Data Glacier

Your Deep Learning Partner

Week 10 deliverables

Bank Marketing (Campaign) - Group Project

Group Name - Bloodhounds,
Batch code - LISUM09,
Specialization: Data science.
Group member details:

- Name - Margarita Prokhorovich,
- email - marusya15071240@gmail.com,
- Country – Thailand,
- Submission date – 9 July, 2022

Problem description

Statement

- ABC Bank wants to sell it's term deposit product to customers and before launching the product they want to develop a model which help them in understanding whether a particular customer will buy their product or not (based on customer's past interaction with bank or other Financial Institution).

- Bank wants to use ML model to shortlist customer whose chances of buying the product is more so that their marketing channel (tele marketing, SMS/email marketing etc) can focus only to those customers whose chances of buying the product is more. This will save resource and their time (which is directly involved in the cost (resource billing))[1].

Data set problem statement

- A big part of customers are convinced of the effectiveness of an individual approach to service. In an Accenture Financial Services global study of nearly 33,000 banking customers spanning 18 markets, 49% of respondents indicated that customer service drives loyalty. By knowing the customer and engaging with them accordingly, financial institutions can optimize interactions that result in increased customer satisfaction and wallet share, and a subsequent decrease in customer churn[2].

- One of the challenges the banks encounter, is following:
How does a bank figure out what its customers specifically think about its services? Are their issues getting resolved? How satisfied are they with the experience? Why can't banks analyze customer care sessions to find real-time information about the customers and their pressing issues? Customer care records are very pointed and specific about the challenges the customer faces. In these cases building a model and detect patterns can help to improve customers' retaining and loyalty and reduce the churn[3].

1. Problem Statement. Data Science:: Bank Marketing (Campaign) -- Group Project. Data Glacier, URL
2. Top 10 Banking Industry Challenges — And How You Can Overcome Them. Hitachi Solutions, URL
3. Why Retaining Customers For Banks Is As Important As Winning New Ones. Forbes, URL

EDA- data cleansing and transformation techniques applied

As we remember, all the data types in the dataframe are correct. However, there are some other issues in the data that we need to eliminate. We will use data cleansing and transformation approaches from the previous project part: will drop the duplicates, transform pdays category, fill unknown values using KNN approach (the fourth one). As for outliers, we prefer to keep them on this stage because removing the outliers can prevent us from identifying patterns in the data more clearly.

We decided to create ‘pdays_categ’ feature with labels 'yes' and 'no' because it's easier to process this feature with other categorical features.

```
#find duplicates and display their quantity
n_duplicates = df.duplicated().sum()
print(f"Number of duplicates - {n_duplicates}.")

#delete all the full duplicates. We can see decreasing in a rows number
df = df.drop_duplicates()
df.shape
```

✓ 0.2s

```
Number of duplicates - 12.

(41176, 21)
```

Drop the duplicates and transform ‘pdays’

```
print(f'\nNumber of examples where a client was not previously contacted - {len(df[df.pdays =
df['pdays_categ'] = ['no' if pday == 999 else 'yes' for pday in df.pdays]
df = df.drop(['pdays'], axis = 1)
```

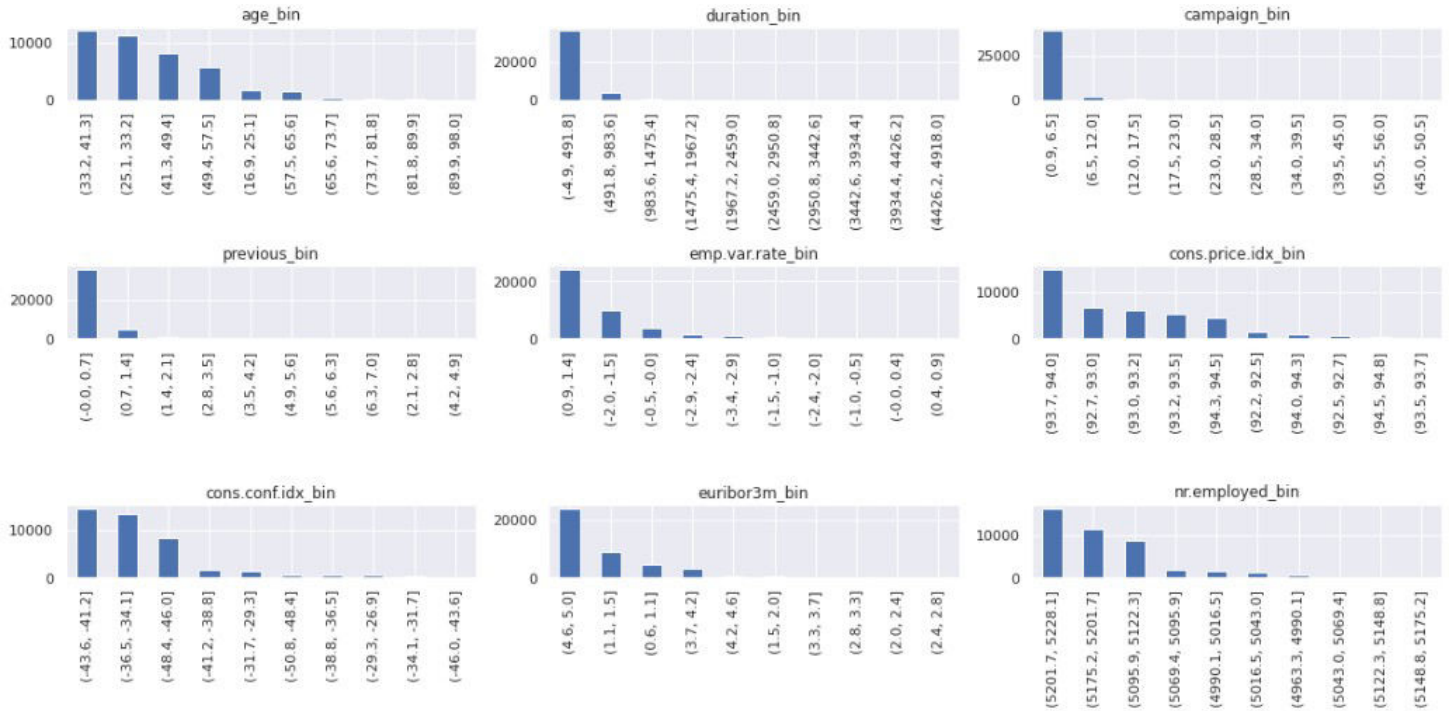
Dataframe columns information

```
Data columns (total 21 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             41188 non-null  int64
 1   job             41188 non-null  object
 2   marital         41188 non-null  object
 3   education       41188 non-null  object
 4   default         41188 non-null  object
 5   housing         41188 non-null  object
 6   loan            41188 non-null  object
 7   contact         41188 non-null  object
 8   month           41188 non-null  object
 9   day_of_week     41188 non-null  object
 10  duration        41188 non-null  int64
 11  campaign        41188 non-null  int64
 12  pdays           41188 non-null  int64
 13  previous        41188 non-null  int64
 14  poutcome        41188 non-null  object
 15  emp.var.rate    41188 non-null  float64
 16  cons.price.idx  41188 non-null  float64
 17  cons.conf.idx   41188 non-null  float64
 18  euribor3m       41188 non-null  float64
 19  nr.employed     41188 non-null  float64
 20  y               41188 non-null  object
dtypes: float64(5), int64(5), object(11)
```

EDA- data cleansing and transformation techniques applied

It's worth noting that when using KNN algorithm, we treated all the categorical variables as nominal ones. However, as for education feature, in fact it's ordinal one, i.e., has certain gradation from lower to higher education level. We encoded this feature as a nominal one because it had an 'unknown' class, so, we couldn't classify it as lower or higher level, also, the model performed better with education encoded as a nominal feature. In next steps we will stick to treating this variable as an ordinal.

