CS 725: Project Presentation

Analysis of performance of various classification models for a binary classification task on Bank marketing dataset

About the Dataset

The dataset we chose to work on is taken from the following link:
UCI Bank Marketing Dataset

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. The output will be if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

The CSV file of the data contains 17 columns, containing 16 attributes and 1 output(whether subscribed or not for the term deposit).

The dataset have in total 45211 samples out of which 39922 are negative samples and 5289 are positive samples.

Attribute Information

The Attributes it have and their detailed description is as follows:

1 - age (numeric)

2 - job : type of job (categorical: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')

3 - marital : marital status (categorical: 'divorced', 'married', 'single', 'unknown')

4 - education (categorical: 'tertiary', 'secondary', 'primary')

5 - default: has credit in default? (categorical: 'no','yes','unknown')

6 - balance: How much account balance a person have if he have an account with the bank

7 - housing: has housing loan? (categorical: 'no','yes','unknown')

8 - loan: has personal loan? (categorical: 'no','yes','unknown')

9 - contact: contact communication type (categorical: 'cellular','telephone')

Attribute Information (Contd.)

10 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')

11 - day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')

12 - duration: last contact duration, in seconds (numeric).

13 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

14 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)

15 - previous: number of contacts performed before this campaign and for this client (numeric)

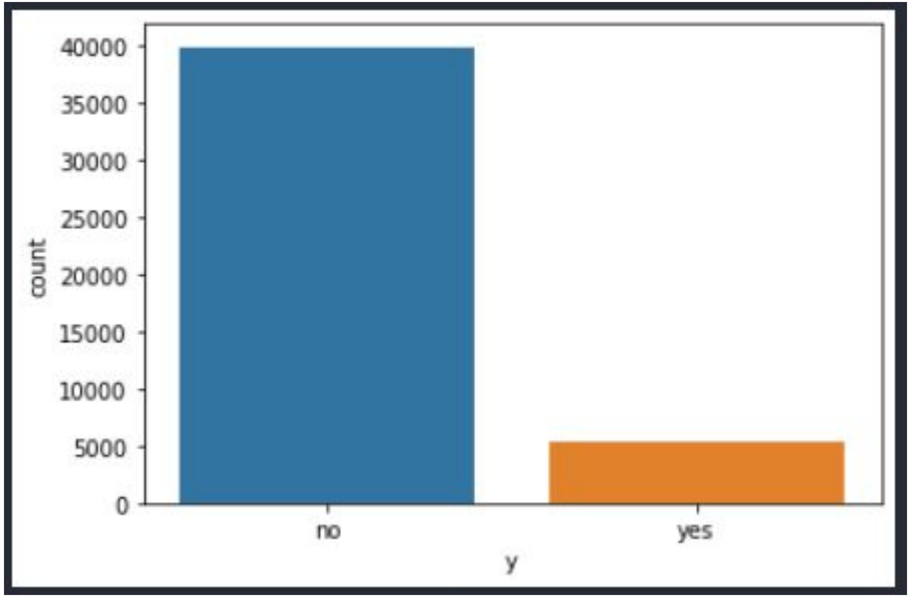
16 - poutcome: outcome of the previous marketing campaign (categorical: 'failure', 'unknown', 'success')

