Purwadhika Digital Technology School

BANK MARKETING CAMPAIGN

Baginda Mirza
Ilham Narendrodhipo
Ziske Noflianrini

BACKGROUND

A marketing campaign is a set of commercial operations all pursuing the same objective, which may concern the improvement of brand awareness and/or sales objectives. Companies use direct marketing strategies when they target customer segments by contacting them to achieve a specific sales campaign.

Direct marketing ultimately aims to establish cost-effective, two-way and one-to-one communications with individual customers, which is not restricted to the internet.

Before we start, let's try to answer this question:

WHAT IS THE PROBLEM?

PROBLEM STATEMENT

As said before, eventhough this direct marketing aims to establish cost effective however if we call all of our customer without knowing either they will place the deposit or not it's not effective at all.

Also, this kind of marketing could also make some customer feel uncomfortable. So, it is important to make the campaign cost as low as possible and prevent from customers complaint.

OBJECTIVE

According to the problem, we have to know what kind of customer that will subscribe a term of deposit and to train the best machine learning model that is able to predict whether the customers will subscribe a term of deposit or not.

FEATURES DESCRIPTION

Euribor3m : 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.

Cons.conf.index: a survey, that measures how optimistic or pessimistic consumers are regarding their expected financial situation.

Attribute	Data Type, Length	Description
Age	Integer	Age of customers
Job	Text	Type of customer's job
Marital	Text	Customer's marital status
Education	Text	Customer's education level
Default	Text	Do the customer credit card default or not?
Housing	Text	Do the customer have housing loan?
Loan	Text	Do the customer have personal loan?
Contact	Text	Contact communication type
Month	Text	Last contact month of year
Day of week	Text	Last contact day of the week
Duration	Integer	Last contact duration
Campaign	Integer	Number of contacts performed during this campaign and for this client
Pdays	Integer	Number of days that passed by after the client was last contacted from a previous cam
Previous	Integer	Number of contacts performed before this campaign and for this client
Poutcome	Text	Outcome of the previous marketing campaign
Emp.var.rate	Float	Employment variation rate - quarterly indicator
Cons.price.idx	Float	Costomer price index - monthly indicator
Cons.conf.idx	Float	Customer confidence index - monthly indicator
Euribor3m	Float	Euribor 3 month rate - daily indicator
Nr.employed	Float	Number of employees - quarterly indicator

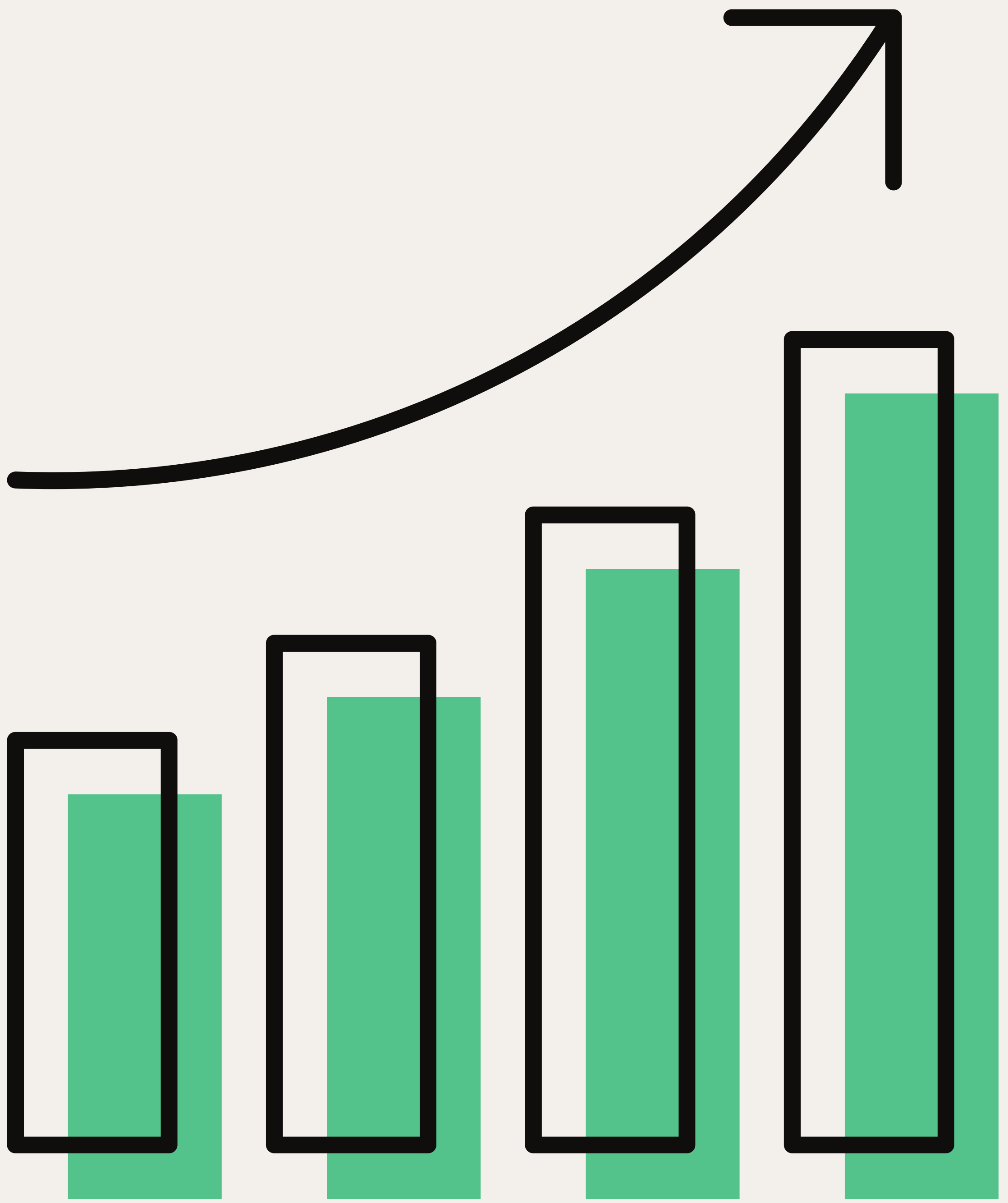
METRIC EVALUATION & CONSEQUENCE

Confusion Matrix Term

- **True Positive (TP)**: customer who are predicted subscribe and actually subscribe
- **True Negative (TN)**: customer who are predicted not subscribe and actually not subscribe
- **False Positive (FP)**: customer who are predicted subscribe and actually not subscribe
- **False Negative (FN)**: customer who are predicted not subscribe and actually subscribe

Consequence

- **False Positive**: The bank will lost money because we contact the wrong customer, we can assume it will cost **500 EUR** for the operational cost.
- **False Negative**: The customer will place the deposit if the bank call them, but the bank doesn't do anything. So, there will be revenue lost, that we can assume around **2000 EUR**.
- We should reduce the number of false negative because the cost from false negative is greater than false positive. We use F-beta score (beta=2) metric because this data is imbalance and we don't want to make a very big gap value in False Negative and False Positive but we want to put more attention on minimizing false negatives than minimizing false positives.



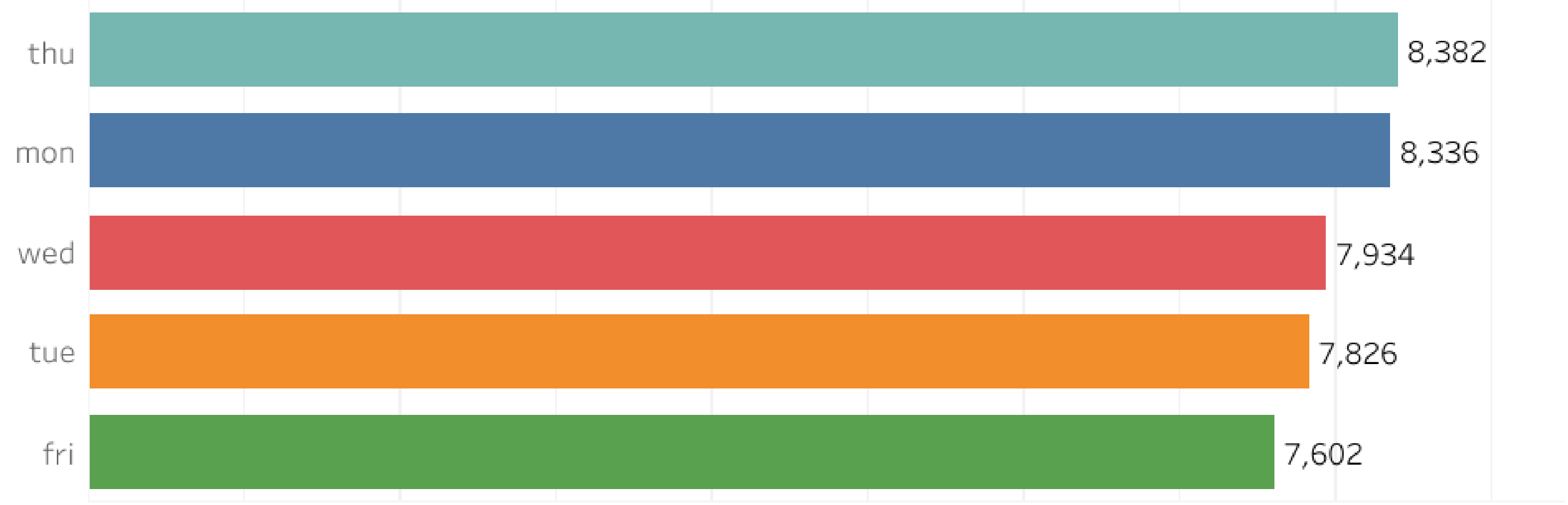
DATA UNDERSTANDING

- Data has 20 features and 1 Target
- Data is Imbalance (88,7% is not subscribe and 11,3% is subscribe)
- There are 12 duplicate data
- There is no missing value but there are "unknown" data that we treat as missing value
- Some numerical features have outliers and we delete it manually so we didn't lost much data
- We removed Duration and Default column

