python project: Personnal Loan Campaign

November 30, 2021

1 Report on the success of a marketing campaign for loans

For our inquiry into the success of a marketing campaign we analyze a data set with 5000 observations. The variables included in the data set can be roughly split into variables of personal characteristics or socio-economic background and variables that are related to the financial situation of an observation.
The variables of socio economic background include:
1. Customers Family size
2. Age
3. Years of professional experience
4. A customers education level ordered into: 1. Undergrad 2. Graduate 3. Advanced/Professional
5. The customers annual income in 1000 Dollars

The variables related to the financial background of a customer include:

1. Average Credit card spending per month in 1000 Dollars
2. The Value of a customers house mortgage if he has one
3. Whether the customer has a Securities account
4. Whether the customer has a Deposit account with the bank
5. Whether the customer uses the online banking of the bank
6. Whether the customer uses a Credit Card issued by the bank

Concerning the analyse of the variable ZIP Code, we choose to get a less precise analysis, at the scale of the departement. We only truncated the ZIP Code, in order to create the new variables "departement". Indeed, by doing a describe, and looking at the maw value, we saw that only departements with two numbers were represented.

Lastly, we are interested if a customer has responded positively on personal loan marketing campaign by taking up the loan offer. This is our target variable.

By looking at the variable experience, the negative values seem to us not coherent. We choose to delete them from the data set.
Already, we can postulate some intuitions for our data. First, we expect that every variable that is positively correlated with trust between the customer and the bank (strong customer relationship) is also strongly correlated with the success of the marketing campaign. This would for instance be the fact that a customer already owns a credit card by the bank or has a deposit account there.

Secondly, loans are only given to costumers that have a strong salary or own property they can use as collateral. Hence, every variable that is positively correlated with income or wealth should also be positively correlated with a success of the marketing campaign. Variables that could fall into this category are annual income, education or years of professional experience.

However, these are only intuitions so let us dive into the facts, shall we?

2 Descriptive analysis: The conversion rate of the marketing campaign

We got a conversion rate of 0.097. The data set is totally unbalanced.
We can see it by looking at the first figure of the figure 1.

