Charles University in Prague
Faculty of Social Sciences
Institute of Economic Studies

Home assignment 4

Marek Chadim, Jakub Strašlipka

Econometrics II – JEB110
Winter semester 2022/2023

Problem 1

We seek to find the maximum likelihood estimates of β_0, β_1 for the simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \ .$$

Assuming normality of the error term, $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$, the likelihood function is a density of a normal distribution with the following form

$$\begin{aligned} p(y|\beta_0, \beta_1, \sigma^2) &= \prod_{i=1}^{n} p(y_i|\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \mathrm{N}(y_i; \beta_0 + \beta_1 x_i, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left[-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right] \\ &= \frac{1}{\sqrt{(2\pi\sigma^2)^n}} \cdot \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right] \ . \end{aligned}$$

Then the log-likelihood function

$$\mathcal{L}(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \ .$$

First, we will find the MLE estimator of β_0 so that we can use it in β_1 estimation. We take a derivative of the log-likelihood function with respect to β_0,

$$\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\beta_0} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) \ .$$

Set it equal to zero and solve for β_0

$$\begin{aligned} \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) &= 0 \\ \beta_0 &= \frac{1}{n}\sum_{i=1}^{n} y_i - \beta_1 \frac{1}{n}\sum_{i=1}^{n} x_i \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \ , \end{aligned}$$

and we arrive to the same result as in the traditional approach.

We follow the same steps for β_1. Take a derivative of the log-likelihood function with respect to β_1,

$$\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\beta_1} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i y_i - \beta_0 x_i - \beta_1 x_i^2) \ .$$

Set it equal to zero and solve for β_1 (we plug in the formula for β_0 obtained above),

$$\begin{aligned} \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i y_i - \beta_0 x_i - \beta_1 x_i^2) &= 0 \\ \sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 &= 0 \end{aligned}$$

$$\sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i + \beta_1 \bar{x}\sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})} \ ,$$

and again, we get the well-known formula.

Problem 2

Data

We use the provided data of 10,127 customers as a starting point. Information on customer loyalty is missing for one customer – because loyalty serves as an explanatory variable in our model, we remove this customer from our dataset. Furthermore, an outlier analysis is conducted. After inspecting distributions of variables using descriptive statistics and plots, we identify two variables for which we eliminate observations due to outliers. They are *Total_Amt_Chng_Q4_Q1* and *Total_Ct_Chng_Q4_Q1*, and we remove all observations that have z-scores below -3 or above 3. This way, we eliminate another 248 observations, so we eventually work with a data sample of 9,878 customers. We also detect the unknown date in case of qualitative variables, further in the text we argue for not eliminating these observations as none of the qualitative variables will be included in the regression. Furthermore, we change and gender to a binary variable to enable numerical computations.

We scan the data and perform a simple analysis to see, if we can build an initial model. We use only simple averages of all variables and compare them between the attrited customers and the whole sample. Variables that we consider for a regression are those, for which the ratio $|\frac{mean_{attr} - mean_{all}}{mean_{all}}| > 0.25$. We identify the following possible regression:

$$\begin{aligned} Loyalty_i = \beta_0 + \beta_1 \cdot Total\_Revolving\_Bal_i + \beta_2 \cdot Total\_Trans\_Amt + \\ + \beta_3 \cdot Total\_Trans\_Ct_i \ + \beta_4 \cdot Avg\_Utilization\_Ratio_i + \varepsilon_i \ . \end{aligned} \tag{1}$$

Exploratory analysis and model description

Now, we expand our exploratory analysis by building a correlation matrix for numerical variables and creating bar charts for qualitative variables, the latter for a) all customers and b) only attrited customers.

Comparison of the plots of four qualitative variables (education, income level, marital status, card category; see Appendix, Figures 2–5) leads us to a conclusion that the are no significant structural differences between the attrited and non-attrited customers in terms of these qualities, thus we do not include any of these variables into our regression. This analysis also enables us not to remove observations with "unknown" as an answer as we argue that the qualitative variables have no or negligible impact on customer attrition.

