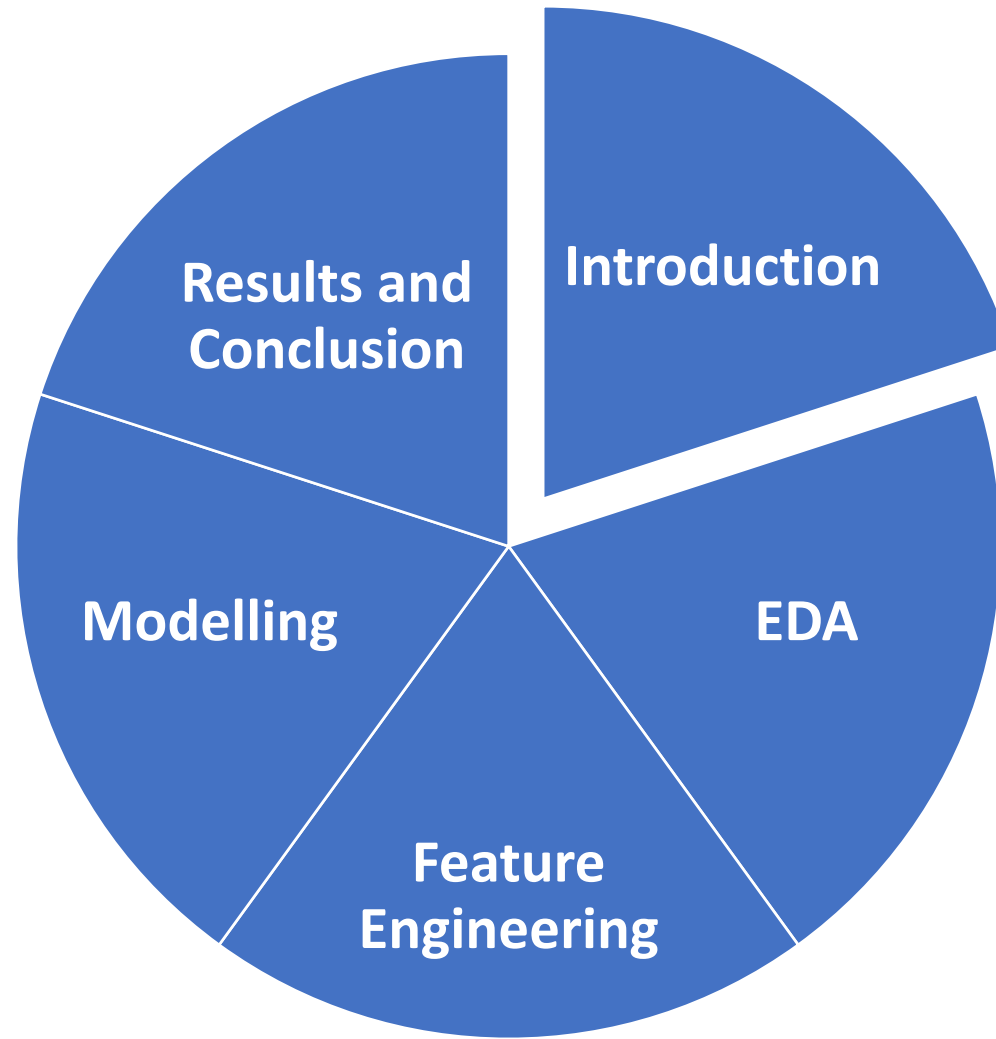


Bank Marketing Campaign Prediction

- **Name:** Xiaoying Yang
- **Email:** xiaoyingyy97@outlook.com
- **Country:** United States
- **College:** University of Illinois at Urbana Champaign
- **Specialization:** Data Science



