STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

CS 513 A
Knowledge Discovery & Data Mining

***FINAL PROJECT* - Banking Dataset Classification**

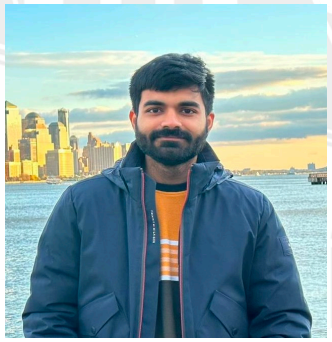
Guided By: Prof. Khashayar Dehnad

GROUP 16:

DEVSHREE PARIKH - 10476941

RAJ BHAGAT - 10477880

VRAJ PATEL - 20009021



AGENDA

- PROBLEM STATEMENT
- DATASET
- DATA PRE-PROCESSING (EDA)
- MACHINE LEARNING MODELS
- MODEL EVALUATION
- CONCLUSION

Problem Statement

The Portuguese Bank has experienced a reduction in earnings and is unsure of the best course of action.

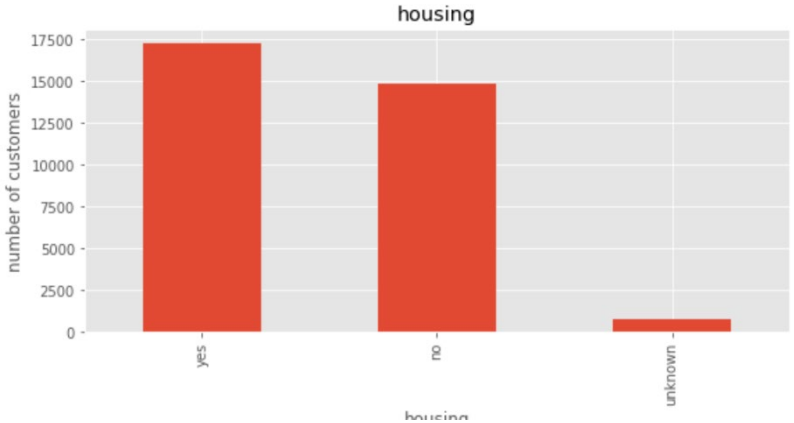
Following an inquiry, they discovered that their client insufficient investment in hard accounts was the main contributing factor.

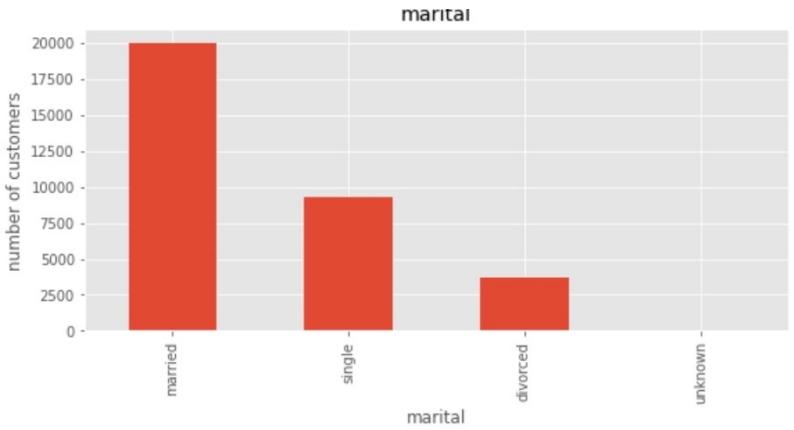
Therefore, the bank wants to identify current clients who are more likely to sign up for a long-term loan and concentrate sales promotion on them.

