Take-Home-Exam: Introduction to Statistical Machine Learning

Name: Nikolaus Schäfer, Matrikelnummer: 3308460

June 21, 2018

1 Problem

The first step of this little project is to find the relevant variables of the data set. We are confronted with a set of 14 predictor variables (I left out 'obs_id' for obvious reasons), and to create a model with all of them would not only result in a high computational demand, but also in a poor interpretability. In this problem (Poblem 1) I will make a pre-selection of the variables, meaning I will try to detect all irrelevant (or uninterpretable) variables, which clearly have little to no connection (or have uninterpretable data) to the probability of default.
After first looking at the data, I noticed, that for the "emp_length" variable exist a few missing values. Computing the percentage of missing values in the training set and in the test set gave about 5.37 % and 5.10% respectively. Figure 1 shows the default percentages for each employment length category. As all categories

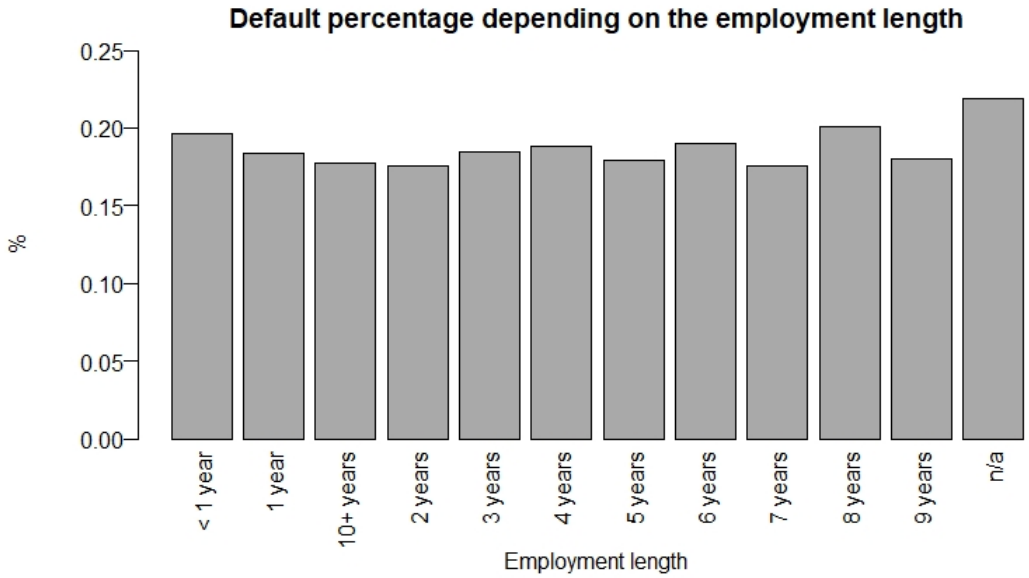

Figure 1: Default % depending on Employment length

seem to have a default percentage between 18 and 20 %, the problem arises in which category one should put the observations with the missing values. Since the "n/a"-category actually has the highest default percentage (see Figure 1), assigning the missing values to the mode of the other categories would clearly distort the relationship between the employment length and the probability of default. Consequently I have decided not to include the variable "emp_length" in my prediction model.
The second variable I have filtered out is the "addr_state"-variable, since it has no causal connection to the probability of default. This is also backed up by some statistical data. Figure 2 shows the Default percentages per State. Most of the default percentages are between 15 and 25 percent. Only Idaho (0 %) and Washington D.C (8.2 %) seem to stand out. This is, however, due to the fact, that in our training data, there is only

