Data Glacier

Your Deep Learning Partner

EDA Presentation and proposed modeling technique

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 – 16 July, 2022

Stages of the project

Problem description

Describing the problem from a business perspective

Models selection

Select a bunch of models that could be applied

Data understanding

Describing the data set and its attributes

Models building

Build the models and compare the results

Exploratory data analysis

Identifying patterns and relationships in data

Results interpretation

Choose the best option and explain the results from business perspective

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

Data understanding

Characteristics

- Data Set Characteristics: Multivariate
- Attribute Characteristics: Real
- Associated Tasks: Classification
- Area: Business
- Date Donated: 2012-02-14

Source

- [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

Description

- The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed[4].

Goal

- The binary classification goal is to predict if the client will subscribe a bank term deposit (variable y).

Data set used

- bank-additional-full.csv with all examples, ordered by date (from May 2008 to November 2010).

- This data set is an updated bank-full.csv data set. The data is enriched by the addition of five new social and economic features/attributes (national wide indicators from a ~10M population country), published by the Banco de Portugal.

- It was found that the addition of the five new social and economic attributes (made available here) lead to substantial improvement in the prediction of a success, even when the duration of the call is not included.

4. Bank Marketing Data Set. UCI Machine learning repository, URL

Types of data

Categorical features

bank client data:

- job : type of job
- marital : marital status
- education
- default: has credit in default?
- housing: has housing loan?
- loan: has personal loan?

related with the last contact of the current campaign:

- contact: contact communication type
- month: last contact month of year
- day_of_week: last contact day of the week

other attributes

poutcome: outcome of the previous marketing campaign

Output variable - y - has the client subscribed a term deposit? (binary)

Numeric features

bank client data:

- age

related with the last contact of the current campaign:

- duration: last contact duration, in seconds

other attributes

- campaign: number of contacts performed during this campaign and for this client
- pdays: number of days that passed by after the client was last contacted from a previous campaign
- previous: number of contacts performed before this campaign and for this client
- poutcome: outcome of the previous marketing campaign

social and economic context attributes

- emp.var.rate: employment variation rate - quarterly indicator
- cons.price.idx: consumer price index - monthly indicator
- cons.conf.idx: consumer confidence index - monthly indicator
- euribor3m: euribor 3 month rate - daily indicator
- nr.employed: number of employees - quarterly indicator[5].

5. Bank Marketing Data Set. UCI Machine learning repository, URL

Social and economic context attributes description

Employment variation rate
Employment rates are defined as a measure of the extent to which available labor resources (people available to work) are being used. They are calculated as the ratio of the employed to the working age population. Employment rates are sensitive to the economic cycle, but in the longer term they are significantly affected by governments' higher education and income support policies and by policies that facilitate employment of women and disadvantaged groups. Employed people are those aged 15 or over who report that they have worked in gainful employment for at least one hour in the previous week or who had a job but were absent from work during the reference week. The working age population refers to people aged 15 to 64. This indicator is seasonally adjusted and it is measured in terms of thousand persons aged 15 and over; and as a percentage of working age population.

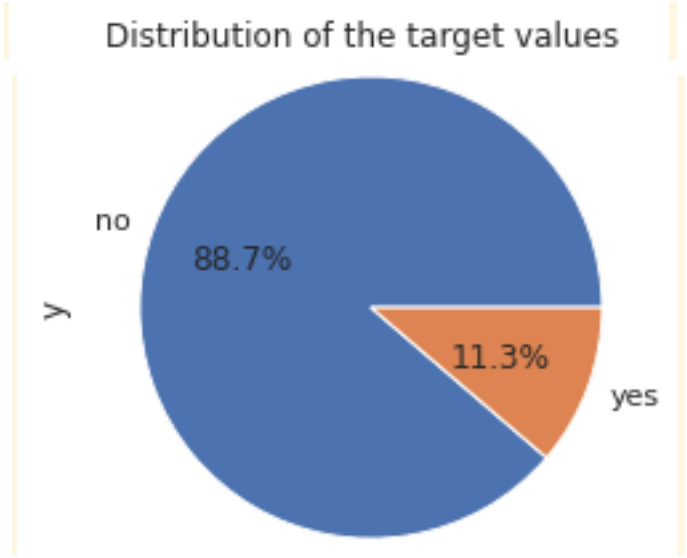
Consumer price index
The Consumer Price Index (CPI) measures the monthly change in prices paid by consumers.

Euribor 3 month rate
The 3 month Euribor interest rate is the interest rate at which a selection of European banks lend one another funds denominated in euros whereby the loans have a maturity of 3 months. Alongside the 3 month Euribor interest rate we have another 14 Euribor interest rates with different maturities (see the links at the bottom of this page). The Euribor interest rates are the most important European interbank interest rates. When the Euribor interest rates rise or fall (substantially) there is a high likelihood that the interest rates on banking products such as mortgages, savings accounts and loans will also be adjusted.

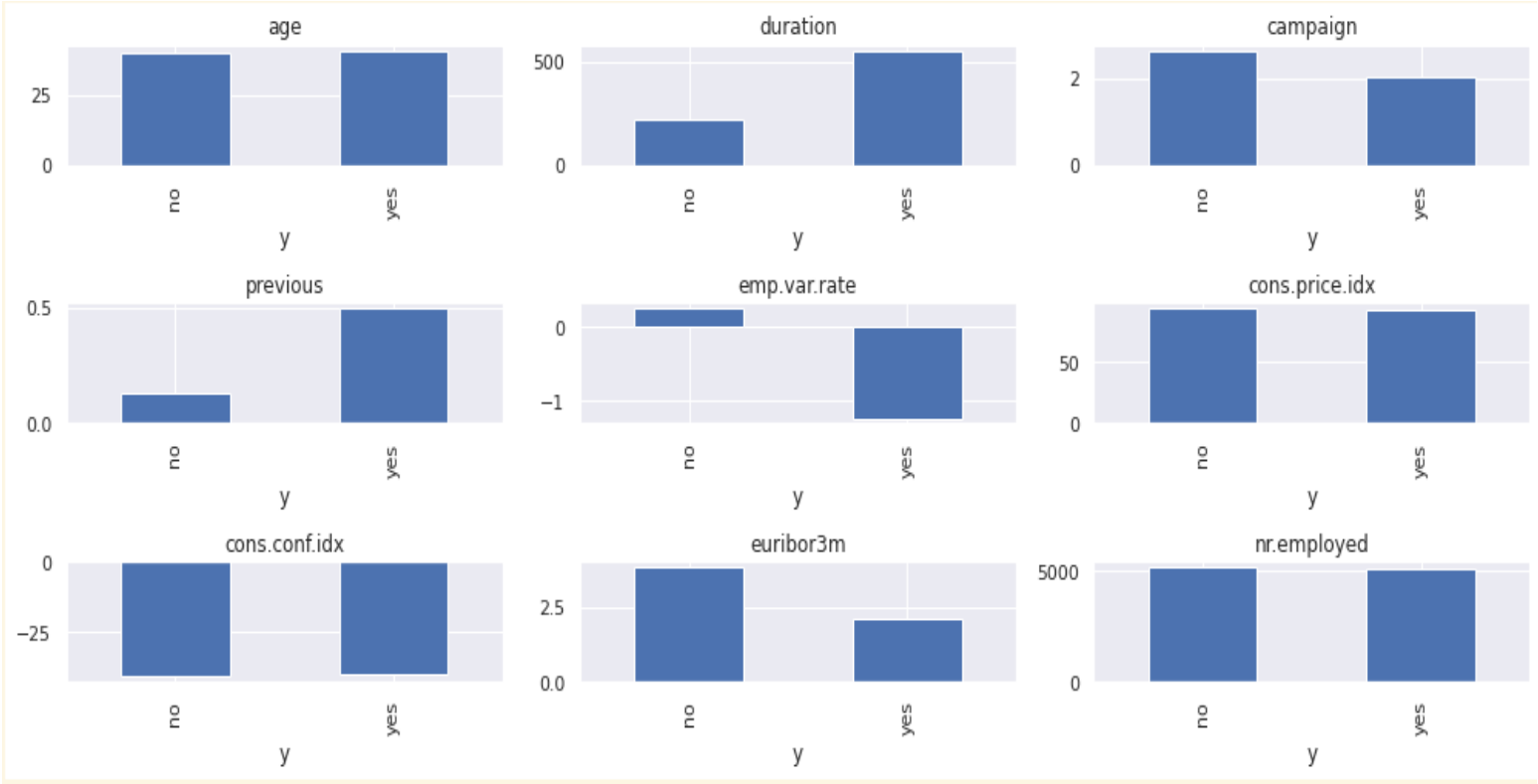
Number of employees
The number of employees is used as a temporary approximation of the number of persons employed.
The number of employees is defined as those persons who work for an employer and who have a contract of employment and receive compensation in the form of wages, salaries, fees, gratuities, piecework pay or remuneration in kind.

EDA – explore relations between target and numeric features

After cleansing and transformation techniques applied to the data set, we can move to identifying relationships between the 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

Visually we can see that several numeric features have noticeable difference.