Figure 1: distribution of the several variables of our data set

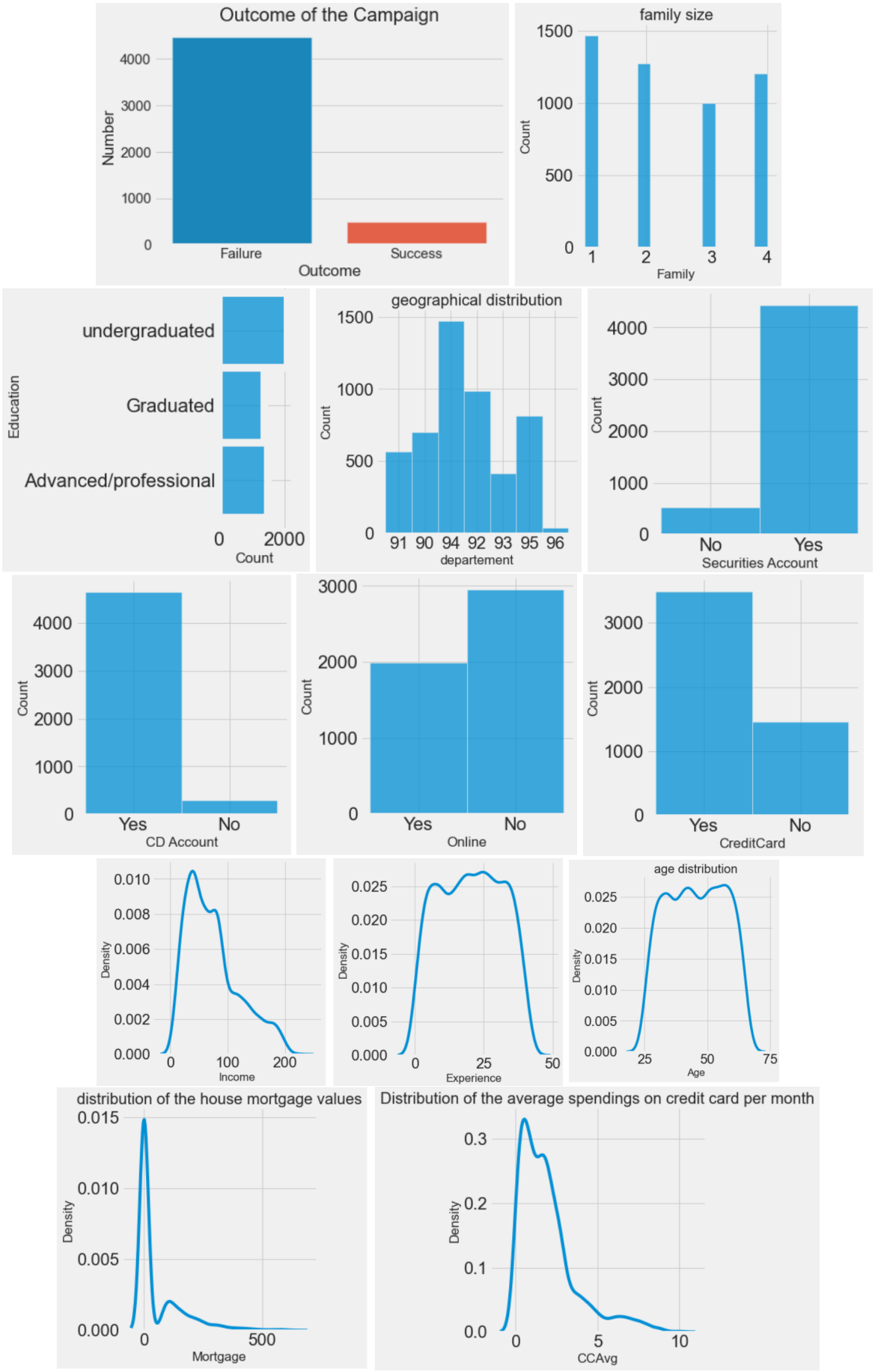

Figure 2: non-parametric test of the distribution of the variables by Securities account

Variable	test-statistic	p-value
Age	1.14234e+06	0.485482
Experience	1.14193e+06	0.480244
Income	1.13363e+06	0.374471
Family	1.1014e+06	0.078269
CCAvg	1.1044e+06	0.101648
Education	1.12886e+06	0.305893
Mortgage	1.14009e+06	0.446788
Personal Loan	1.11886e+06	0.0590829
Securities Account	0	0
CD Account	857694	7.88284e-112
Online	1.11391e+06	0.128717
CreditCard	1.11443e+06	0.115497
departement	1.1386e+06	0.435775

3 How are the different variables distributed in the data set?

We are now going answer the questions asked for the evaluation and explore the key distributions in our sample.

All of our variable distributions are represented in the figure 1.

We can observe that the undergraduate category is the most represented in our data set, and that people who are alone are the most represented.

Only 6 departements are represented (the 91, the 90, the 94, the 92, the 93, the 95, and the 96), that most the 94 and the 92 are the most represented departements, and that really few people come from the 96.

We can notice that most of people in our data set have a security account, a certificate of deposit account with the bank (CD account figure), use the online facilities provided by the bank, a credit card.

People have generally an income between 20 000 and 100 000 euros per year, and spend between 100 and 3000 euros per month.

Age and experience seem to be pretty homogeneously distributed, except for the low and the high values.

Finally, most of people have no house mortgage.

4 How do the different variables interact with each other?

4.1 Grouped analysis:Are age, income, education, etc. distributed similarly for customers who have security accounts and those who don't?

The non-parametric test bellow (figure 2) suggests that there is a significant difference between the distributions of CD Accounts for the group by Securities Account ownership. Furthermore, Average Credit Card : The spending seems to differ in distributions as well as the family size.
We see that the difference is very small for the family size(figure 3.a). Only the probability to have no children is higher for people without a securities account. Plus the probability to have a family of four is slightly higher for people that have a savings account.