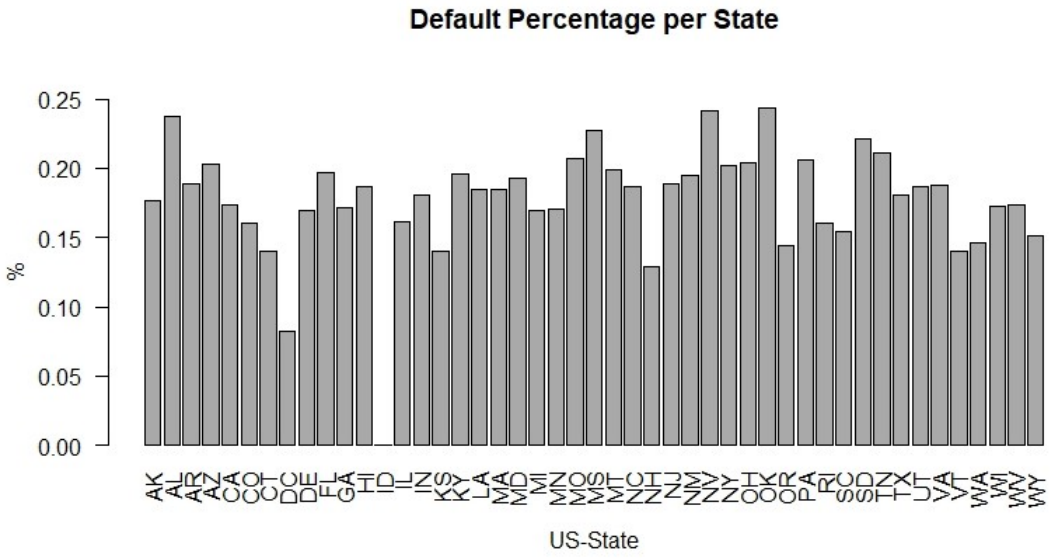

Figure 2: Default % per State

one observation for Idaho and only 124 observations for Washington D.C. If one compares this to the number of observations in Texas (4053 obs.) for example, it becomes evident fairly quickly that an inclusion of the "addr_state"-variable makes no sense at all.

Moreover, the categorical variable "verif_status" seems have a counter-intuitive relationship to the default probability. As can be seen in Figure 3, costumers with verfied income were actually more likely to default than customers with a non-verified income. However, one would actually expect it to be the other way around, since

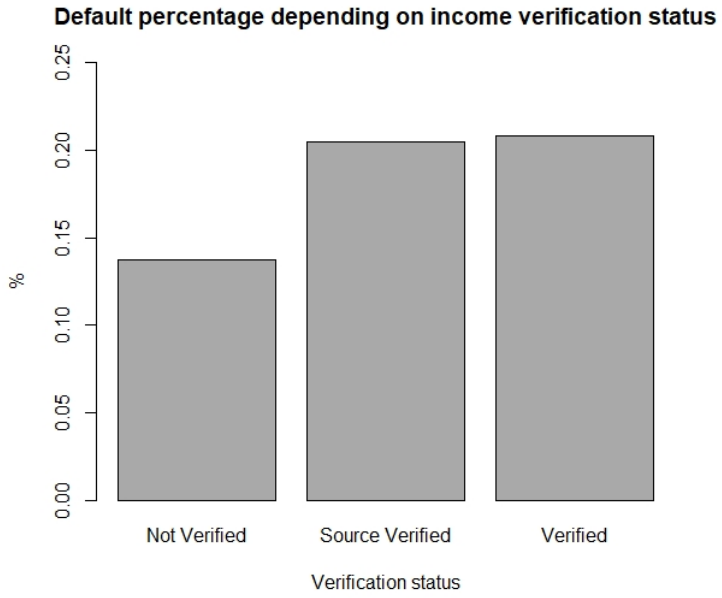

Figure 3: Default % depending on verification status

some might try to lie about their income in order to receive a credit they would not have received with a their actual income. As a result, since I do not believe, that this counter-intuitive relationship to be equally valid for the test set I will not include the "verif_status" variable in my model.