- For example, duration of a call for negative answers is much less, customers who answered 'no', were contacted more often (campaign).
- Vice 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 (euribor 3 months).
- We don't see any significant difference for the rest of the variables.

Besides graphical and tabular forms analyzing, it could be helpful to conduct statistical tests that can prove or refute a hypothesis that means for positive and negative outcomes are not significantly different.

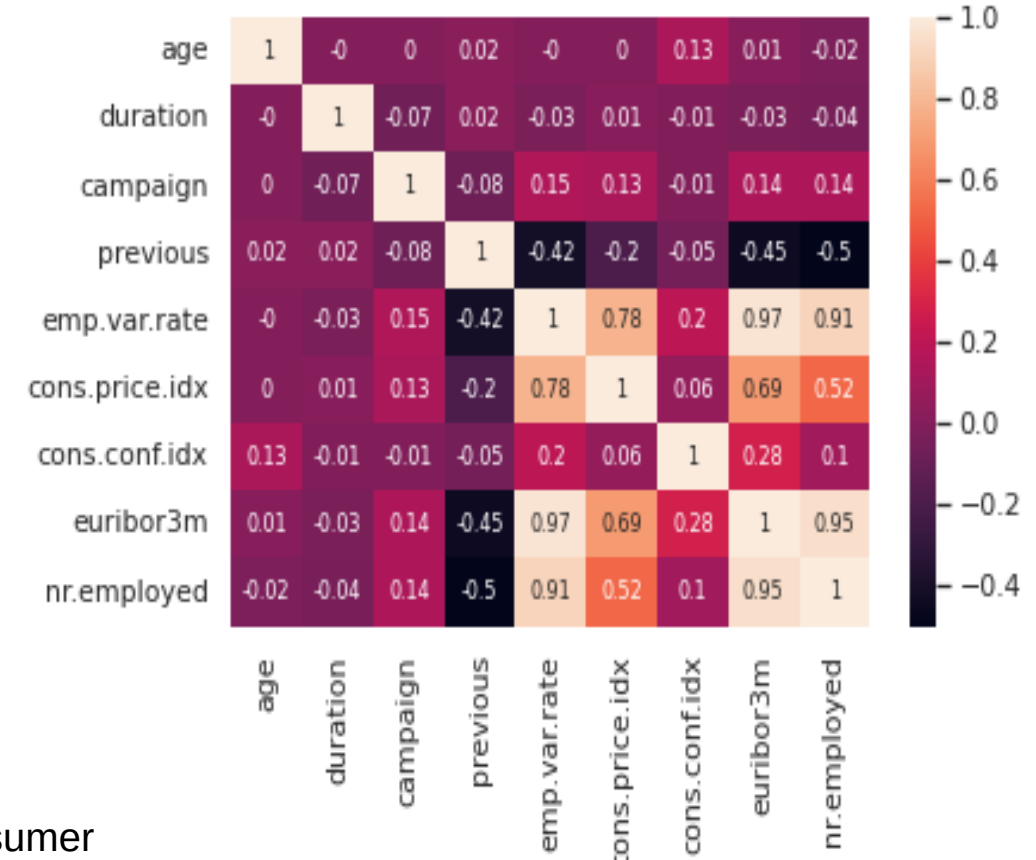
After conducting Student t-test and ANOVA test we can conclude that for all the numeric features there is a significant difference in means for customers, who answered ‘yes’ and for customers, who answered ‘no’.

EDA – explore relations between numeric features

- Also, to estimate the relations between the numeric features, we can build a correlation matrix.
- This statistical instrument shows in what extend the features behave in the same way.
- The more is value on the features intersection, the more co-directional they are.
- This kind of analysis is helpful since it helps to understand the relations between factors that can affect a customer final decision.
- Moreover, it can help to exclude some features that are too interconnected from analysis and make clearer predictions.

In combination with other statistical tools, like VIFs analysis, we can conclude, that features euribor3m and emp.var.rate are candidates for exclusion in some models.

Correlation matrix between numeric features



Employment variance rate ↻ Number of employees

Employment variance rate ↻ Consumer price index

Employment variance rate ↻ Euribor 3 month

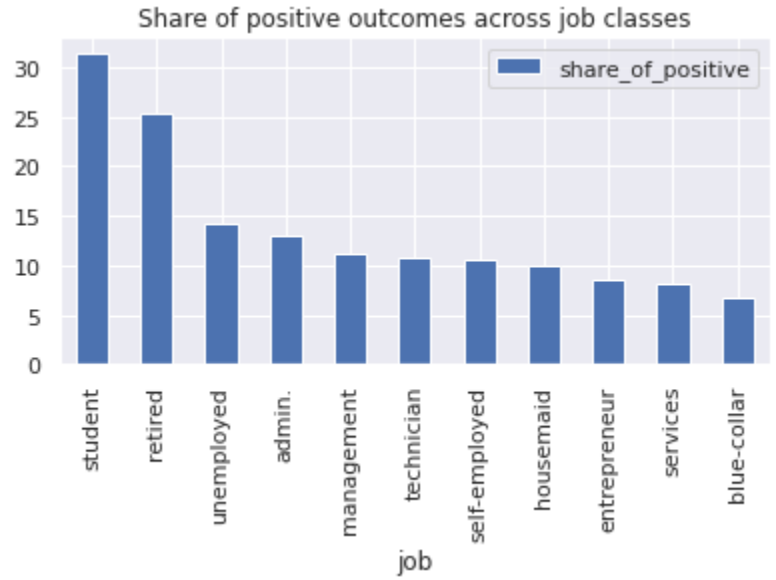
Euribor 3 month ↻ Number of employees

Euribor 3 month ↻ Consumer price index

EDA – explore relations between target and categorical features

Job feature – types of job

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.

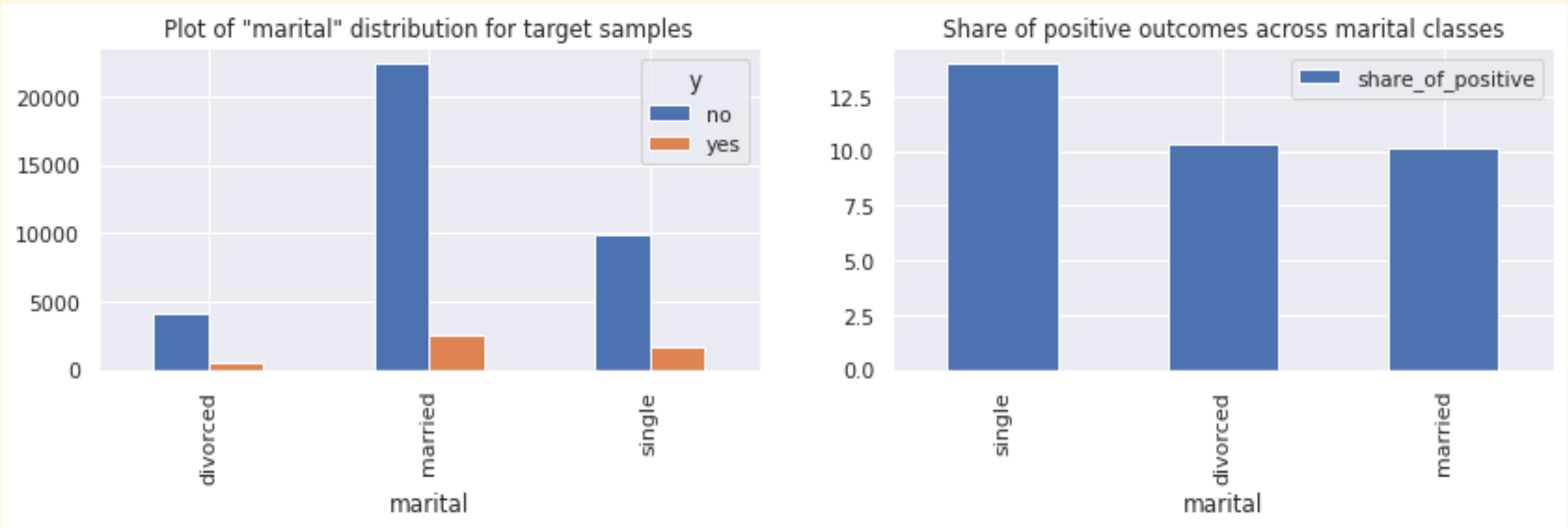


job	share_of_positive
student	31.4286
retired	25.3619
unemployed	14.2012
admin.	13.0385
management	11.2666
technician	10.8300
self-employed	10.4856
housemaid	9.8973
entrepreneur	8.5165
services	8.1366
blue-collar	6.8375

As we can see, ratio between positive and negative answers varies for different job types. 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 – customer's marital status