Distributions of numeric features after binning




```
Model accuracy for job is 47.18 %.
Model accuracy for marital is 62.16 %.
Model accuracy for education is 48.04 %.
Model accuracy for housing is 51.78 %.
Model accuracy for loan is 84.31 %.
Model accuracy for default is 99.99 %.
Number of "unknown" occurrences:
- feature -  default , number -  8596 , percentage - 20.8762 %
```

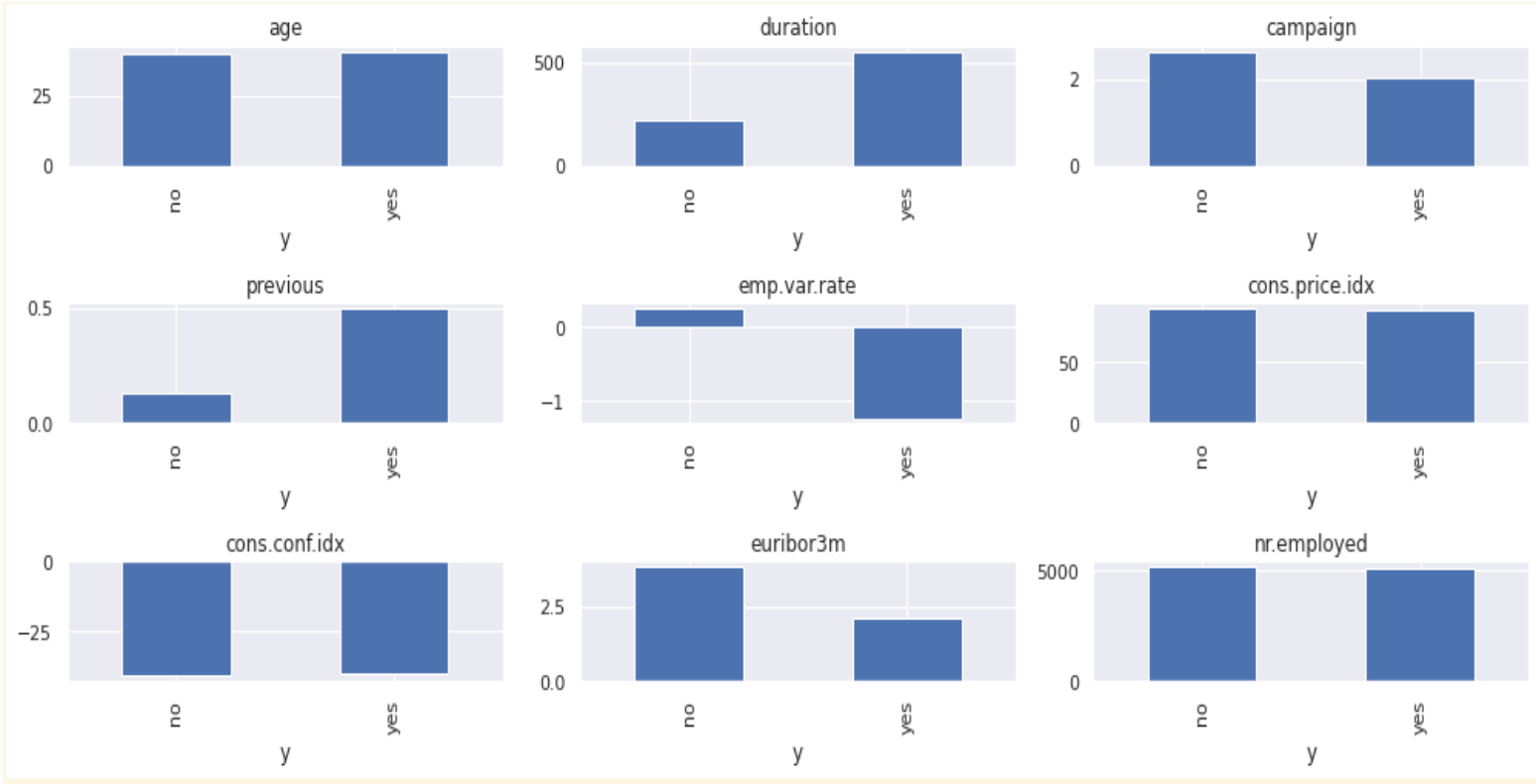
Result of applying KNN algorithm to predict unknown values

Moreover, we could consider one more technique for handling skewed distributions. in the previous week project part we tried to use log function. We can also try to use binning. Using binning can help us to make distributions smoother. We can try to use the processed numeric features for model predictions in the future.

EDA – explore relations between target and numeric features

Let's move to the output variable and start to explore it's relation to the other variables in the data set.
We plot a ratio between positive and negative answers (subscribed a term deposit or not) and see that our data is imbalanced We need to take this fact into account when building the models.

Now we are going to explore relationships between our output variable and numeric input variables. First, we could visually look if means for y = no and y = yes are different for each numeric feature.



y	age	duration	campaign	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
no	39.910994	220.868079	2.633385	0.132414	0.248885	93.603798	-40.593232	3.811482	5176.165690
yes	40.912266	553.256090	2.051951	0.492779	-1.233089	93.354577	-39.791119	2.123362	5095.120069

EDA – explore relations between target and numeric features

We can see that several numeric features have a visualy significant difference in means. For example, duration of a call for negative answers is much less, customers who answered 'no', were contacted more often (campaign). Vise versa, previously they were contacted less often than customers, who answered 'yes' (previous). Negative answers subgroup has a higher employment variance rate but lower short-term lending rates (eurlibor 3 months). We don't see any significant difference for the rest of the variables.