After having eliminated a few variables, which seem to be irrelvant or have a strange relationship to the probability of default, I also want to look for variables which should definitely be included in the model (e.g. are highly correlated to the default probability). Logically, one would expect the "grade"-variable to be strongly

correlated with the default probability. Indeed, Figure 4 verifies this expectation. Customers with loan grade 'A' (highest creditworthiness) defaulted significantly less (5.4 %) than curstomers with loan grade 'G' (lowest creditworthiness, 47.3 %). Therefore, one would expect the regression coefficients for the different grade levels

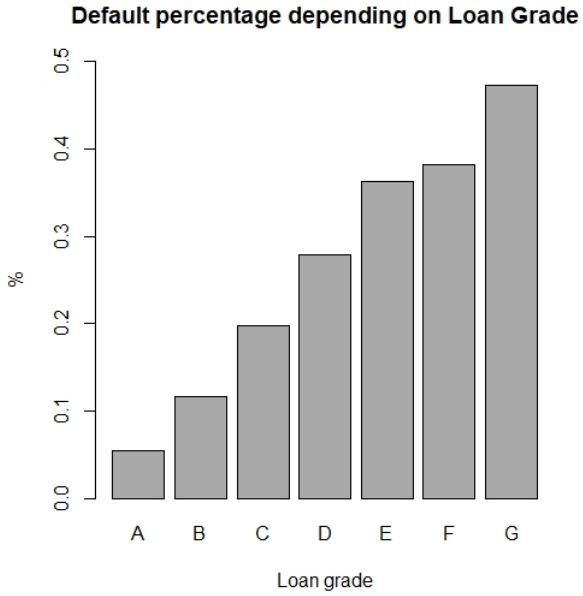

Figure 4: Default % depending on loan grade

to be fairly high. Another important variable that needs to be considered is the "int_rate"-variable. Intuitively, a higher interest rate on the credit should also lead to a higher probability of default, since the amount the customer has to pay back to the bank is much larger in comparison to customers with lower interest rates. Figure

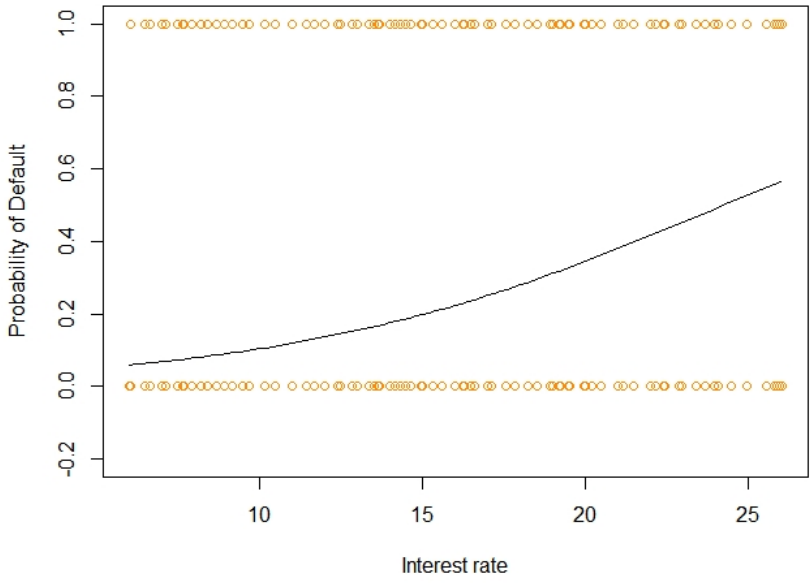

Figure 5: Default % depending on the Interest Rate

5 displays a logistic regression of the interest rate on the probability of default. The graph clearly confirms my expectation of a higher default probability with increasing interest rates. For that reason, as one can see in Problem 2 (!!Spoiler!!), these two variables will be of great significance in my prediction model.

2 Problem

In this problem I will use the remaining set of variables, that were not dropped in problem 1 to find a statistical model for predicting defaults.

Firstly, I divided the training data set into my own training and test data set. I used a 2/3 to 1/3 split (meaning I used 2/3 of the data for training and 1/3 for testing). To compare my results when evaluating my test data with the predictions, I created two more (and different) training and test data sets with the same split.