If you have a CD Account, chances are you also own a Securities account(figure 3.b). This result is confirmed by the frequency table further below.

The probability of spending little on Credit Cards on average has a spike close to zero for the people without a Securities account(figure 3.c). Otherwise, the distributions look rather similar. This reflects the fact that people without an account usually also do not have the means to own credit card.

Figure 3: The distributions of the family size, of customer with a certificate deposited to the bank, and of the average : The spending on credit card, by security account

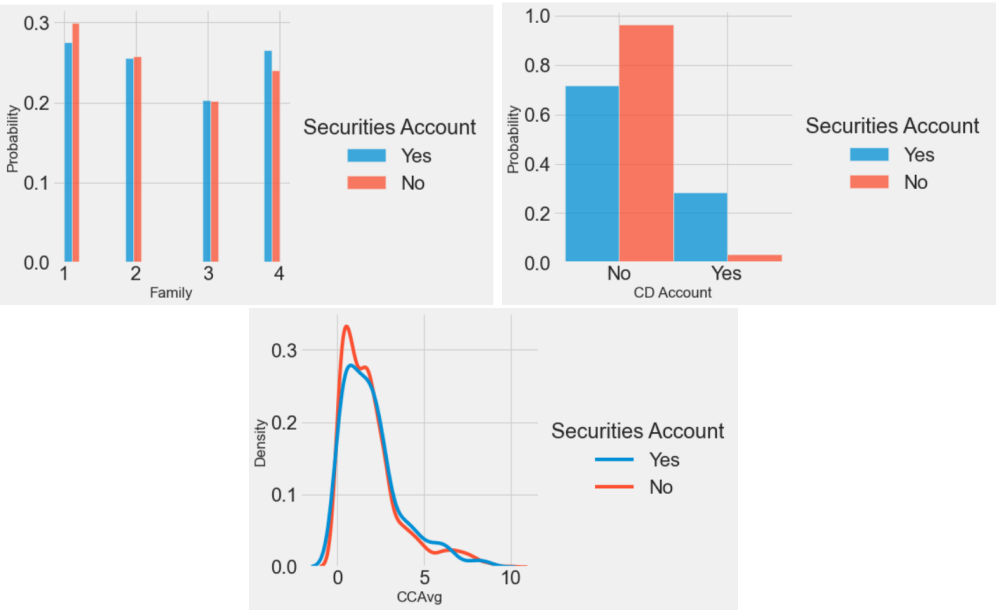

Figure 4: non parametric test of the distribution of the variables by CD Account

Variable	test-statistic	p-value
Age	696978	0.424688
Experience	694244	0.380707
Income	461747	1.03995e-23
Family	676216	0.137966
CCAvg	511367	1.29611e-15
Education	677986	0.147806
Mortgage	624088	4.14049e-05
Personal Loan	427666	1.38782e-109
Securities Account	415784	7.88284e-112
CD Account	0	0
Online	447458	8.65885e-36
CreditCard	327491	9.8946e-87
departement	666692	0.069368

Remarque: The graphics of the other variables are in the code.

4.2 Grouped analysis: Are age, income, education, etc. distributed similarly for customers who have CD accounts and those who don't?

Here (figure 4) the picture changes. Income is of course distributed differently: the less income you have, the more likely you are to not need a CD Account. As we already know from before, owning a Securities Account correlates strongly with owning a CD account (figure 5). Anything related to bank services also has a vastly different distribution for CD Account owners. That makes sense. A CD Account is often a prerequisite to open a Securities account, to get a mortgage or obtain a credit card by your bank.

Striking here is the skewedness of the mortgage distribution for the observations without a CD account. Apparently it is really hard to get a mortgage if you do not have CD Account. Furthermore, it is interesting to note that the income is larger in the observations with a CD Account.