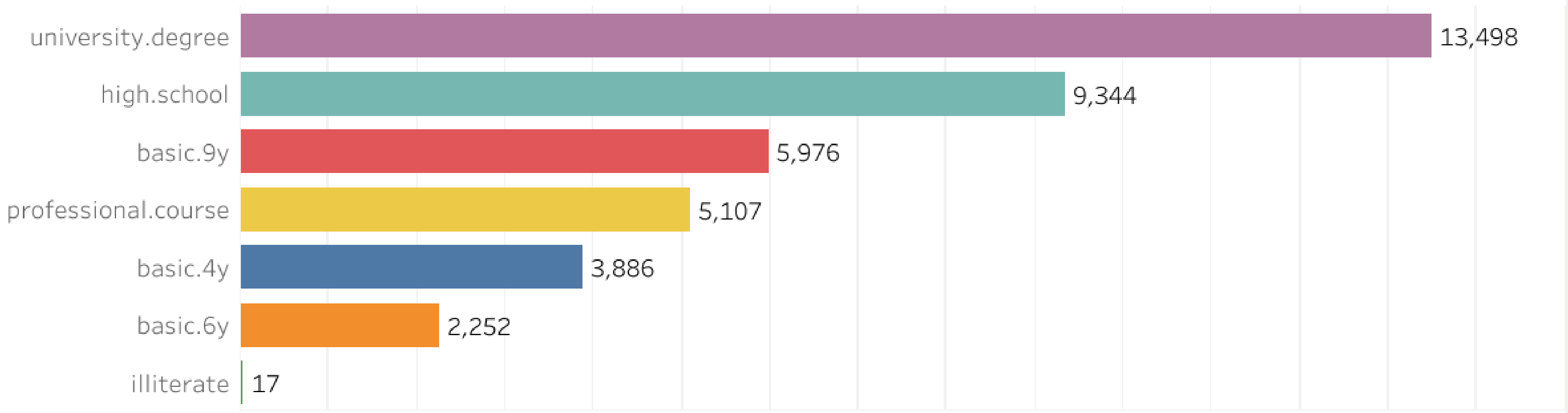
BUSINESS QUESTION

- Does duration and campaign (amount of call) on this campaign have correlation with the customer decision to subscribe ?
- How does social dan economic context related to the customer Subscription?
- Which job type tend to open the term subscribe ?
- Does marital status relate on customer's decision to subscribe ?
- How customer's last education relate on Subscription?
- Does personal loan and housing loan feature has a relationship with the customer decision to subscribe?
- Whether cellphones or telephones should be used to improve consumer interest to subscribe ?
- Do customers tends to subscribe if the outcome on previous campagin is success?
- When is the best time to attract customers, so They decide to subscribe?

Which day is the most busy?



Demography Based on Education



KPI 1

39.56

Avg Age of Target Customers

KPI 2

2.589

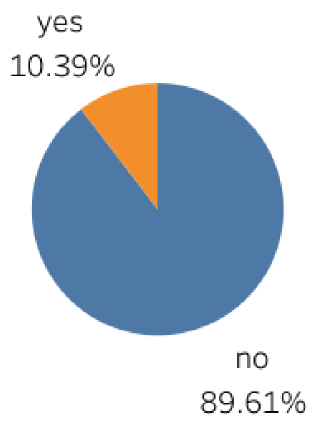
Avg of Telemarketing to Customer

KPI 3

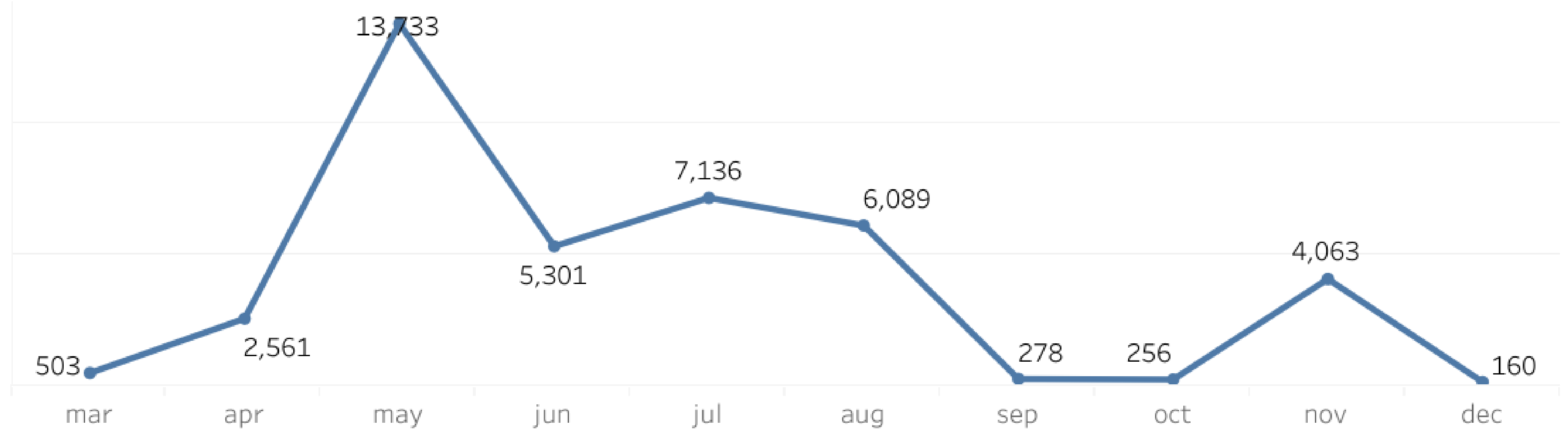
93.60

Average Consumer Price Index

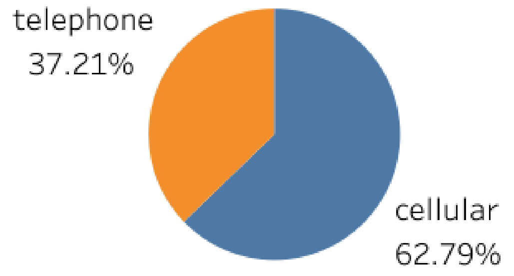
Target (Y)