Data Cleaning And analysis

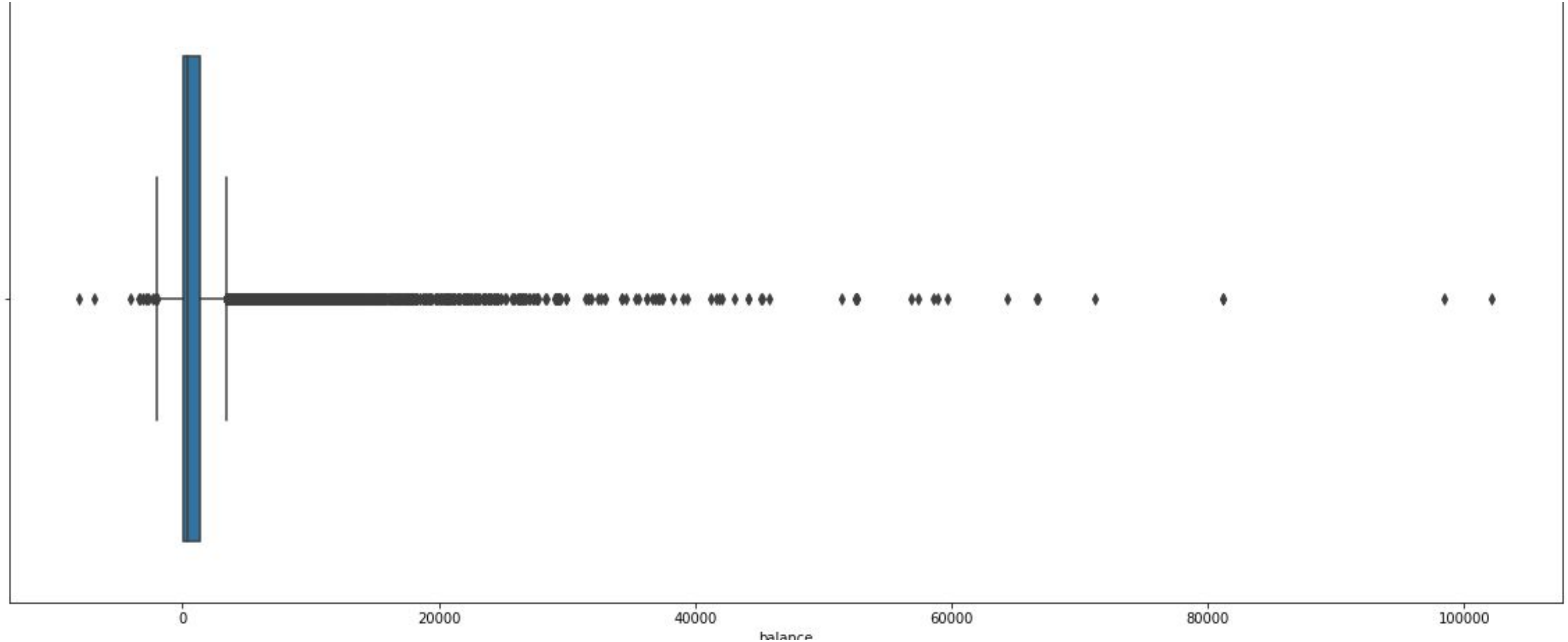
1) In balance column , We have dropped the outlier values which are more than 2.5 standard deviations away from the mean.

2) We Dropped Contact field and changed duration of call from seconds to minutes and change month from words to integers.

