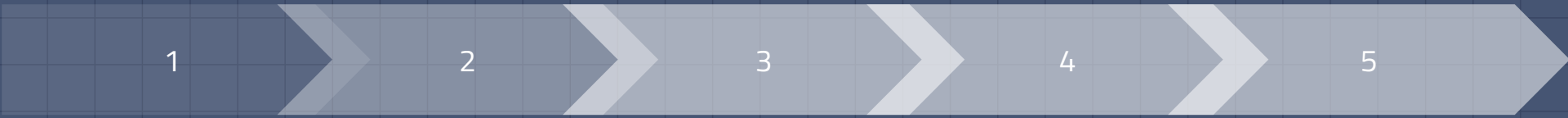
Bank Marketing Campaign Analysis

BA820 Team 4

Ziyan Pei, Youming Qiu, Xiaoyang Xu,

Dongzhe Zhang (David), Peng Yuan

Agenda for Today



1. Dataset Overview & Business Problems
2. Exploratory Data Analysis
3. Unsupervised Machine Learning
4. Supervised Machine Learning
5. Conclusions & Recommendations

Dataset Overview & Business Problems

1

Dataset Summary

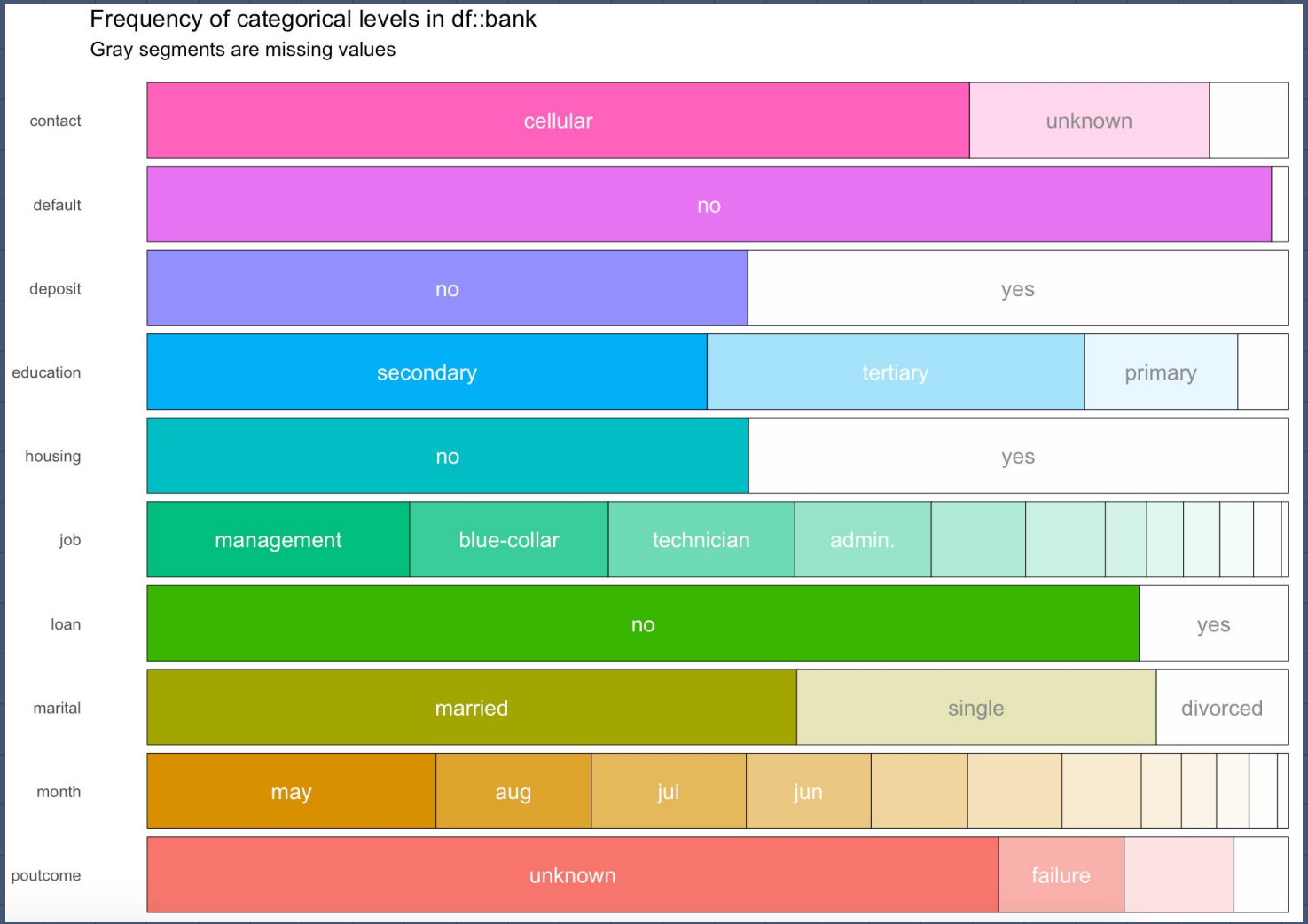
- **“Bank Marketing Dataset”**
- We have **11,162 rows, 17 columns**
  - 10 categorical variables
  - 7 numerical variables
- There is **NO** missing data!
- We only have **two data types** in this dataset
  - factors and integers

Business Problems & Goals

- How can the financial institution have greater effectiveness for future marketing campaigns of term deposits?
- Our project goals:
  - Build clusters and models that will find a specific cluster of customers who will be most likely to open the term deposit.
  - Recommend the bank about which part of the customers should most likely be sent the message of the next marketing campaign.