4.3 How many customers have one account only? How many have multiple accounts (i.e., both security and CD accounts)?

Next, we see thanks to a cross table (figure 6) that 4277 observations in our sample have both accounts. Meanwhile 155+369 = 524 people have only on of the two accounts. Among those, the ones with only a CD account are more numerous.

Figure 5: cross table of the variables with a predictive power with CD Account

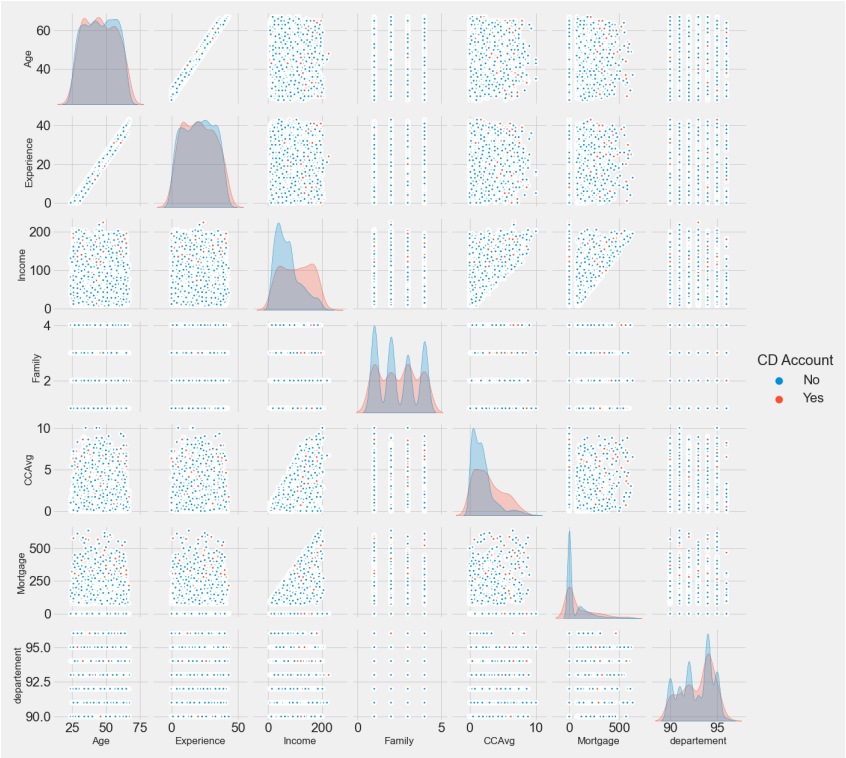

Figure 6: cross table of the account type

CD Account	No	Yes
Securities Account		
No	4277	155
Yes	369	147

Variable	test-statistic	p-value
Age	1.04268e+06	0.159364
Experience	1.04304e+06	0.162345
Income	167168	8.94985e-204
Family	943904	3.95268e-06
CCAvg	411768	1.10695e-109
Education	795806	1.55762e-23
Mortgage	964237	4.44569e-06
Personal Loan	0	0
Securities Account	1.04772e+06	0.0590829
CD Account	798440	1.38782e-109
Online	1.06135e+06	0.332031
CreditCard	1.06774e+06	0.422615
departement	1.07113e+06	0.483755

Figure 7: non parametric predictive test

5 predictive analysis:What are the most important factors that lead to customers responding favorably to the marketing campaign?

We want to find what differentiates the customers that picked up the loan from the ones that resisted.
The figure 7 reports the results of a two-sided non-parametric Mann-Whitney U test on the difference of the distributions of all variables in the success and the failure group.

Here, the non parametric test can help us again to find out where to look. We see that: 1. Income 2. Family Size 3. Credit card spending 4. Education 5. The Size of someones mortgage 6. CD account are all variables that have significantly different distributions in among the people that picked up on the marketing campaign and those that did not pick it up. The distribution for Experience and Age is relatively similar in both groups. It is likely that the effect of those variables is then negligible with respect to the success of the marketing campaign.

In general, we can say that on average, the size of a family of a customer is higher in the success group relative to the failure group(figure 8). An interpretation could be that kids are an investment and require a lot of things such that parents might prefer to take out a loan today and pay it off later when the kids have left the house and have found a job.