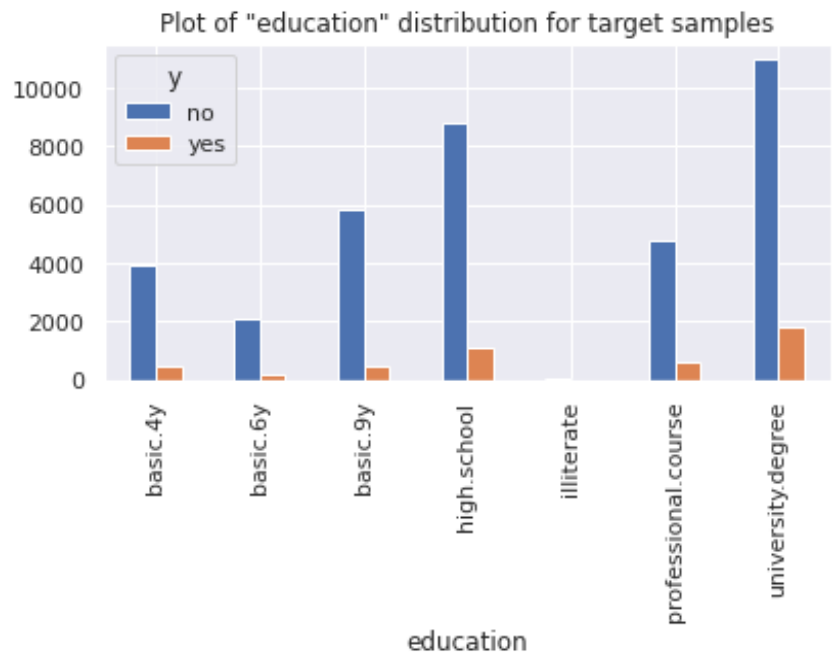


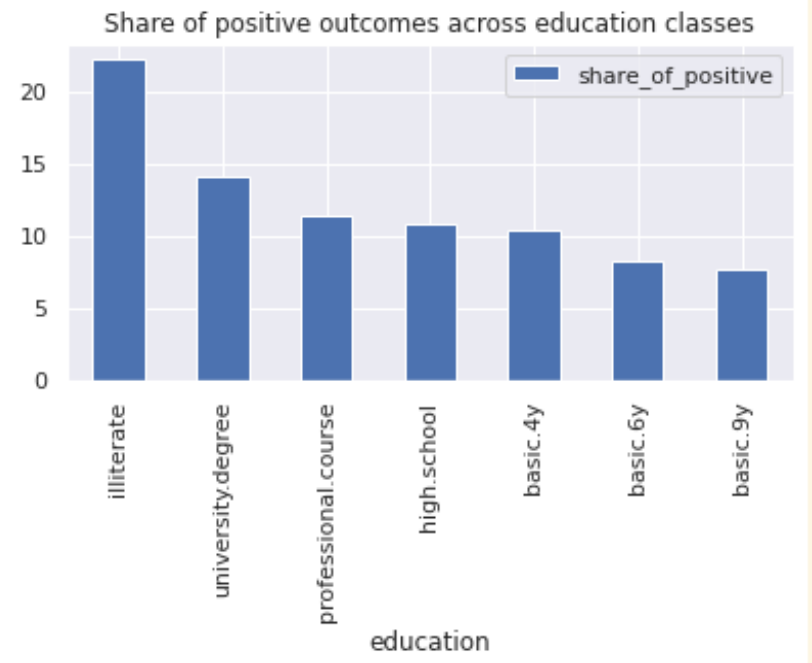
Marital status is also connected with y because single customers are more inclined to give a positive answer and subscribe to the term deposit. However there are much less divorced customers than married ones, shares of positive outcomes is almost equal.

marital	share_of_positive
single	14.0093
divorced	10.3359
married	10.1669

EDA – explore relations between target and categorical features

Education feature – customer's education level



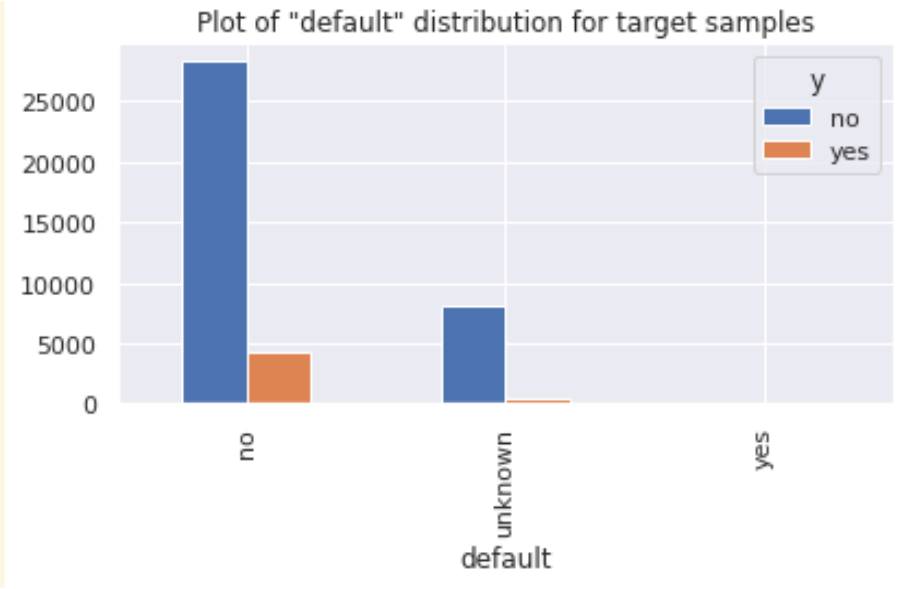


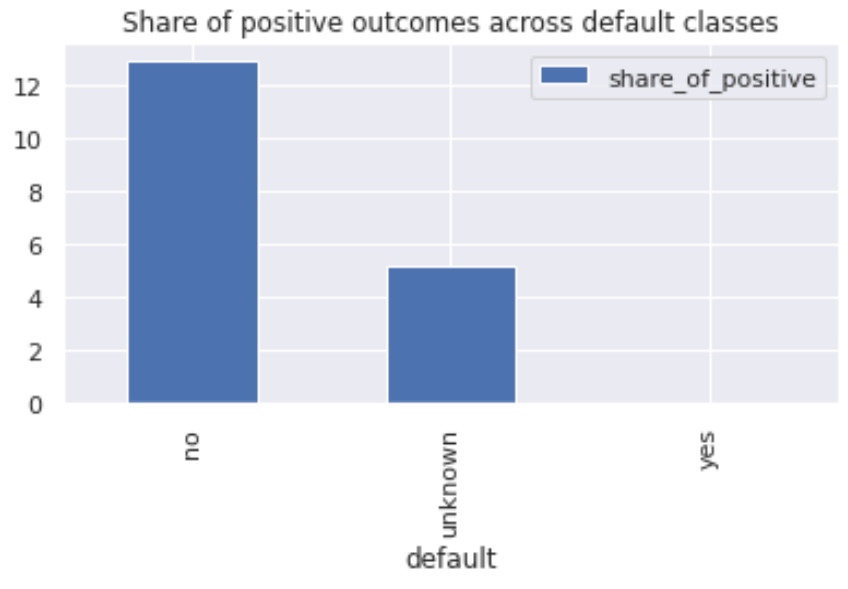
education	share_of_positive
illiterate	22.2222
university.degree	14.1359
professional.course	11.3800
high.school	10.8955
basic.4y	10.4138
basic.6y	8.1861
basic.9y	7.6621

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 – if a customer has a credit in default or not



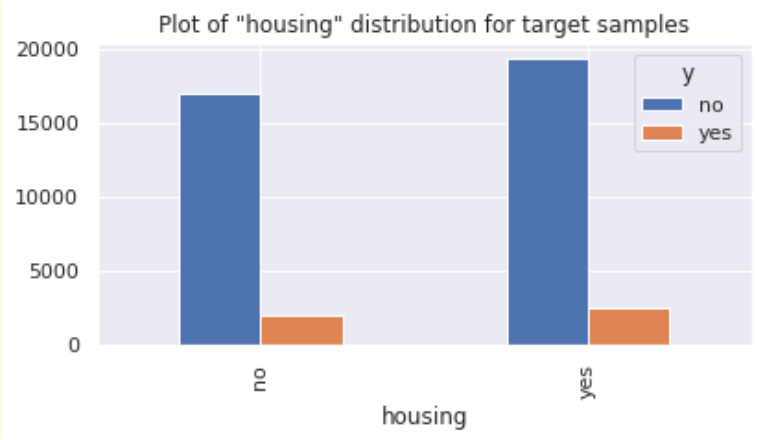


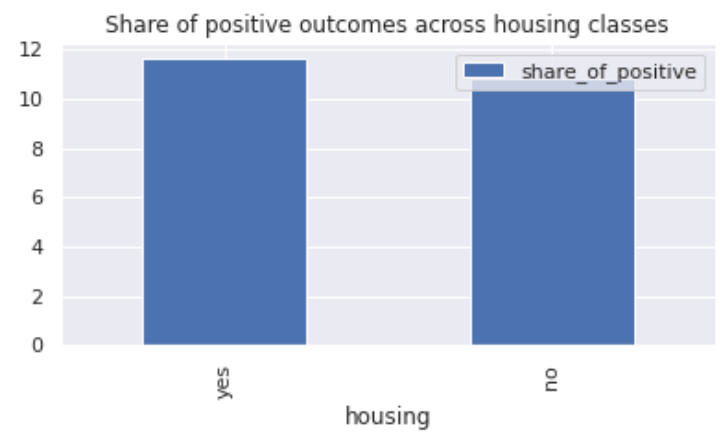
default	share_of_positive
no	12.8803
unknown	5.1536
yes	NaN