Focusing on the correlation matrix (Appendix, Figure 1), our impression is that all variables in equation (1) may explain customer loyalty (column 3, row 4 in the matrix) as they all have reasonably high correlation coefficient (in absolute terms). We furthermore identify an additional variable that will be included in our regression,

$Total\_Ct\_Chng\_Q4\_Q1$, that has correlation coefficient with attrition of 0.32 (in absolute value).

Since we currently consider five explanatory variables out of twelve possibilities, in order not to make our model overspecified, it is decide not to add more variables (all of which have $|Corr| \leq 0.15$) Hence, our model is

$$\begin{aligned} Loyalty_i = \beta_0 + \beta_1 \cdot Total\_Revolving\_Bal_i + \beta_2 \cdot Total\_Trans\_Amt + \\ + \beta_3 \cdot Total\_Trans\_Ct_i + \beta_4 \cdot Avg\_Utilization\_Ratio_i + \\ + \beta_5 \cdot Total\_Ct\_Chng\_Q4\_Q1 + \varepsilon_i \ . \end{aligned} \tag{2}$$

Regression results and coefficient discussion

We estimate (2) by the logit and probit models.

	Logit	Probit
Intercept	-5.392***	-2.958***
	(0.161)	(0.087)
$Total\_Revolving\_Bal$	0.00106***	0.000533***
	(0.0000598)	(0.0000318)
$Total\_Trans\_Amt$	-0.000472***	-0.000264***
	(0.0000200)	(0.0000107)
$Total\_Trans\_Ct$	0.104***	0.0589***
	(0.00323)	(0.00170)
$Avg\_Utilization\_Ratio$	-0.250	-0.0805
	(0.188)	(0.100)
$Total\_Ct\_Chng\_Q4\_Q1$	3.059***	1.651***
	(0.1783)	(0.0965)
Observations	9,878	9,878
McFadden pR^2	0.3660	0.3664

Note: *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

Table 1: Regression Results

First, notice that coefficients in both models imply the same direction of effects on customer loyalty. The three positive coefficients were anticipated.

β_1: It seems natural that high revolving balance increases customer loyalty as more money provided likely has positive impact on customers' perception of the bank.

β_3: The more often a customer pays with the bank's credit card, the less likely their attrition is. This also is in line with expectation.

β_5: If a customer starts using the bank's credit card more frequently, it implies higher loyalty, which can be expected.

On the other hand, the two negative coefficients are rather surprising.

β_2: The more a customer spends on their credit card, the more likely their attrition is. Although we expected a positive effect (similar reasoning as for β_1, β_3 above), it might be that high total spendings lead to high payments back to the bank which has a detrimental effect on the bank's reputation from the customer's perspective – "bank appears to be greedy" (while it is due to customer's spending). Moreover, when the customer does not repay his debt (which is more likely when they owe more money), the bank typically initiates a legal process against the customer that has negative effect on the relationship of the two counterparties.

β_4: Although statistically insignificant, the regression results indicate negative effect of credit card utilization ratio on loyalty. Since credit card limits are set by the bank to ensure customer's debt repayment, approaching these limits by customers may endanger their repayment ability, thus leading to trouble and a deterioration of the bank's reputation in their view (argumentation as for β_2).

Both the logit and probit regressions have a reasonably high pseudo R-squared at 0.37 with the probit model being a marginally better fit.

Marginal effects

Table 2 in Appendix presents values of the average partial effects (APE) and the partial effect at average (PEA) for both the logit and probit regressions. The effects do not vastly differ between the two models (apart from the statistically insignificant utilization ratio), therefore, accounting also for the same R-squared, we conclude that use of any of the two models is reasonable.

Inspecting the values of APE in more detail, an increase in revolving balance of \$100 on average leads to an increase in loyalty probability by 0.8–0.9% and so does an extra payment in a year (+0.9%). Similarly, doubling quarter-to-quarter payment count increases chances of loyalty by 26%. On the contrary, spending \$100 extra in a year results in a 0.4% drop in loyalty probability. Lastly, becasuse the utilization ratio is statistically not different from zero, we beware of any comments on its effects.