Another way to find out how the input features are related to the target variable - use statistical tests. For this purpose we will use 2 samples Student t-test. For each numeric feature we will compare samples, where y is equal to 'yes' and 'no' respectively When conducting the t-test, we assume that samples have equal variances and that data is normally distributed (for big samples it's not a strict assumption). As a rule of thumb, we can assume the populations have equal variances if the ratio of the larger sample variance to the smaller sample variance is less than 4:1. Hypothesis:

- H0 - positive and negative y samples means are equal
- H1 - positive and negative y samples means are not equal

Feature	Ratio between variances	P-value (t-test)
age	1.95	7.00324e-10
duration	3.75	0
campaign	0.34	2.04343e-41
previous	4.42	1.68378e-161
emp.var.rate	1.2	0
cons.price.idx	1.46	1.62223e-169
cons.conf.idx	1.95	9.13218e-29
euribor3m	1.13	0
nr.employed	1.84	0

EDA – explore relations between target and numeric features

Obtained results show that almost all features have equal variance except for previous feature. For this feature we pass an argument that variances are not equal. As for check for normality of distributions, all p-values are less than 0.05. That means we need to reject null hypothesis - features are normally distributed. However, since we have a large data set, violation of this assumption is not critical.

Finally, for each numeric features Student t-test shows significant difference for two samples means (subscribed/didn't subscribe a term deposit). Each p-value is less than 0.05 - accepted significance level. We can conclude that there is relationship between the target feature values and the input features values.

Additionally, we decided to conduct ANOVA (analysis of variance) test. Typically, a one-way ANOVA is used when you have three or more categorical, independent groups, but it can be used for just two groups.

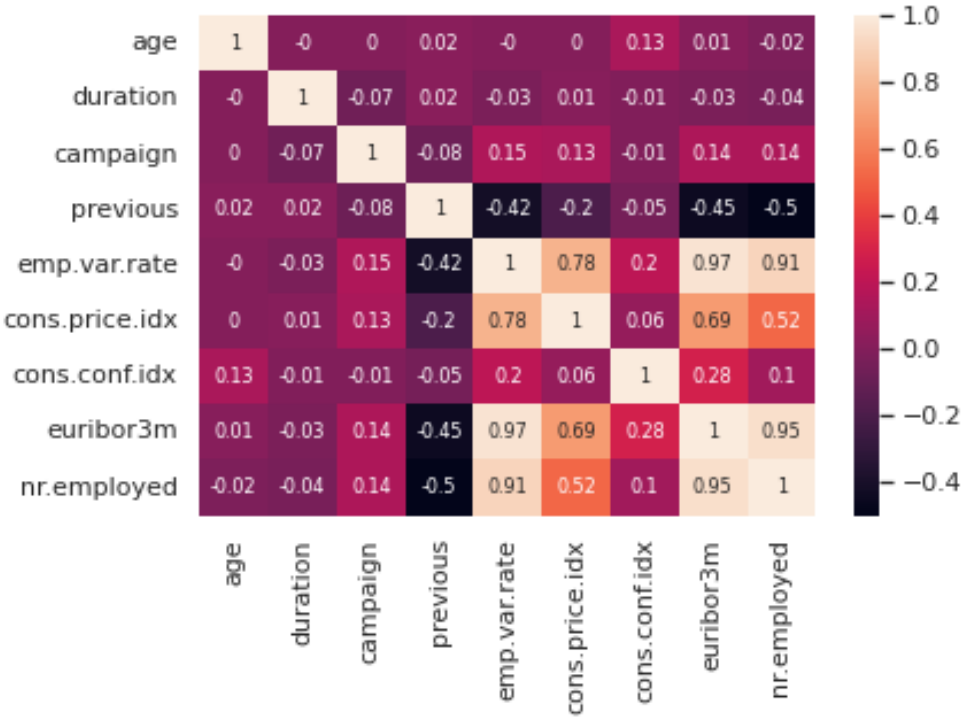
ANOVA results are consistent with t-test results.

Feature	P-value (ANOVA)
age	7.0032438456849 08e-10
duration	0
campaign	2.0434309097339834e-41
previous	0
emp.var.rate	0
cons.price.idx	1.6222328681832695e-169
cons.conf.idx	9.1321757745501 33e-29
euribor3m	0
nr.employed	0

EDA – explore relations between numeric features

Next step is to look at correlations between numeric features themselves.

We can see that several features related to social and economic context are highly correlated. This fact can cause a problem of multicollinearity in the models. We also can check the features for multicollinearity by calculating VIFs (variance inflation factors).
VIFs do not have any upper limit. The lower the value the better. VIFs between 1 and 5 suggest that the correlation is not severe enough to warrant corrective measures.



	Variables	VIF
0	const	528303.388424
1	age	1.018790
2	duration	1.008052
3	campaign	1.038421
4	previous	1.349387
5	emp.var.rate	33.063173
6	cons.price.idx	6.314483
7	cons.conf.idx	2.617565
8	euribor3m	64.331181
9	nr.employed	31.636555

Indeed, several features have extremely high VIF. We should remove some of them to avoid the multicollinearity. Only if we would plan to build a neural network, we can keep highly correlated features. Let's look how the picture can change if we remove 2 variables with the highest VIFs.

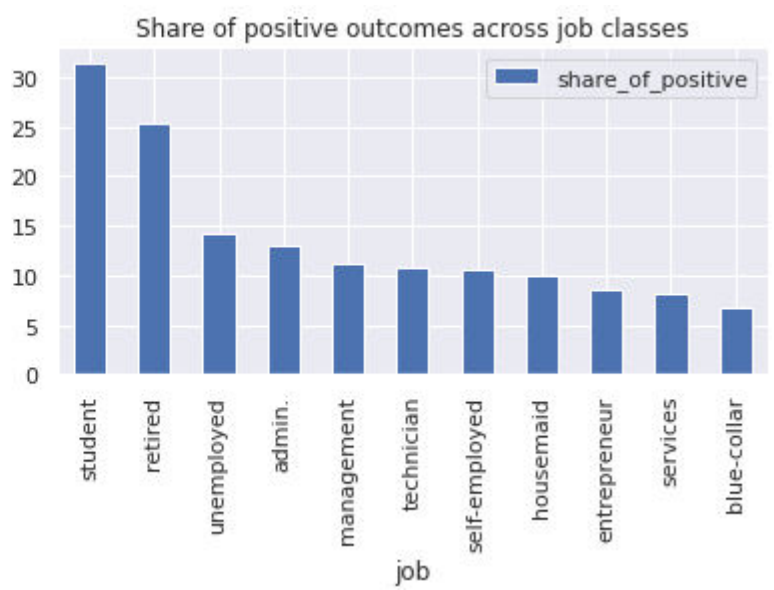
	Variables	VIF
0	const	27962.675266
1	age	1.018470
2	duration	1.007907
3	campaign	1.031206
4	previous	1.344990
5	cons.price.idx	1.390942
6	cons.conf.idx	1.029060
7	nr.employed	1.795510

Now we can see that there's no VIFs greater than 5, so, the multicollinearity problem could be eliminated.

EDA – explore relations between target and categorical features

Job feature

Let's move to analysis of relation between the target variable and input categorical variables. We can plot number of positive and negative answers for each class in categorical features. Also we will plot a percentage of positive answers and provide tabular data.

