Interaction	0.0386	0.0116	3.33	0.0009
Log Likelihood	-440.720		Null Deviance	898.10
AIC	889.44		Model Deviance	881.44

Table 6: Summary table for step-wise model

Step-wise method automates the analysis conducted in the section above, yet it takes AIC as a criterion of significance. Nevertheless, using AIC for single-variable-at-a-time stepwise selection is (at least asymptotically) equivalent to stepwise selection using a cut-off for p-values of about 15.7%. To be consistent with previous 2 models, we add the penalty for a 5% significance level so that models can be

compared in the same significance level. As we have transformed all the variables to have 1 degree of freedom, we can guarantee that the χ^2 test yield the correct values. We implement step-wise selection based on both: forward and backward methods and the final result - the suggested model - is reported in the Table 6.

4.5 LASSO

LASSO performs variable selection by maximizing the log likelihood subject to the sum of the absolute value of the coefficients being less than a positive constant λ. The maximum likelihood estimator is defined as follows:

$$\boldsymbol{\beta}_{LASSO} = \underset{\beta}{\operatorname{argmax}} \sum_{i=1}^{N} y_i log(p_i) + (1 - y_i) log(1 - p_i) - \lambda \sum_{j=1}^{p} |\beta_j|$$

where p_i is an associated probability $P(Y_i = 1)$ and p is the total number of variables.

The optimal λ can be obtained by k-fold cross validation. In k-fold cross validation, data are split into k folds and 1 fold is selected for validation while the rest folds are used to train model. The above procedure is repeated until we get an evaluation metric (here we use binomial deviance) on each fold, and the final performance of the model is given by the average binomial deviance on k folds. By this method, we can compare the model performance with different λ on one dataset and choose the one with least binomial deviance.

In order to reproduce the result, we set the seed to 200 and find that the λ that minimises the binomial deviance is equal to 0.001363988, leaving 4 variables and 1 intercept. Another potential optimal λ is the one that is 1 standard error from minimum λ, but it shrinks all coefficients to zero, so we will not consider it.

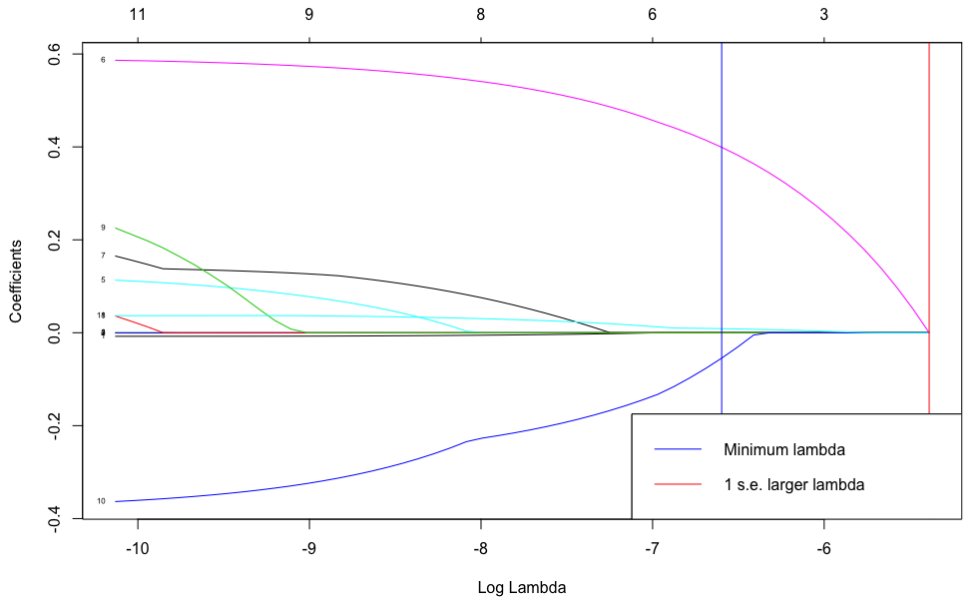


Figure 3: LASSO trace plot

A obvious drawback of LASSO in our case is that the results are heavily dependent on the way of splitting folds, which suggests that the model is unstable. This is not surprising if taking into account the extremely small target rate. By splitting the data into 10 sub-samples, there is a high chance that some of those samples will have two or three more/less counts of positive Gold responses, which is a small value

but high proportion of total positive responses in the sample. Consequently, the reliability and suitability of the LASSO model for this dataset is weak. Nevertheless, the coefficients estimated using the seed of 200 as well as the trace map are reported in the Table 7 and Figure 3.