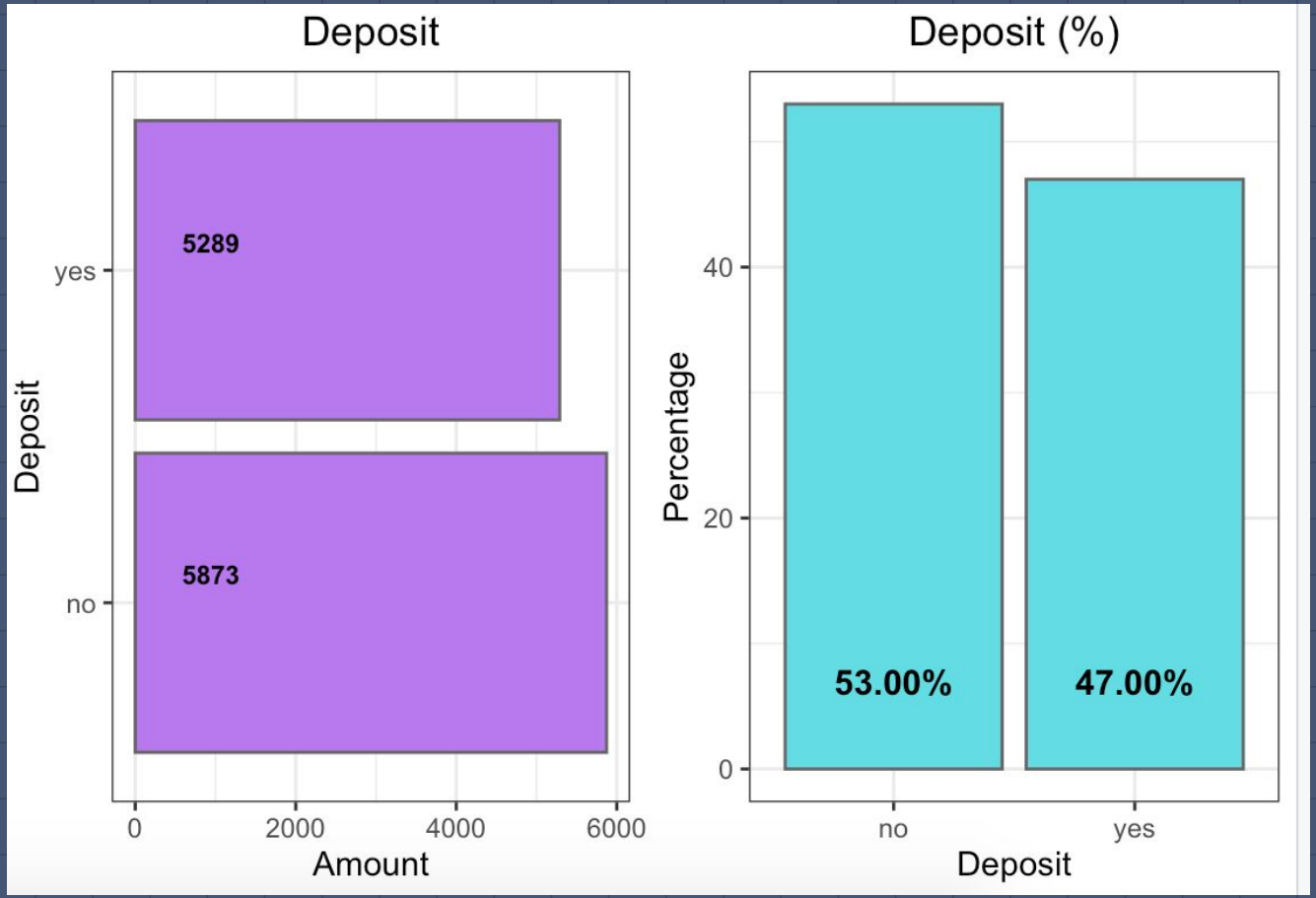
Exploratory Data Analysis

2

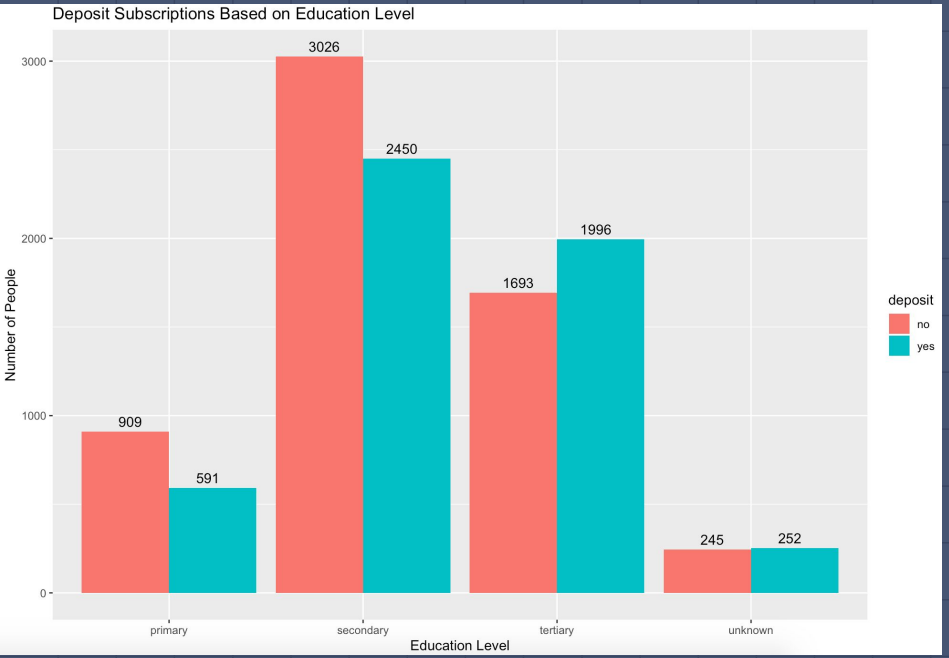


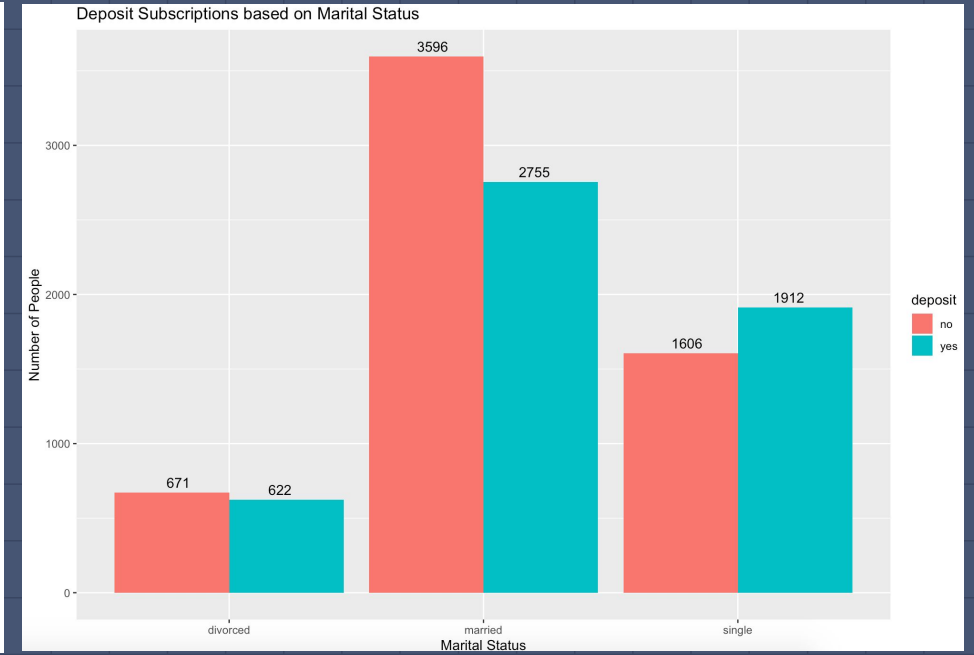
Overview of Categorical Variables

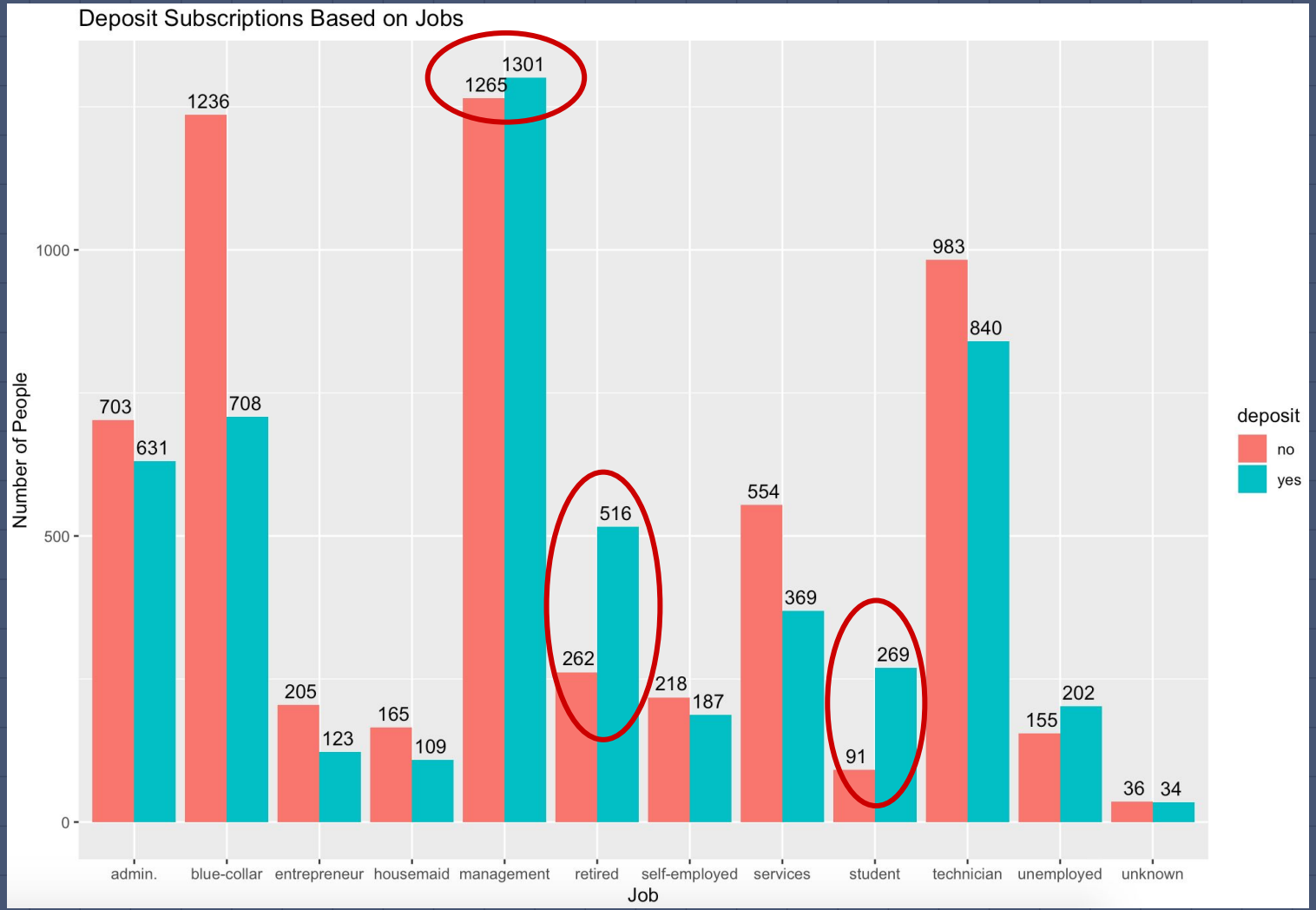
Distributions of “deposit” in numbers and percentages