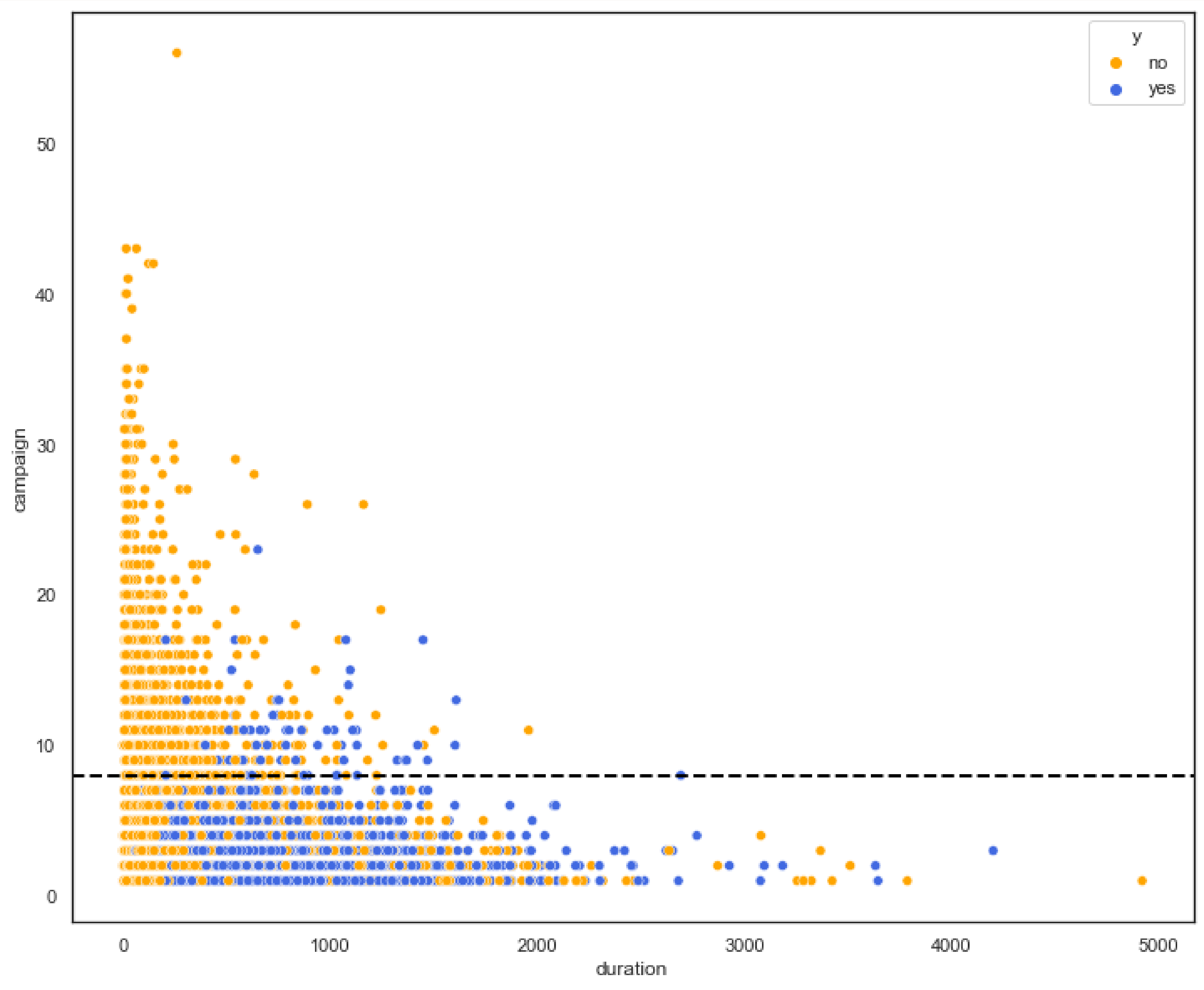
Trend based on month



Contact



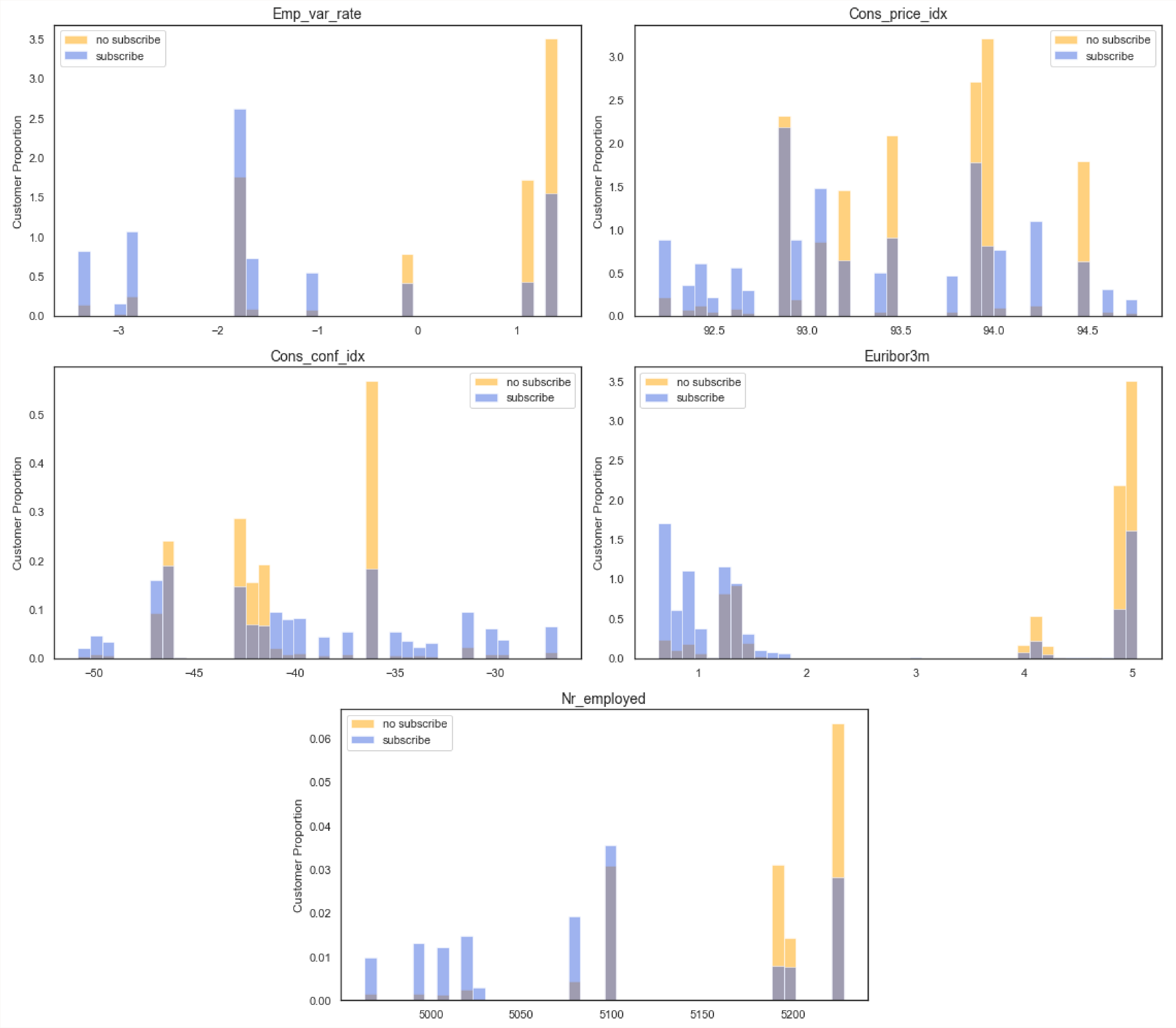
SUBSCRIPTION BASED ON DURATION AND CAMPAIGN



Insight:

- Customers are likely to not subscribe if the call duration is 164 seconds or lower
- Customer has a higher chances to subscribe when the number of contacts performed (campaign) is lower than 8.
- After eight campaign calls, clients are more likely to reject the subscription unless the duration is quite high

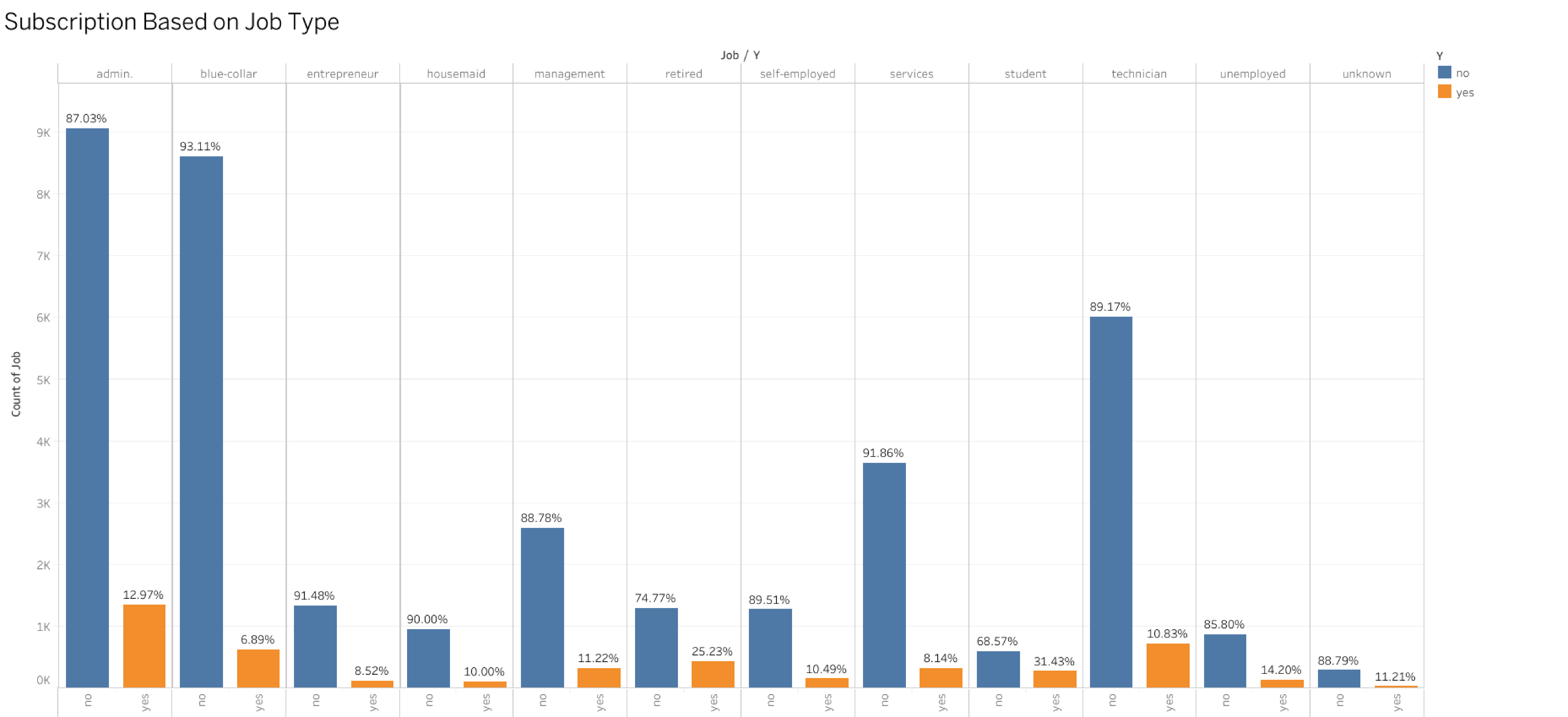
SUBSCRIPTION BASED ON SOCIOECONOMIC CONTEXT



Insight:

- In this dataset, the customer tend to not subscribe with the increases of emp_var_rate, cons_price_idx, euribor3m, nr_employed. Meanwhile, The customer tend to not subscribe with the decrease of cons_conf_idx or there is no certain tendency.