	Estimate
Intercept	-4.19513047468
User	-0.05468243586
Cash withdrawals	-0.00002496219
Male	0.39914800487
Interaction	0.00873877750

Table 7: LASSO output for minimum λ

5 Selected Model Analysis

Step-wise and deviance model are identical. Considering the fact that step-wise is in both direction, more automated and easy to implement, we are inclined to step-wise model when comparing these two models. When it comes to the comparison of full model, step-wise model and LASSO, step-wise model is still the optimal, the reasons are as follows i) lowest AIC, ii) parsimony and iii) stability. Whilst neither of the models showed very significant improvement from the full-model, step-wise model managed to lower AIC, yet not deviance. Together with the fact that, step-wise model is more parsimonious, it dominated the full model. Finally, even though LASSO arguably deals better with feature selection, it proved to be unstable - coefficients would change substantially depending on how folds are split. Thus, the step-wise/deviance model is chosen over LASSO.

5.1 Interpretation

The final model is described by the following formula.

$$log\left(\frac{\boldsymbol{p}}{\boldsymbol{1-p}}\right) = -4.4665 - 0.0007 * \boldsymbol{mcardwdl} + 0.5715 * \boldsymbol{sex\_m} + 0.0386 * \boldsymbol{Interaction} + \boldsymbol{\epsilon}$$

where $\boldsymbol{p} = (p_1, p_2, \ldots, p_n)$, $p_i = P(Y_i = 1)$ is an associated probability of being issued with a Gold card, $\boldsymbol{mcardwdl}$, $\boldsymbol{sex\_m}$ and $\boldsymbol{Interaction}$ are variable vectors, $\boldsymbol{\epsilon}$ is an error vector term.

To interpret the model, consider a logistic regression with 3 variables, the fitted value of the linear predicotr at a particular value of $\boldsymbol{x_i} = (x_{i1}, x_{i_2}, x_{i3})$ is

$$\hat{\eta}(x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_1 x_{i2} + \hat{\beta}_1 x_{i3}$$

Lets say we increase x_{i1} by one unit while keep other variables the same, the fitted value at $x_i^{'}$ is

$$\hat{\eta}(x_i^{'}) = \hat{\beta}_0 + \hat{\beta}_1 (x_{i1} + 1) + \hat{\beta}_1 x_{i2} + \hat{\beta}_1 x_{i3}$$

Notice the $\hat{\eta}(x_i)$ is just the log-odds, therefore the difference in two fitted values is

$$\begin{aligned} \hat{\eta}(x_i^{'}) - \hat{\eta}(x_i) &= \ln(odds_{x_i+1}) - \ln(odds_{x_i}) \\ &= \ln(\frac{odds_{x_i+1}}{odds_{x_i}}) \\ &= \hat{\beta}_1 \end{aligned}$$

By taking antilogs, we can get odds ratio

$$\hat{O}_R = \frac{odds_{x_i+1}}{odds_{x_i}} = e^{\hat{\beta}_1}$$

The odds ratio can be interpreted as the estimated increase in the probability of success associated with a one - unit change in the value of the predictor variable. In our case, if the average credit card withdrawals decrease 1000 Kč, the odds ratio for mcardwdl is $\hat{O}_R = e^{\hat{\beta}_1} = e^{0.7} = 0.4966$. This implies that every additional 1000 Kč credit card withdrawals decreases the odds of being issued with gold card by about 50%. Similarily, male has 77% higher odds of being issued with gold card compared to female, and if a person uses credit card, he/she will get extra 4% of odds of being issued with gold card each year.

5.2 Prediction

Given the characteristics provided, a person who is aged 42, has a mean card withdrawal of 500 Kč, is Female, has used their credit card, his/her associated probability of getting the Gold card is 4.00%. To turn this probability to a label we need to set a cutoff where probability greater than this cutoff will be labelled as 1 and probability less than it will be labelled as 0. This cutoff can be changed according to the balance of the cost of false negative cases and false positive cases, and we will give further discussion in the next section.

6 Further Model validation

We choose to examine further the model obtained with the automated step-wise selection. We validate our model with two plots, added-variable plot and precision-recall plot.