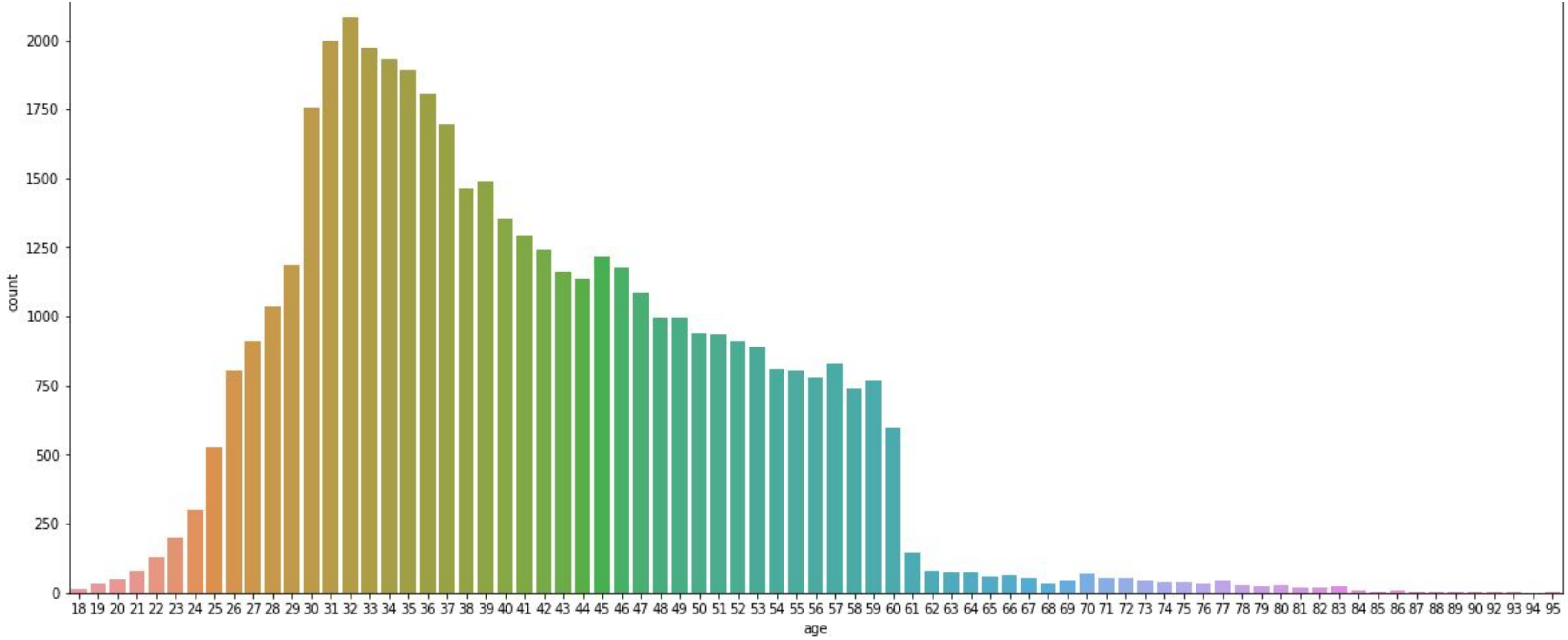
3) We dropped calls that are less than 10s and we dropped rows with education, job, poutcome stated as other.
4) We have also dropped Senior Citizens acting as outliers (age more than 65).