My plan was to use a generalized linear model (glm() in R) to conduct the predictions. Since we are confronted with a binary dependent variable, I used the binomial error family. Here I have the options of choosing the logistic model (can be specified with (link = logit) in R, but is actually the default of the binomial error family), the complementary log-log link function (can be specified with (link = cloglog)) or the probit model (can be specified with (link = probit)).

Now, only two questions remain: Which of the three models work best on our specific data set and which of the 11 remaining predictor variables should be included in the final model. In order to answer these questions I performed each of the three regressions (logit, probit and cloglog) on all three data sets first for all 11 variables and then 11 times for 10 variables, each time dropping a different variable of the original 11. The results are shown in Figure 6. The evaluation of the predictions on the test data was done with the evaluation function

	All 11 Var.	w/o loan_amnt	w/o term	w/o int_rate	w/o instment	w/o grade	w/o h.-ownership	w/o ann_inc	w/o purpose	w/o dti	w/o open_acc	w/o total_acc
Set 1: logit	-0.423299	-0.423477	-0.425829	-0.424112	-0.423569	-0.424191	-0.425329	-0.424008	-0.423264	-0.424509	-0.423926	-0.42376
Set 1: probit	-0.423475	-0.423679	-0.425956	-0.424362	-0.42377	-0.424019	-0.425472	-0.42413	-0.423476	-0.42466	-0.424077	-0.42392
Set 1: clog	-0.423209	-0.423367	-0.4258	-0.423956	-0.423455	-0.424672	-0.425243	-0.423955	-0.42312	-0.424451	-0.423851	-0.423696
Set 2: logit	-0.434847	-0.43505	-0.437258	-0.435093	-0.435139	-0.436158	-0.436628	-0.435548	-0.434949	-0.435604	-0.435232	-0.435138
Set 2: probit	-0.434963	-0.435195	-0.437362	-0.435273	-0.43528	-0.435827	-0.43673	-0.435604	-0.435071	-0.435707	-0.435332	-0.435239
Set 2: clog	-0.434736	-0.434915	-0.437167	-0.434954	-0.435002	-0.436724	-0.436519	-0.435477	-0.434853	-0.43552	-0.435137	-0.435044
Set 3: logit	-0.432384	-0.434147	-0.436403	-0.434359	-0.434237	-0.435585	-0.43577	-0.434623	-0.434229	-0.434751	-0.434362	-0.434279
Set3: probit	-0.432479	-0.434265	-0.436474	-0.434508	-0.434352	-0.435221	-0.435854	-0.434654	-0.434331	-0.43484	-0.43443	-0.434358
Set3: clog	-0.432327	-0.434065	-0.436387	-0.434254	-0.434153	-0.436208	-0.435698	-0.434612	-0.434164	-0.4347	-0.434322	-0.434237
Maximum Set 1	clog	clog	clog	clog	clog	probit	clog	clog	clog	clog	clog	clog
Maximum Set 2	clog	clog	clog	clog	clog	probit	clog	clog	clog	clog	clog	clog
Maximum Set 3	clog	clog	clog	clog	clog	probit	clog	clog	clog	clog	clog	clog
mean: clog	-0.430091	-0.430782	-0.433118	-0.431055	-0.43087	-0.432535	-0.432487	-0.431348	-0.430712	-0.431557	-0.431103	-0.430992
Accuracy: clog	0.823664	0.823064	0.821804	0.823964	0.822884	0.822224	0.823364	0.823004	0.823184	0.822584	0.822704	0.823004
AUC: clog	0.720707	0.720664	0.71667	0.718868	0.720508	0.720389	0.716595	0.718854	0.720825	0.718771	0.719276	0.719796

Figure 6: Finding the optimal model

given in Problem 3. However, for comparison reasons, I divided the output by the number of input values. Additionally, the rows 'Maximum Set 1', 'Maximum Set 2' and 'Maximum Set 3' each show the maximum value of the evaluation of the three models on their particular data set [1]. The 'mean clog' row gives the mean of the cloglog evaluation on all 3 data sets for each variable set. The rows 'Accuracy: clog' and 'AUC: clog' indicate the accuracy of the cloglog evaluation with a threshold of 0.5 and the area under the ROC-curve of set 1 respectively [2].