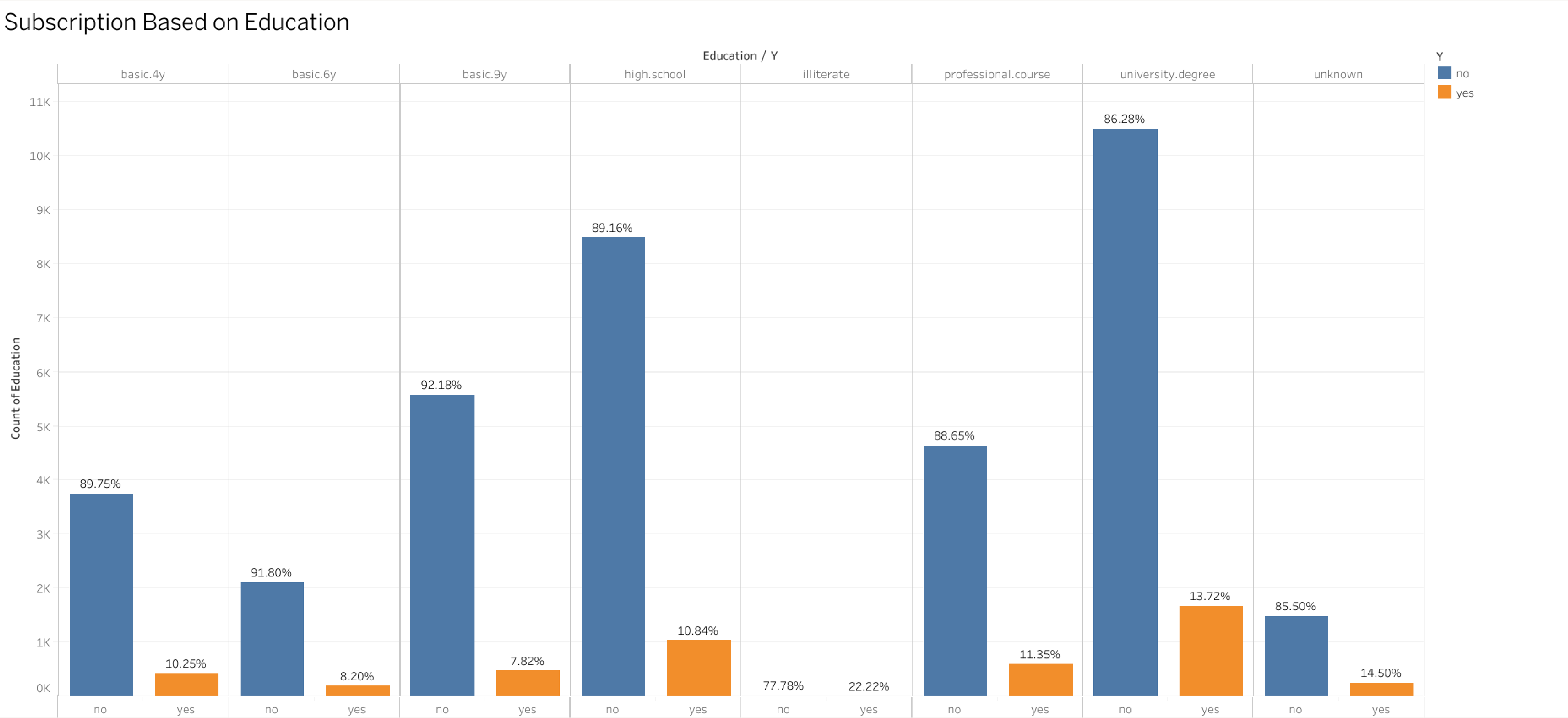
SUBSCRIPTION BASED ON JOB TYPE



Insight:

- Most of the customer that agree to subscribe have a admin type of work (3% of the total data). However, if it is compared to the non-subscribed customer who are working as admin (22% of the total data), this number is really small.
- Job type Student are top job type of customer who tend to agree to subscribe.

SUBSCRIPTION BASED ON EDUCATION



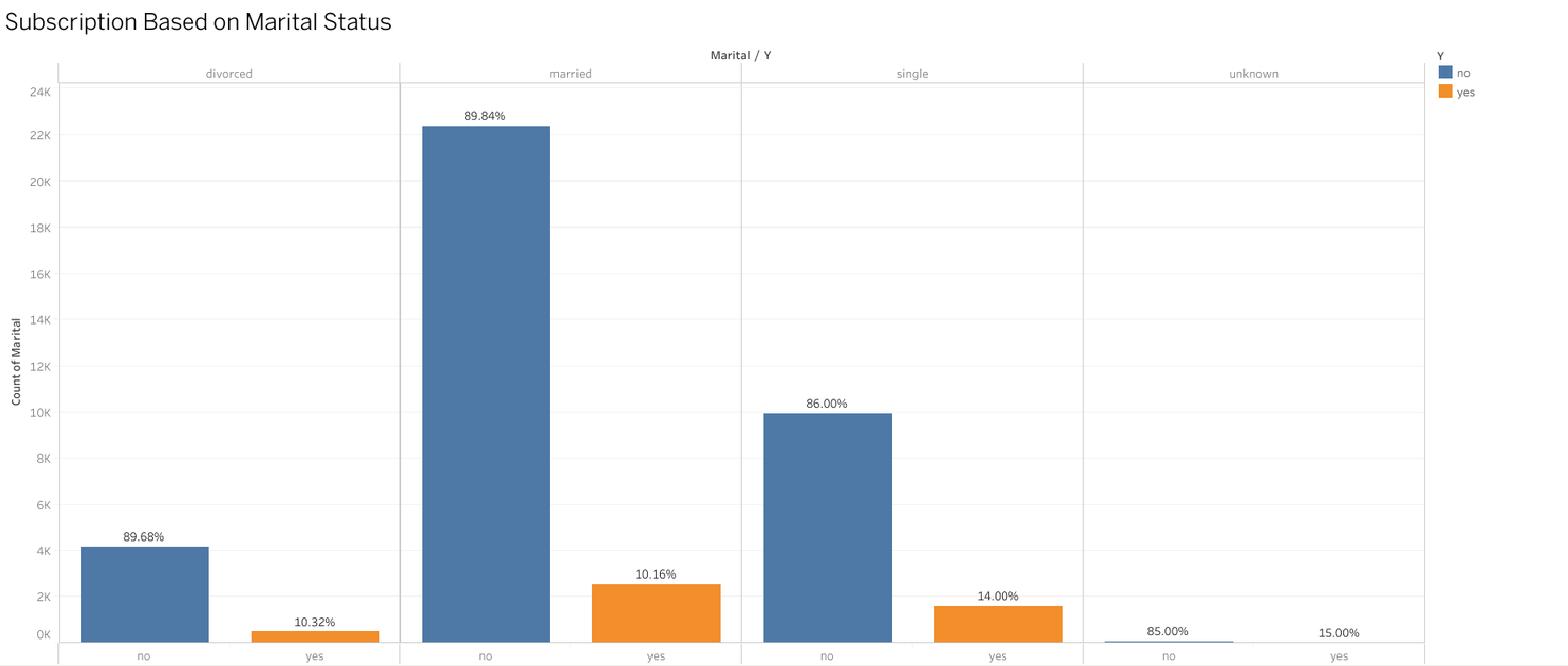
Insight:

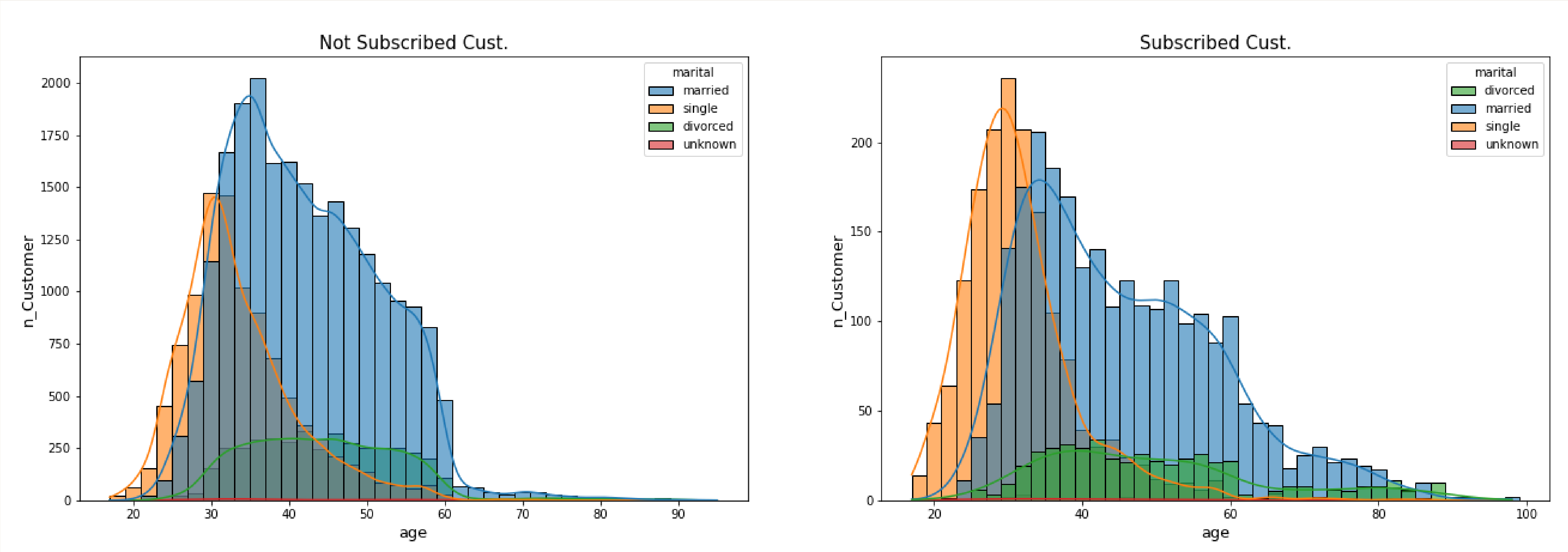
- Customer who are graduated form university has the highest percentage of total customer data.
- Illiterate tend to agree to open subscribe term compare to the other last education. However, customers who the last education is illiterate only taking a small account (0.04%) from total customer data