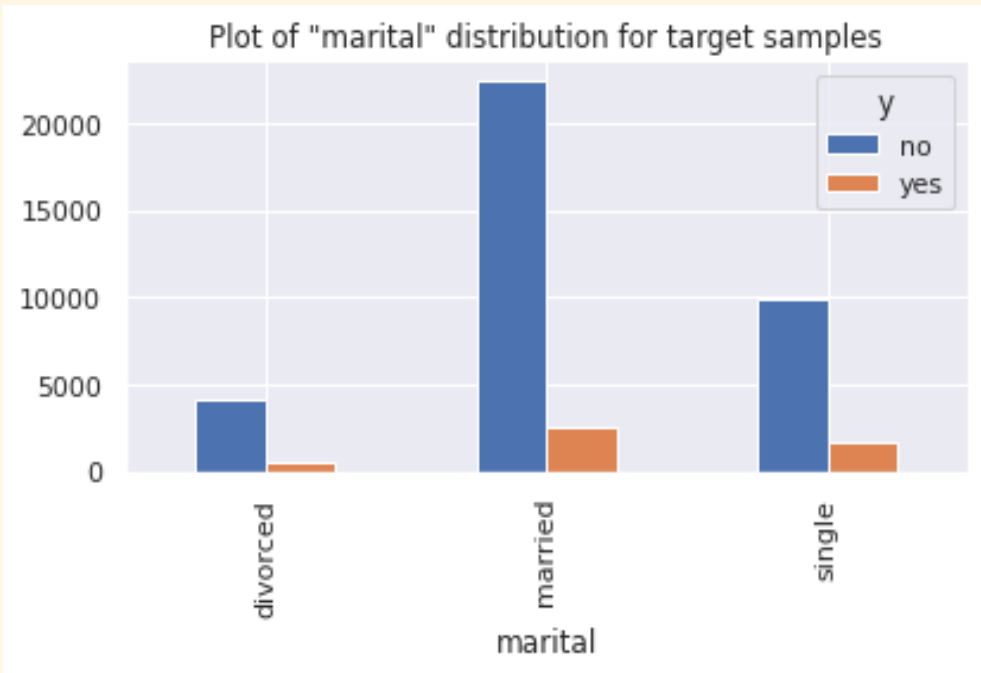


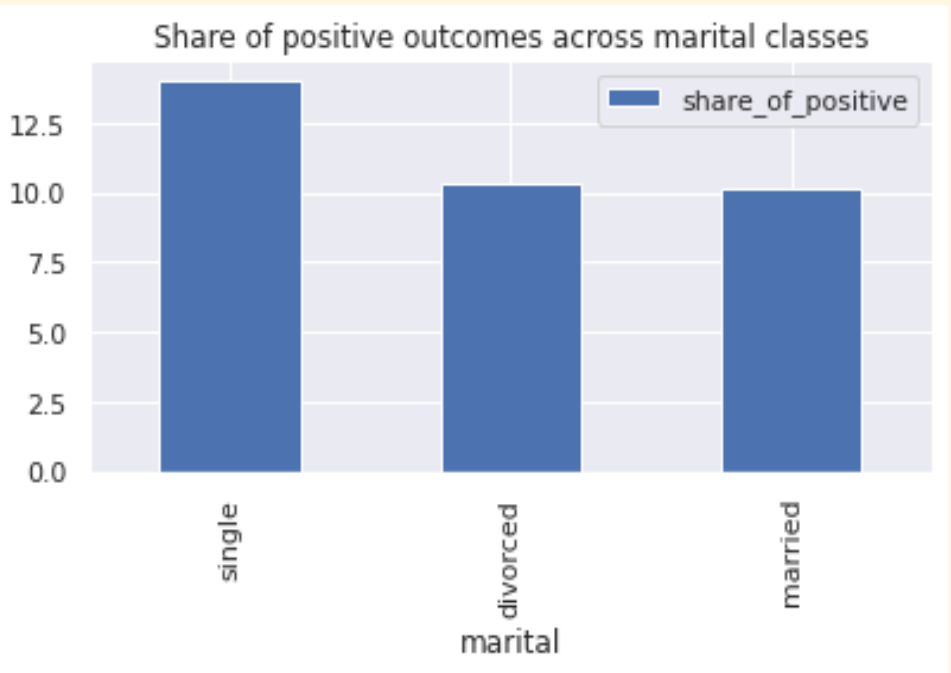
job	share_of_positive
student	31.4286
retired	25.3619
unemployed	14.2012
admin.	13.0385
management	11.2666
technician	10.8300
self-employed	10.4856
housemaid	9.8973
entrepreneur	8.5165
services	8.1366
blue-collar	6.8375

As we can see, ratio between positive and negative answers varies for different job classes. It's quite expectable that for each class share of negative answers is higher. We can see that job feature has no low , top three classes that have the biggest share of y = 'yes' - student, retired, unemployed. Since we can see relation between the target variable and type of job, so, it's needed to keep this feature for a further analysis.