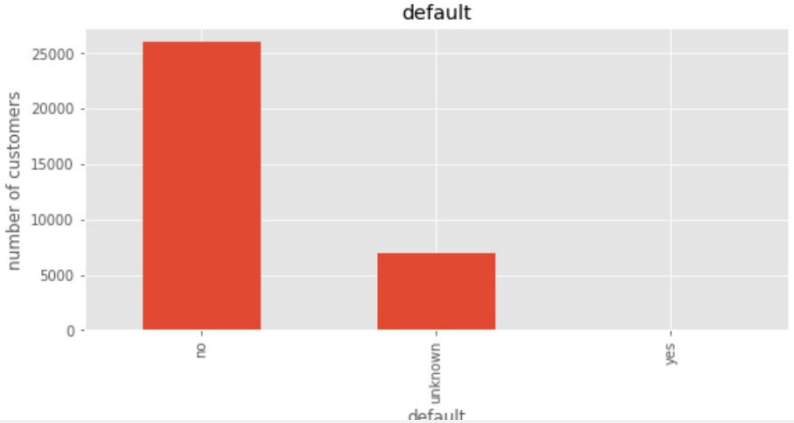
Dataset

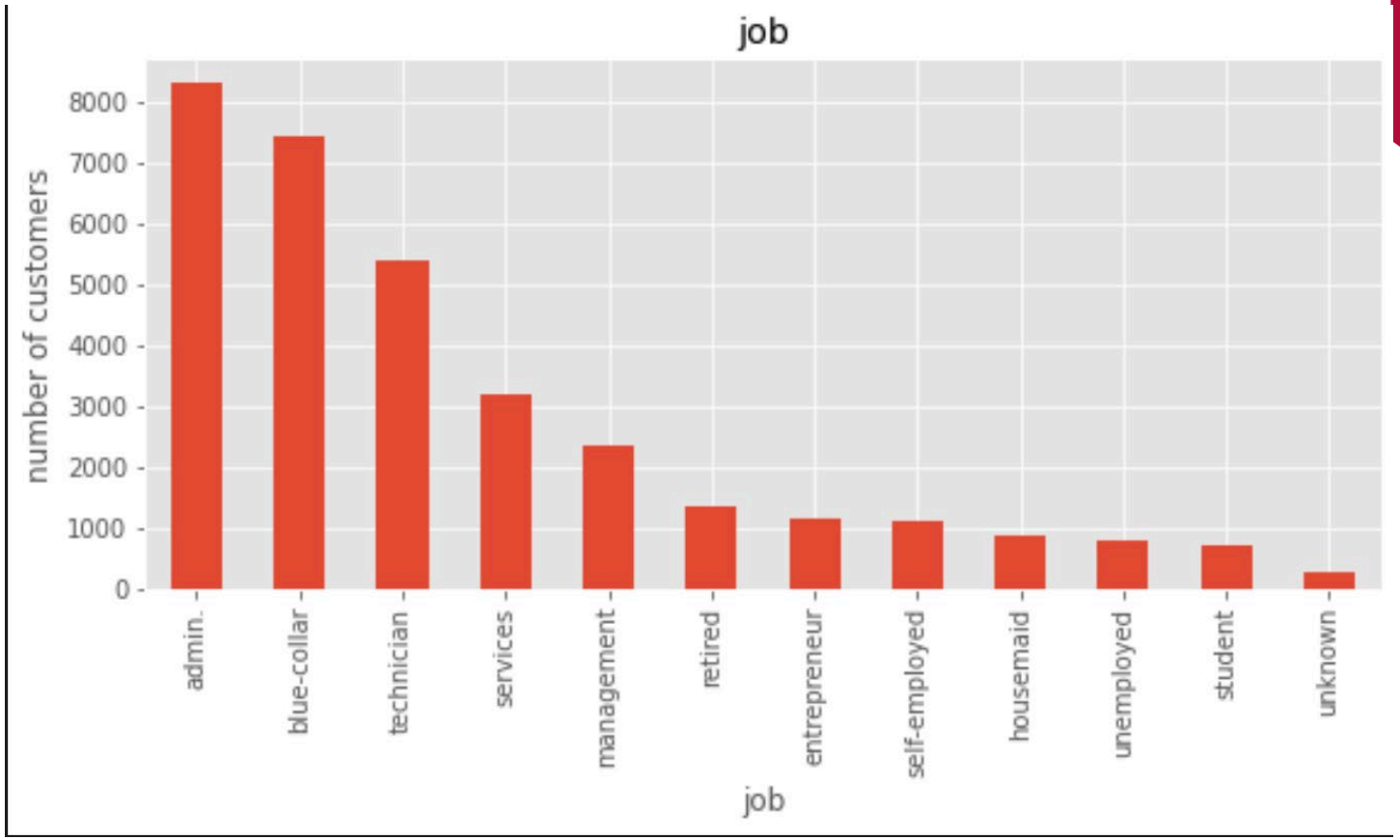
Direct marketing campaigns is being used for getting data for the Portuguese banking institution. The marketing campaign was based on telephone calls.