Deposit Subscriptions Based on Education Level & Marital Status

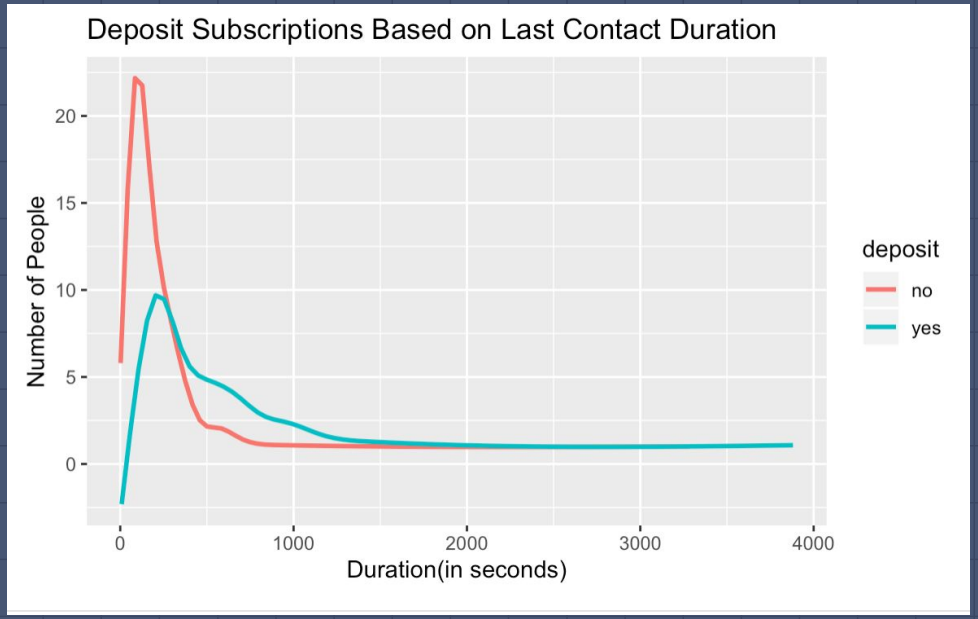


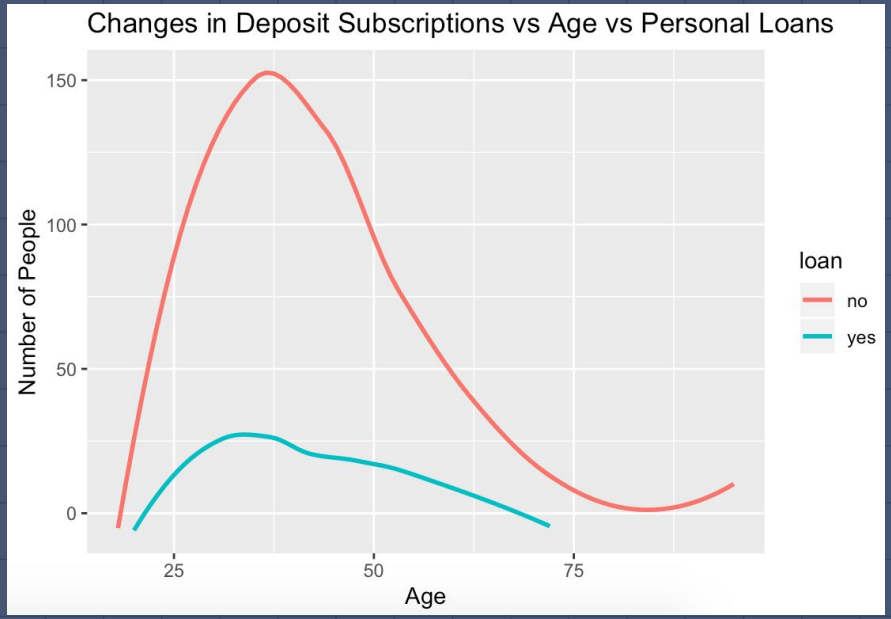




Deposit Subscriptions Based on Jobs

Deposit Subscriptions Based on Last Contact Duration vs. Age vs. Personal Loans





Some Assumptions from EDA process

- The decision of opening a term deposit might be related to these factors:
  - **Education level** (Secondary and tertiary)
  - **Marital Status** (Married)
  - **Job Type** (Management level, Students, Retired)
  - **Age** (Ranges from 25 to 50 years old)
  - People who spent **less time on calls** are less likely to open a term deposit
  - People who have **personal loans** are less likely to open a term deposit