EDA – explore relations between target and categorical features

Marital feature



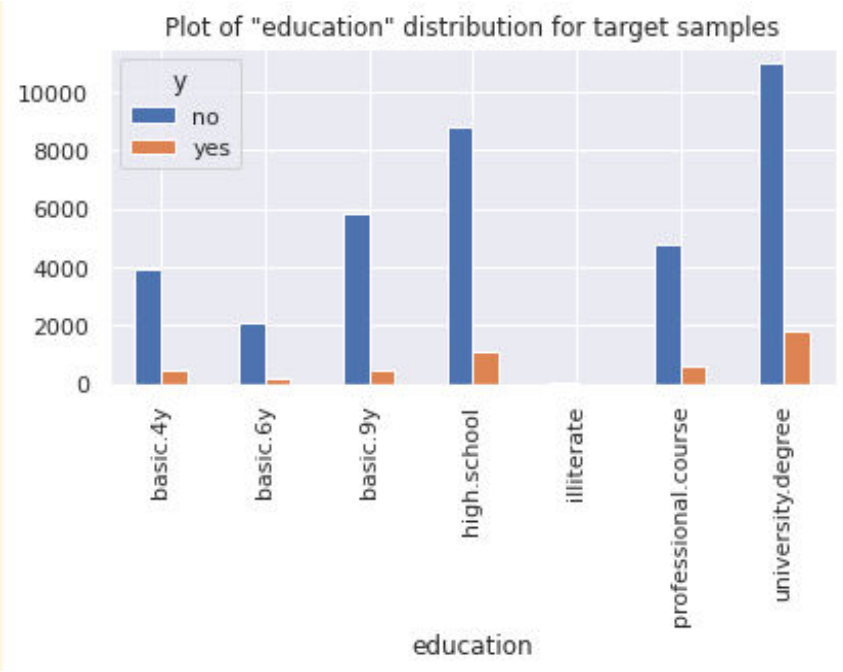


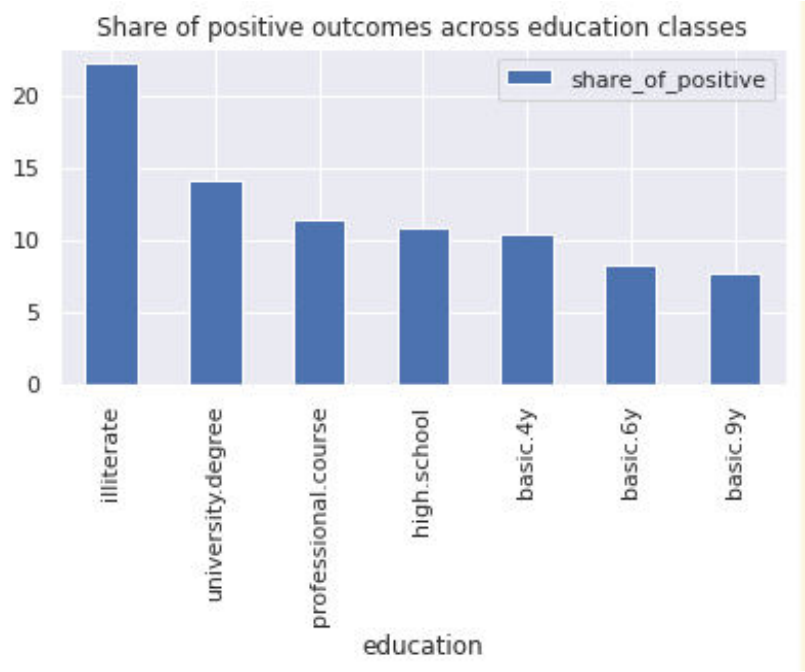
Marital status is also connected with y because single customers are more inclined to give a positive answer and subscribe to the term deposit. However there are much less divorced customers than married ones, shares of positive outcomes is almost equal.

marital	share_of_positive
single	14.0093
divorced	10.3359
married	10.1669

EDA – explore relations between target and categorical features

Education feature



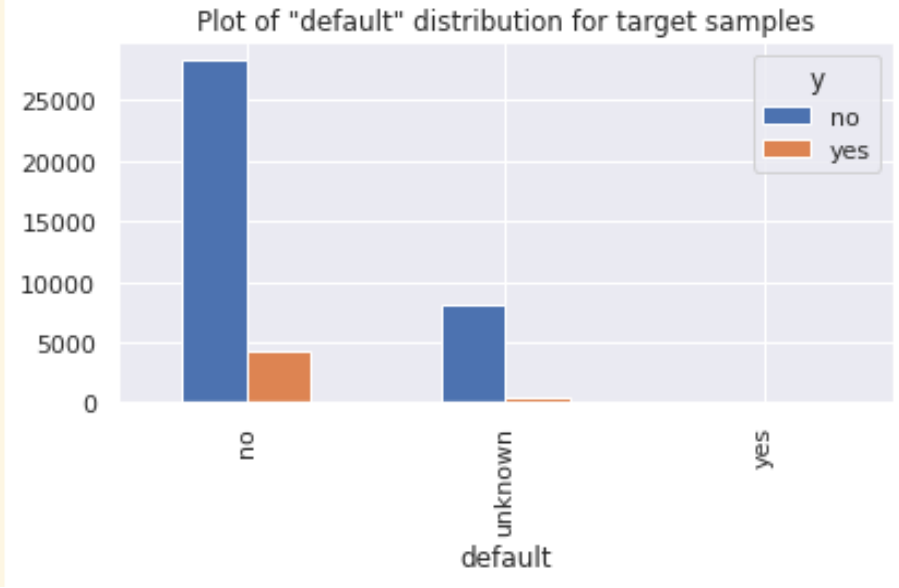


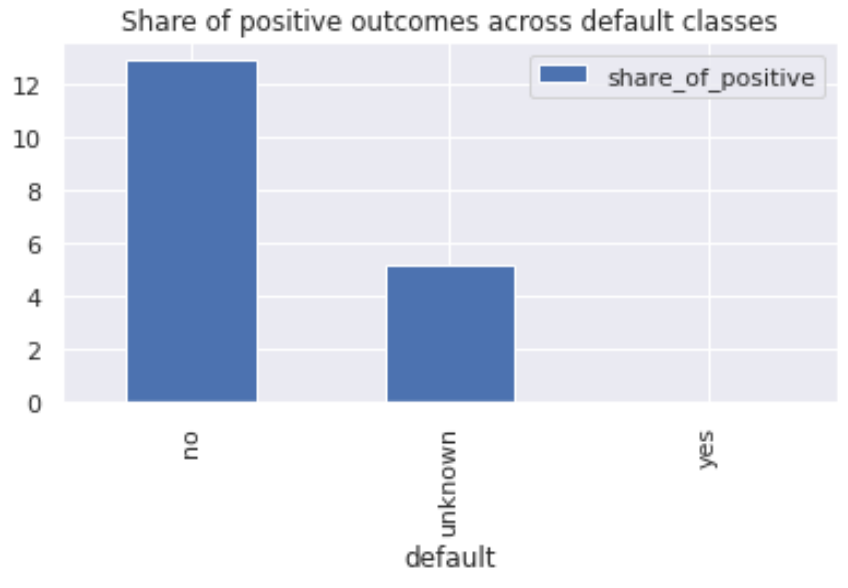
education	share_of_positive
illiterate	22.2222
university.degree	14.1359
professional.course	11.3800
high.school	10.8955
basic.4y	10.4138
basic.6y	8.1861
basic.9y	7.6621

Although a share of illiterate customers is extremely small in the entire data set, this class has the highest share of positive outcomes. It's interesting that there's no obvious relation between illiteracy rate and share of positive outcomes because the second place belongs to customers with university degree, the third one - to customers who completed some professional courses. Anyway, this feature might have an impact on the target feature.