Graph of “Yes”/”No’ with Frequency

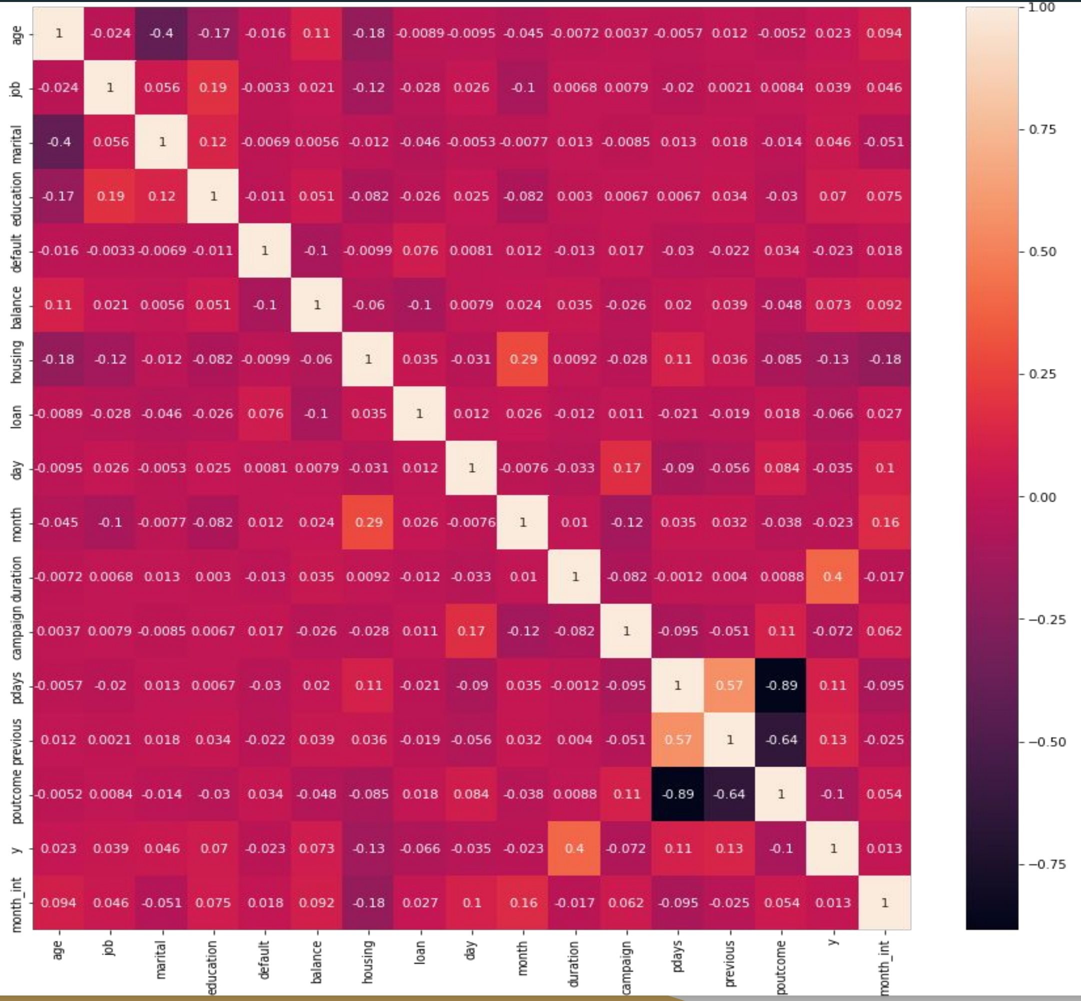


Graph for “Balance” Distribution



Graph of “Age” Distribution

Correlation Matrix of the data



	age	job	marital	education	default	balance	housing	loan	day	month	duration	campaign	pdays	previous	poutcome	y	month_int
age	1	-0.024	-0.4	-0.17	-0.016	0.11	-0.18	-0.0089	-0.0095	-0.045	-0.0072	0.0037	-0.0057	0.012	-0.0052	0.023	0.094
job	-0.024	1	0.056	0.19	-0.0033	0.021	-0.12	-0.028	0.026	-0.1	0.0068	0.0079	-0.02	0.0021	0.0084	0.039	0.046
marital	-0.4	0.056	1	0.12	-0.0069	0.0056	-0.012	-0.046	-0.0053	-0.0077	0.013	-0.0085	0.013	0.018	-0.014	0.046	-0.051
education	-0.17	0.19	0.12	1	-0.011	0.051	-0.082	-0.026	0.025	-0.082	0.003	0.0067	0.0067	0.034	-0.03	0.07	0.075
default	-0.016	-0.0033	-0.0069	-0.011	1	-0.1	-0.0099	0.076	0.0081	0.012	-0.013	0.017	-0.03	-0.022	0.034	-0.023	0.018
balance	0.11	0.021	0.0056	0.051	-0.1	1	-0.06	-0.1	0.0079	0.024	0.035	-0.026	0.02	0.039	-0.048	0.073	0.092
housing	-0.18	-0.12	-0.012	-0.082	-0.0099	-0.06	1	0.035	-0.031	0.29	0.0092	-0.028	0.11	0.036	-0.085	-0.13	-0.18
loan	-0.0089	-0.028	-0.046	-0.026	0.076	-0.1	0.035	1	0.012	0.026	-0.012	0.011	-0.021	-0.019	0.018	-0.066	0.027
day	-0.0095	0.026	-0.0053	0.025	0.0081	0.0079	-0.031	0.012	1	-0.0076	-0.033	0.17	-0.09	-0.056	0.084	-0.035	0.1
month	-0.045	-0.1	-0.0077	-0.082	0.012	0.024	0.29	0.026	-0.0076	1	0.01	-0.12	0.035	0.032	-0.038	-0.023	0.16
duration	-0.0072	0.0068	0.013	0.003	-0.013	0.035	0.0092	-0.012	-0.033	0.01	1	-0.082	-0.0012	0.004	0.0088	0.4	-0.017
campaign	0.0037	0.0079	-0.0085	0.0067	0.017	-0.026	-0.028	0.011	0.17	-0.12	-0.082	1	-0.095	-0.051	0.11	-0.072	0.062
pdays	-0.0057	-0.02	0.013	0.0067	-0.03	0.02	0.11	-0.021	-0.09	0.035	-0.0012	-0.095	1	0.57	-0.89	0.11	-0.095
previous	0.012	0.0021	0.018	0.034	-0.022	0.039	0.036	-0.019	-0.056	0.032	0.004	-0.051	0.57	1	-0.64	0.13	-0.025
poutcome	-0.0052	0.0084	-0.014	-0.03	0.034	-0.048	-0.085	0.018	0.084	-0.038	0.0088	0.11	-0.89	-0.64	1	-0.1	0.054
y	0.023	0.039	0.046	0.07	-0.023	0.073	-0.13	-0.066	-0.035	-0.023	0.4	-0.072	0.11	0.13	-0.1	1	0.013
month_int	0.094	0.046	-0.051	0.075	0.018	0.092	-0.18	0.027	0.1	0.16	-0.017	0.062	-0.095	-0.025	0.054	0.013	1