In order to determine if the service (Long term deposit) will be enrolled ('yes') or not ('no'), it was frequently necessary to make multiple contacts with a single consumer.


housing
number of customers
17500
15000
12500
10000
7500
5000
2500
0
yes
no
unknown
housing


marital
number of customers
20000
17500
15000
12500
10000
7500
5000
2500
0
married
single
divorced
unknown
marital


default
number of customers
25000
20000
15000
10000
5000
0
no
unknown
yes
default


job
number of customers
8000
7000
6000
5000
4000
3000
2000
1000
0
admin.
blue-collar
technician
services
management
retired
entrepreneur
self-employed
housemaid
unemployed
student
unknown
job

EDA

Classification Algorithm Implemented

KNN	Naive Bayes	Decision Tree	Random Forest
ANN	Logistic Regression	SVM	Multi-Layer Perceptron

K-Nearest Neighbors Algorithm

ACCURACY

0.8882144663631766

CONFUSION MATRIX

[[8532 191]

[914 248]]

AREA UNDER CURVE

0.595764962187724

F1 – SCORE

0.3098063710181137

MEAN ABSOLUTE ERROR

0.11178553363682348

MEAN SQUARED ERROR

0.11178553363682348

Naïve Bayes

ACCURACY

0.860394537177541 8

CONFUSION MATRIX

[[8135 588]

[792 370]]

AREA UNDER CURVE

0.625504260700784 5

F1 – SCORE

0.349056603773584 4

MEAN ABSOLUTE ERROR

0.139605462822458 28

MEAN SQUARED ERROR

0.139605462822458 28

Decision Tree

ACCURACY

0.836722306525038

CONFUSION MATRIX

[[7859 864]
[750 412]]

AREA UNDER CURVE

0.627756304528969

F1 – SCORE

0.3379819524200164

MEAN ABSOLUTE ERROR

0.16327769347496207

MEAN SQUARED ERROR

0.16327769347496207

Random Forest

ACCURACY

0.8990389479008599

CONFUSION MATRIX

[[8597 126]
[872 290]]

AREA UNDER CURVE

0.617562567789706

F1 – SCORE

0.367553865652725

MEAN ABSOLUTE ERROR

0.1009610520991401 1

MEAN SQUARED ERROR

0.1009610520991401 1

ARTIFICIAL NEURAL NETWORK

ACCURACY

0.8965

CONFUSION MATRIX

[[5227 551]
[414 3366]]

AREA UNDER CURVE

0.7479

F1 – SCORE

0.29032587

MEAN ABSOLUTE ERROR

0.1708

MEAN SQUARED ERROR

0.0842

LOGISTIC REGRESSION

ACCURACY

0.894790085988872

CONFUSION MATRIX

[[8614 109]
[931 231]]

AREA UNDER CURVE

0.5931497398513003

F1 – SCORE

0.3075898801597864

MEAN ABSOLUTE ERROR

0.10520991401112798

MEAN SQUARED ERROR

0.10520991401112798

MULTILAYER PERCEPTRON

ACCURACY

0.890642387455741

CONFUSION MATRIX

[[8523 200]
[881 281]]

AREA UNDER CURVE

0.6094482744196352

F1 – SCORE

0.3420572124163115

MEAN ABSOLUTE ERROR

0.10935761254425898

MEAN SQUARED ERROR

0.10935761254425898

SUPPORT VECTOR MACHINES

ACCURACY

0.896206373292868

CONFUSION MATRIX

[[8610 113]
[913 249]]