6.1 Added-value plot

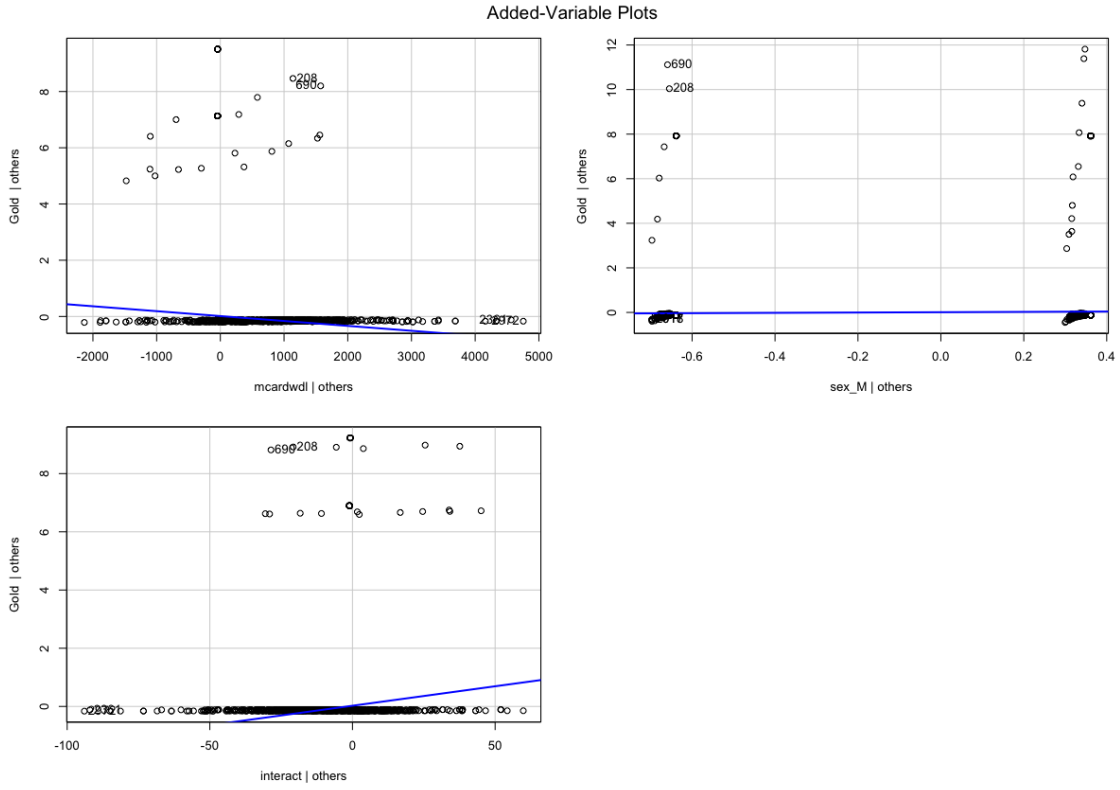


Figure 4: Added variable plots

The added-variable plot in logistic regression is an extension of that in linear regression. In logistic regression, we can get the added-variable plot on variable x_j by refitting the model with x_j removed and extracting the residuals from this fit. Then x_j is regressed on the other x values except x_j. Finally, the two sets of residuals are plotted against each other. The significance testing for the coefficient $\beta_j = 0$ under the original logistic regression and the *slope* $= 0$ in a fitted linear model in add-value plot is the computational equivalent [1]. Based on this property, we can easily find influential points or outlier in the data.

In our case (Figure 4), a obvious deviation of fitted line and points is observed in mcardwdl and interaction plots. In addition, a group of outliers are above the line in three plots. These points show a completely different pattern from the majority of the points, and have strong affect on slope of the fitted line, which means they will also affect the testing accuracy of coefficients in the original logistic regression. Due to the large number of outliers, we may consider doing a careful checking on the data or trying other nonlinear classification model.

6.2 Precision-Recall Plot

In our case, we are curious about how meaningful is a positive result from logistic regression given the small target distribution (only 1.64% is Gold card holders in the dataset). To answer it, we introduce two metrics: precision and recall. Under a certain threshold, where record with predicted probability larger than this threshold will be predicted as 1, precision is defined as:

$$Precision = \frac{\#True\,Positive}{\#True\,Positive + \#False\,Positive}$$

It represents the share of actually positive records out of all those which were *predicted* to be positive. Another metric - Recall - is defined as:

$$Recall = \frac{\#True\,Positive}{\#True\,Positive + \#False\,Negative}$$

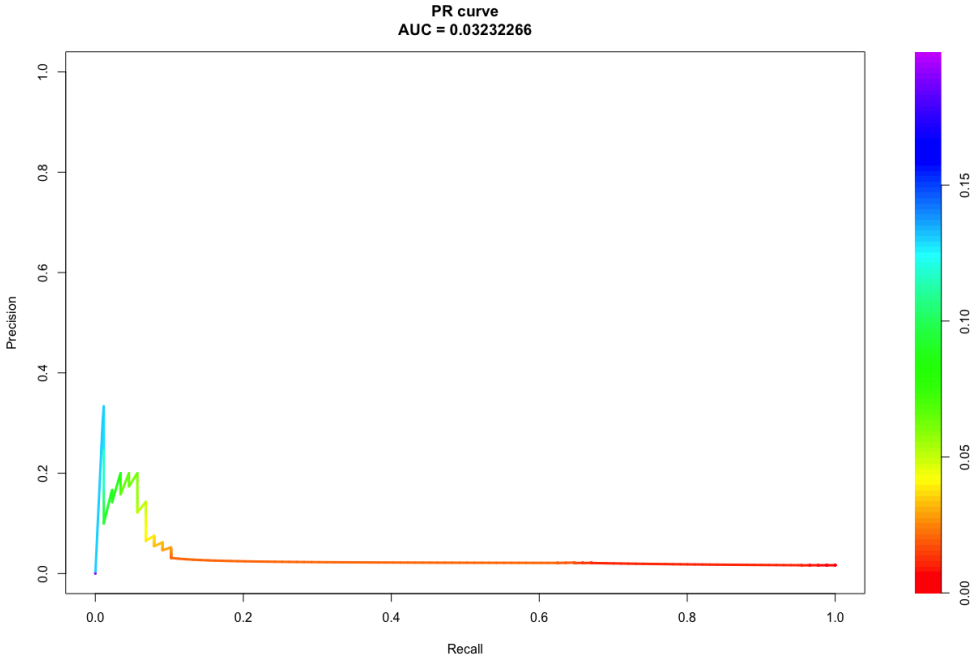


Figure 5: Precision-Recall plot

It shows the fraction of records that are correctly labelled as positive, out of all the records that actually are positive. Ideally, a good model should have both high precision and high recall, but in practice, a trade-off between them is always needed, and a precision-recall plot can help us intuitively know the model performance and find the balance point.

Figure 5 plots precision against recall of step-wise model under different threshold, the value of threshold is indicated by colour and is showed in the legend. AUC (area under curve) is calculated to measure the overall goodness of the model (larger AUC intuitively indicates that the curve is closer to the upper-right corner which is the perfect operating point).

In this precision-recall plot we find that with threshold decreasing from 0.20 to about 0.13, precision increases from 0 to about 0.35 while recall is almost 0. As the threshold keeps dropping, the decrease of precision is accompanied by the increase in the value of recall. When threshold is less than 0.05, the precision is almost 0 whereas recall keeps rising as the threshold tends to 0. The plot indicates that our model does not capture useful pattern of the gold card customers and the predicted probabilities of target and non-target are quite small.

7 Further analysis

Several things deserve exploration:

- Regenerate the data. Current data only contains variables that have small relation with the target both in mathematics and business. We would like to include variables that are more relevant to bringing revenue to the card issuer and thus relevant to issuing an gold card. Variables may contain but not limit to transaction amount by credit card, frequency and plan of instalments, etc.
- Split data into training set and validation the set, use under-sampling/over-sampling to increase the target rate in training set, then evaluate model performance on the validation set.
- As the simple logistic regression essentially uses a linear classification boundary to separate target from non-target, we may consider doing transformation on variables to include some non-linear pattern in the model. Or we may try other tree-structure classification models such as a decision tree.
- If the interpretation of the model is in the highest priority, we can try ensemble methods (e.g. Boosting) which combines several weak classifier into a strong classifier to achieve better model performance.

8 Conclusion

The model suggested by step-wise and deviance model was chosen due to its parsimony, lower AIC and stability. However further analysis revealed that our model performs poorly as a predictive model and has an abundance of outliers. A better data source with more relevant variables and a larger target rate is needed to construct better predictive models. Furthermore, other classification models than the logistic regression can be implemented in order to fit non-linear behaviour of the relationships.

References

[1] PC Wang. Adding a variable in generalized linear models. *Technometrics*, 27(3):273–276, 1985.