Since education is strongly correlated with education, we are not surprised to see that a customer with a worse education responded less to the marketing campaign than one with a strong professional education (figure 9).
Money makes the world go round. If you have a strong income that works as a collateral and allows you to take out a loan. It is therefore not surprising to see that the median income is much higher in the Success group than in the failure group(figure 10)
Spending a sum between 3 and 10 seem to be a sign for someone to respond positively to the campaign (figure 11). Intuitively, we can suppose it is link to the fact that the higher your income is, the more you spend.
Lastly, we can say that having a Securities account and a CD Account at the same time and is a strong sign for someone to respond positively to the marketing campaign (figure 12).

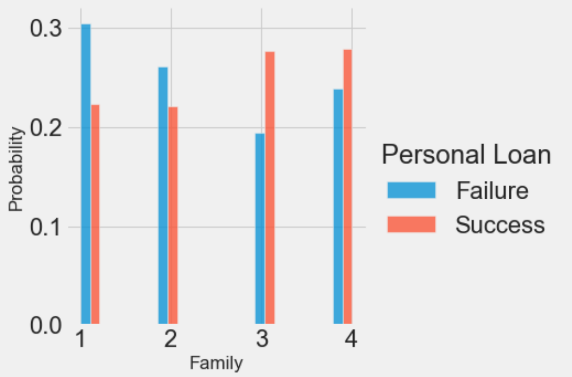

Figure 8: frequency of failure and success depending on the different family size

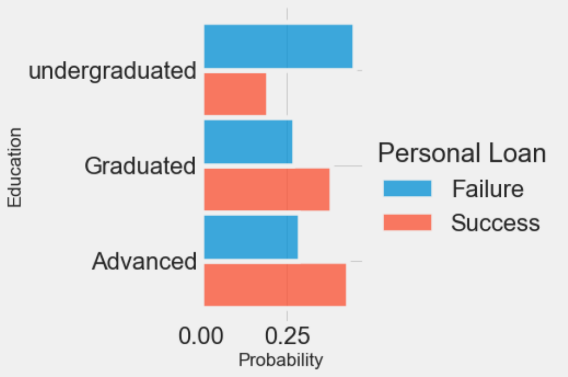

Figure 9: frequency of failure and success depending on the educational level

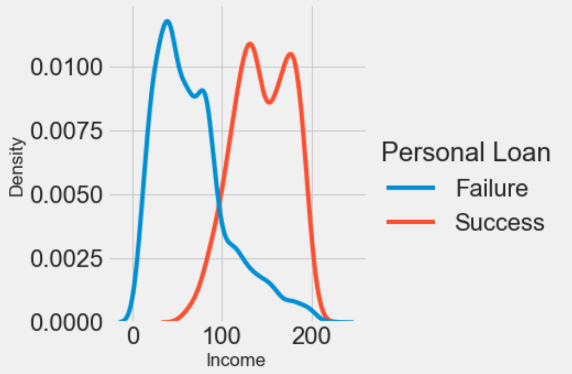

Figure 10: frequency of failure and success depending on the income per year

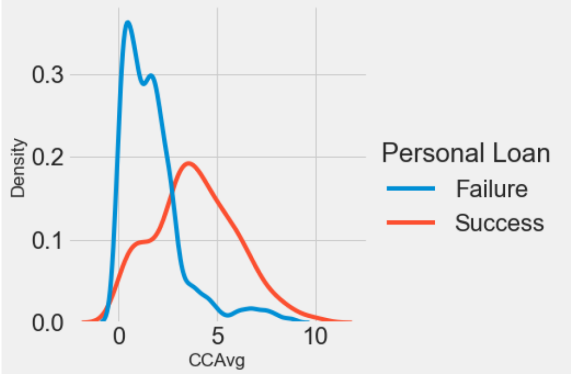

Figure 11: frequency of failure and success depending on the spending with credit card per month