Regarding PEA, it holds for an average customer that \$100 increse in revolving balance means approximatelly 0.7% higher probability of being loyal to the bank. Furthermore, an additional payment in a year raises the probability by 0.65–0.8% and doubling the q/q payment count increases it by 19–22%. Again, increasing yearly spending by \$100 impacts the loyalty chances adversely by about 0.3%.

Predictive ability of the model

In addition, we perform an anaysis on the predictive ability of our model. We construct a confusion matrix (Appendix, Table 3) and obtain values for specificity, sensitivity and accuracy. The accuracy rate of our model is over 88%, sensitivity of estimating loyalty is 96%, and specificity 49%. The first two values imply a solid quality of our model but

lower specificity might be a weakness when modelling loyalty. The corresponding ROC curve is enclosed in the Appendix.

Conclusion and recommendation

We identified five candidate variables that affect customer attrition. Our recommendations are the following. We recommend increasing revolving balance on credit cards. Furthermore, supporting a more frequent credit card usage should be beneficial for the bank. At the same time, while enhancing usage and revolving balance, the bank should monitor the amounts spent by customers as excessive spending might lead to a higher probability of attrition. In rephrase it into a business plan, we recommend supporting usage of the credit card for small regular purchases such as grocery shopping.

Appendix

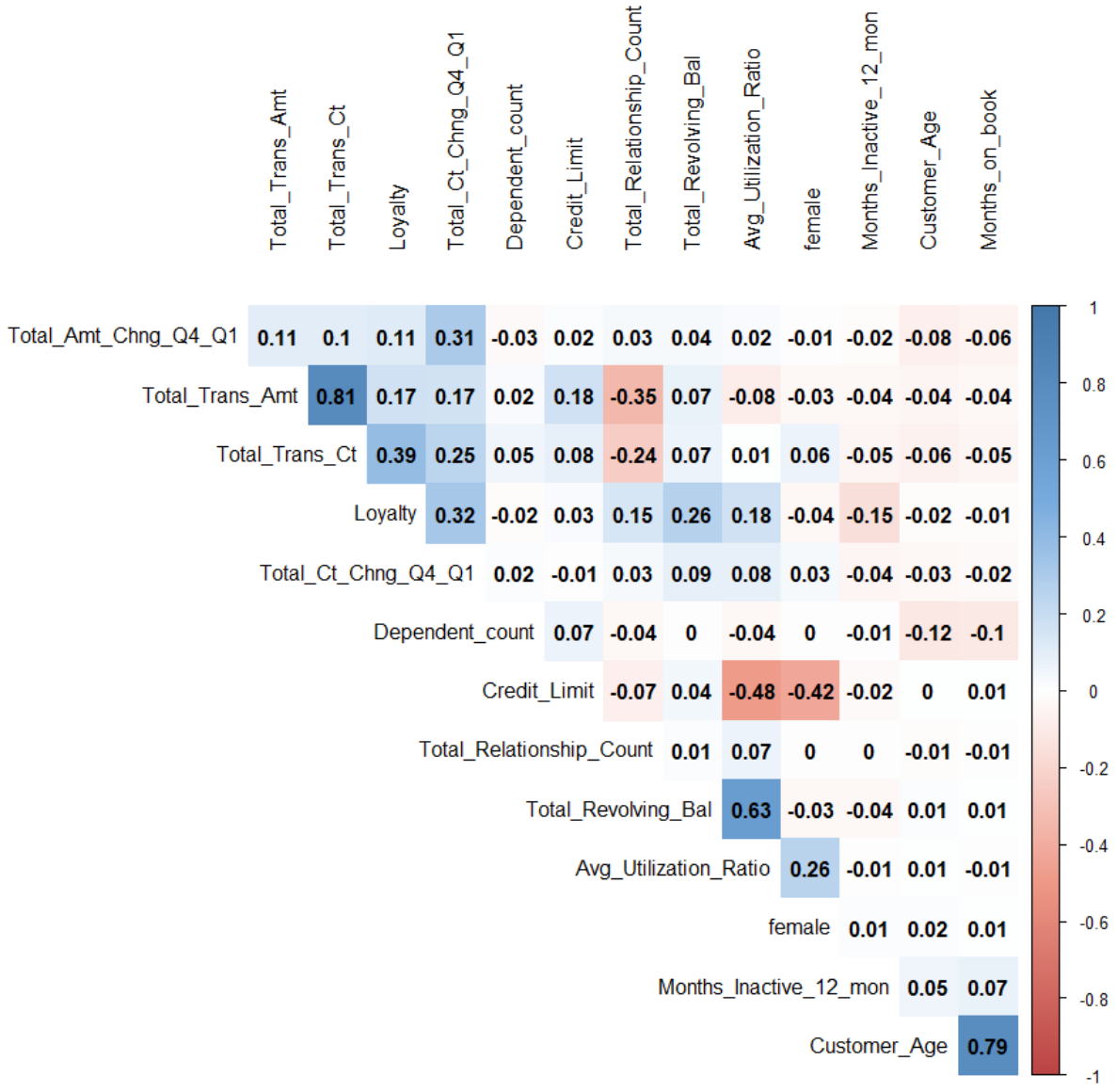

Figure 1: Correlation matrix

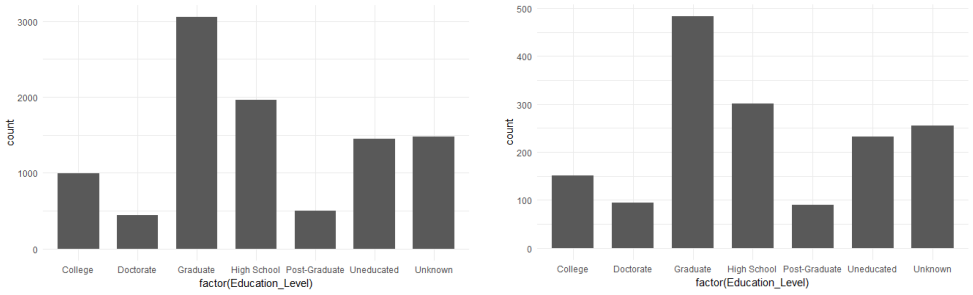

Figure 2: Education Level (left: all customers, right: attrited customers)

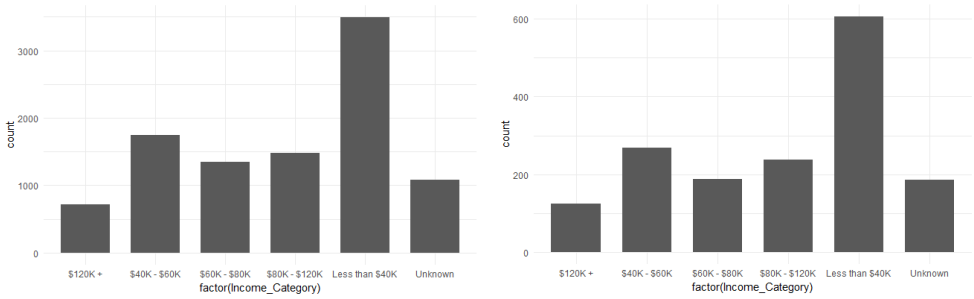

Figure 3: Income Category (left: all customers, right: attrited customers)

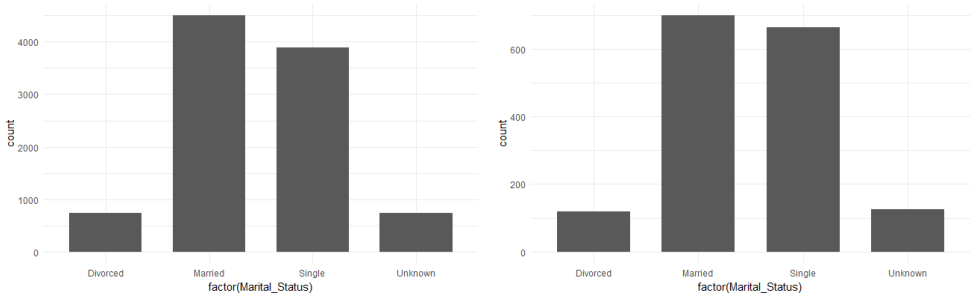

Figure 4: Marital Status (left: all customers, right: attrited customers)