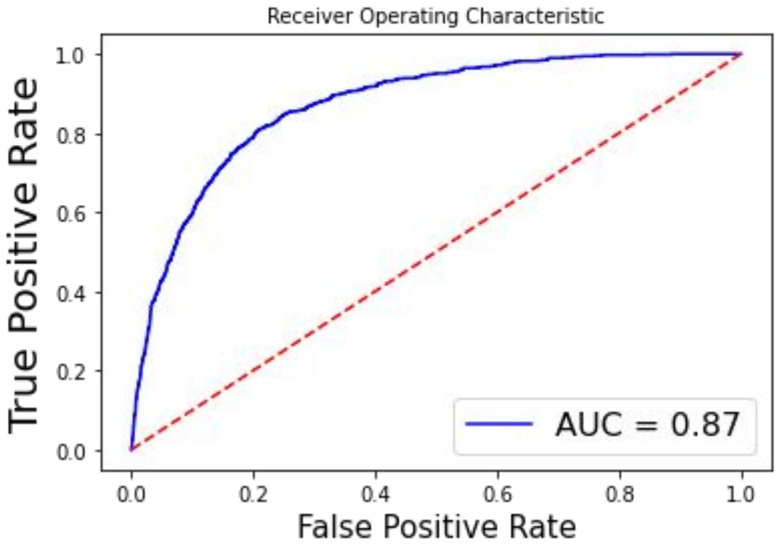
Techniques used

We used the following techniques for the chosen Binary Classification task:

1) Logistic Regression
2) Decision Trees
3) Support Vector Machines
4) Neural Networks

Logistic Regression

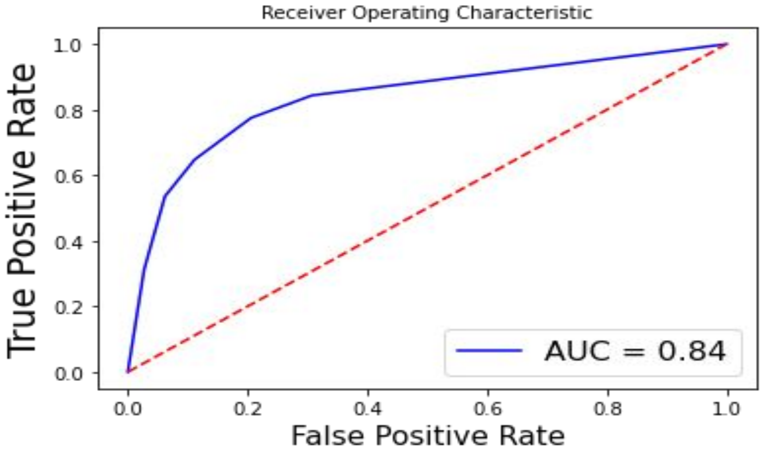
```
Confusion matrix

[[7073  117]

 [ 716  187]]

Accuracy = 90.0

Classification Report

     precision    recall  f1-score    support

0        0.91       0.98       0.94       7190

1        0.62       0.21       0.31        903
```



Decision Trees

Entropy as criterion is used .