SUBSCRIPTION BASED ON MARITAL STATUS

Insight:

- Customer who subscribe the most is married customer but this does not mean that this marital status have a higher tendency to subscribe compare to the other marital status.
- Unknown will be treated as missing value. Thus, marital status that tend to subscribe is single (14% of the total data).
- Single customer who are 30 y/o tend to subscribe. On the other hand, Single customer who are 32 y/o tend not to subscribe.
- Married customers who are 43 y/o tend to subscribe. On the other hand, Married customers who are 41 y/o tend not to subscribe.

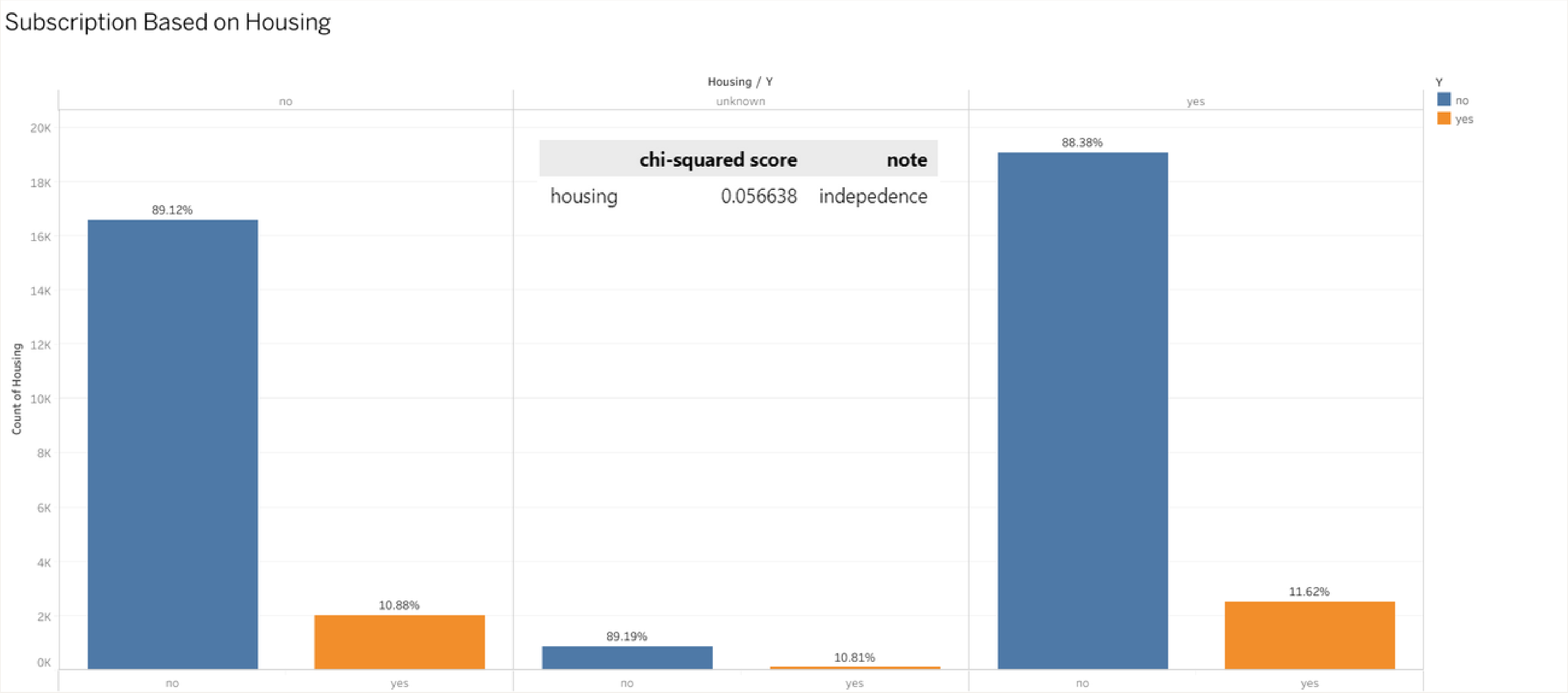


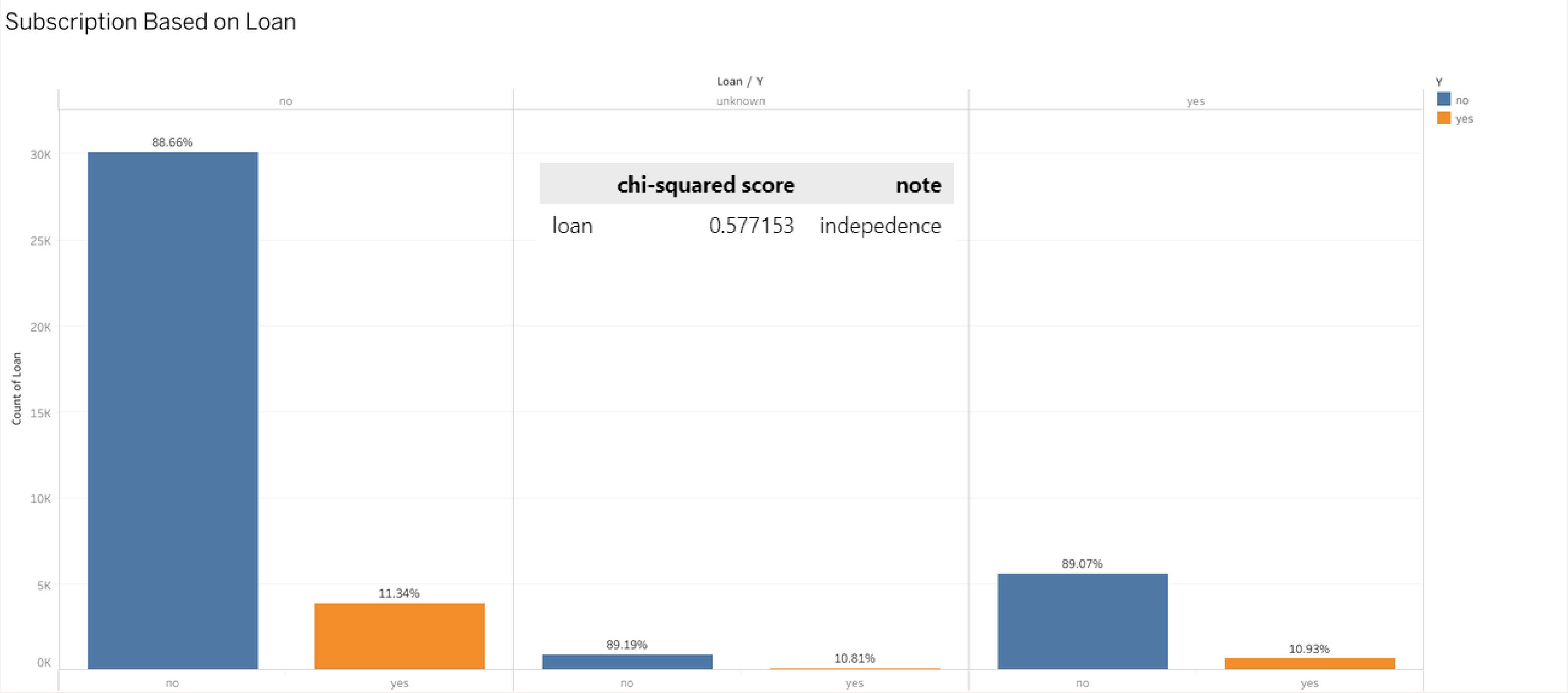


SUBSCRIPTION BASED ON HOUSING AND LOAN

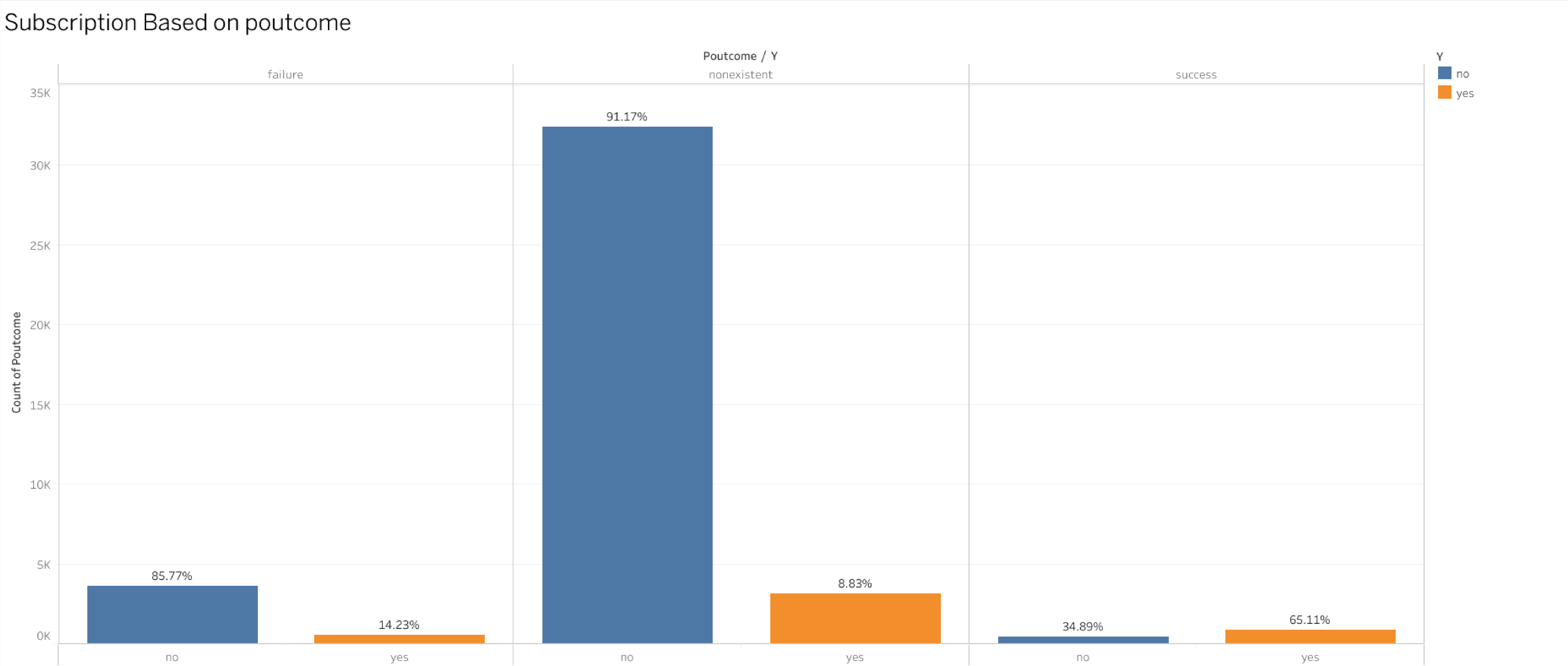
Insight:

- Loan and housing feature are independence towards the target (according to chi-squared test). Thus, this imply that loan and housing feature don't have relationship to the customer decision to subscribe.
- However, it does not mean that we will drop these feature later for the modeling. The reason is because we afraid that this feature will be important if it is combine with other feature.





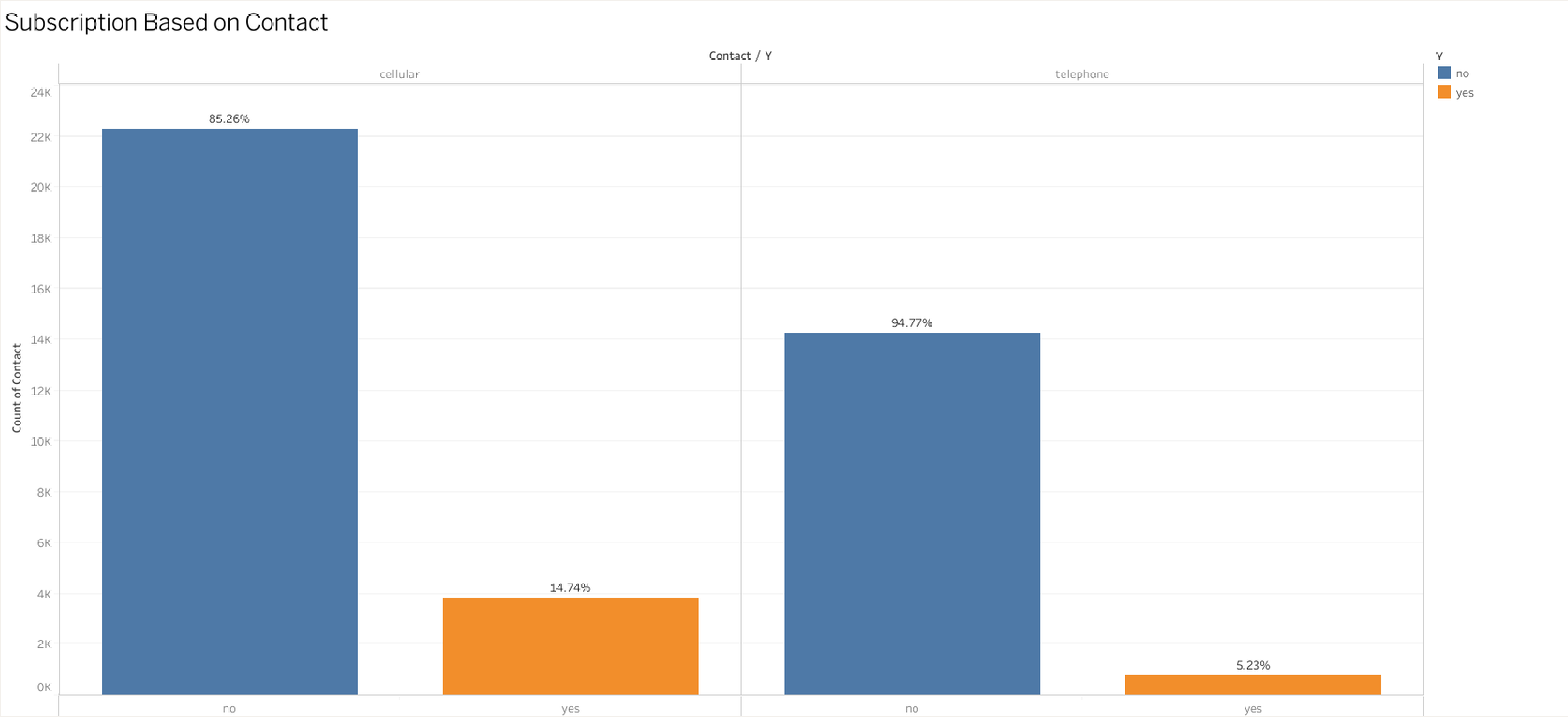
SUBSCRIPTION BASED ON POUTCOME



Insight:

- Customers who were a success in previous campaign are likely to subscribe than a person who was a failure
- 35 % customer campaign who are success in previous campaign decide not to subscribe again

SUBSCRIPTION BASED ON CONTACT



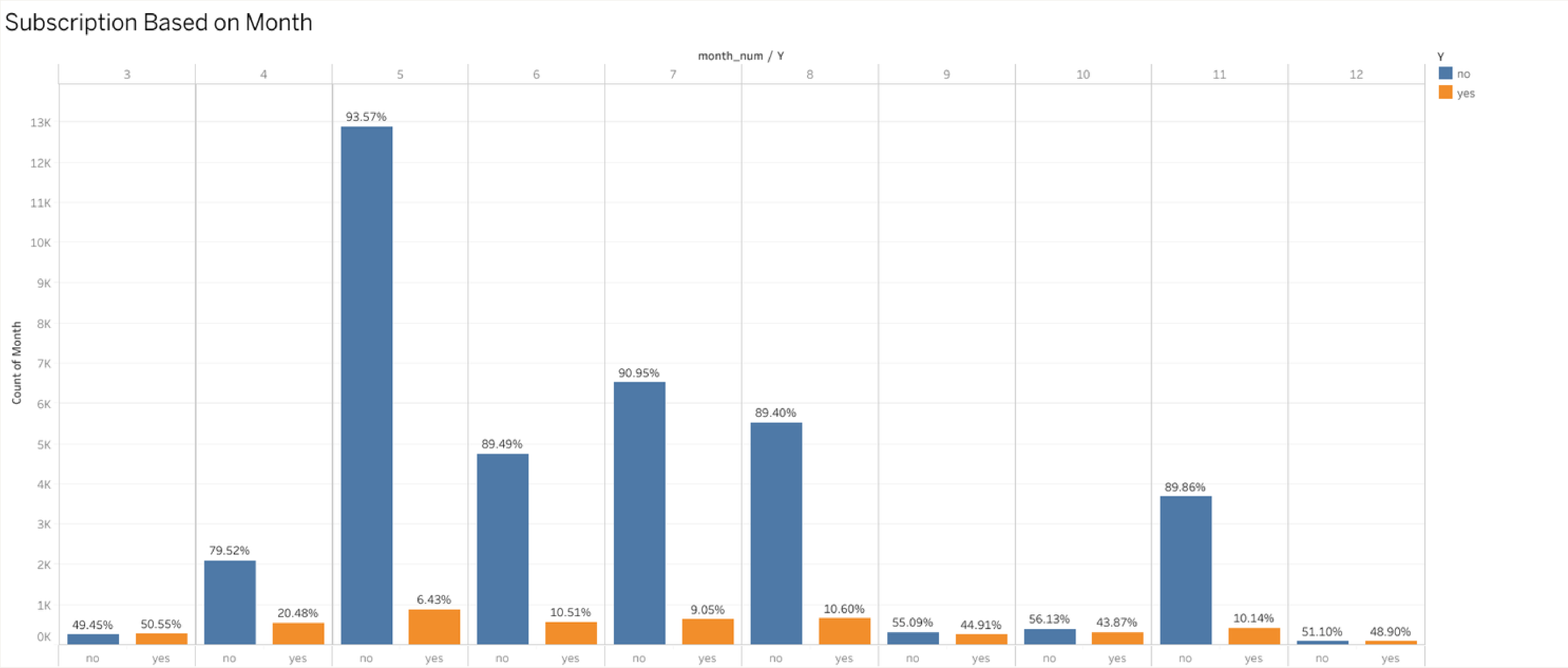
Insight:

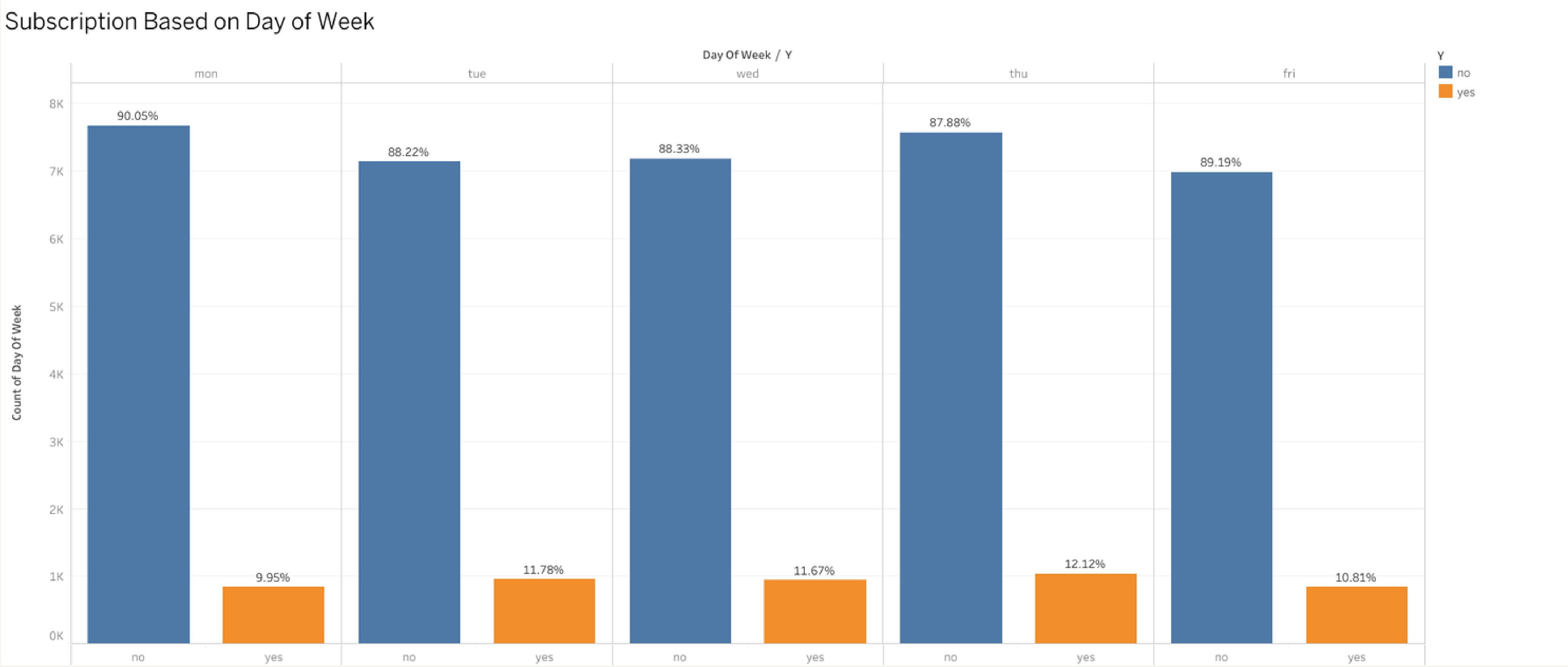
- Campaign mainly use cellular to promote the product
- Most of the customer agree and have a higher tendency to subscribe if they are contacted by cellular rather than telephone.

SUBSCRIPTION BASED ON MONTH AND DAY

Insight:

- According to month, the highest percentage customer that tend to subscribe in the current campaign is on March, followed by December and September.
- According to day, The highest percentage customer that tend to subscribe in the current campaign is on Thursday, followed by Tuesday and Wednesday.





MACHINE LEARNING

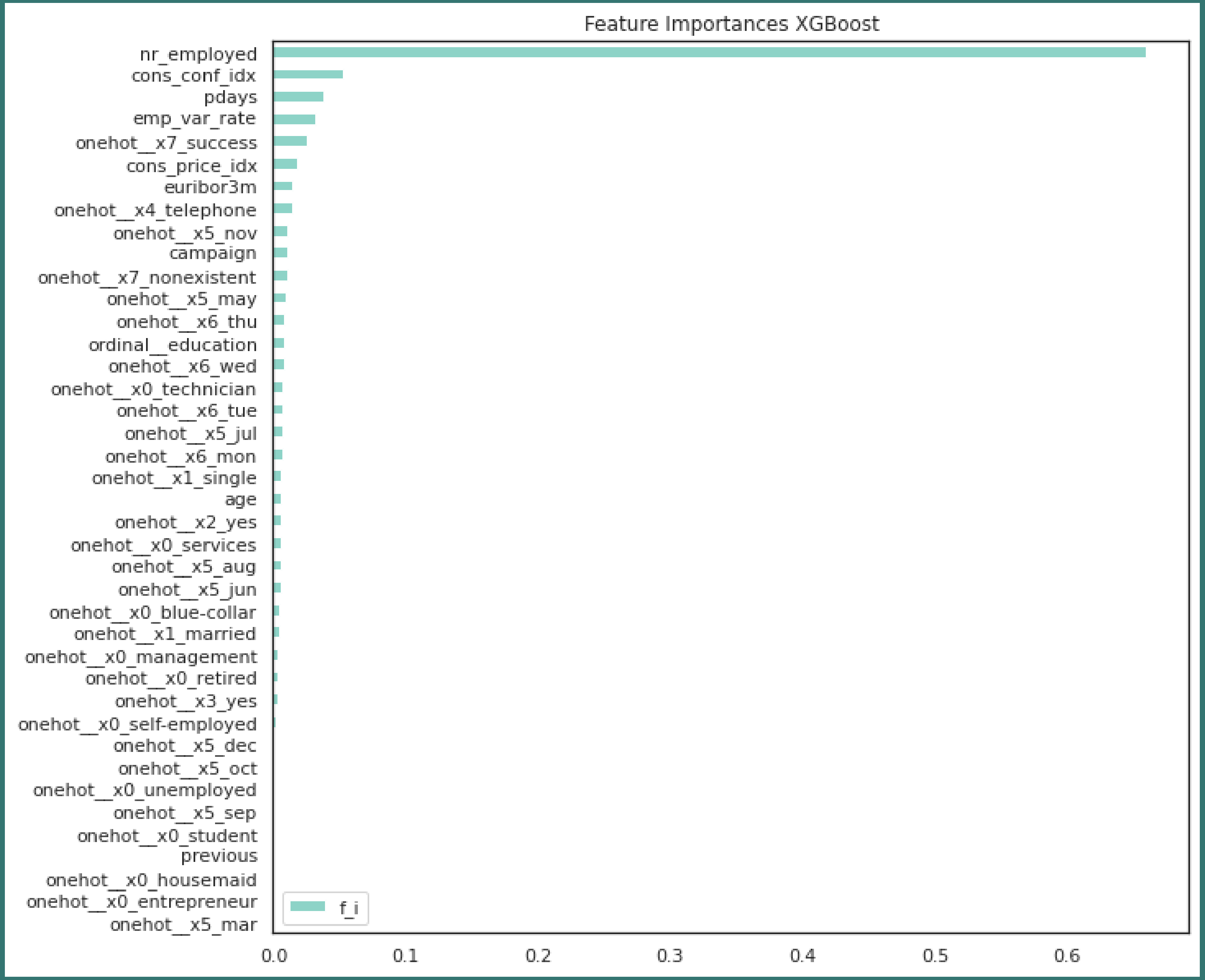
- All categorical columns are encode with one hot encoder, but Education column is encode with Ordinal encoder
- We use RandomUnderSampler for resampler
- Best Models after tuning based on F2 score are
  - XGBoost (0.5309)
  - Random Forest (0.5307)
  - Gradient Boost (0.5275)
- We use XGBoost for the final model

MACHINE LEARNING

- We already trying feature engineering, such as polynomial (orde = 2) and result is not quite different without polynomial.
- We try to use resampler with oversampling (SMOTE) and the result score is lower than using undersampling (RamdomUnderSampler) that we use.
- We try to use some feature importance and the result is not quite different from using all features. So, we decide to use 8 feature importance