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

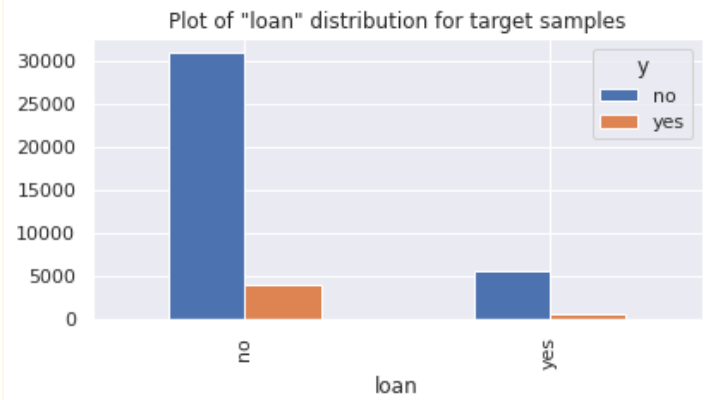
Housing and loan features – if a customer has housing loan/ personal loan or not

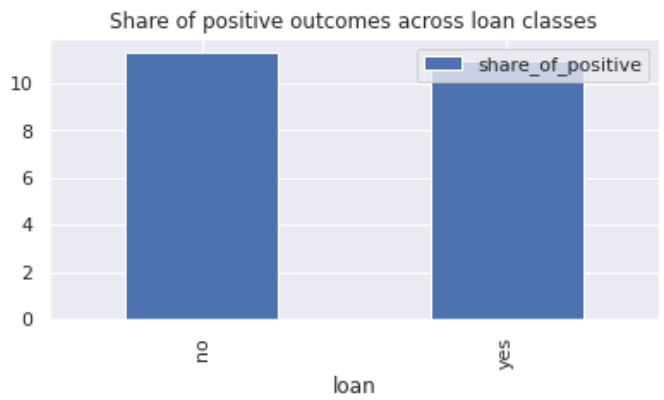




housing	share_of_positive
yes	11.6631
no	10.8110

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.



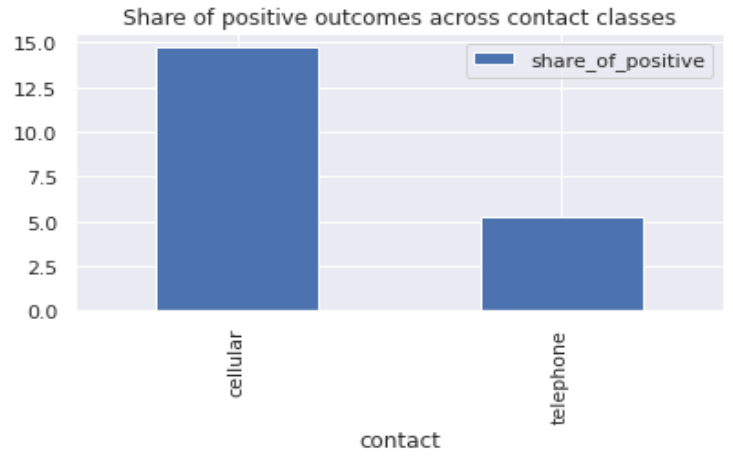


loan	share_of_positive
no	11.3268
yes	10.9280

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.

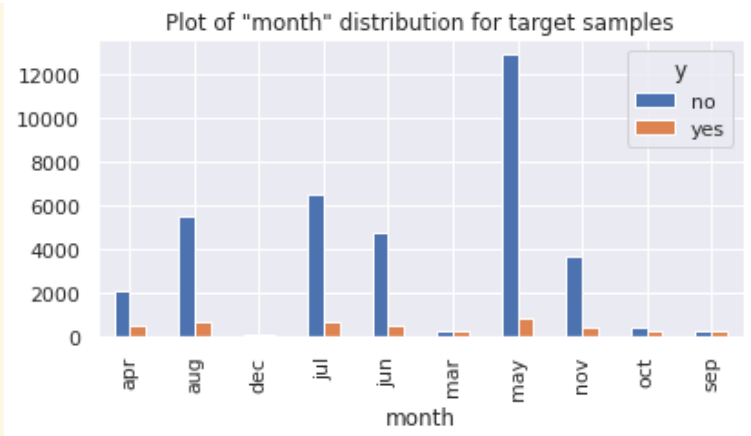
EDA – explore relations between target and categorical features

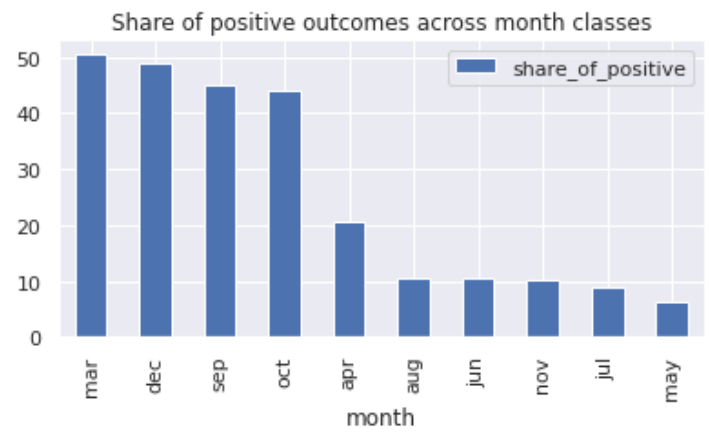
Contact and month features – devise used for contact and a month of a last contact



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.



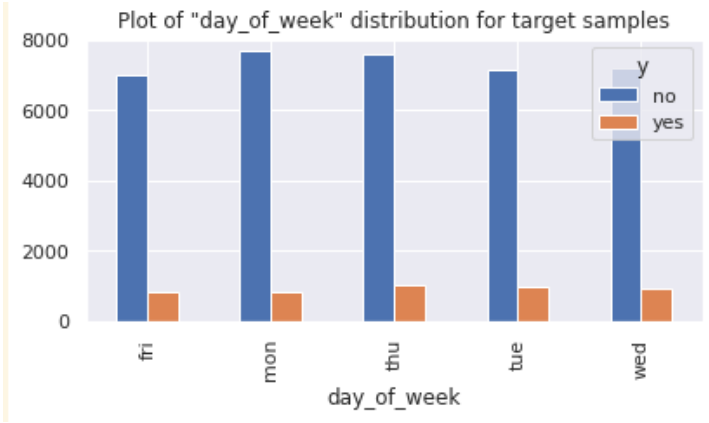


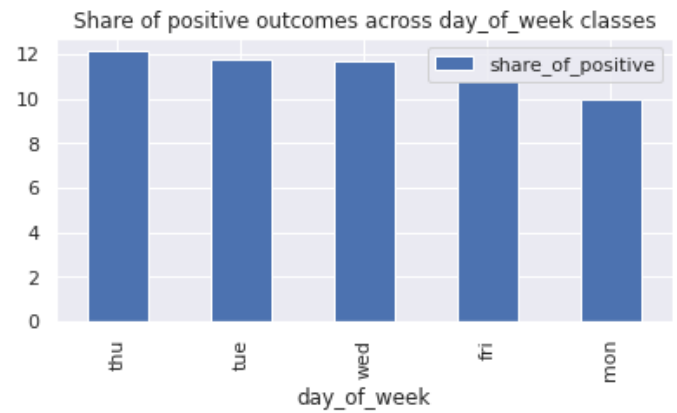
month	share_of_positive
mar	50.5495
dec	48.9011
sep	44.9123
oct	43.9331
apr	20.4865
aug	10.6056
jun	10.5115
nov	10.1463
jul	9.0389
may	6.4357

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

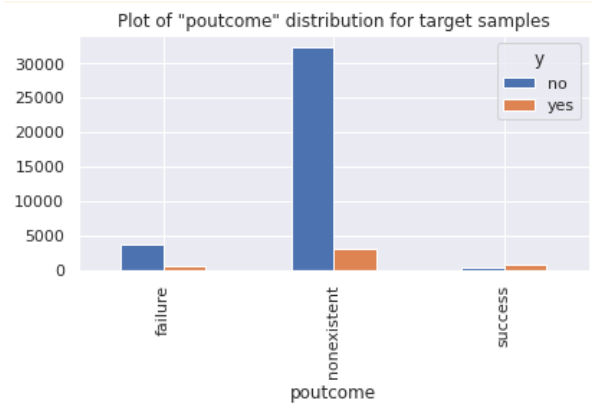
Day of week and poutcome features - last contact day of the week and result of previous marketing campaign

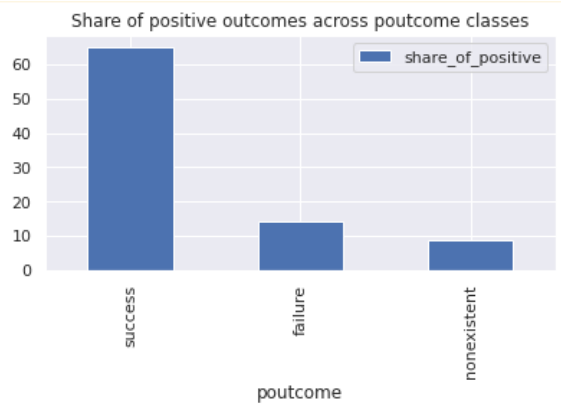




day_of_week	share_of_positive
thu	12.1142
tue	11.7858
wed	11.6671
fri	10.8101
mon	9.9507

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.