Unsupervised Machine Learning

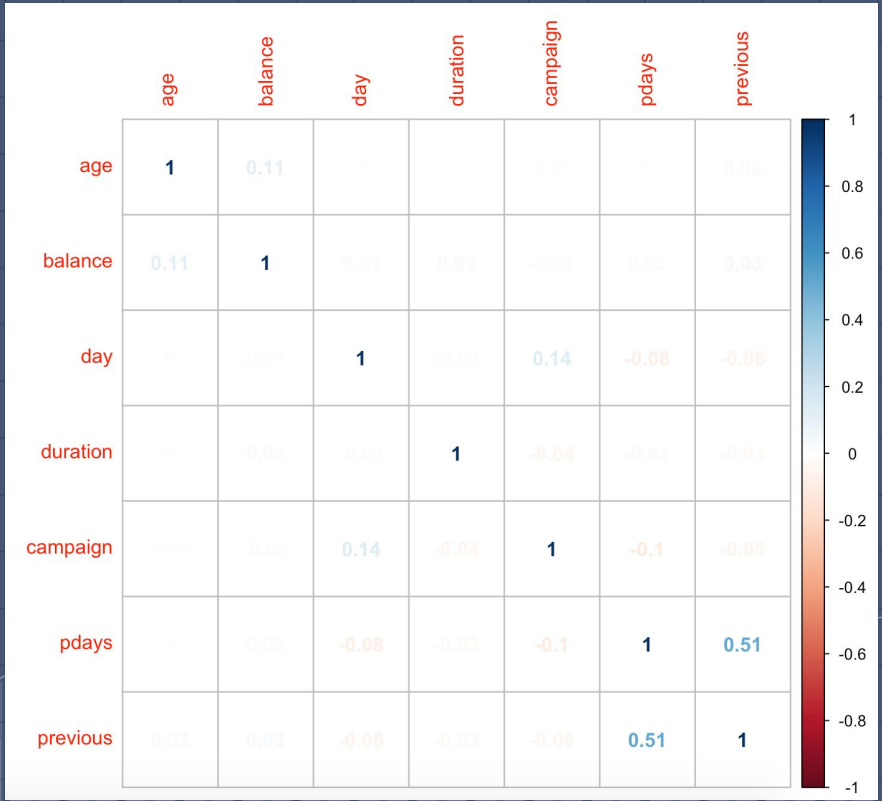
3

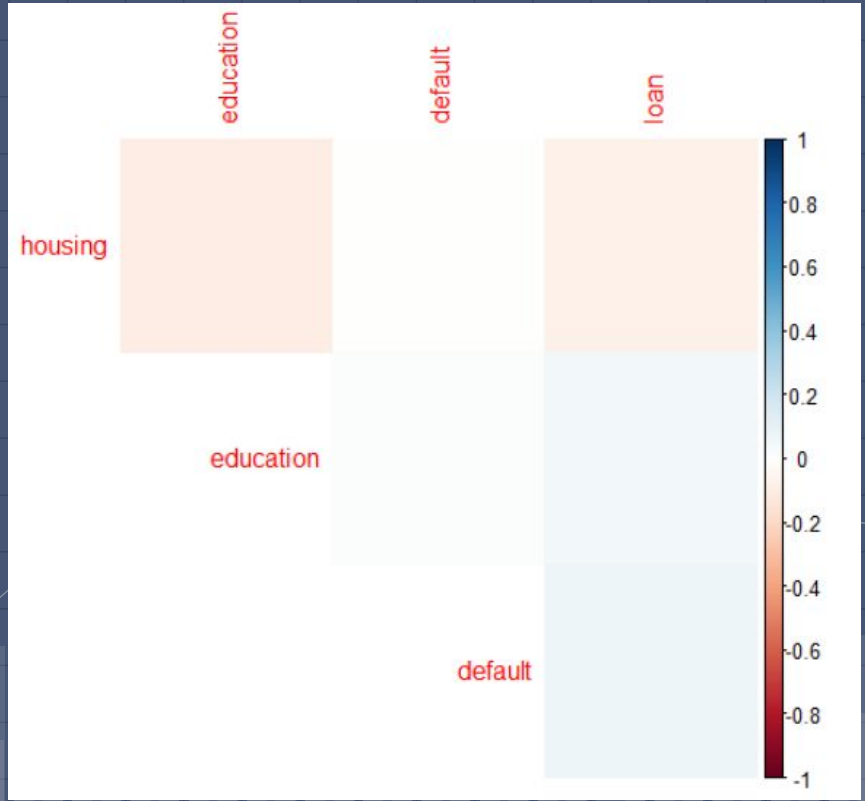
Data Cleaning Process

- Convert all categorical variables into **dummy variables**
- **Rename** some columns to a more standardized format
- **Delete** unnecessary columns
  - “month” and “day” => “pdays”
- Now we have **11,090 rows and 27 variables**

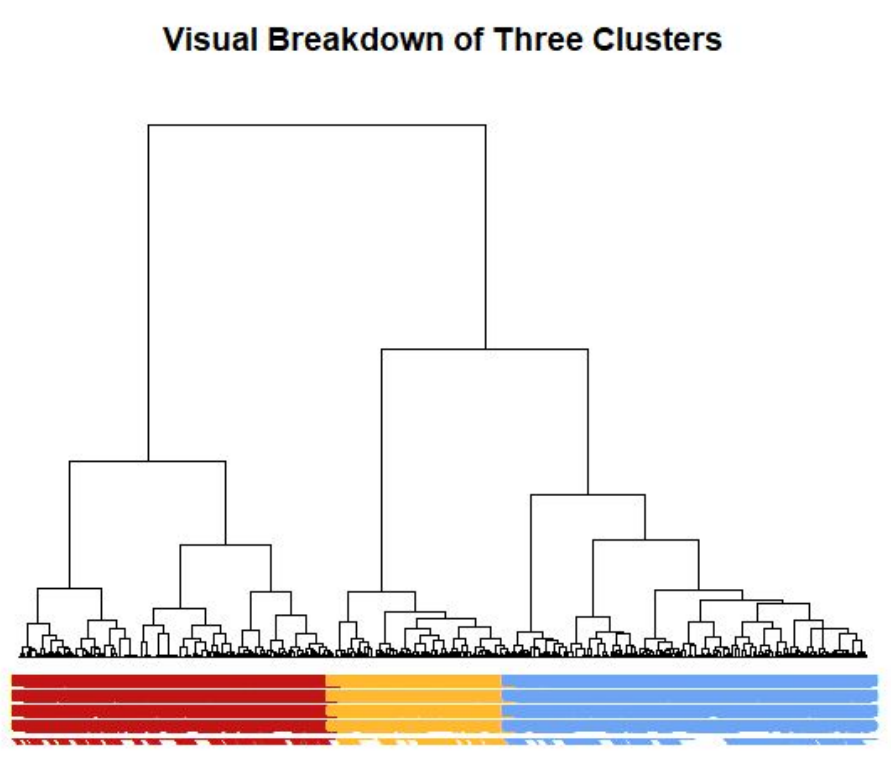
```
> glimpse(bank_clean)
Observations: 11,090
Variables: 27
$ age                 <int> 59, 56, 41, 55, 54, 42, 56, 60, 37, 28,
$ balance             <int> 2343, 45, 1270, 2476, 184, 0, 830, 545,
$ duration            <int> 1042, 1467, 1389, 579, 673, 562, 1201,
$ campaign            <int> 1, 1, 1, 1, 2, 2, 1, 1, 1, 3, 1, 2, 4,
$ pdays               <int> -1, -1, -1, -1, -1, -1, -1, -1, -1, -1,
$ previous            <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
$ job_blue_collar     <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,
$ job_entrepreneur    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
$ job_housemaid       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
$ job_management      <int> 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1,
$ job_retired         <int> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,
$ job_self_employed   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
$ job_services        <int> 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0,
$ job_student         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
$ job_technician      <int> 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,
$ job_unemployed      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
$ marital_married     <int> 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 0, 1, 1,
$ marital_single      <int> 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0,
$ education_secondary <int> 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 0,
$ education_tertiary  <int> 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1,
$ education_unknown   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
$ default_yes         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
$ housing_yes         <int> 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1,
$ loan_yes            <int> 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1,
$ contact_telephone   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
$ contact_unknown     <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
$ deposit_yes         <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
```