EDA – explore relations between target and categorical features

Default feature



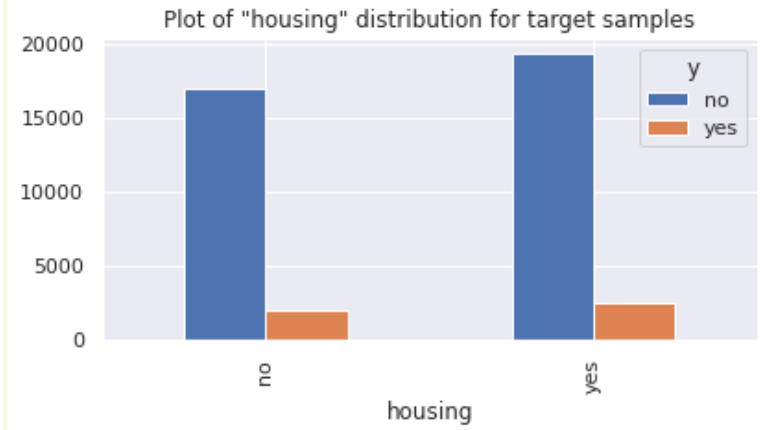


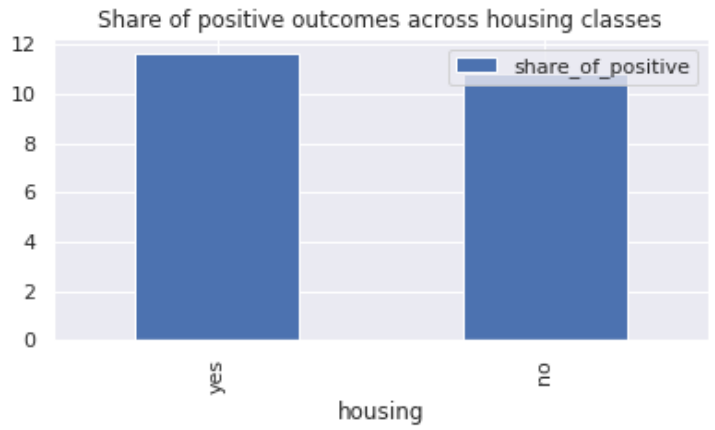
default	share_of_positive
no	12.8803
unknown	5.1536
yes	NaN

As for default feature, share of customers who has credit in default is very low in general as expected. Looking at tabular data we can see that there's no customers with credit in default who subscribed the term deposit. We can consider this feature in model building to find out, how the fact that customer doesn't want to reveal information if he has credit in default or not can affect the target variable. However, in general, this feature seems to be not extremely informative because we have only two classes for further analysis and one of them could potentially belong entirely to the second class ('unknown' to 'no'). In this case we could have zero entropy what isn't good for predicting the target.

EDA – explore relations between target and categorical features

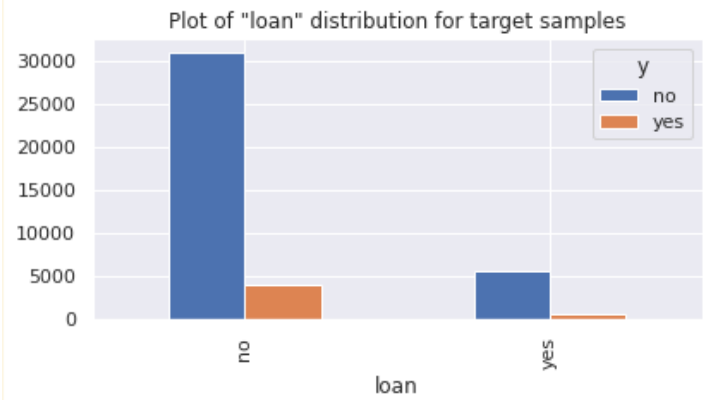
Housing and loan features

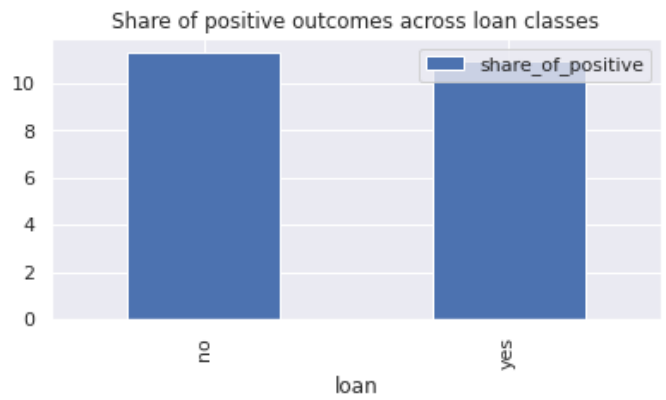




housing	share_of_positive
yes	11.6631
no	10.8110

We can see that there's almost no difference for two y classes, either in number of positive and negative outcomes, or in shares of positive outcomes. So, this feature tends to have low predictive power.



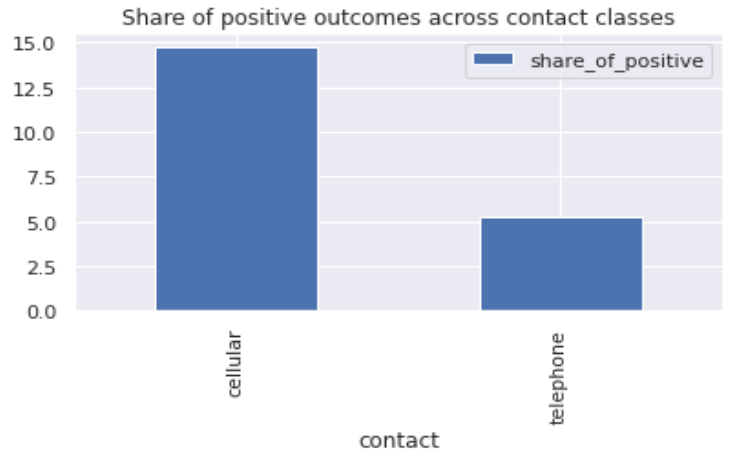


loan	share_of_positive
no	11.3268
yes	10.9280

The situation with loan feature is quite similar to housing feature, except that number of positive outcomes in much lower. However, shares of positive outcomes across two classes are almost equal. So, this feature also seems to be not very informative.

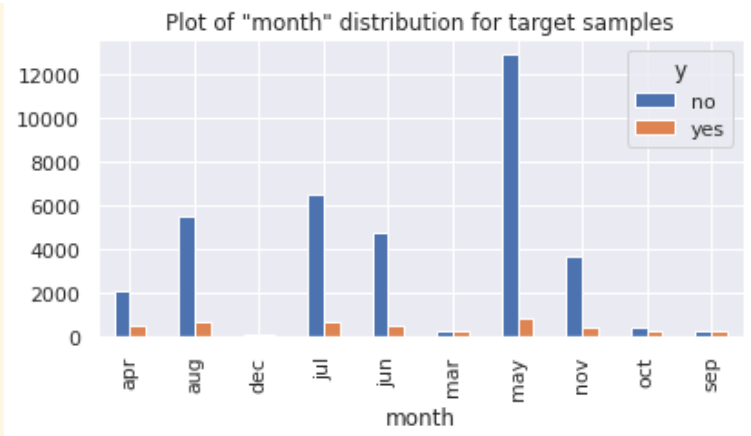
EDA – explore relations between target and categorical features

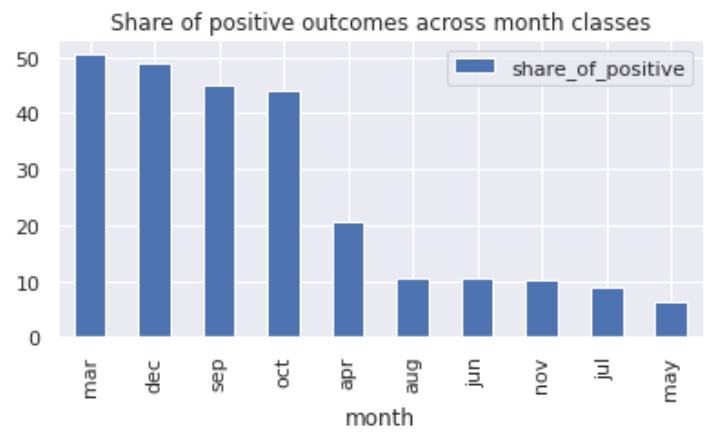
Contact and month features



contact	share_of_positive
cellular	14.7389
telephone	5.2324

As for type of contact, we can see that relationship exists, so we plan to keep this feature for further analysis.