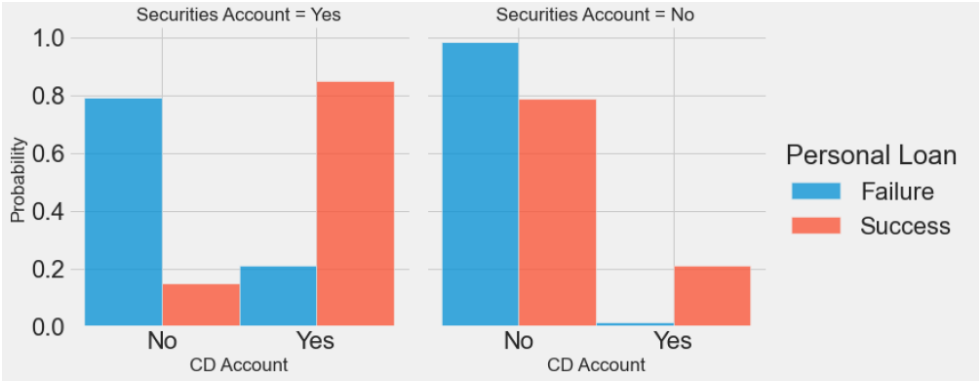

Figure 12: frequency of failure and success depending on the possession of a kind of account in the bank

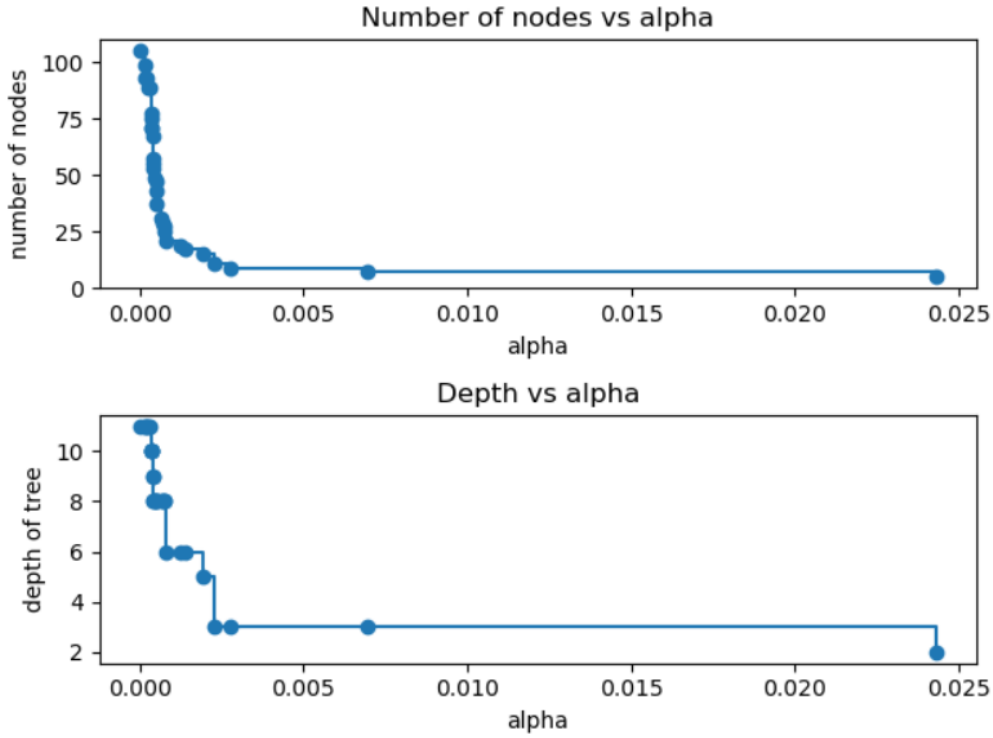

Figure 13: parameter alpha depending on the depth of three, and the number of the nodes

6 Machine learning algorithm

First of all, we want to program a simple machine learning algorithm using a DecisionTreeClassifier. For the basic validation of our result, we resort to a standard split sample approach. We then apply post estimation cost-complexity-pruning two find the optimal restriction on the complexity of our decision tree.