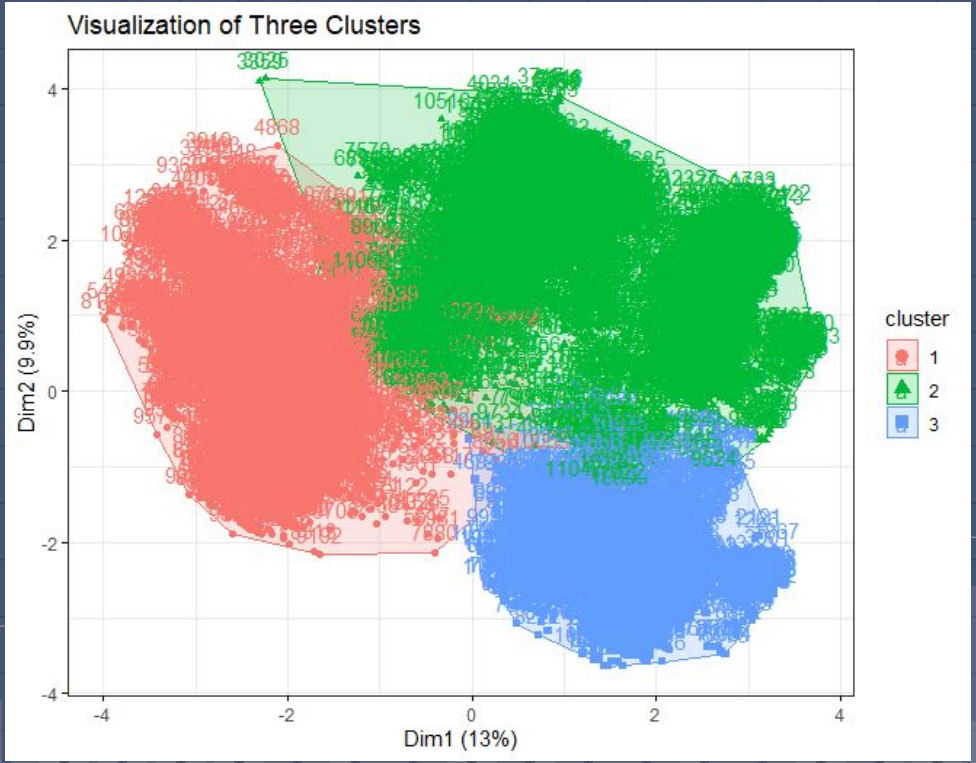
Dimensionality Reduction: PCA & EFA



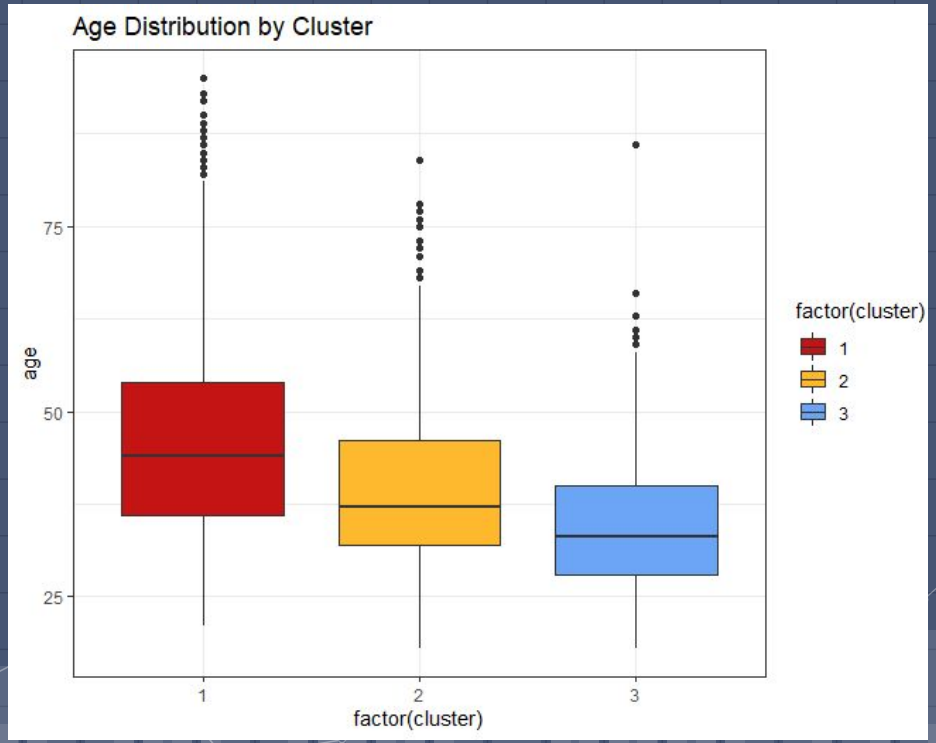


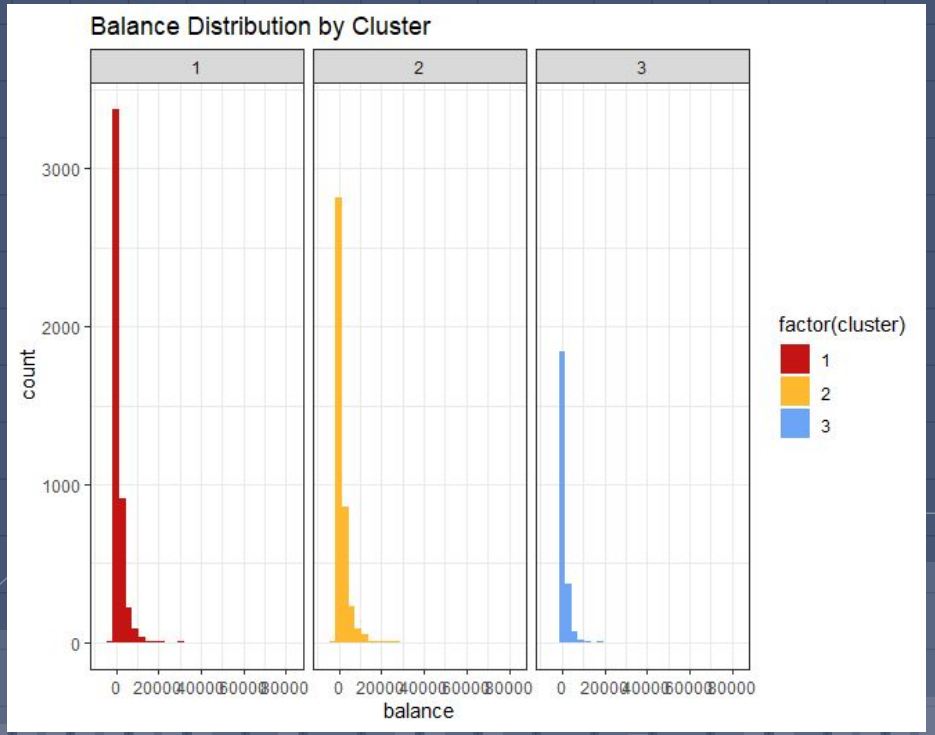
Hierarchical Clustering

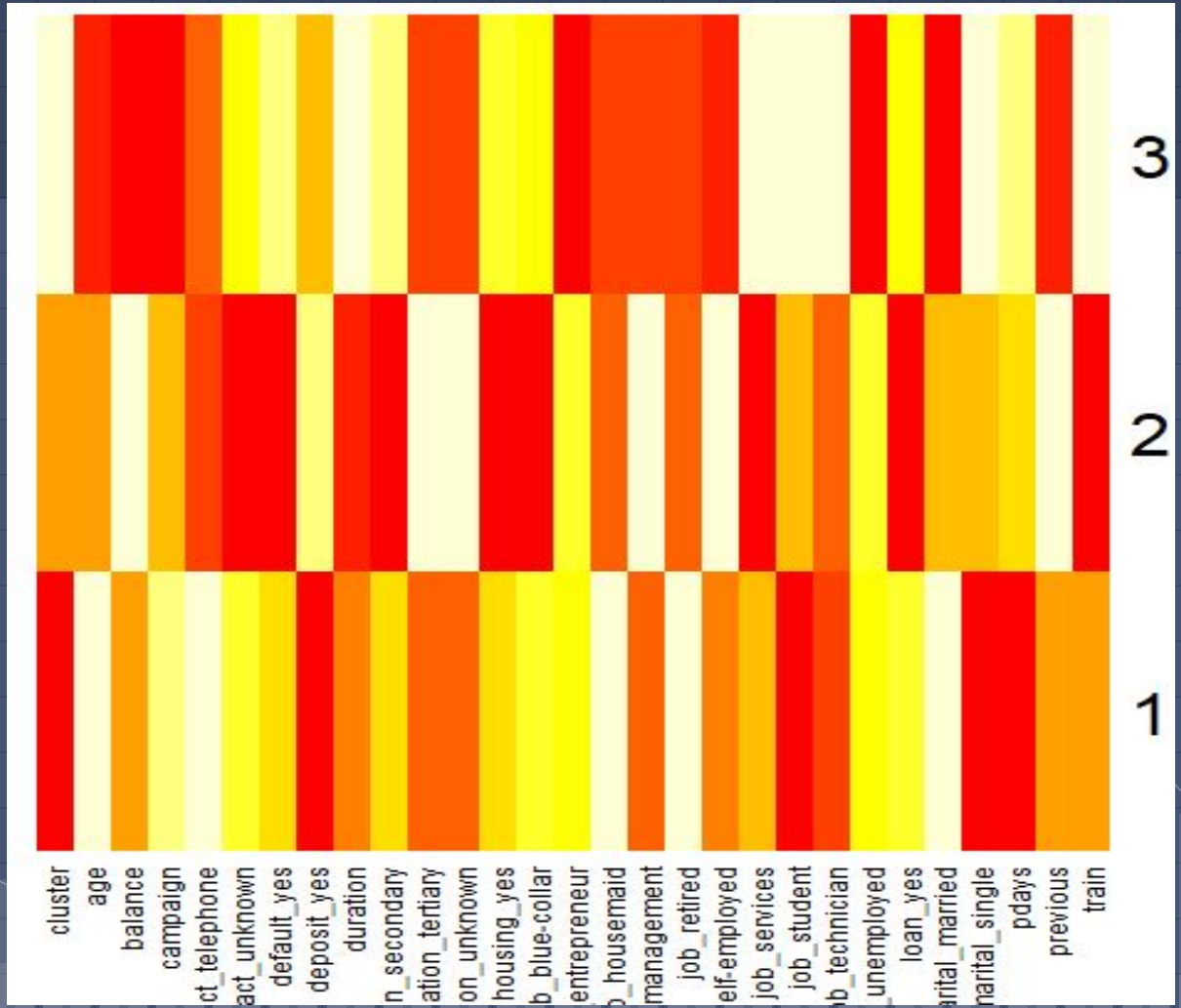




Cluster Characteristics







- **Cluster 1** is more likely the elders with the most balance.
- **Cluster 3** seems to be the younger generation, with the least balance.
- **Cluster 2** is between them.