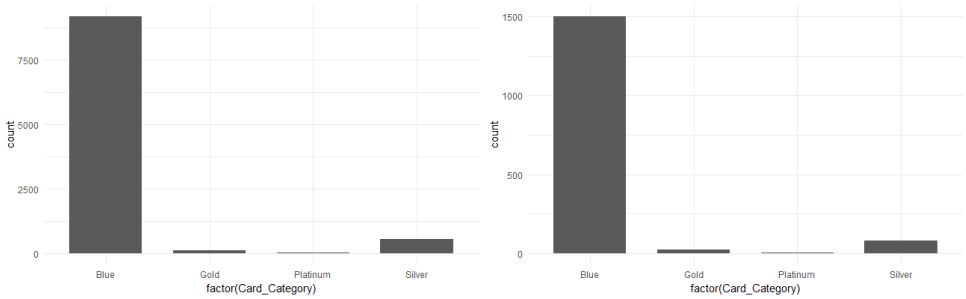

Figure 5: Card Category (left: all customers, right: attrited customers)

	Logit		Probit	
	APE	PEA	APE	PEA
$Total\_Revolving\_Bal$	0.0000909***	0.0000665***	0.0000829***	0.0000723***
$Total\_Trans\_Amt$	-0.0000403***	-0.0000297***	-0.0000410***	-0.0000358***
$Total\_Trans\_Ct$	0.00892***	0.00656***	0.00915***	0.00799***
$Avg\_Utilization\_Ratio$	-0.0214	-0.0157	-0.0125	-0.0109
$Total\_Ct\_Chng\_Q4\_Q1$	0.262***	0.192***	0.257***	0.224***
Observations	9,878		9,878	

Note: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

Table 2: Partial effects

		Predicted	
		Attrited	Loyal
		467	2,540
Reality	Attrited 332	229	103
	Loyal 2,675	238	2,437
Sensitivity		0.9594	
Specificity		0.4904	
Accuracy		0.8866	
Training sample		6,871	
Test sample		3,007	

Table 3: Confusion Matrix

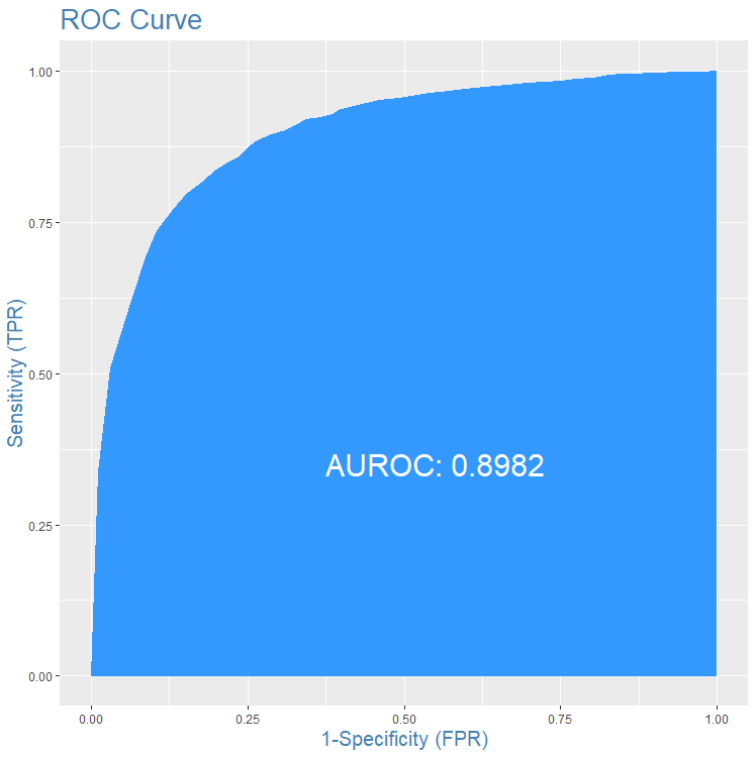

Figure 6: ROC Curve