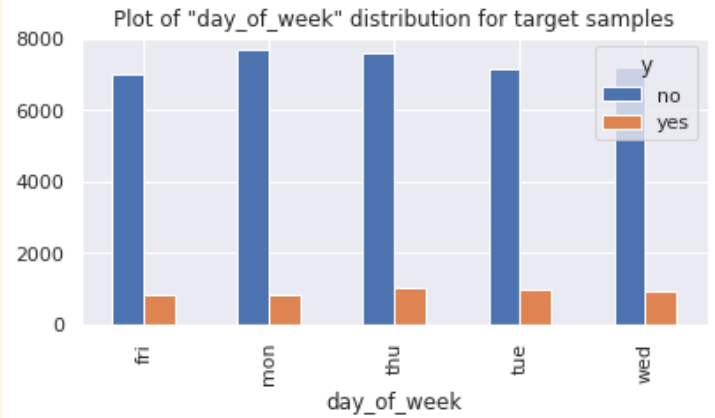


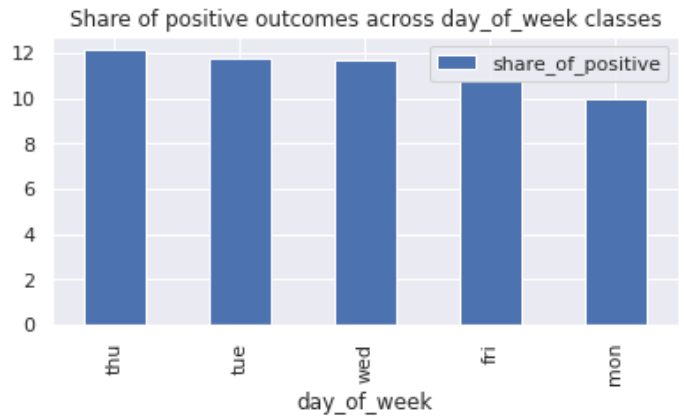
month	share_of_positive
mar	50.5495
dec	48.9011
sep	44.9123
oct	43.9331
apr	20.4865
aug	10.6056
jun	10.5115
nov	10.1463
jul	9.0389
may	6.4357

Last contact month of year can affect the target, as wee can see. However, the relationship isn't obvious - the first three places are taken by months from different year seasons. So, maybe there's no any regularity here.

EDA – explore relations between target and categorical features

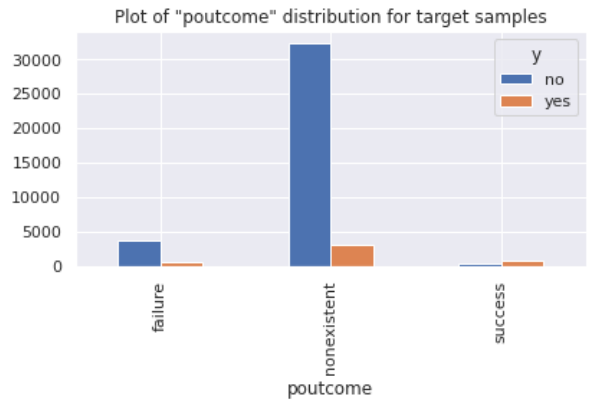
Day of week and poutcome features

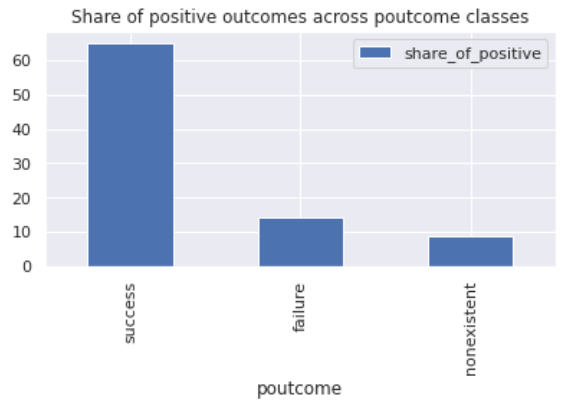




day_of_week	share_of_positive
thu	12.1142
tue	11.7858
wed	11.6671
fri	10.8101
mon	9.9507

As for day of the week, we cannot see any strong relationship. On Mondays and Fridays share of positive outcomes is a little bit less. Maybe due to this slight difference we shouldn't exclude this feature from the further analysis and check it's impact on the target.