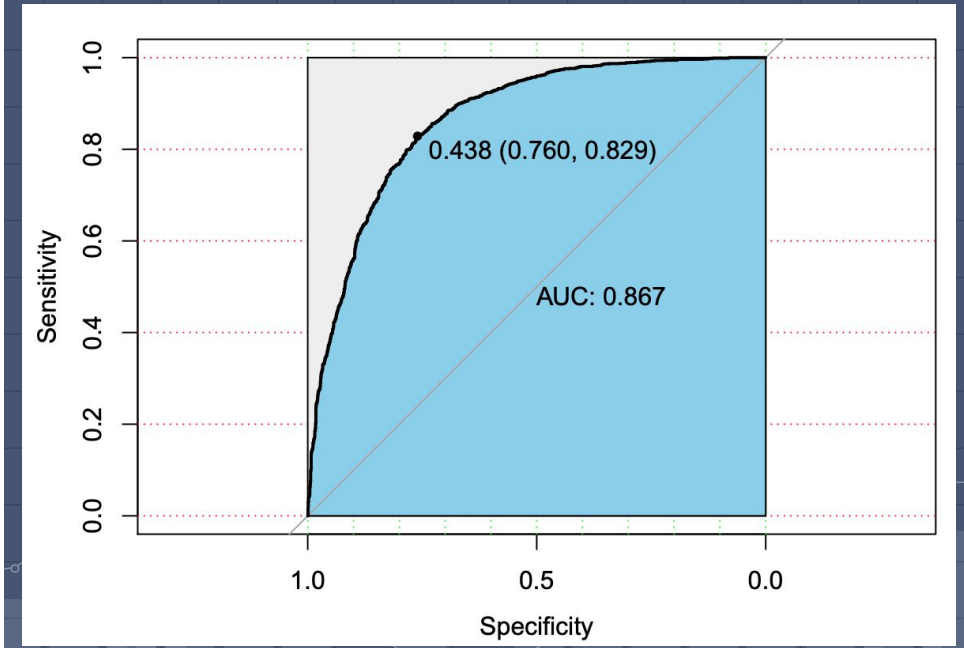
Supervised Machine Learning

4

Logistic Regression

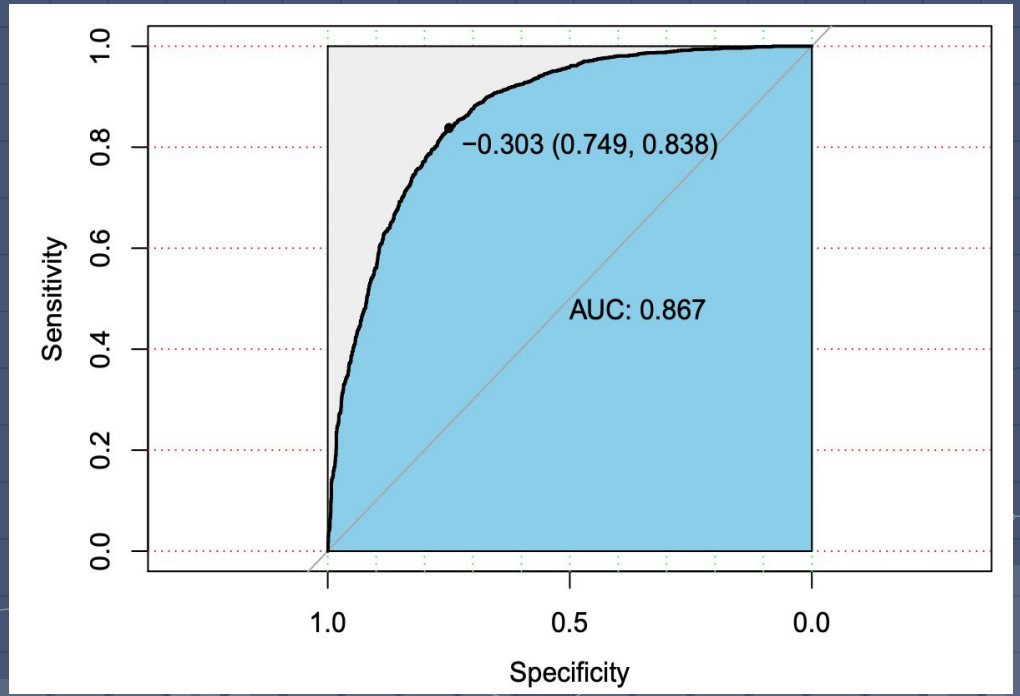
	Est.	S.E.	z val.	p
(Intercept)	0.56	0.16	3.50	0.00
age	0.07	0.04	1.57	0.12
balance	0.06	0.03	2.06	0.04
duration	1.82	0.05	36.64	0.00
campaign	-0.40	0.04	-9.17	0.00
pdays	0.22	0.03	6.43	0.00
previous	0.24	0.04	6.14	0.00
job_blue_collar	-0.56	0.12	-4.82	0.00
job_entrepreneur	-0.57	0.20	-2.91	0.00
job_housemaid	-0.59	0.21	-2.89	0.00
job_management	-0.38	0.12	-3.26	0.00
job_retired	0.21	0.16	1.31	0.19
job_self_employed	-0.53	0.18	-2.99	0.00
job_services	-0.50	0.14	-3.74	0.00
job_student	0.47	0.19	2.43	0.02
job_technician	-0.24	0.11	-2.18	0.03
job_unemployed	-0.13	0.19	-0.69	0.49
marital_married	0.01	0.10	0.08	0.94
marital_single	0.31	0.11	2.77	0.01
education_secondary	0.31	0.10	2.95	0.00
education_tertiary	0.62	0.12	5.03	0.00
education_unknown	0.35	0.17	2.06	0.04
default_yes	-0.27	0.25	-1.05	0.29
housing_yes	-0.88	0.06	-13.56	0.00
loan_yes	-0.73	0.09	-7.77	0.00
contact_telephone	-0.08	0.12	-0.67	0.50
contact_unknown	-1.34	0.09	-14.42	0.00

Continuous predictors are mean-centered and scaled by 1 s.d.