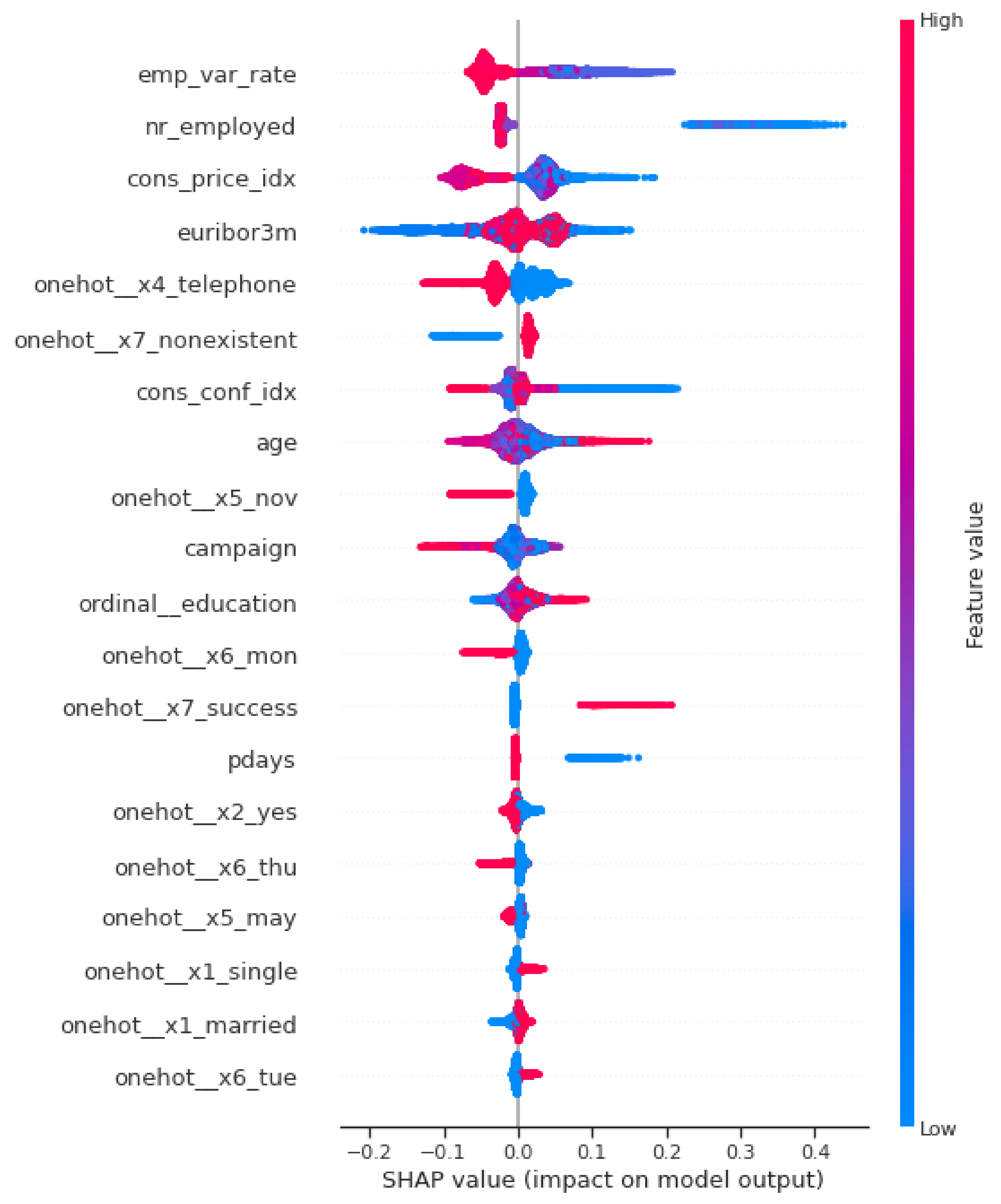
FEATURE IMPORTANCE

- nr_employed
- emp_var_rate
- onehot_x7_success
- cons_conf_idx
- onehot_x5_may
- pdays
- euribor3m
- onehot_x5_oct

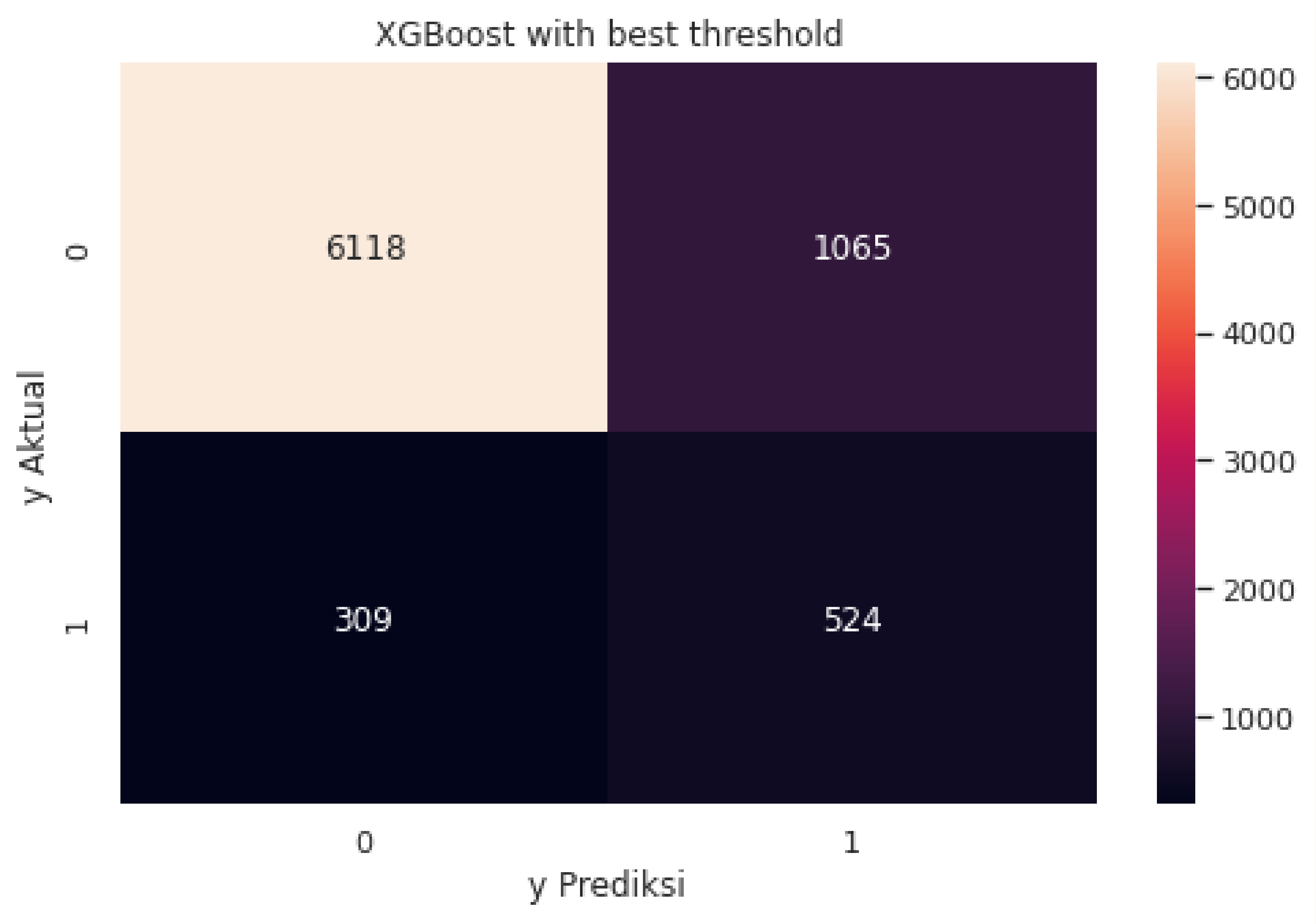


SHAP

- SHAP is used to see the relation from features value and target
- **emp_var_rate**, **nr_employed, cons_price_idx**, the lower the value, the higher tendency for a customer to subscribe.
- **euribor3m**, the lower the value, the customer either subscribe or not and the higher the value there is no tendency
- **Telephone in contact**, if the contact is not telephone (cellular), the customer tend to subscribe
- **Nonexistent in Poutcome**, if the contact is not nonexistent (success or failure), the customer tend to not subscribe



CONFUSION MATRIX



From 20% data test, we can predict:

- True Negative = 6118
- False Positive = 1065
- False Negative = 309
- True Positive = 524

BUSINESS CONCLUSION & RECOMMENDATION

- The best time to approach customer is March, September, and December, especially on Thursday.
- If it is possible call customer by cellular.
- The bank should approach more single customers who are thirty years old .
- The bank should explore more why students and illiterate have high proportion of subscription in job and last education feature.

MODEL CONCLUSION

- The best F2 score is 0.532.

- If we call all of our customer without machine learning, assume that we contact 8016 customer (20% from our data) and we just have 2 results, False Positive and True Positive. So, the total error cost that we lost if we call all the customer:
    - Total cost due to false positive (customer that we call but didn't subscribe) = 3,555,000 EURO
    - Total Customer that have potential to subscribe but didn't subscribe => 0 customer (because we call all of the customer)

- If we use machine learning, we know how many customer that is predicted subscribe and actually not subscribe (False Positive) and customer who are predicted not subscribe and actually subscribe (False Negative). So, the total cost from our error:
    - Error cost = 1,151,500 EURO

- So, we can reduce the error cost:

  Reduce error cost = 2,403,500 EURO or around 67.6%

MODEL RECOMMENDATION

- Our data is highly imbalance, so it is better to gather more positive class data to reduce the imbalance data.
- Add more features that likely have correlation to the target to improve the metrics score.
- Try to use another machine learning algorithm (such as SVM) and hyperparameter tuning again.

THANK YOU!