Content

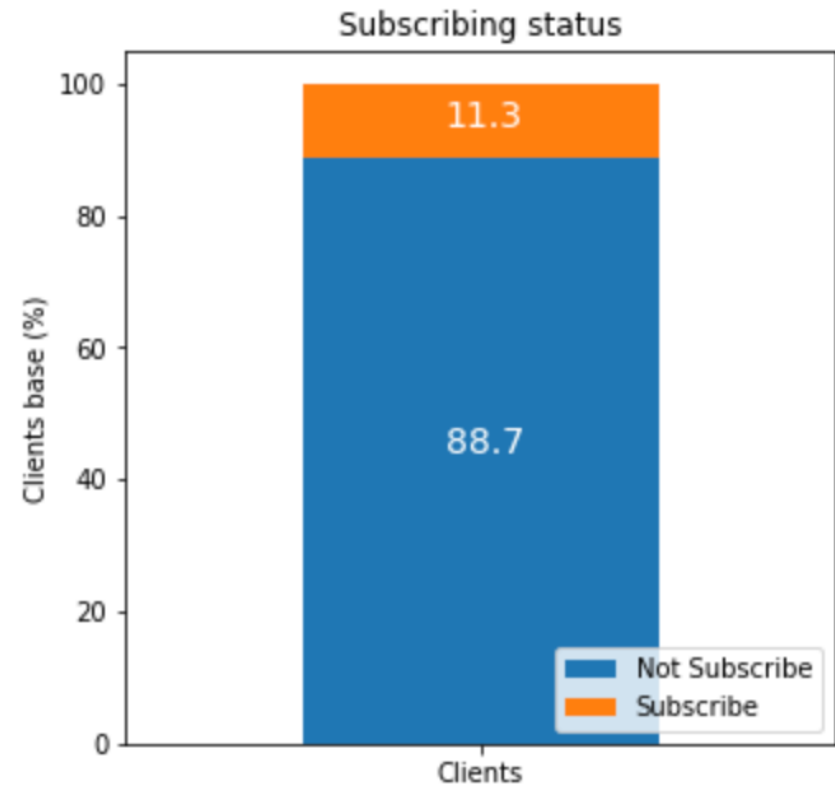


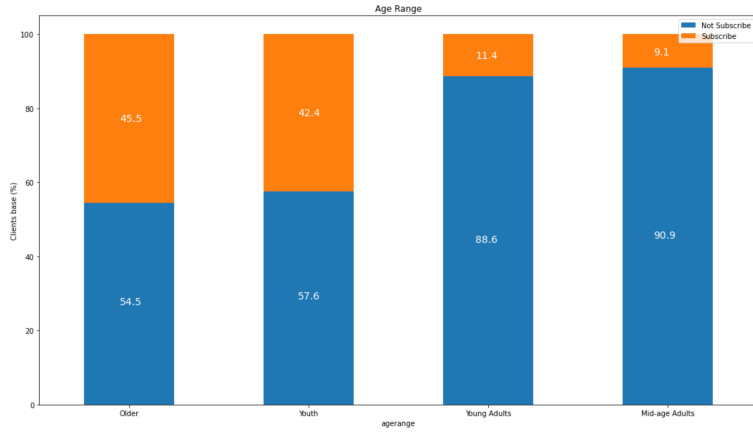
Introduction

- **Problem description**
 - ABC Bank wants to sell its 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).
- **Github Repo link**
 - <https://github.com/Shinuing/Bank-Marketing-Campaign-Prediction>