Figure 6 clearly shows, that the best model for this case (or rather the seed 567 in R) is the cloglog model performed on all 11 variables. I changed the seeds numerous times to see if the conclusion of choosing this model would stay the same (I haven't depicted all the tables due to space reasons). I actually attained slightly

[1] In other words: The rows show the model which performed best on the particular data set

[2] The last two rows I calculated purely out of fun. They are not really used for picking the model

different results for some seeds, where another model or another set of variables performed a little better than this one. However, on average the cloglog model together with all 11 variables achieved the best results. Therefore, I decided to use this model to predict the probability of default.
Figure 7 shows summary of this regression, this time, performed on the whole training data. As mentioned in

```
Call:
glm(formula = default ~ ., family = binomial(link = cloglog),
    data = dat_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6765  -0.6620  -0.4965  -0.3162   2.8481

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)               -5.451e+00  1.871e-01 -29.142  < 2e-16 ***
loan_amnt                 -3.679e-05  8.981e-06  -4.097 4.19e-05 ***
term                       3.531e-02  2.354e-03  15.001  < 2e-16 ***
int_rate                   7.768e-02  1.018e-02   7.634 2.28e-14 ***
installment                1.425e-03  2.856e-04   4.989 6.07e-07 ***
gradeB                     3.587e-01  6.510e-02   5.510 3.58e-08 ***
gradeC                     5.538e-01  8.370e-02   6.616 3.68e-11 ***
gradeD                     5.495e-01  1.093e-01   5.026 5.01e-07 ***
gradeE                     4.966e-01  1.395e-01   3.561 0.000369 ***
gradeF                     1.755e-01  1.788e-01   0.981 0.326355
gradeG                     1.477e-01  2.054e-01   0.719 0.472090
home_ownershipOWN          1.583e-01  3.722e-02   4.252 2.12e-05 ***
home_ownershipRENT         2.972e-01  2.349e-02  12.650  < 2e-16 ***
annual_inc                -2.847e-06  3.544e-07  -8.033 9.53e-16 ***
purposecredit_card         4.704e-01  1.499e-01   3.138 0.001700 **
purposedebt_consolidation  4.418e-01  1.485e-01   2.975 0.002929 **
purposehome_improvement    4.307e-01  1.554e-01   2.771 0.005587 **
purposehouse               1.210e-02  2.313e-01   0.052 0.958282
purposemajor_purchase      3.359e-01  1.717e-01   1.956 0.050493 .
purposemedical             5.283e-01  1.757e-01   3.006 0.002644 **
purposemoving              3.347e-01  1.968e-01   1.701 0.088980 .
purposeother               2.842e-01  1.559e-01   1.823 0.068231 .
purposerenewable_energy    5.428e-01  4.011e-01   1.353 0.175955
purposesmall_business      6.623e-01  1.714e-01   3.863 0.000112 ***
purposevacation            3.014e-01  2.068e-01   1.458 0.144929
purposewedding            -8.038e+00  8.075e+01  -0.100 0.920704
dti                        1.421e-02  1.432e-03   9.924  < 2e-16 ***
open_acc                   1.689e-02  2.742e-03   6.158 7.36e-10 ***
total_acc                 -6.597e-03  1.262e-03  -5.227 1.73e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 47806  on 49999  degrees of freedom
Residual deviance: 43226  on 49971  degrees of freedom
AIC: 43284

Number of Fisher Scoring iterations: 9
```

Figure 7: cloglog Regression

Problem 1, the ”grade” variable and the interest rate seem to be of great importance. This can be seen by the low p-values and the relatively high coefficient estimates for these variables.