A Appendix: R code

```
##################################
#  Check and install packages  #
##################################
library(car)
library(carData)
library(MASS)
library(leaps)
library(glmnet)
library(PRROC)
library(ggplot2)
library(GGally)
library(gridExtra)
library(plyr)


#####################################
# Exploratory analysis of the data #
#####################################


summary(czechgold)
# gold to factor
# string to dummies

gold <- czechgold
gold$Gold <- factor(gold$Gold)

gold$sex <- as.factor(gold$sex)
gold$second <- as.factor(gold$second)
gold$frequency <- as.factor(gold$frequency)
gold$type <- as.factor(gold$type)
gold$carduse <- as.factor(gold$carduse)
summary(gold)
str(gold)


# Plot1
p1 <- ggplot(gold, aes(x=age,color=Gold,fill=Gold)) +
  geom_histogram(bins = 40, alpha=.5, position = 'identity')+
  scale_color_manual(values=c("steelblue", "orange"))+
  labs(x = "Age") +
  scale_fill_manual(values=c("steelblue", "orange"))+
  facet_grid(Gold ~ .,scales="free_y") +
  theme(legend.position='none',
        axis.text.y= element_text(size=6),
        axis.title.y = element_blank())

p2 <- ggplot(gold, aes(x=mcardwdl,color=Gold,fill=Gold)) +
```

```
  geom_histogram(bins = 40, alpha=.5, position = 'identity')+
  scale_color_manual(values=c("steelblue", "orange"))+
  scale_fill_manual(values=c("steelblue", "orange"))+
  labs(x = "Card withdrawals") +
  facet_grid(Gold ~ .,scales="free_y")+
  theme(legend.position='none',
        axis.text.y= element_text(size=6),
        axis.title.y = element_blank())

p2.1 <- ggplot(gold[gold$mcardwdl>0,], aes(x=mcardwdl,color=Gold,fill=Gold)) +
  geom_histogram(bins = 40, alpha=.5, position = 'identity')+
  scale_color_manual(values=c("steelblue", "orange"))+
  scale_fill_manual(values=c("steelblue", "orange"))+
  labs(x = "Card withdrawals (no zeros)") +
  facet_grid(Gold ~ .,scales="free_y")+
  theme(legend.position='none',
        axis.text.y= element_text(size=6),
        axis.title.y = element_blank())

p2.all <- grid.arrange(p2,p2.1,nrow = 1)

p3 <-ggplot(gold, aes(x=mcashcr,color=Gold,fill=Gold)) +
  geom_histogram(bins = 40, alpha=.5, position = 'identity')+
  scale_color_manual(values=c("steelblue", "orange"))+
  scale_fill_manual(values=c("steelblue", "orange"))+
  facet_grid(Gold ~ .,scales="free_y")+
  theme(legend.position='none',
        axis.text.y= element_text(size=6),
        axis.title.y = element_blank()) +
  labs(x = "Cash credit")

p3.1 <-ggplot(gold[gold$mcashcr>0,], aes(x=mcashcr,color=Gold,fill=Gold)) +
  geom_histogram(bins = 40, alpha=.5, position = 'identity')+
  scale_color_manual(values=c("steelblue", "orange"))+
  scale_fill_manual(values=c("steelblue", "orange"))+
  facet_grid(Gold ~ .,scales="free_y")+
  theme(legend.position='none',
        axis.text.y= element_text(size=6),
        axis.title.y = element_blank())+
  xlab("Cash credit (no zeros)")

p3.all <- grid.arrange(p3,p3.1,nrow = 1)

p4 <-ggplot(gold, aes(x=mcashwd,color=Gold,fill=Gold)) +
  geom_histogram(bins = 40, alpha=.5, position = 'identity')+
  scale_color_manual(values=c("steelblue", "orange"))+
  scale_fill_manual(values=c("steelblue", "orange"))+
  facet_grid(Gold ~ .,scales="free_y")+
  theme(legend.position='none',
        axis.text.y= element_text(size=6),
```

```
        axis.title.y = element_blank())+
  xlab("Cash withdrawals")

p5 <-ggplot(gold, aes(x=second)) +
  geom_bar(alpha=.5, position = 'identity',
           color=c("steelblue","steelblue","orange","orange"),
           fill=c("steelblue","steelblue","orange","orange"))+
  facet_grid(Gold ~ .,scales="free_y")+