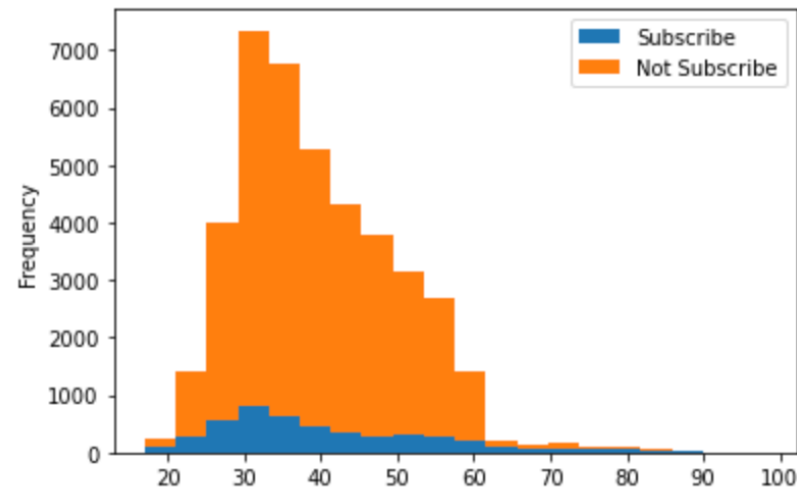
EDA - subscribing

Around 11% clients subscribe, make sense



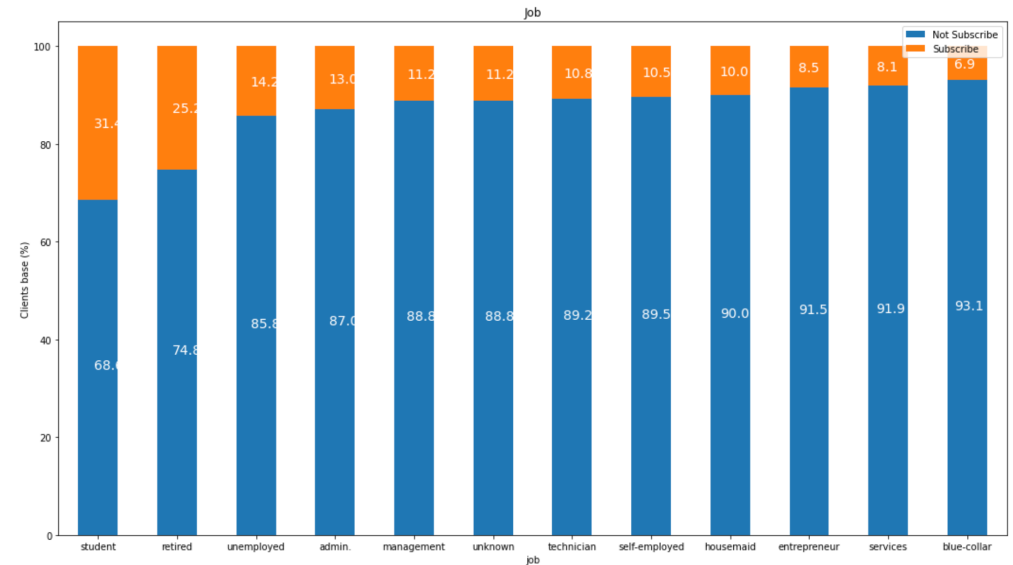


EDA – client age



older people and youth are more likely to subscribe

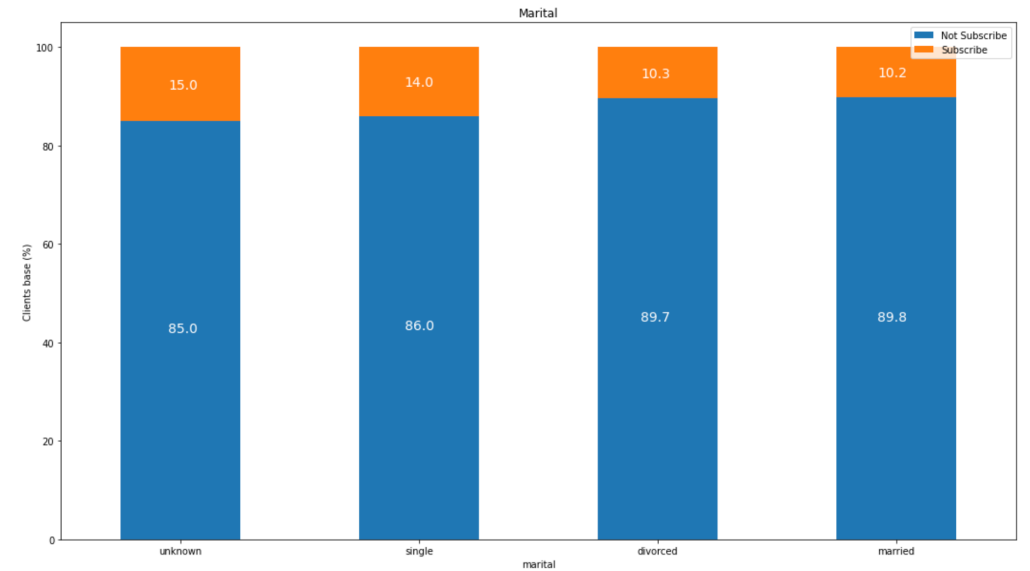
EDA – client job



similar to age variables, students and retired are more likely to subscribe

EDA – client marital

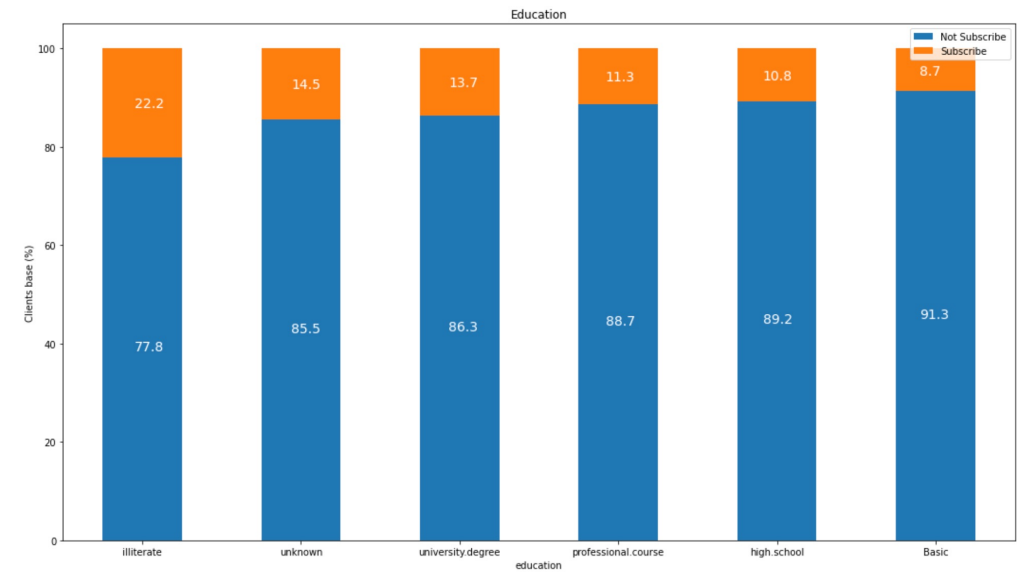
similar rate
noticed that unknown is more likely to subscribe



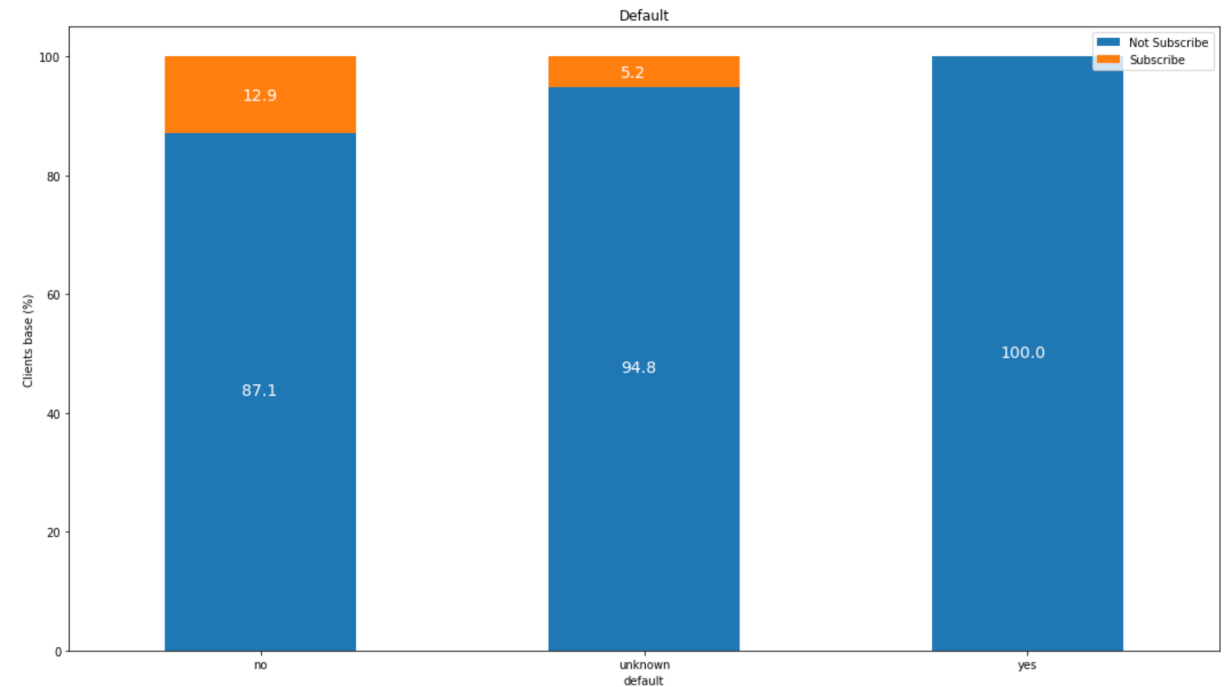
EDA – client education

illiterates are more likely to purchase, probably they earn more money?