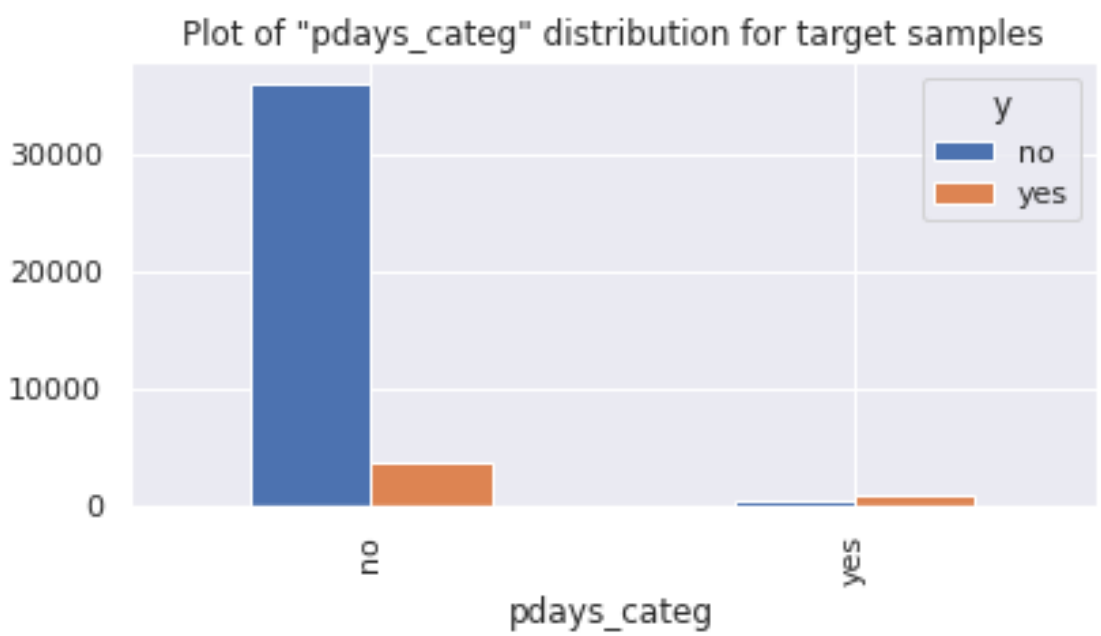


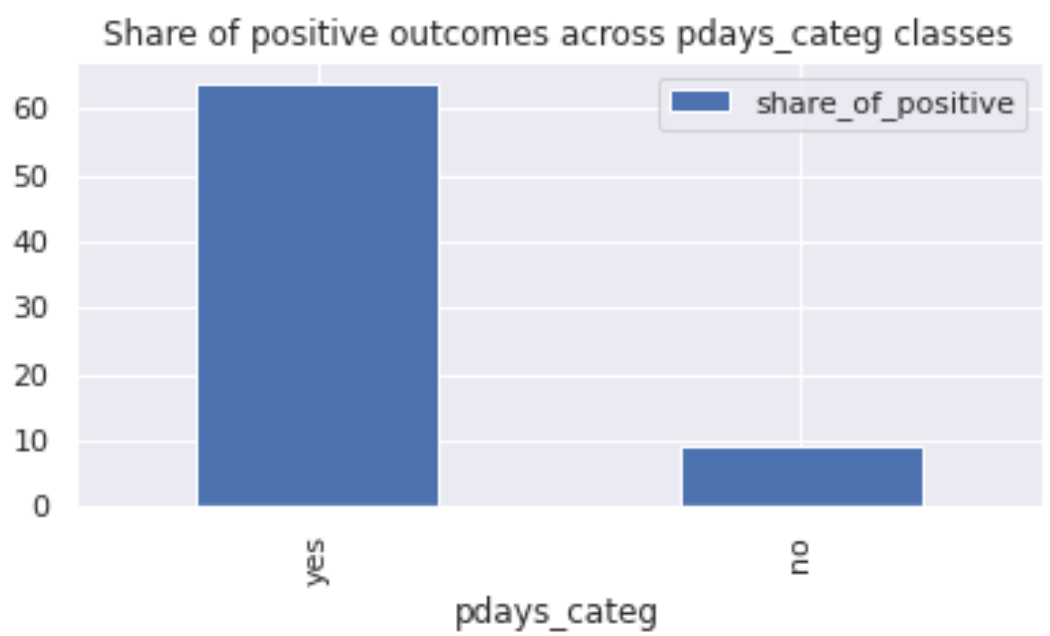
poutcome	share_of_positive
success	65.1129
failure	14.2286
nonexistent	8.8324

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 – if customer was contacted before or not





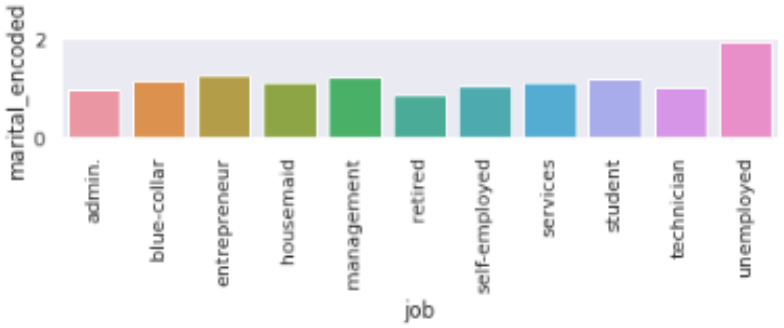
Finally, the fact that the client was contacted earlier also affects the target. This created feature is quite close to the previous one in meaning, that's why we should try to build the model with this feature and without it to see its personal impact.

pdays_categ	share_of_positive
yes	63.8284
no	9.2585

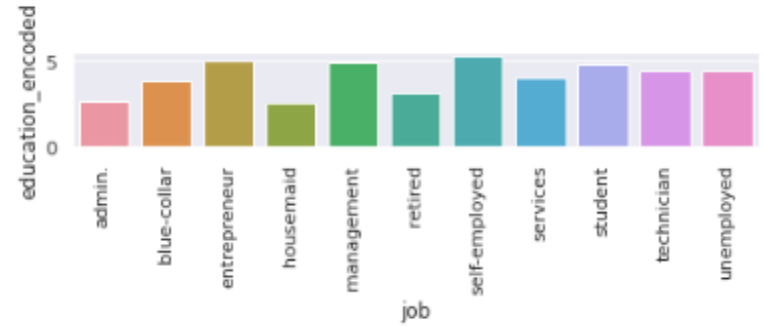
EDA – explore relations between categorical features

Job feature – types of job

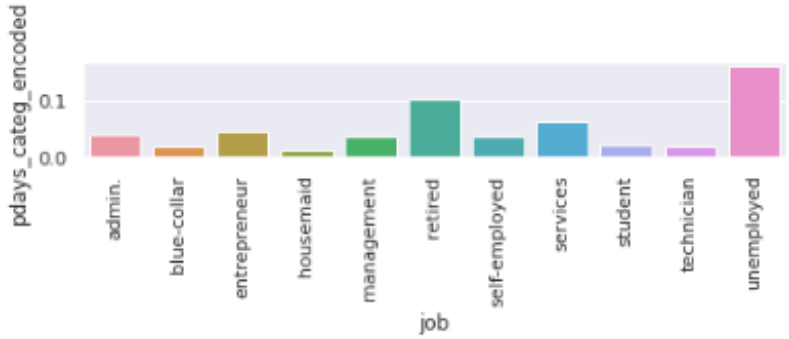
Additionally to the main analysis of relationships between the features and the target we also can explore relationships between the categorical features using label encoding and visualization. We will explore just several features not to be repetitive.