  theme(legend.position='none',
        axis.text.y= element_text(size=6),
        axis.title.y = element_blank())+
  xlab("Second card")

p6 <-ggplot(gold, aes(x=sex)) +
  geom_bar(alpha=.5, position = 'identity',
           color=c("steelblue","steelblue","orange","orange"),
           fill=c("steelblue","steelblue","orange","orange"))+
  facet_grid(Gold ~ .,scales="free_y")+
  theme(legend.position='none',
        axis.text.y= element_text(size=6),
        axis.title.y = element_blank())+
  xlab("Sex")

p7 <-ggplot(gold, aes(x=frequency)) +
  geom_bar(alpha=.5, position = 'identity',
           color=c("steelblue","steelblue","steelblue",
           "orange","orange","orange"),
           fill=c("steelblue","steelblue","steelblue",
           "orange","orange","orange"))+
  facet_grid(Gold ~ .,scales="free_y")+
  theme(legend.position='none',
        axis.text.y= element_text(size=6),
        axis.title.y = element_blank())+
  xlab("Frequency of statements")

p8 <-ggplot(gold, aes(x=carduse)) +
  geom_bar(alpha=.5, position = 'identity',
           color=c("steelblue","steelblue","orange","orange"),
           fill=c("steelblue","steelblue","orange","orange"))+
  facet_grid(Gold ~ .,scales="free_y")+
  theme(legend.position='none',
        axis.text.y= element_text(size=6),
        axis.title.y = element_blank())+
  xlab("Card user")

p9 <-ggplot(gold, aes(x=type)) +
  geom_bar(alpha=.5, position = 'identity',
           color=c("steelblue","steelblue","orange","orange"),
           fill=c("steelblue","steelblue","orange","orange"))+
  facet_grid(Gold ~ .,scales="free_y")+
```

```
  theme(legend.position='none',
        axis.text.y= element_text(size=6),
        axis.title.y = element_blank())+
  scale_x_discrete(labels=c("OWNER" = "O", "USER" = "U"))+
  xlab("Type")


lay <- rbind(c(1,1,2,2,2,2),
             c(3,3,4,4,4,4),
             c(5,6,7,7,8,9))
grid.arrange(p1,p2.all,p4,p3.all,p5,p6,p7,p8,p9,layout_matrix = lay)

#plot 2

attach(czechgold)
myemplogit <- function(yvar=y,xvar=x,maxbins=30,sc=1, xlabel){
  yvar.fac <- as.factor(yvar)
  breaks  <- unique(quantile(xvar, probs=0:maxbins/maxbins))
  levs  <- (cut(xvar, breaks, include.lowest=FALSE))
  num <- as.numeric(levs)
  c.tab <- do.call(data.frame,aggregate(yvar~addNA(levs)
  ,na.action=na.pass,FUN=function(x)
    c(length(x),sum(x))))
  mean.tab <- aggregate(xvar~addNA(levs),na.action=na.pass,FUN=mean)
  c.tab <- cbind(c.tab,mean.tab[,2])
  colnames(c.tab) <- c("levs","n","y","mean")
  # c.tab <<- count(num, 'levs')
  c.tab$levs <- factor(c.tab$levs, levels = c.tab$levs, labels =
                         c(levels(c.tab$levs)[-length(c.tab$levs)],
                         paste("[",min(xvar),"]",sep="")),
                       exclude = NULL)
  c.tab <- c.tab[c(nrow(c.tab),1:nrow(c.tab)-1),]

  c.tab$odds <-  (c.tab$y+0.5)/(c.tab$n - c.tab$y + 0.5)
  c.tab$logit <- log(c.tab$odds)

  ggplot(c.tab, aes(x=mean, y=logit)) +
    geom_point(aes(size=n),color="steelblue")+
    geom_smooth(method="lm",se=F,color="orange")+
    xlab(deparse(substitute(xvar))) + ylab("Log(Odds)")+xlab(xlabel)+
    theme(legend.position='none')
}

p1 <- myemplogit(Gold,age,maxbins=40, xlabel= "Age")
p2 <- myemplogit(Gold,mcardwdl,maxbins=40, xlabel= "Card withdrawals")
p3 <- myemplogit(Gold,mcashcr,maxbins=40, xlabel= "Cash credit")
p4 <- myemplogit(Gold,mcashwd,maxbins=40, xlabel= "Cash withdrawals")

Ratio.plot <- function(variable, var.lab){
  formula <- paste("Gold ~ ",variable,sep = "")
```

```
  df <- aggregate(as.formula(formula),data=czechgold,FUN=mean)