note that unknowns are also likely to subscribe

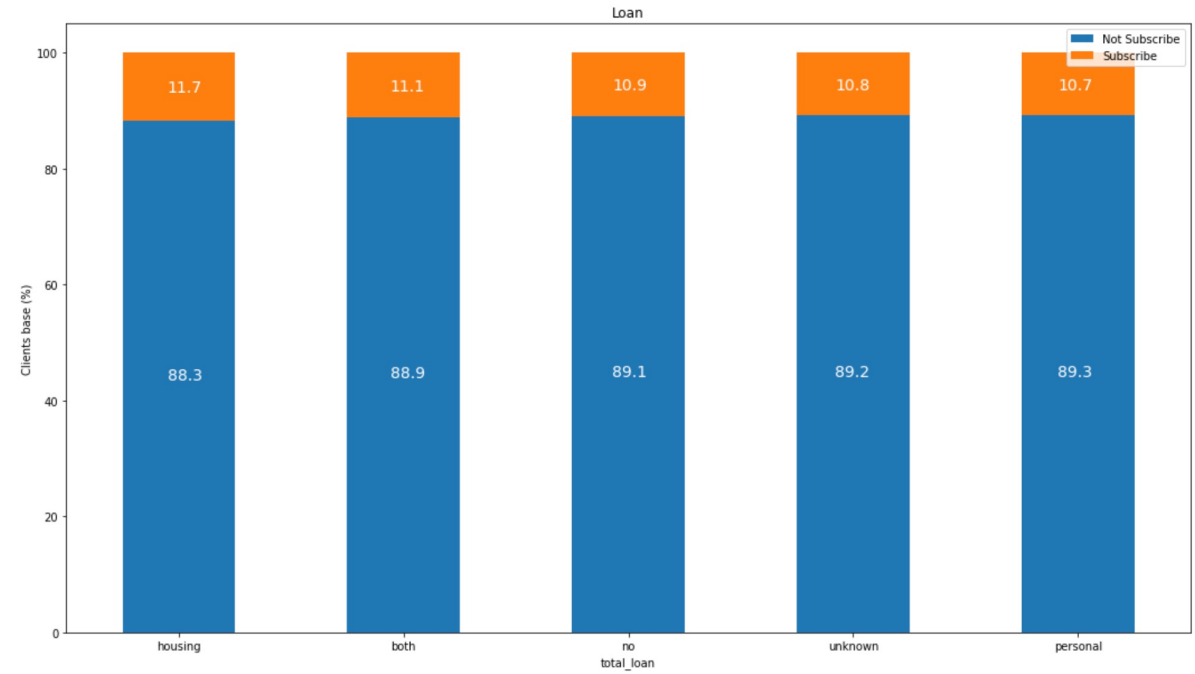
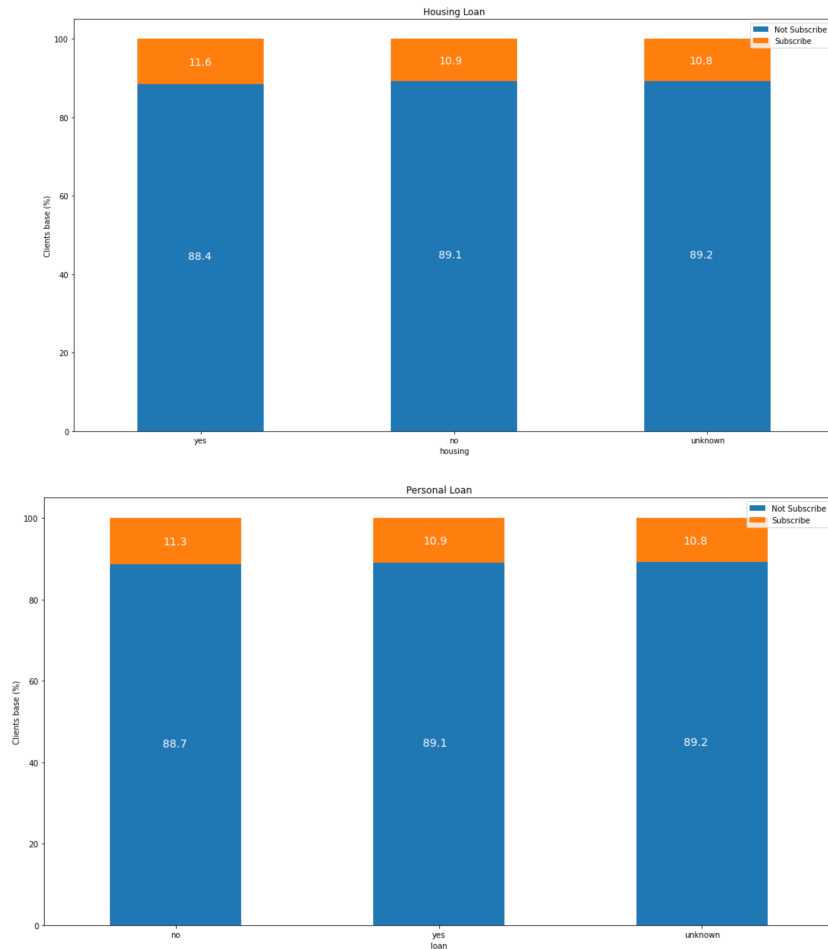