We calculate the accuracy and the confusion matrix of the basic estimator for different specifications of the cost-complexity pruning parameter alpha. The results are plotted in the code. We show the number of nodes and the depth of the tree in relation to the parameter alpha (figure 13). The more restricted the mdoel becomes, the less deep it is and the less nodes it has.

The figure 14 shows the accuracy for different alphas on our training and our test set. We conclude that the alpha should be set very low for a good performance on the test set.

Now we try a different approach. We want to implement a grid search for the optimal 'min sample split', 'max depth' and 'min sample leaf' meta-parameters. The problem with the previous approach is that de-

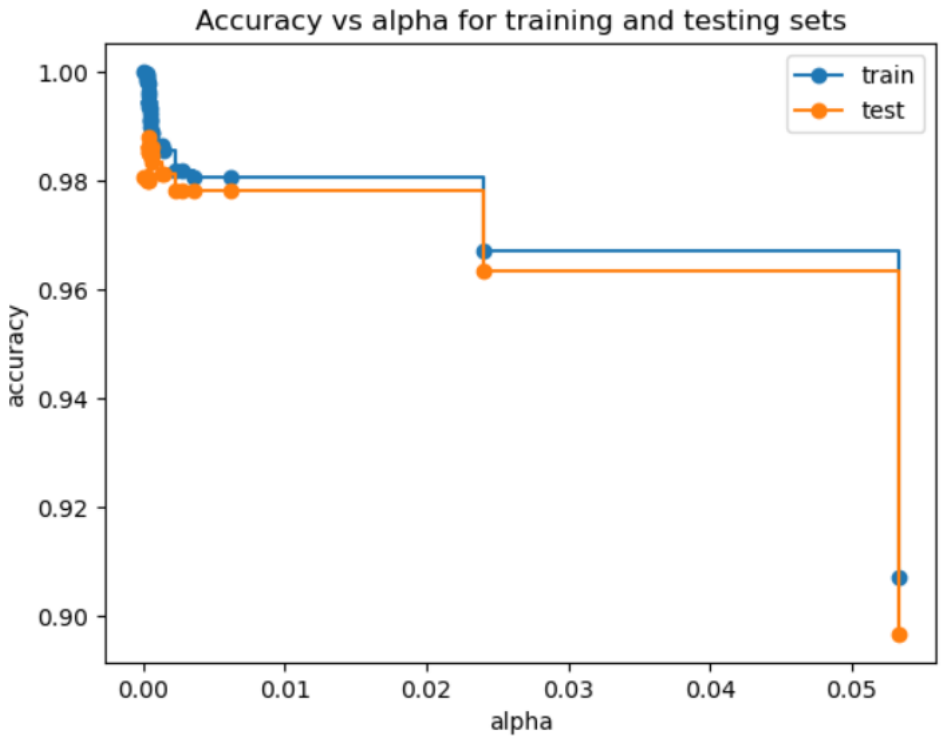

Figure 14

```
DecisionTreeClassifier(max_depth=7, min_samples_leaf=4, min_samples_split=9,
                       random_state=0)
```

Figure 15: best model

pending on the split you randomly draw with 'train test split' function, this can give vastly different results. To reduce this variability we are going to refine our resampling methods by using kfold cross-validation and a grid search for the best algorithm. We optimize the accuracy of our DecisionTreeClassifier over a different model specifications given by the parameter dictionary Dt CV params. The search produces as a model specification a classifier with maximum depth of seven, specific minimum number of samples required to split an internal node of 9 and a minimum number of samples required to be at a leaf node of 4 (figure 15).

We find an accuracy of 99.2 per cent on our test set which is even higher than the best score wee found in the cross validated grid search. Furthermore, we report the confusion matrix for our estimator on the test set (figure 16).

```
[[1351    0]
 [  12  137]]
0.992 0.9858
```

Figure 16: confusion matrix and accuracy of the estimator on the test set