  colnames(df)[2] <- "Ratio"
  if (variable!="frequency"){
    return(ggplot(df, aes_string(x=variable, y="Ratio")) +  xlab(var.lab) +
             geom_bar(stat="identity", color=c("steelblue","orange"),
                      fill=c("steelblue","orange"))) + xlab(var.lab)
  } else {
    return(ggplot(df, aes_string(x=variable, y="Ratio")) + xlab(var.lab) +
             geom_bar(stat="identity", color=c("steelblue","lightblue","orange"),
                      fill=c("steelblue","lightblue","orange")))
  }
}

col.fac <- c("second","sex","frequency","carduse","type")
lab.fac <- c("Second card", "Sex", "Frequency of the statements",
"Card user", "Type")
for (i in 1:length(col.fac)){
  assign(paste("p",as.character(i+4),sep=""), Ratio.plot(variable = col.fac[i],
  var.lab = lab.fac[i]))
}

grid.arrange(p1,p2,p3,p4,p5,p6,p7,p8,p9,nrow = 3)

detach(czechgold)



##turn off scientific display of numbers
options(scipen = 999)
options(digits= 7)


#################################
## Full model plus age:carduse #
#################################

#transform Data
get_dummy <- function(df, col){
  col.fac <- factor(df[[col]])
  dummy <- model.matrix(~col.fac+0)
  colnames(dummy) <- gsub("col.fac",paste(col,"_",sep=""),
  colnames(dummy),fixed = TRUE)
  df <- cbind(df, dummy)
  return(df)
}

czechgold1 <- czechgold
for (i in 6:10){
  cols <- colnames(czechgold1)
  czechgold1 <- get_dummy(czechgold1, cols[i])
```

```
}
czechgold1$interact <- czechgold1$age * czechgold1$carduse_Yes

col.slt <- colnames(czechgold1)[-c(6:10,11,13,15,18,20)]
czechgold2 <- czechgold1[,col.slt]
#using transformed data in all model

full.fit <- glm(Gold~., family = binomial, data = czechgold2)
summary(full.fit)

########################
# 2b                   #
# Analysis of deviance #
########################
a <- glm(Gold~., family = binomial, data = czechgold2)
Anova(a, type = 2)
# /Weekly
b_min1 <- glm(Gold~.-frequency_Weekly, family = binomial, data = czechgold2)
Anova(b_min1, type = 2)
# /mcashcr
b_min2 <- glm(Gold~.-frequency_Weekly-mcashcr, family = binomial,
data = czechgold2)
Anova(b_min2, type = 2)
# /carduse
b_min3 <- glm(Gold~.-frequency_Weekly-mcashcr-carduse_Yes, family = binomial,
data = czechgold2)
Anova(b_min3, type = 2)
# /Monthly
b_min4 <- glm(Gold~.-frequency_Weekly-mcashcr-frequency_Monthly-carduse_Yes,
family = binomial, data = czechgold2)
Anova(b_min4, type = 2)
# /second_Yes
b_min5 <- glm(Gold~.-frequency_Weekly-mcashcr-frequency_Monthly-carduse_Yes
-second_Y, family = binomial, data = czechgold2)
Anova(b_min5, type = 2)
# /type_USER
b_min6 <- glm(Gold~.-frequency_Weekly-mcashcr-frequency_Monthly-carduse_Yes
-second_Y-type_USER, family = binomial, data = czechgold2)
Anova(b_min6, type = 2)
# /age
b_min7 <- glm(Gold~.-age-frequency_Weekly-mcashcr-frequency_Monthly-carduse_Yes
-second_Y-type_USER, family = binomial, data = czechgold2)
Anova(b_min7, type = 2)
# /mcashwd
b_min8 <- glm(Gold~.-age-mcashwd-frequency_Weekly-mcashcr-frequency_Monthly
-carduse_Yes-second_Y-type_USER, family = binomial, data = czechgold2)
Anova(b_min8, type = 2)

summary(b_min8)
# STOP - b_min8 is the chosen model.
```

```
####################
# 2c               #
# Step-wise method #
####################

k.quantile <- qchisq(0.05, 1, lower.tail = FALSE)
# k=3.814146

stepwise <- step(a, direction = "both", test = "Chisq", k = k.quantile )
# Identical to the deviance model.

#########
# 2d    #
# LASSO #
#########

x <- czechgold2[,c(2:12)]
x <- as.matrix(x)

y <- czechgold2[,1]
y <- ifelse(y==1,1,0)

# model unstable - seed required
set.seed(200)