EDA – default



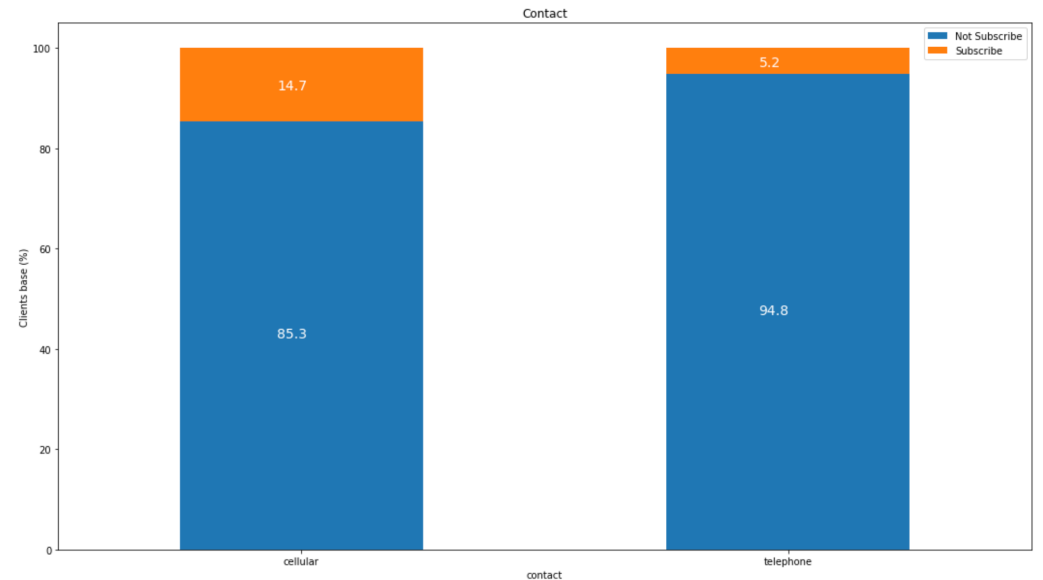
clients who has confirmed credit has NO subscribe at all
No credits are more likely to subscribe

EDA – loan



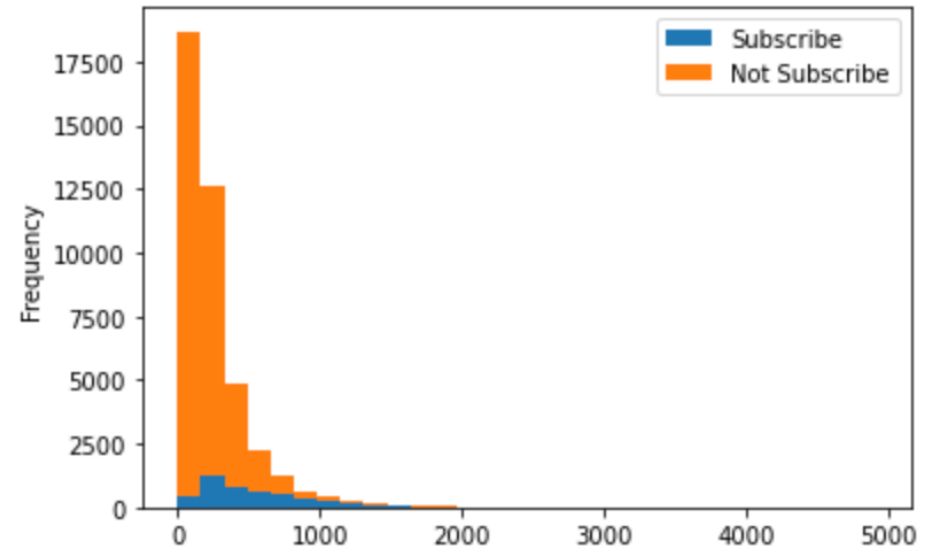
Not too much difference in housing or personal loan clients or both

EDA – contact



clients who contacted with cellular are more likely to subscribe probably because when calling telephone, people are not around