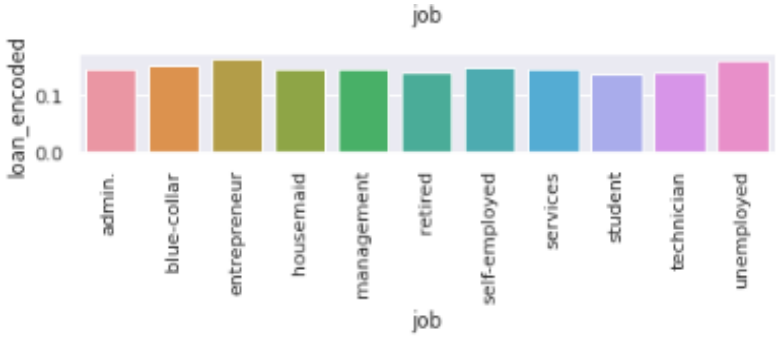
more unemployed tend to be singles



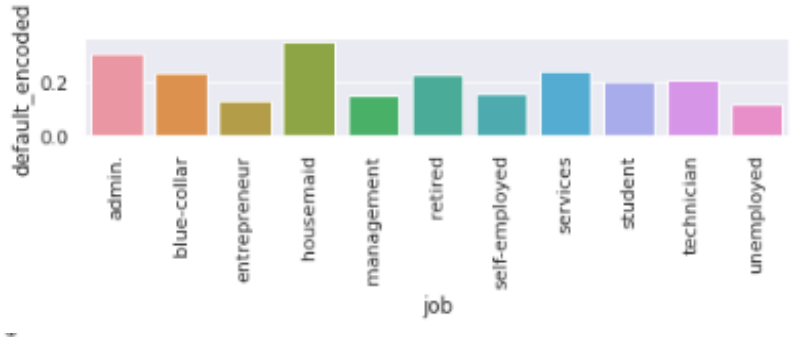
entrepreneurs, management and self-employed have a higher education level



unemployed and retired were contacted before more often



entrepreneurs and unemployed have a little higher probability to have a loan

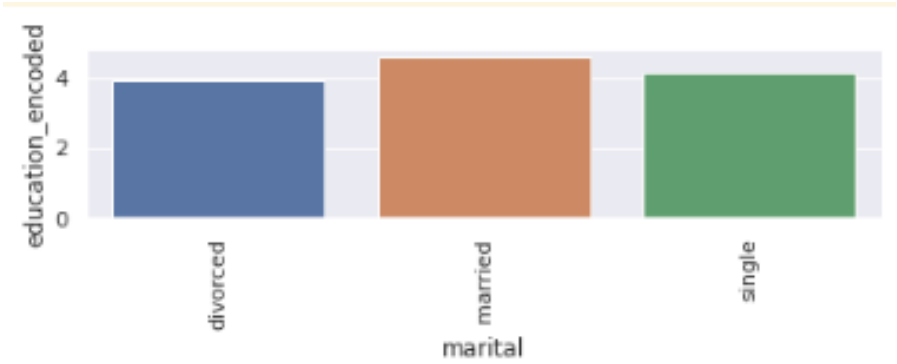


housemaids are more prone to have a credit in default

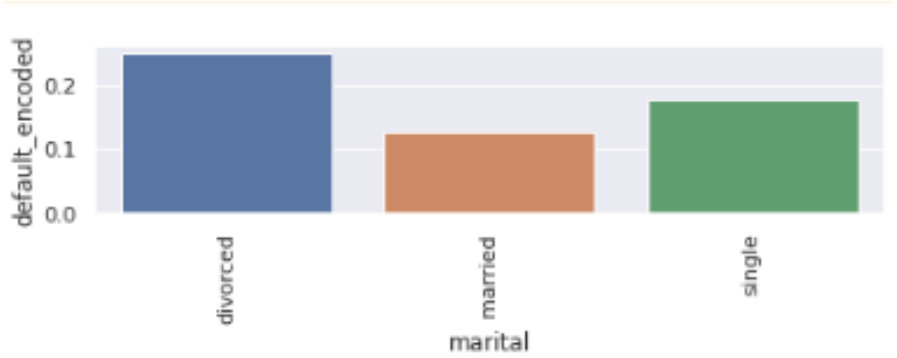
EDA – explore relations between categorical features

Marital feature – customer's marital status

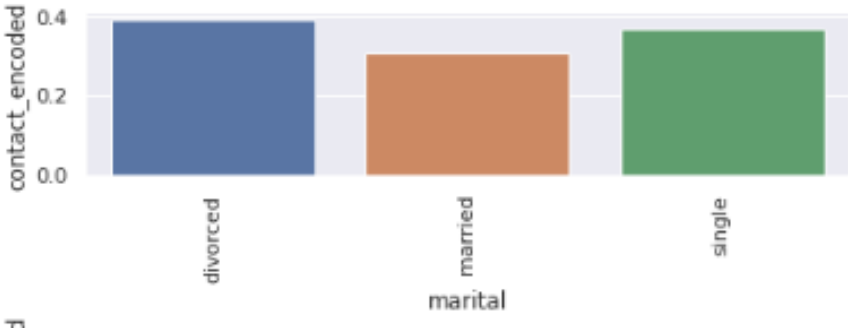
Additionally to the main analysis of relationships between the features and the target we also can explore relationships between the categorical features using label encoding and visualization. We will explore just several features not to be repetitive.



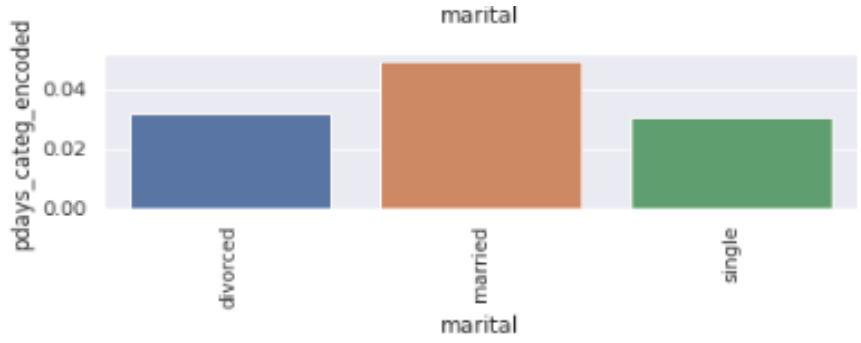
married in average have a little bit higher education level



married are less prone to have a credit in default



married are contacted by cellular more often

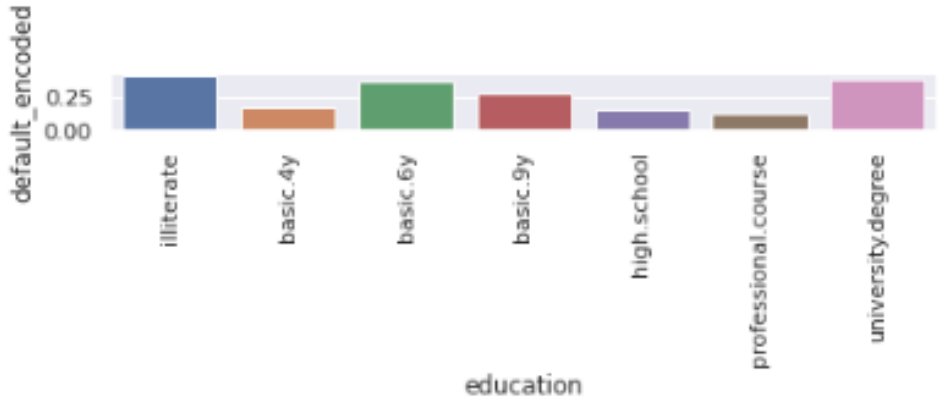


married were previously contacted more often

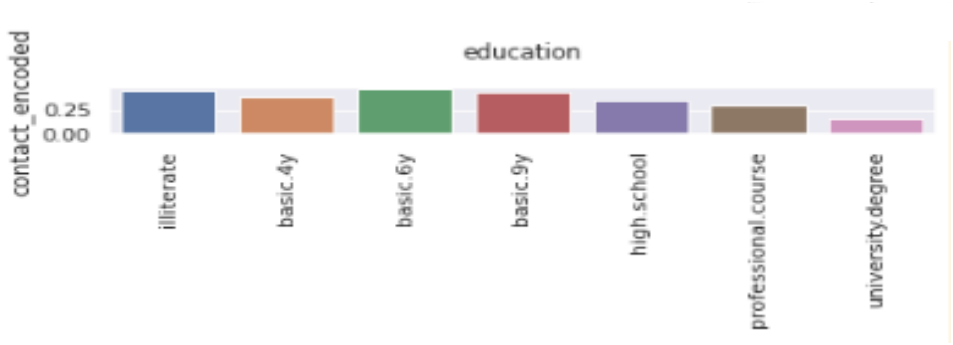
EDA – explore relations between categorical features

Education feature – customer's education level

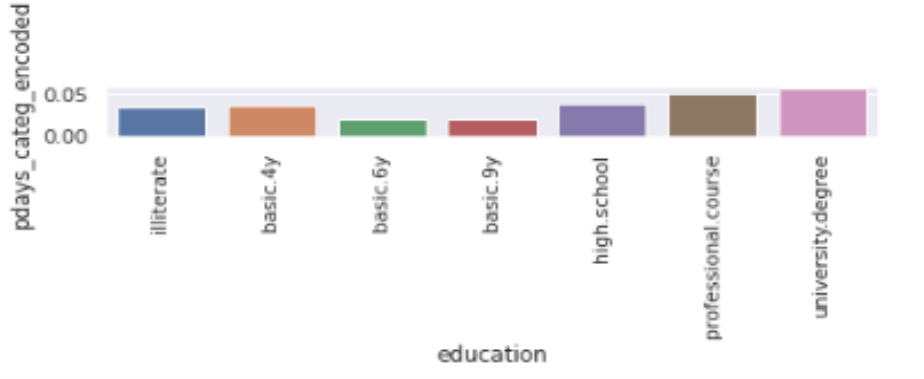
Additionally to the main analysis of relationships between the features and the target we also can explore relationships between the categorical features using label encoding and visualization. We will explore just several features not to be repetitive.