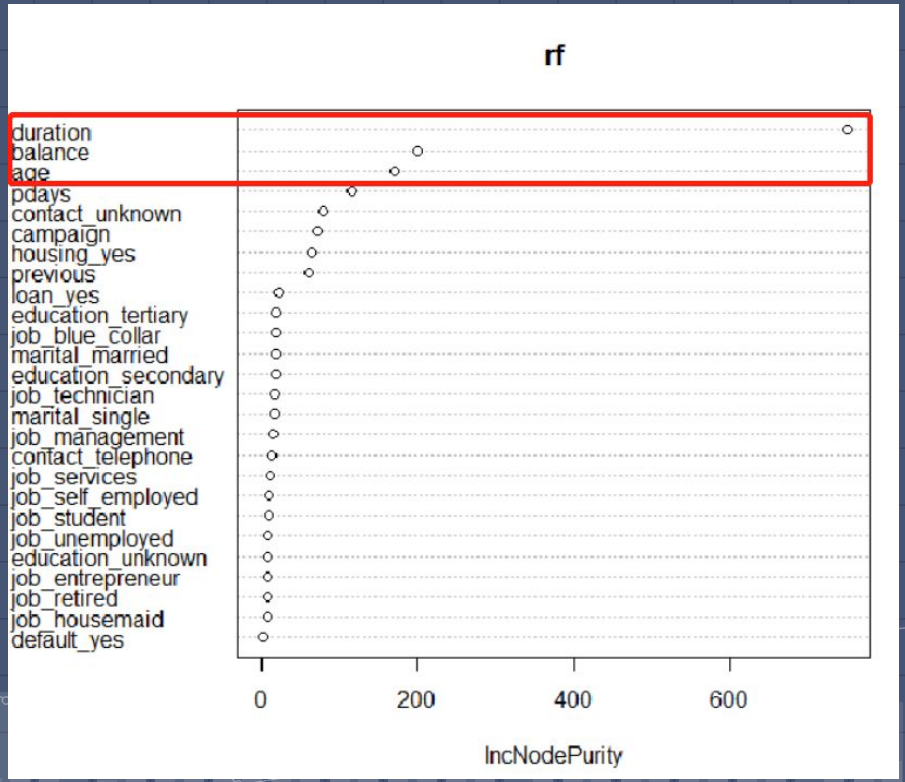


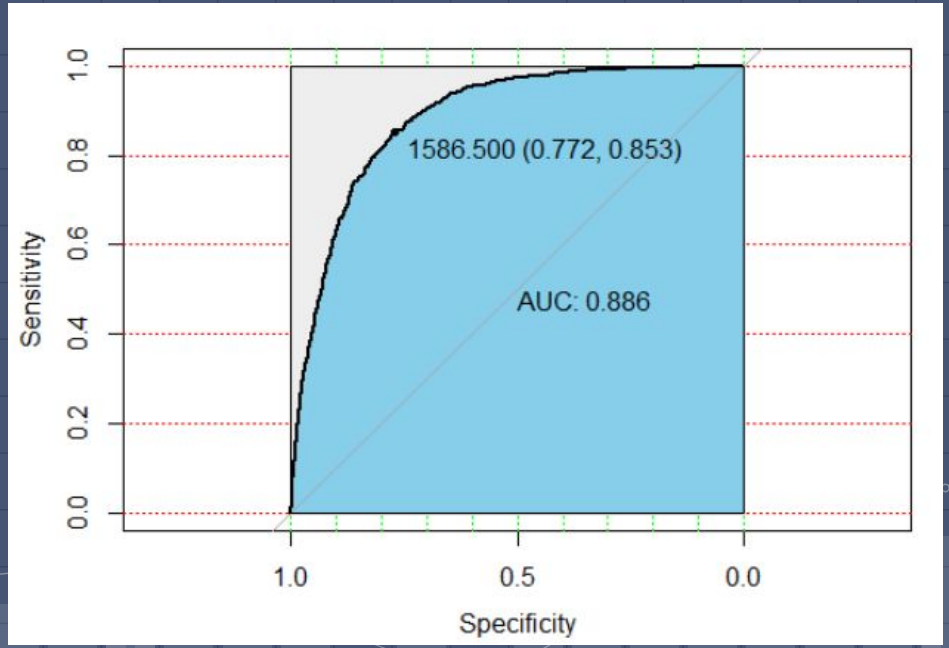
Lasso Regression

(Intercept)	-1.0225185862255
age	.
balance	0.0000006825567
duration	0.0042975890944
campaign	-0.0860905747594
pdays	0.0016527179663
previous	0.0715021021798
job_blue_collar	-0.2291130619449
job_entrepreneur	.
job_housemaid	.
job_management	.
job_retired	0.2708840407153
job_self_employed	.
job_services	-0.0240362812941
job_student	0.3894146975155
job_technician	.
job_unemployed	.
marital_married	.
marital_single	0.1629271476710
education_secondary	.
education_tertiary	0.1252117759733
education_unknown	.
default_yes	.
housing_yes	-0.7252610490219
loan_yes	-0.4975236045817
contact_telephone	.
contact_unknown	-1.1111908731681