EDA – duration



this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. should be removed

EDA – date

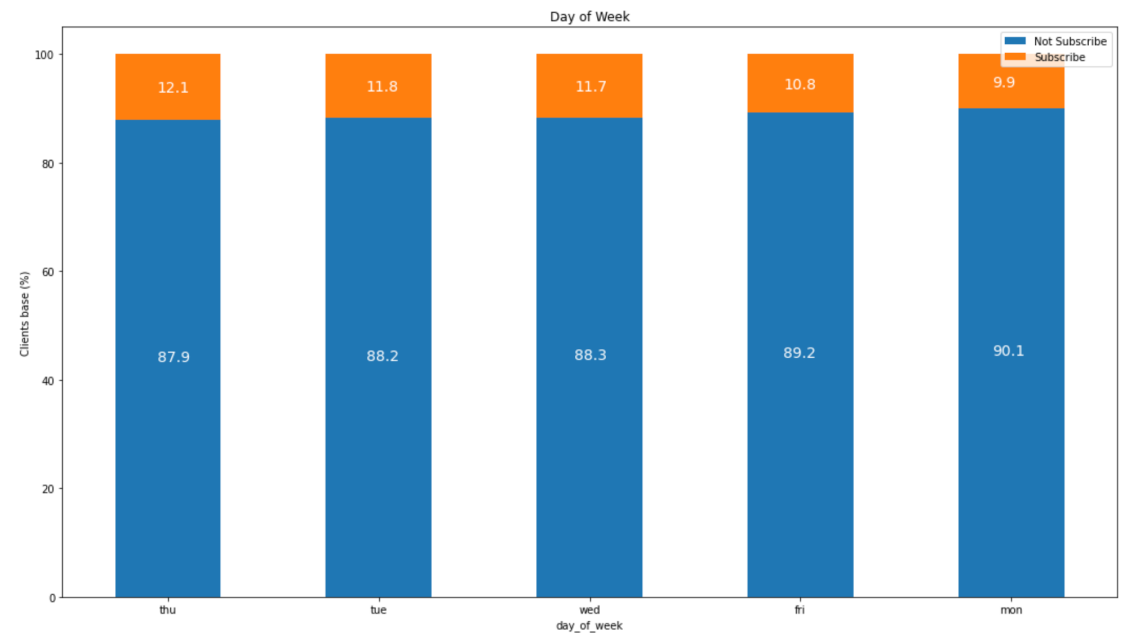
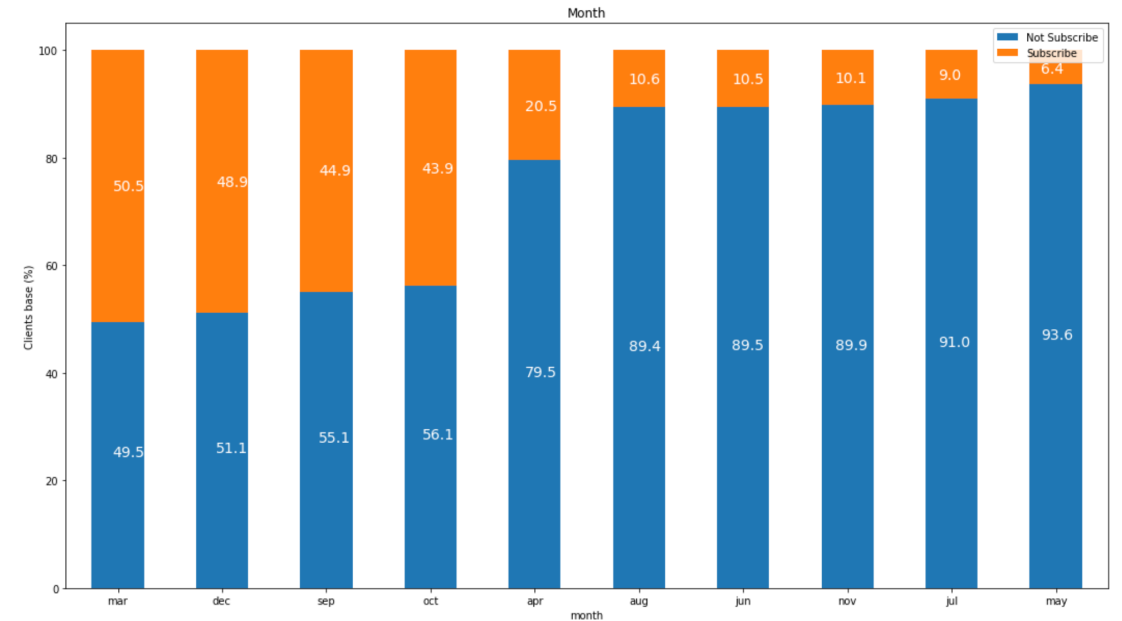
Month:

It looks like there are more subscribers at the beginning (March) and towards the end (Dec., Sep, Oct.) of the year

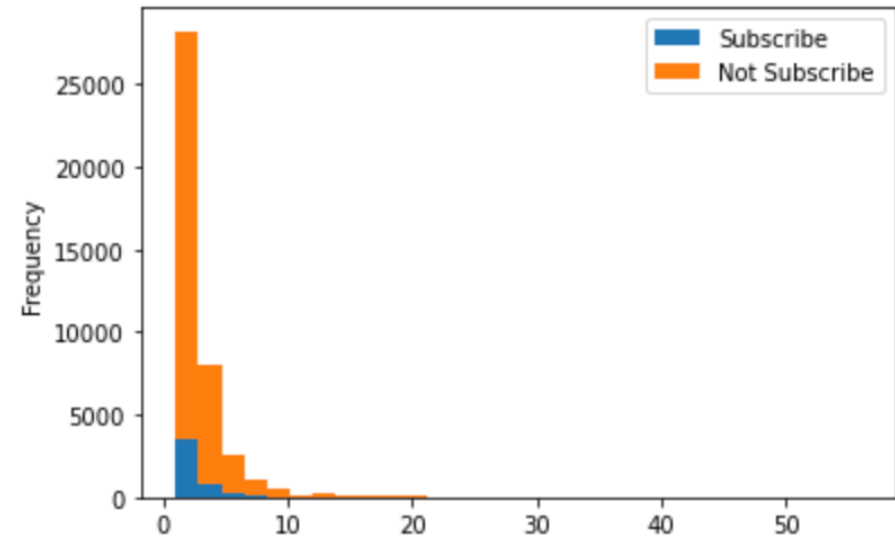
Not for sure because lacking Jan and Feb.

Day of week:

Not too much difference

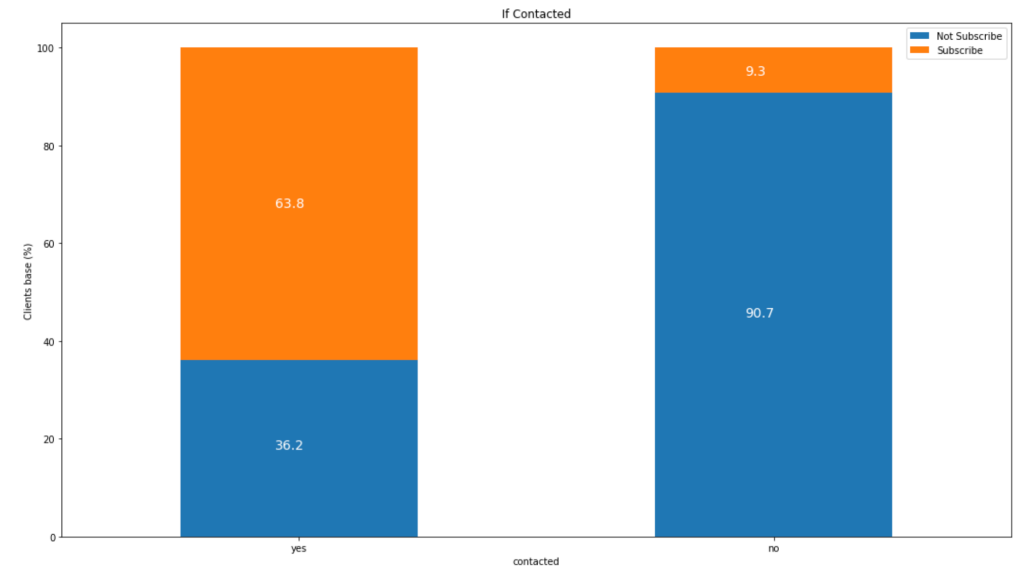


EDA – campaign



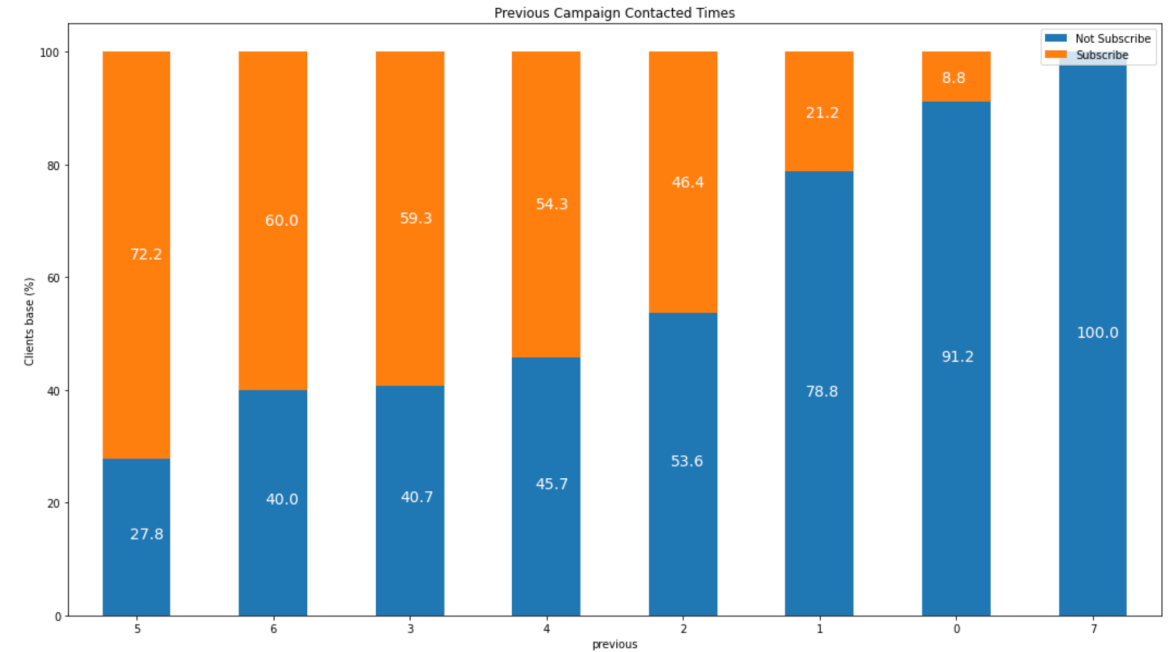
lower than subscribe ratio -> more campaign contacted,
less subscribe