# Fit lasso regression
fit1 <- glmnet(x,y, family="binomial", alpha=1)
par(mfrow=c(1,1))
#Crude plot trace
plot(fit1)
# plot of log(lambda) and label traces
plot(fit1,"lambda",label = T)
# all coefficients shrank to zero

# CROSS VALIDATION

cvfit1 = cv.glmnet(x,y, family="binomial",alpha=1)
# Plot this
plot(cvfit1)
#  find optimum
cvfit1$lambda.min
# 0.001363988
cvfit1$lambda.1se
# 0.004571546

coef(cvfit1, s = "lambda.min")
# mcashwd, sex, type, interaction remain
coef(cvfit1, s = "lambda.1se")
```

```
# all zeros

plot(fit1,"lambda",label = T)
abline(v=log(cvfit1$lambda.1se),col="red")
abline(v=log(cvfit1$lambda.min),col="blue")
legend("bottomright",legend=c("Minimum lambda", "1 s.e. larger lambda")
,lty=c(1,1),col=c("blue","red"), ins=0.00)

###################################
# Interpretation and predictions #
###################################

###### Predictions

betas <- coefficients(stepwise)
betas <- as.matrix(betas)
values <- c( 1, 500, 0, 42)
values <- as.matrix(values)

logistic <- function(X, beta) {
  p <- exp(X %*% beta)/(1+exp(X%*%beta))
  p
}

logistic(X=t(values), beta = betas)
# 4 %

means <- c(1, 359.3, 1, 46.28)
credit <- c(1, 1359.3, 1, 46.28)
female <- c(1, 359.3, 0, 46.28)
age <- c(1, 359.3, 1, 56.28)

# change for taking 1000 more card withdrawls
logistic(X=t(means), beta = betas) - logistic(X=t(credit), beta = betas)
# change for being female
logistic(X=t(means), beta = betas) - logistic(X=t(female), beta = betas)
# change for being 10 years older given being a card user.
logistic(X=t(means), beta = betas) - logistic(X=t(age), beta = betas)


###################################
# Further Validation Plots #
###################################


avPlots(stepwise)

y.pred <- predict(stepwise,type="response")

pr <- pr.curve(scores.class0 = y.pred, weights.class0 = y.train, curve=T)
```

```
plot(pr)

###################################
# EXTRA: data without duplicates #
###################################

#################################################
##for accounts which have users                 #
##if user has gold card||owner has gold card    #
##set both user and owener has gold card(gold=1)#
##delete data of user                           #
##delete 'type'                                 #
#################################################
data1 = czechgold[,-10]
repeat_data = data1[duplicated(data1[,-1]),]
user_gold = repeat_data[(repeat_data$Gold == 1),]
data1[as.numeric(row.names(user_gold))-1,]$Gold = 1
deleted_data = data1[!duplicated(data1[,-1]),]

deleted_data$Gold <- factor(deleted_data$Gold)
deleted_data$sex <- as.factor(deleted_data$sex)
deleted_data$second <- as.factor(deleted_data$second)
deleted_data$frequency <- as.factor(deleted_data$frequency)
deleted_data$carduse <- as.factor(deleted_data$carduse)


deleted_data$Af_tr <- ifelse(deleted_data$frequency=="After_Tr", 1,0)
deleted_data$Weekly <- ifelse(deleted_data$frequency=="Weekly", 1,0)
deleted_data$Monthly <- ifelse(deleted_data$frequency=="Monthly", 1,0)
deleted_data$interaction <- ifelse(deleted_data$carduse=="Yes",
deleted_data$age, 0)
fit.deleted <- glm(Gold~.-Af_tr, family = binomial,data=deleted_data)
fit.step<- step(fit.deleted, direction = "both", test = "Chisq",k=3.841)
y_DEL <- deleted_data[,1]
y_DEL <- ifelse(y_DEL==1,1,0)
summary(fit.step)


coefs = summary(fit.step)$coef[,1]
customer <- c(1,500,0,0,42)
 p = customer*coefs
 p = sum(customer*coefs)
 p

exp(p) / (1 + exp(p))
##0.04054554
 y.pred_DEL <- predict(fit.step,type="response")
 roc_DEL <- roc.curve(scores.class0 = y.pred_DEL, weights.class0 = y_DEL,
 curve=T)
 plot(roc_DEL)
```