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

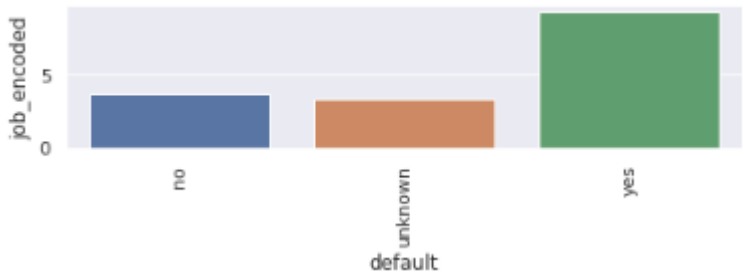


customers with university degree and professional course were contacted before more often

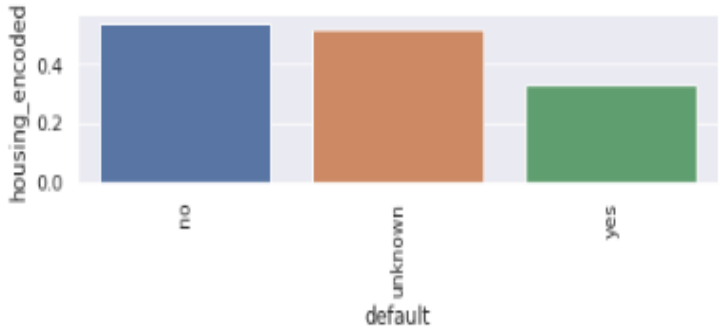
EDA – explore relations between categorical features

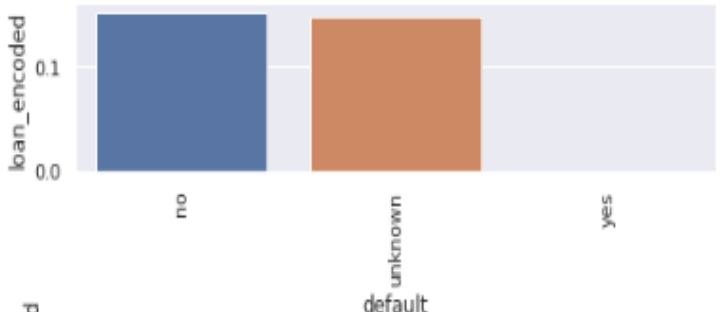
Default feature – has a credit in default or not

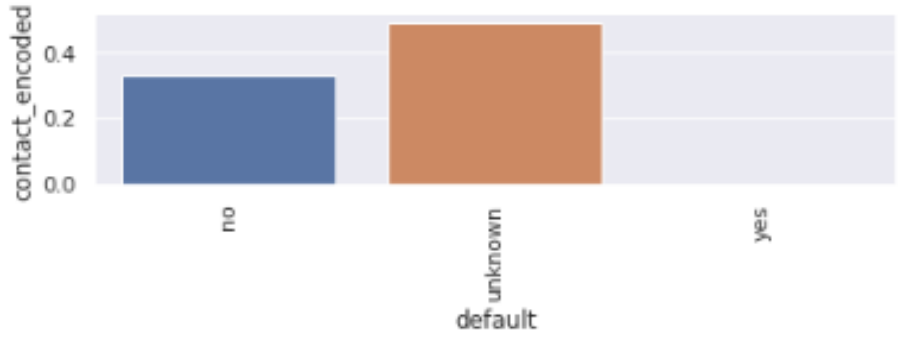
Additionally to the main analysis of relationships between the features and the target we also can explore relationships between the categorical features using label encoding and visualization. We will explore just several features not to be repetitive.

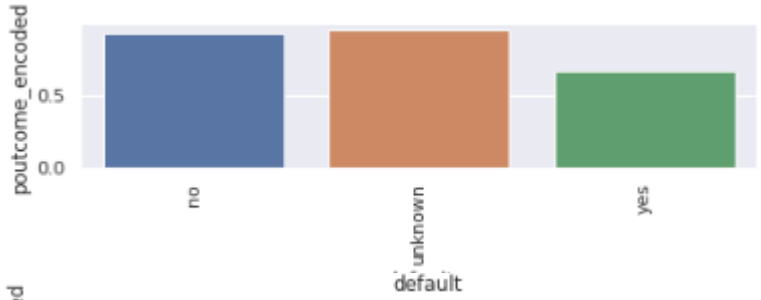


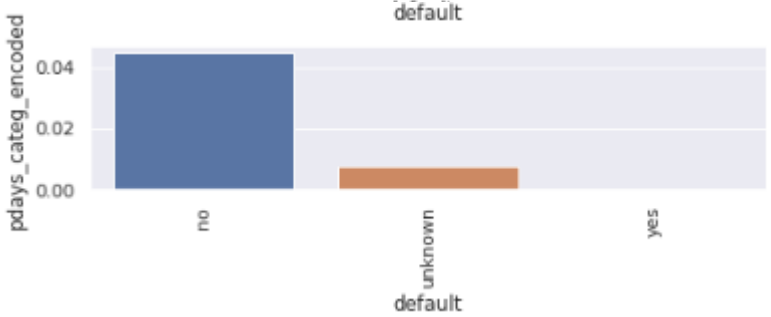
unemployed have a higher probability to have a credit in default











people with a credit in default had a successful outcome in a previous marketing campaign less often

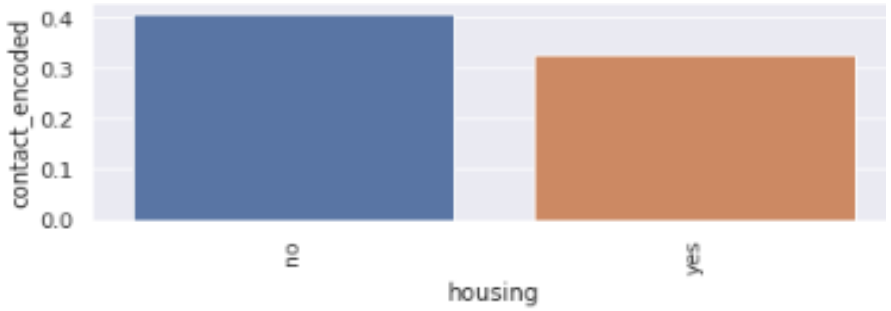
people with a credit in default weren't contacted before

people with a 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

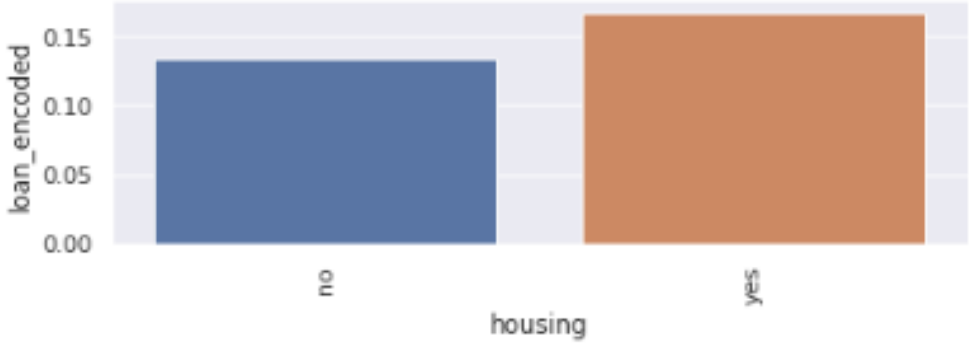
EDA – explore relations between categorical features

Housing feature – if a customer has housing loan or not

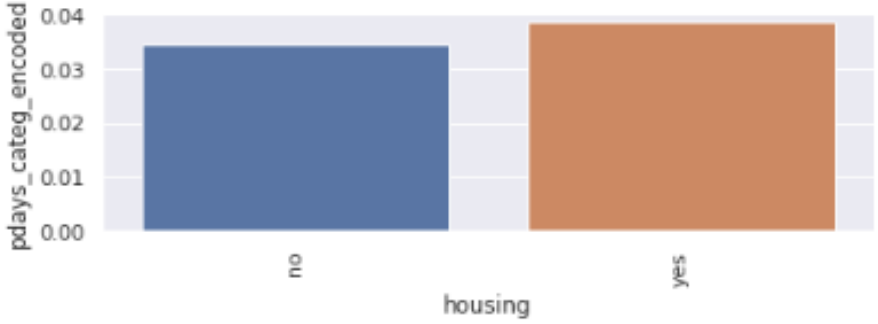
Additionally to the main analysis of relationships between the features and the target we also can explore relationships between the categorical features using label encoding and visualization. We will explore just several features not to be repetitive.