AREA UNDER CURVE

0.600665727714908

F1 – SCORE

0.3267716535433076

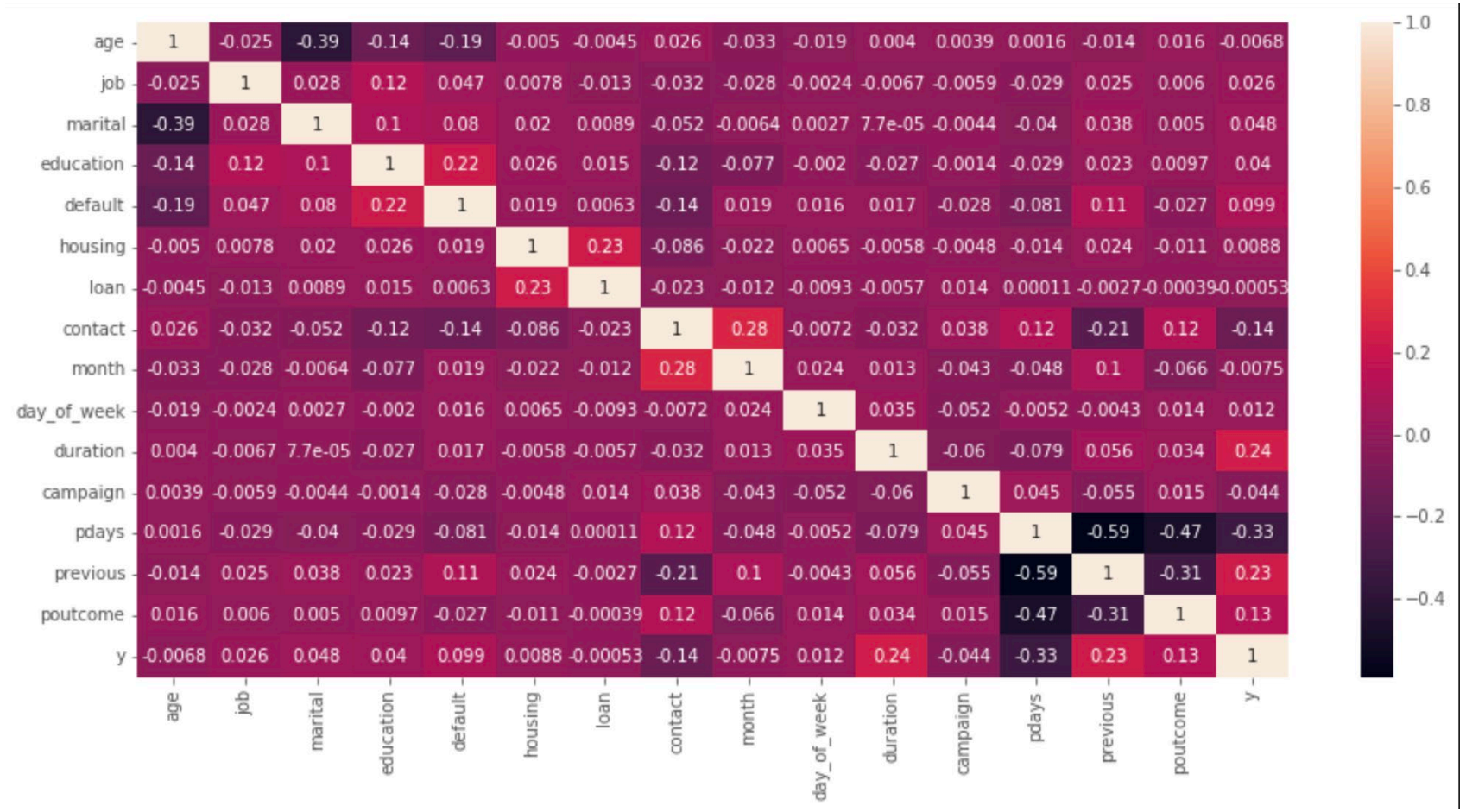
MEAN ABSOLUTE ERROR

0.10379362670713202

MEAN SQUARED ERROR

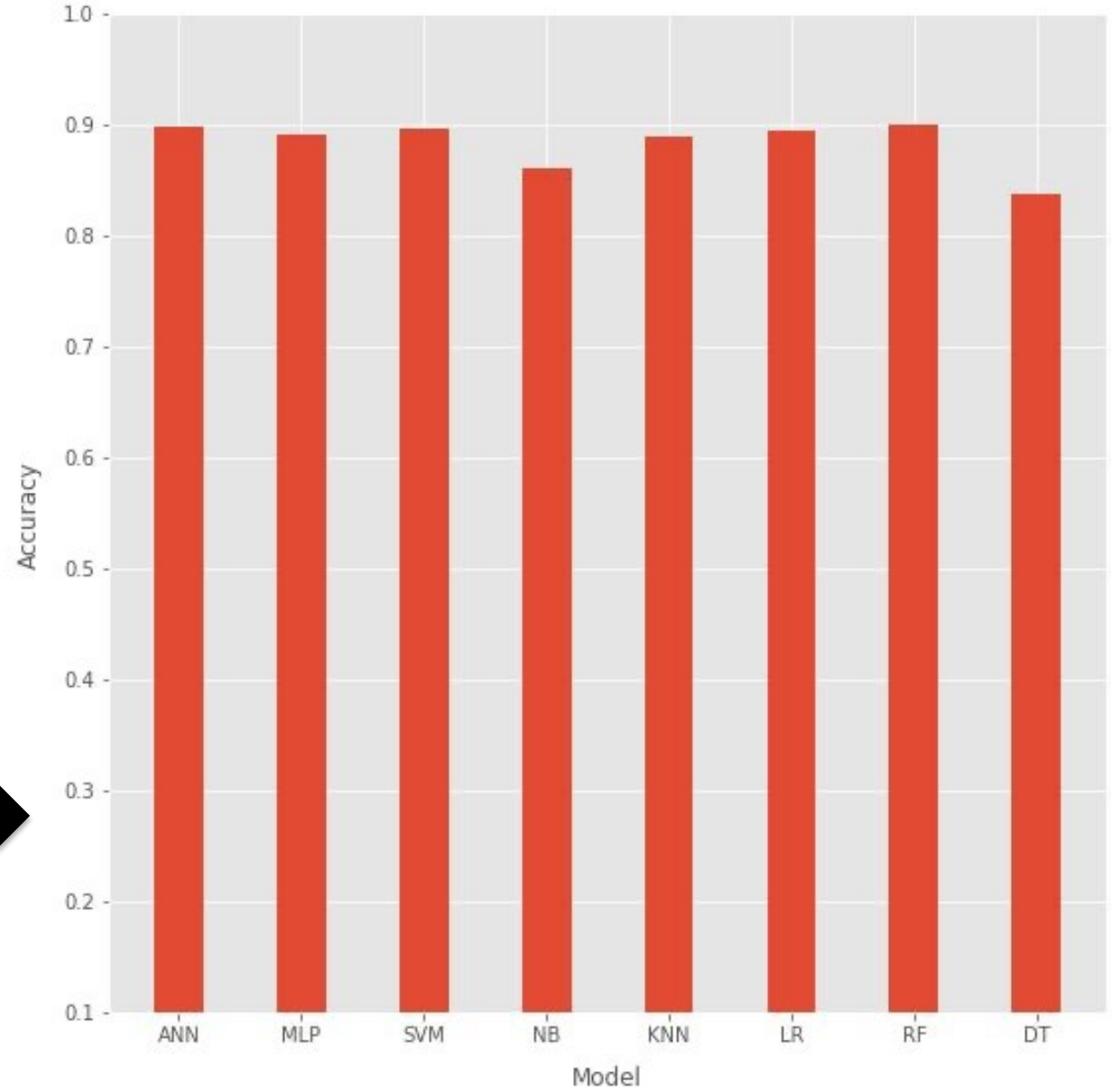
0.10379362670713202

Correlation Heat Map



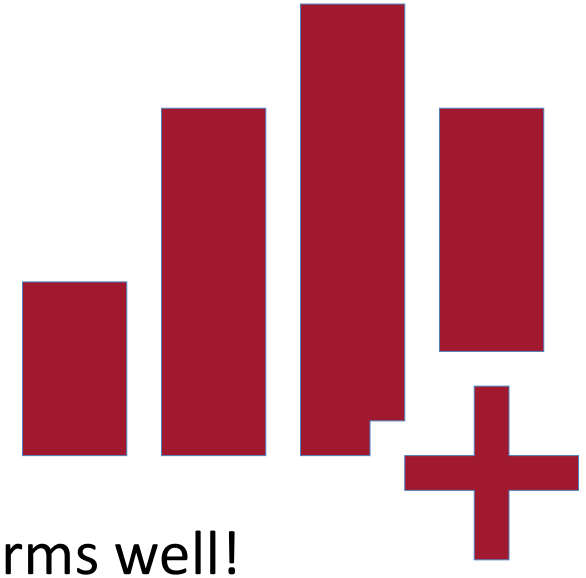
Comparing All the models

```
Accuracy of ANN is: 0.8975214958190918
Accuracy of MLP is: 0.890642387455741
Accuracy of SVM is: 0.896206373292868
Accuracy of NB is: 0.8824481537683359
Accuracy of KNN is: 0.8882144663631766
Accuracy of LR is: 0.894790085988872
Accuracy of RF is: 0.8990389479008599
Accuracy of DT is: 0.836722306525038
```



WHICH MODEL TO CHOOSE ??

CONCLUSION



Based on **Accuracy** as Evaluation Metric , ***RANDOM FOREST*** performs well!

STEVENS

INSTITUTE *of* TECHNOLOGY

THE INNOVATION UNIVERSITY®

stevens.edu