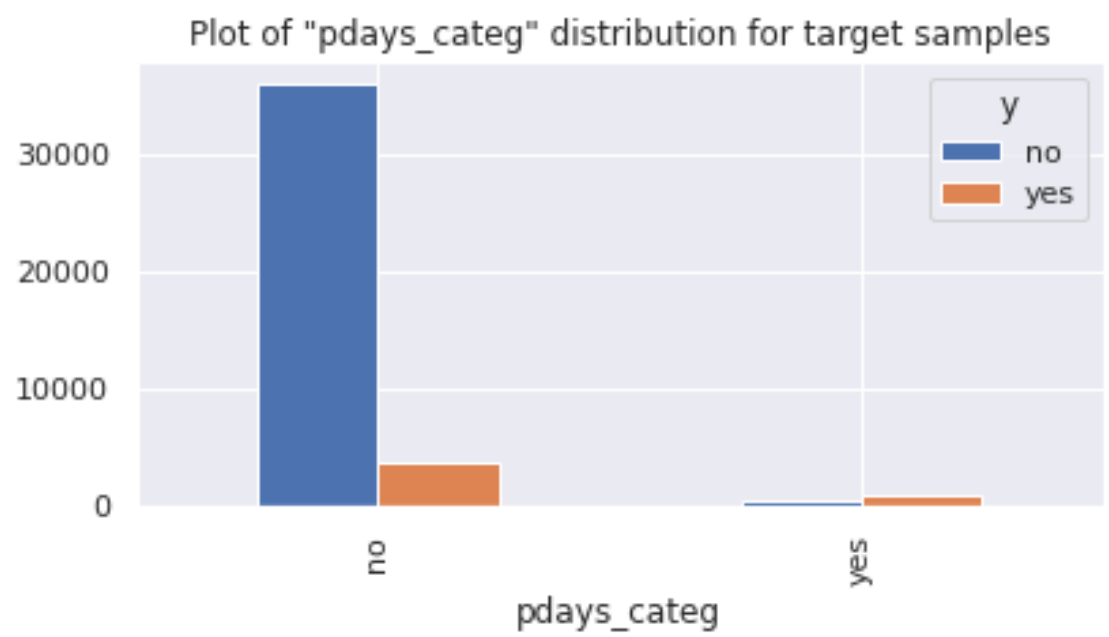


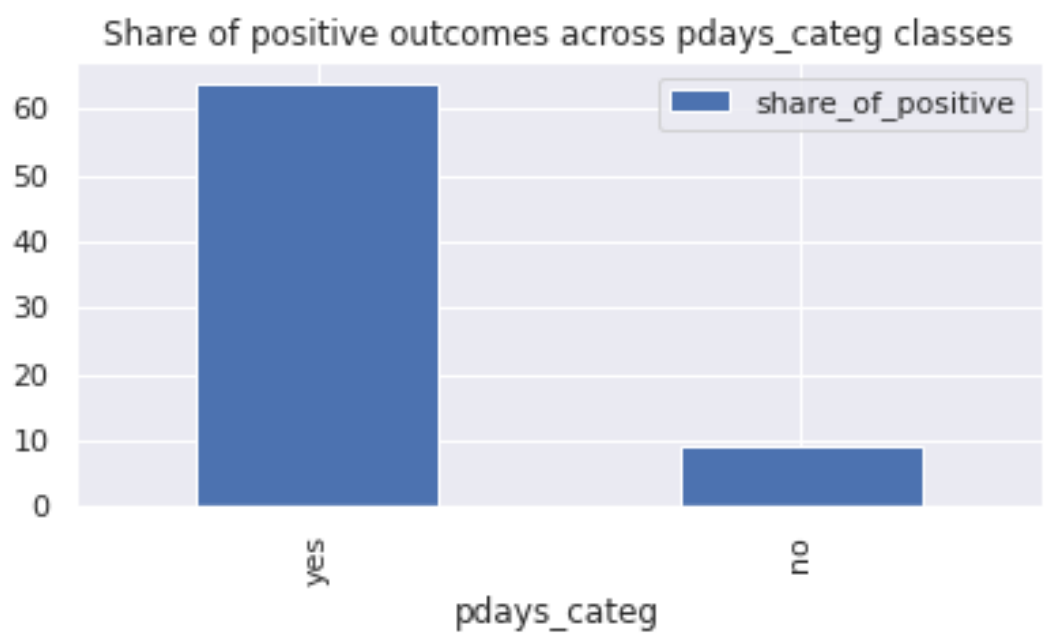
poutcome	share_of_positive
success	65.1129
failure	14.2286
nonexistent	8.8324

Outcome of the previous marketing campaign can influence the outcome of the current marketing campaign. We should keep this feature.

EDA – explore relations between target and categorical features

Pdays_categ feature





Finally, the fact that the client was contacted earlier also affects the target. This created feature is quite close to the previous one in meaning, that's why we should try to build the model with this feature and without it to see its personal impact.

pdays_categ	share_of_positive
yes	63.8284
no	9.2585

EDA – explore relations between target and categorical features

To support our assumptions, we can use a Chi-square test for categorical features.

Hypothesis:

- H0: two categorical variables being compared are independent of each other.
- H1: two categorical variables being compared are dependent of each other.

Also we could use SelectKBest method to select the categorical features which have the highest level of dependence with the target but we'd like to have a look at all the features.

As a result, we can see that all the features except housing, month and loan have p-values less than 0.05. This means that there's dependence between these features and the target. As for three features with p-value more than 0.05, their usefulness for analysis is being questioned. We could compare predictive power of models with and without them.

Chi-square scores and p-values

	ftr	score	pval
0	pdays_categ	4186.8682	0.0000
1	contact	547.7785	0.0000
2	default	321.8911	0.0000
3	poutcome	98.2633	0.0000
4	job	92.9062	0.0000
5	education	86.2995	0.0000
6	marital	27.0094	0.0000
7	day_of_week	10.2336	0.0014
8	housing	3.4657	0.0627
9	month	1.9178	0.1661
10	loan	0.7154	0.3977

EDA – explore relations between categorical features

Additionally to the main analysis of relationships between the features and the target we also can explore relationships between the categorical features using label encoding and visualization. We will explore just several features not to be repetitive.

- more unemployed tend to be singles
- entrepreneurs, management and self-employed have a higher education level
- housemaids are more prone to have a credit in default
- entrepreneurs and unemployed have a little higher probability to have a loan
- unemployed and retired were contacted before more often

- married in average have a little bit higher education level
- married are less prone to have a credit in default
- married are contacted by cellular more often
- married were previously contacted more often

- illiterate, basic 6y and university degree are more prone to have a credit in default
- customers with university degree are contacted by cellular more often
- customers with university degree and professional course were contacted before more often

- unemployed have a higher probability to have a credit in default
- people with a credit in default are less prone to have a housing, they don't have loans and weren't contacted neither by cellular, nor by telephone
- people with a credit in default had a successful outcome in a previous marketing campaign less often
- people with a credit in default weren't contacted before

- people with a housing loan were contacted by cellular more often
- people with a housing loan have a personal loan with a higher probability
- people with a housing loan were contacted before more often

EDA – explore relations between categorical features

Additionally to the main analysis of relationships between the features and the target we also can explore relationships between the categorical features using label encoding and visualization. We will explore just several features not to be repetitive.

- people with a higher education level are contacted by telephone more often
- people who has a credit in default are contacted by cellular more often
- people who was contacted by cellular in average were contacted before less often

- singles had a successful outcome of the previous campaign more often
- people with a higher education level had a successful outcome of the previous campaign more often
- people with a credit in default had an unsuccessful outcome of the previous campaign more often
- people contacted by telephone had an unsuccessful outcome of the previous campaign more often
- people with a successful campaign outcome were contacted more often

Final recommendations

Technical part

Data cleansing and transformation:

- Drop duplicates.
- Transform pdays feature to categorical.
- Fill unknown values using KNN approach.
- Keep outliers (but also consider a z-score approach as a benchmark).
- Keep skewed distribution (but also consider a binning option as a benchmark).

Technical part

Features selection:

- Remove numerical features – euribor3m and emp.var.rate to avoid a multicollinearity problem. In case of building a neural network we can try to keep them.
- Consider models with and without duration feature, as stated in he task.
- Categorical features that might be excluded from the analysis – housing, month, loan, day of week, default. Need to explore their impact on the target by building models with and without these features.

Business part

At the first glance we can conclude, that customers tend to subscribe to a term deposit if:

- They are students, retired or unemployment
- They are single
- They either illiterate or have university degree/professional courses
- They have no a credit in default
- They rather have a housing loan
- They rather don’t have a personal loan
- They were contacted by a cellular
- They rater were contacted in March, December, September or October
- They participated in a previous marketing campaign and subscribed

Bank Marketing (Campaign) - Group Project

Data Glacier

Your Deep Learning Partner

Thank You