EDA – pdays ->
converted to if
contacted



Contacted clients would more likely to subscribe

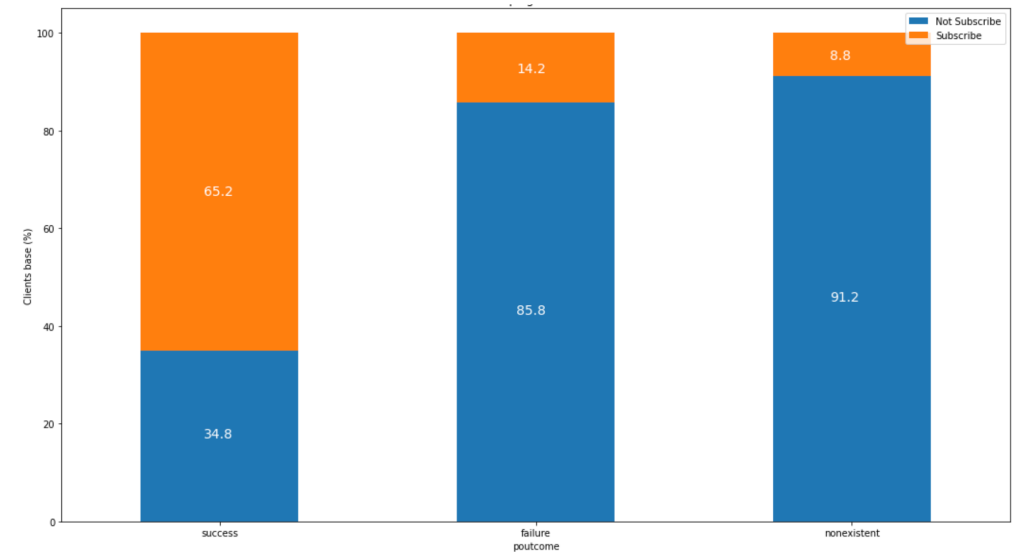
EDA – Previous Campaign Contacted Times



more contacted -> relatively more subscribe

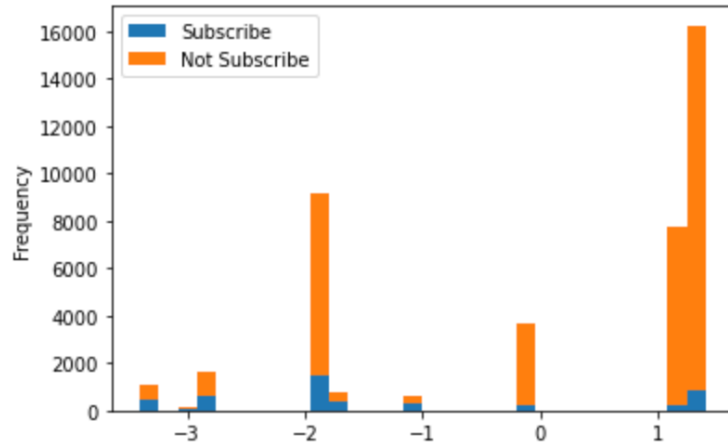
EDA – previous campaign subscription

Clients who attended last campaign are more likely to subscribe

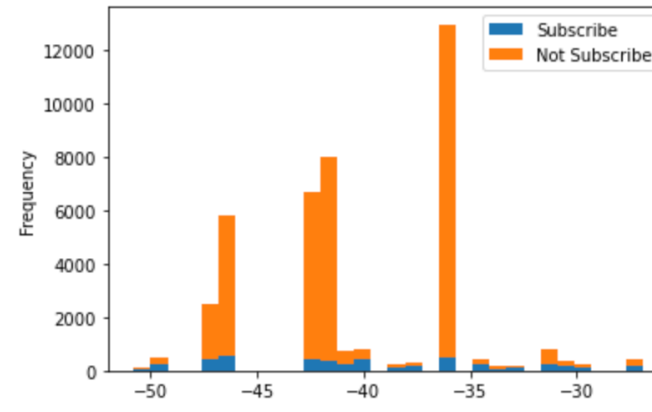


EDA – social and economic context attributes

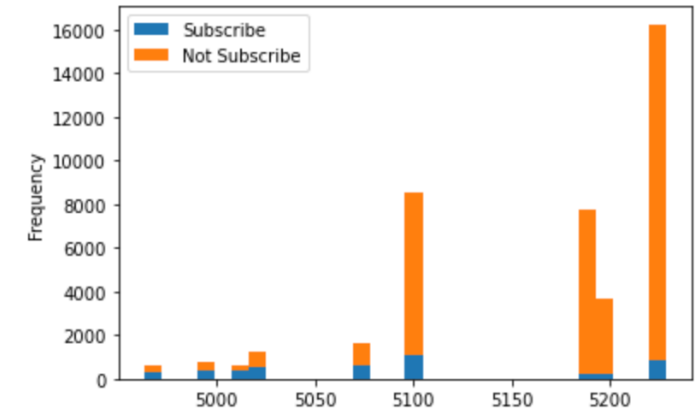
emp.var.rate



cons.conf.idx



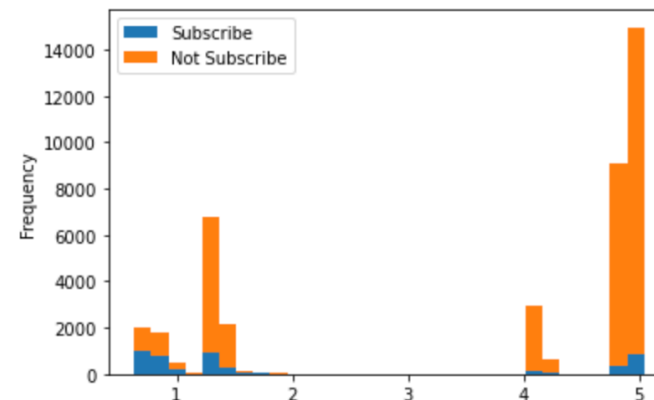
nr.employed



cons.price.idx



euribor




why number of employee was float?

looks like less nr.employed yields more subscribe



EDA Conclusion

- Youth and Elder (students and retired) are more likely to subscribe
 - unknown, single > married, divorced
 - higher education level yield higher rate of subscription
 - No credits default > yes (but only 3 people)
 - loan, regardless personal or housing, has no much influence on the subscribe
 - clients who contacted with cellular are more likely to subscribe, probably because when calling telephone, people are not around
 - There are more subscribers at the beginning (March) and towards the end (Dec., Sep, Oct.) of the year, but no much difference based on each day of week
 - lower than subscribe ratio -> more contacted in this campaign, less subscribe
 - Client who was last contacted subscribe more, the more contacts performed before this campaign, the more chance the clients will subscribe and clients who attended last campaign are more likely to subscribe
- 

Feature Engineering

Removed unrelated variables

- 'loan', 'housing', 'total_loan', 'day_of_week'

Skewed variables

- 'campaign' and 'previous' -> log transformation

Create dummy variables

- 'job', 'marital', 'education', 'default', 'contact', 'month', 'poutcome', 'agerange', 'contacted'

Dealing with correlation variables

- Remove 'emp.var.rate' with high correlation with many other features

Imbalanced data

- Used random resampling and SMOTE technology

Modeling

	Logistic Regression	Random Forest	ANN
Resampling	73.48%	94.91%	72.95%
SMOTE	89.22%	92.98%	88.09%

- Comparing with different resampling methods performance, overall SMOTE method highly increased the test accuracy of logistic regression and ANN model, while SMOTE method slightly decreased the accuracy of random forest model
- Due the bagging method and better performance, the random forest model with random resampling method was chosen to be the model for use

Conclusion



Random forest model successfully predicted around 95% accuracy for the clients' subscription of banking campaign



Clients' subscription is more related to their last time performance, the more they were contacted, more chance to subscribe. Thus, the stuff could increase customer engagement by strengthening contact with customers, especially via cellular.



For people with low subscription rate, such as middle-aged people, low education, etc., you can offer appropriate discounts to increase the subscription rate