Random Forest

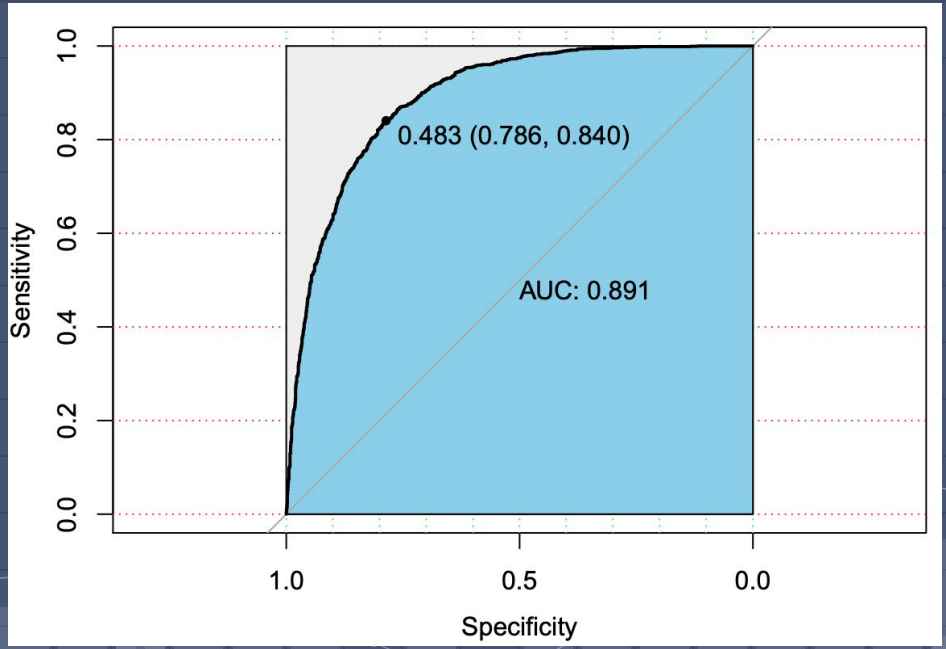




Ntrees - 500

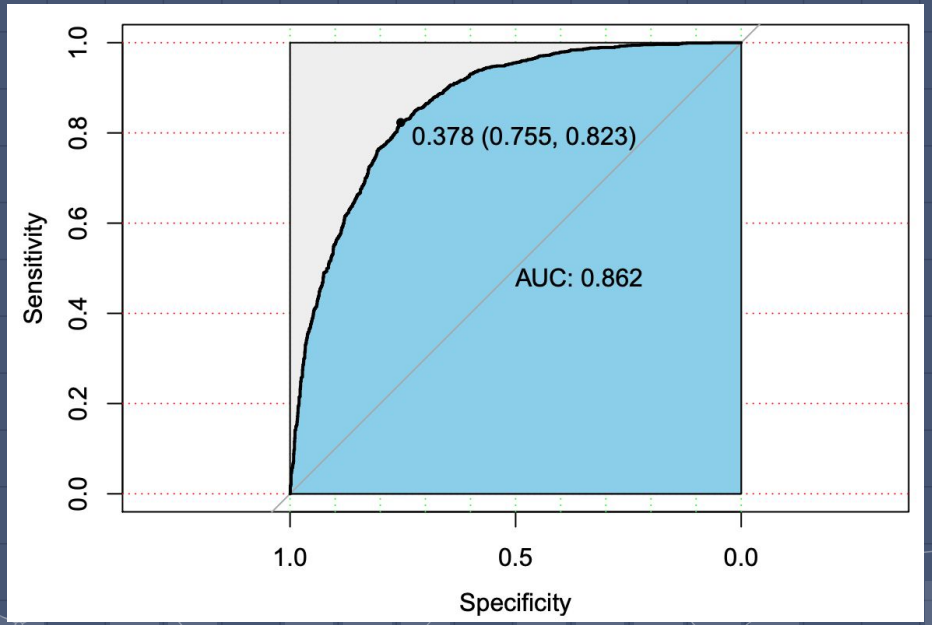
Boosting

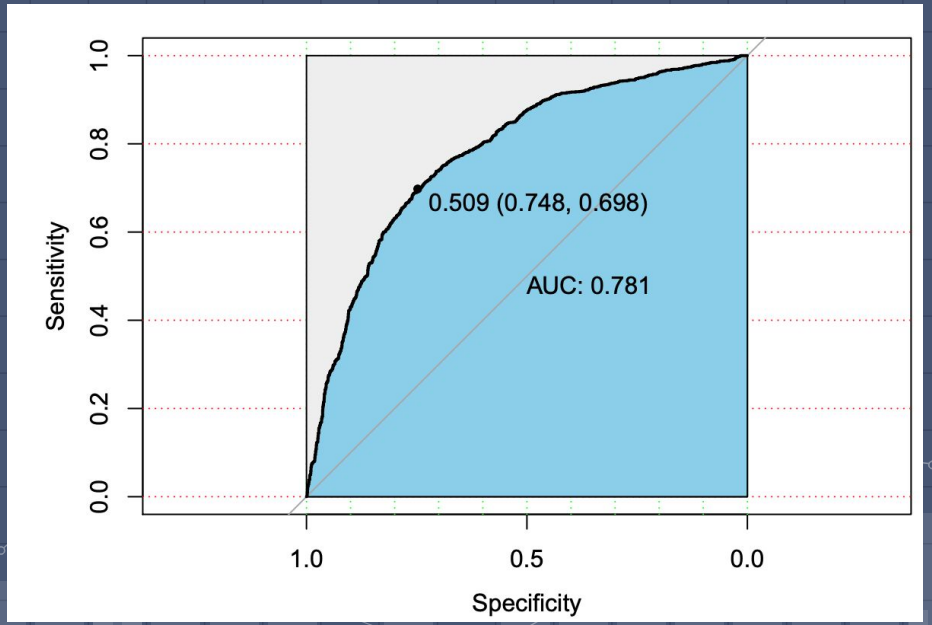
Feature <chr>	Gain <dbl>	Cover <dbl>	Frequency <dbl>
duration	0.4911812335	0.322859511	0.239474988
pdays	0.1220835553	0.132937802	0.108249408
contact_unknown	0.1009319300	0.052992731	0.023884211
balance	0.0753869197	0.140630183	0.192256893
age	0.0681678834	0.092486277	0.151037367
housing_yes	0.0440847417	0.029051951	0.035055858
campaign	0.0302851328	0.064903846	0.067882890
loan_yes	0.0140503916	0.022057848	0.017197733
job_blue_collar	0.0092682186	0.015865050	0.010153541
previous	0.0062205090	0.017441911	0.023719113
education_tertiary	0.0054996245	0.024802358	0.018298388
marital_single	0.0051730615	0.004694015	0.012905179
marital_married	0.0048564684	0.005602240	0.018408453
contact_telephone	0.0043780592	0.011136414	0.014280997
education_secondary	0.0039976699	0.006118441	0.015244070
job_services	0.0025589436	0.005144349	0.007952232
job_technician	0.0019661613	0.001545940	0.006301249
job_student	0.0018686572	0.004822256	0.004870398
job_self_employed	0.0017088203	0.011226249	0.006879093
education_unknown	0.0014385862	0.003337817	0.004237521
job_management	0.0012926053	0.002141573	0.006053602
job_housemaid	0.0012208947	0.012847137	0.006301249
default_yes	0.0007354041	0.004969440	0.002669088
job_unemployed	0.0006335633	0.000819868	0.002228826
job_entrepreneur	0.0006268113	0.005021223	0.002091244
job_retired	0.0003841535	0.004543569	0.002366408



Learning rate 0.01

Support Vector Machines & Naive Bayes Model

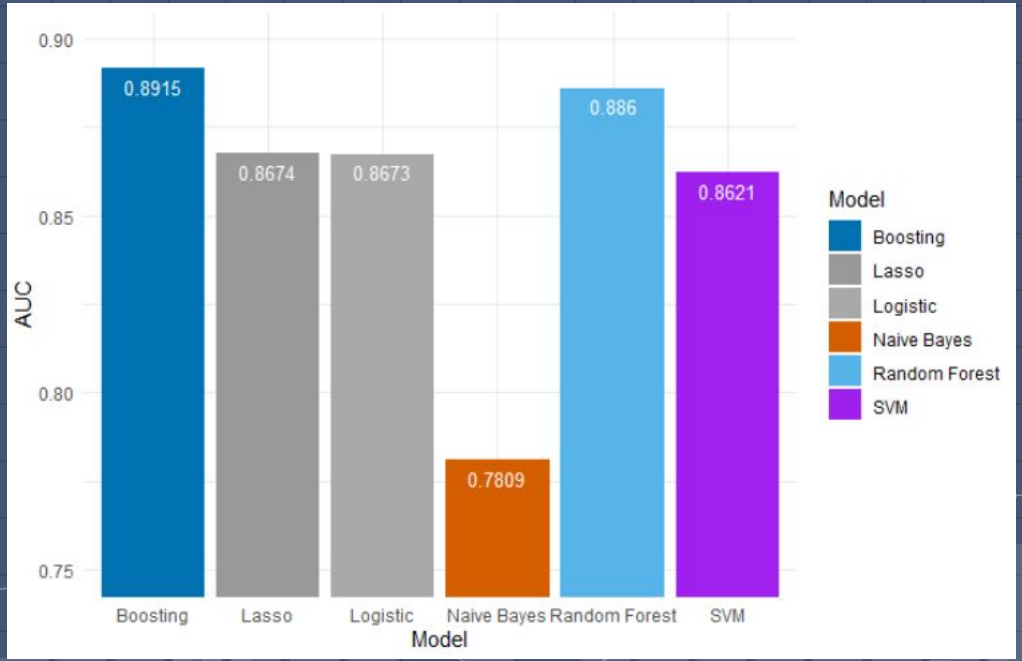




Models Comparison

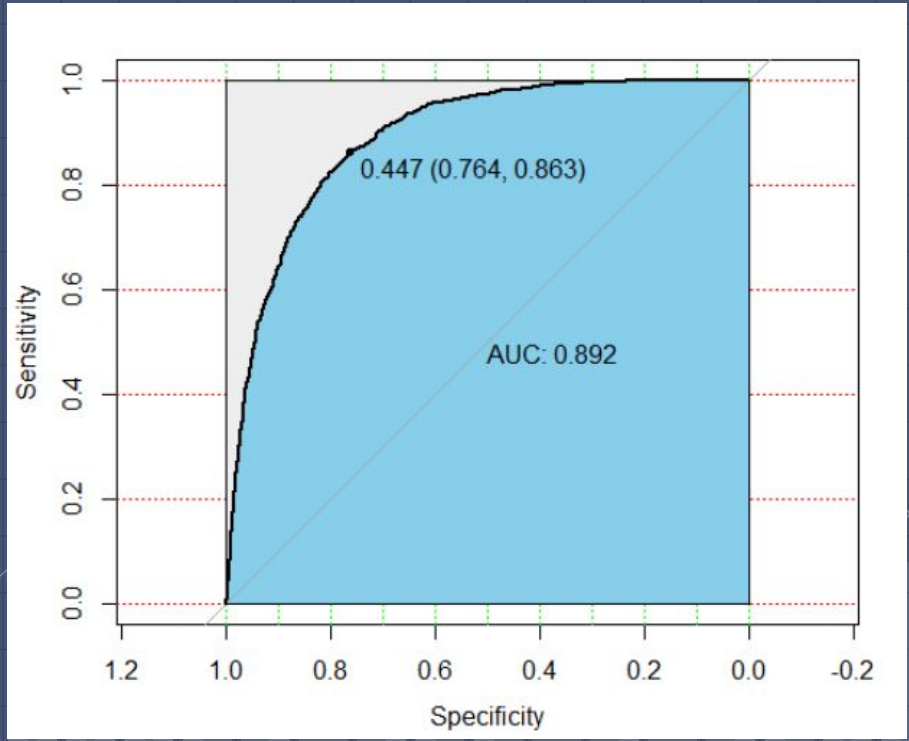
- Model Measurement:
  - Tuning to optimal parameters
  - **AUC** (The area under the curve)

- Performance:
  - Best model: **Boosting** - 0.8915



Trade-off between Precision and Recall

- As precision gets higher, recall gets lower;
- Here we want to cover most "actual" customers with affordable extra efforts.
  - Threshold: 0.447
  - Cover 86.3% term deposit customers with a 23.5% false-positive rate.



Conclusions & Recommendations

5

Conclusions

- Based on our models, the following variables are the most important factors for the decision of opening a term deposit:
  - Last Contact Duration
  - Balance
  - Age
  - Education Level
  - Number of Previous Contacts

Recommendations

- Use our classification model to contact clients with "deposit" label.
- Keep in contact with clients *(especially Cluster 1).*
- Take clients' balance, age and job information into consideration for future marketing campaign.
- Customer-relationship managers could try to talk about term deposit opportunities when clients with these characteristics come into bank

THANKS!

Any questions?