people with a housing loan were contacted by cellular more often



people with a housing loan have a personal loan with a higher probability

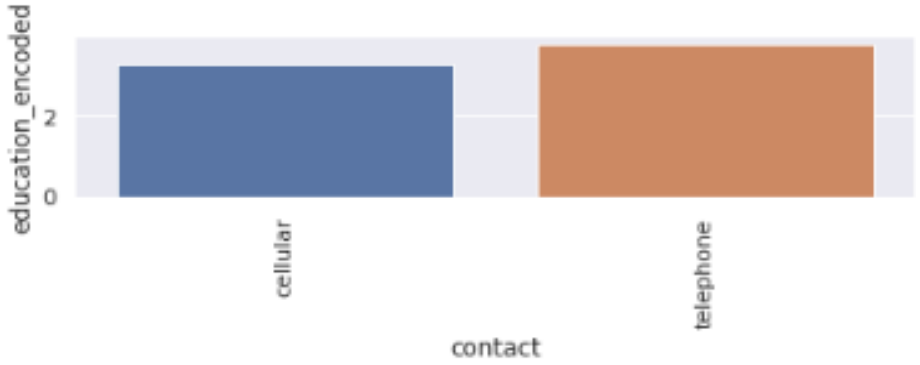


people with a housing loan were contacted before more often

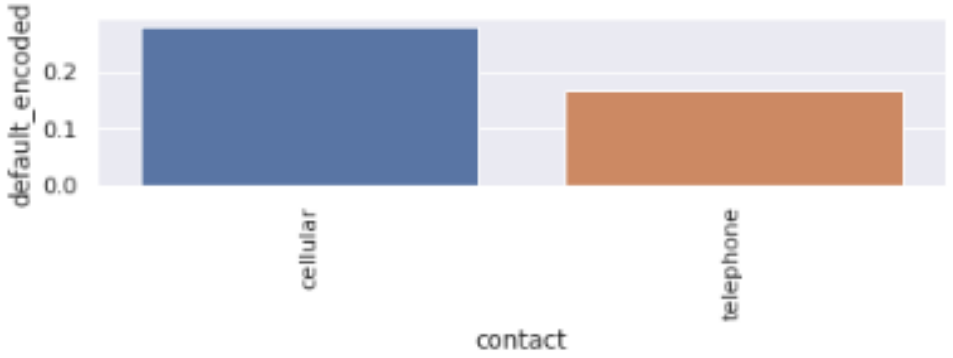
EDA – explore relations between categorical features

Contact feature – devise used for contact

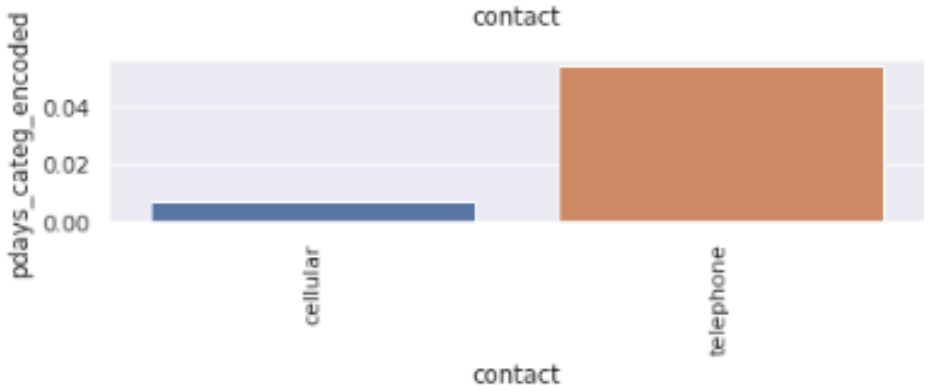
Additionally to the main analysis of relationships between the features and the target we also can explore relationships between the categorical features using label encoding and visualization. We will explore just several features not to be repetitive.



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

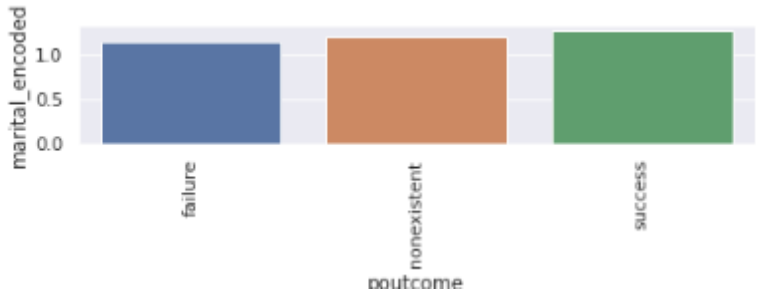


people who was contacted by cellular in average were contacted before less often

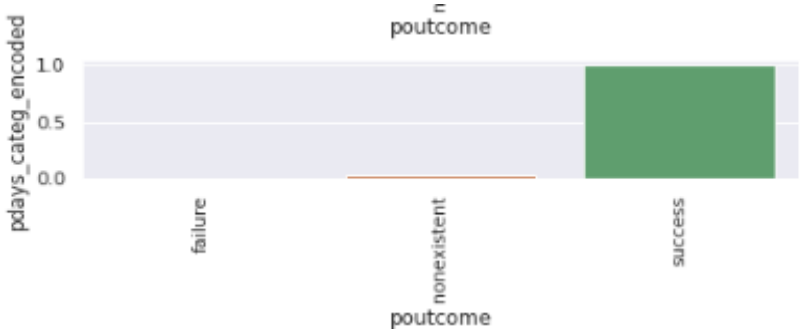
EDA – explore relations between categorical features

poutcome feature - result of previous marketing campaign

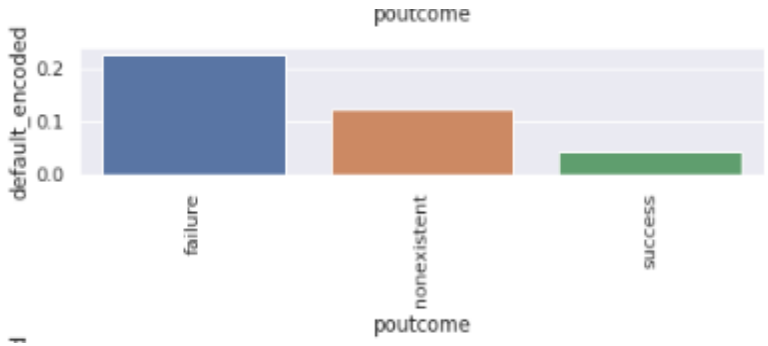
Additionally to the main analysis of relationships between the features and the target we also can explore relationships between the categorical features using label encoding and visualization. We will explore just several features not to be repetitive.



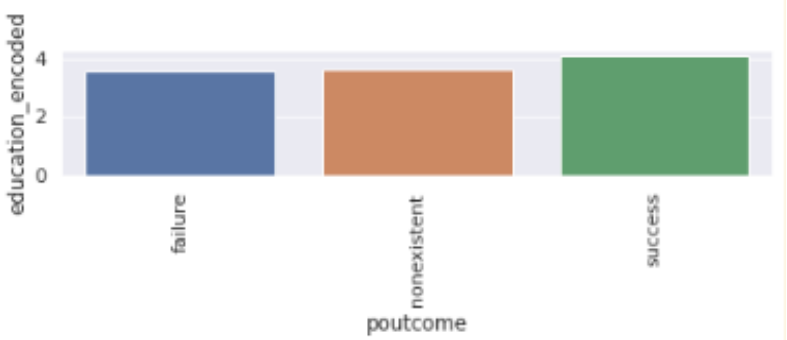
singles had a successful outcome of the previous campaign more often



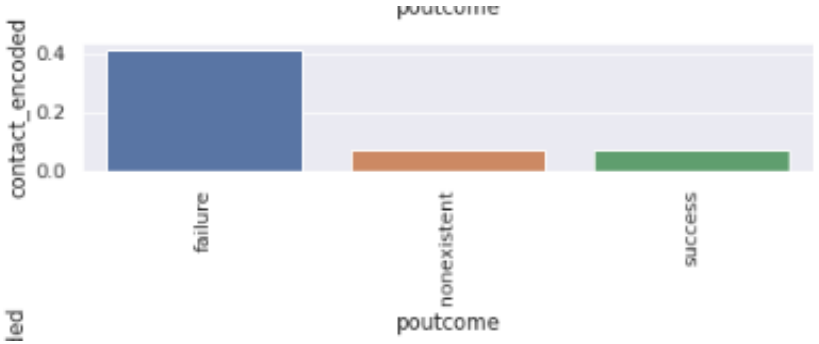
people with a successful campaign outcome were contacted more often



people with a credit in default had an unsuccessful outcome of the previous campaign more often



people with a higher education level had a successful outcome of the previous campaign more often



people contacted by telephone had an unsuccessful outcome of the previous campaign more often

Preliminary findings

- Duration of a call for negative answers is much less, customers who answered 'no', were contacted more often.
- Vice versa, previously they were contacted less often than customers, who answered 'yes'.
- Negative answers subgroup has a higher employment variance rate but lower short-term lending rates.

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.

Recommended models

- Type of problem – classification.
- Type of classification – binary classification.
- Type of learning – supervised learning.

Logistic regression

Estimates the probability of an event occurring. Since the outcome is a probability, the dependent variable is bounded between 0 and 1.

Decision Tree

Creates a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

Random Forest

Operates by constructing a multitude of decision trees at training time. For classification tasks, the output is the class selected by most trees.

Neural Network

K Nearest Neighbors

Attempts to determine what group a data point is in by looking at the data points around it.

Support Vector Machine

Finds a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points.

Naive Bayes

Applies Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable.

Bank Marketing (Campaign) - Group Project

Data Glacier

Your Deep Learning Partner

Thank You