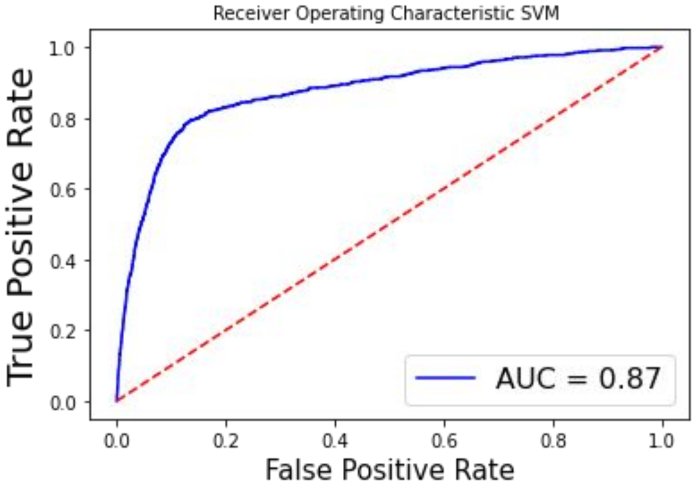
```
Confusion matrix

[[6995  195]

 [ 622  281]]

Accuracy = 90.0

Classification Report

              precision    recall  f1-score   support

           0       0.92      0.97      0.94      7190

           1       0.59      0.31      0.41       903
```



Support Vector Machines

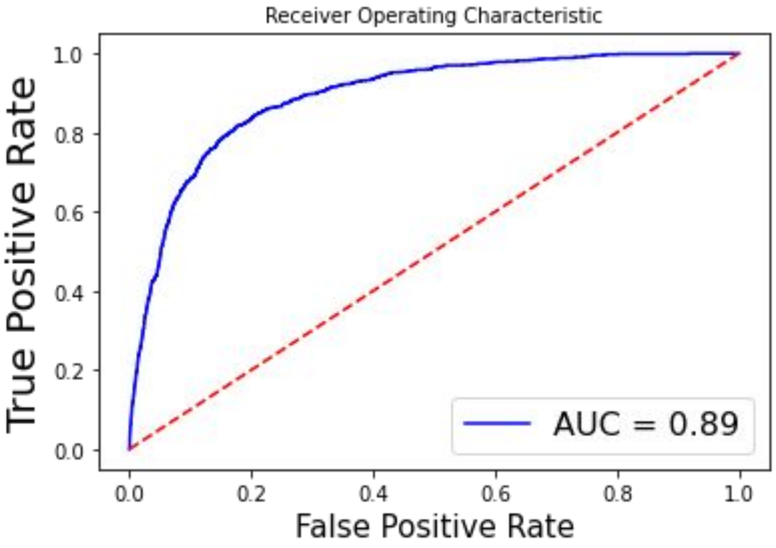
1. RBF used as basis function
2. Reducing the regularization parameter has no effect

```
Confusion matrix

[[7047  142]

 [ 615  289]]

Accuracy = 91.0 %

Classification Report

        precision   recall  f1-score   support

0         0.92       0.98     0.95       7189

1         0.67       0.32     0.43        904
```



Neural Networks

```
Confusion matrix

[[6707  483]

 [ 368  535]]

Accuracy = 89.0

Classification Report

             precision    recall  f1-score   support

          0       0.95      0.93      0.94      7190

          1       0.53      0.59      0.56       903
```



Balancing Techniques

Model	Oversampling minority	Undersampling Majority	SMOTE	Unbalanced
Logistic Regression	80.0	80.0	82.0	90.0
Decision Trees	80.0	81.0	81.0	90.0
SVM	84.0 (AUC 0.91)	82.0	85.0	91.0
Neural Net.	88.0	79.0	88.0	89.0

Ablation Study

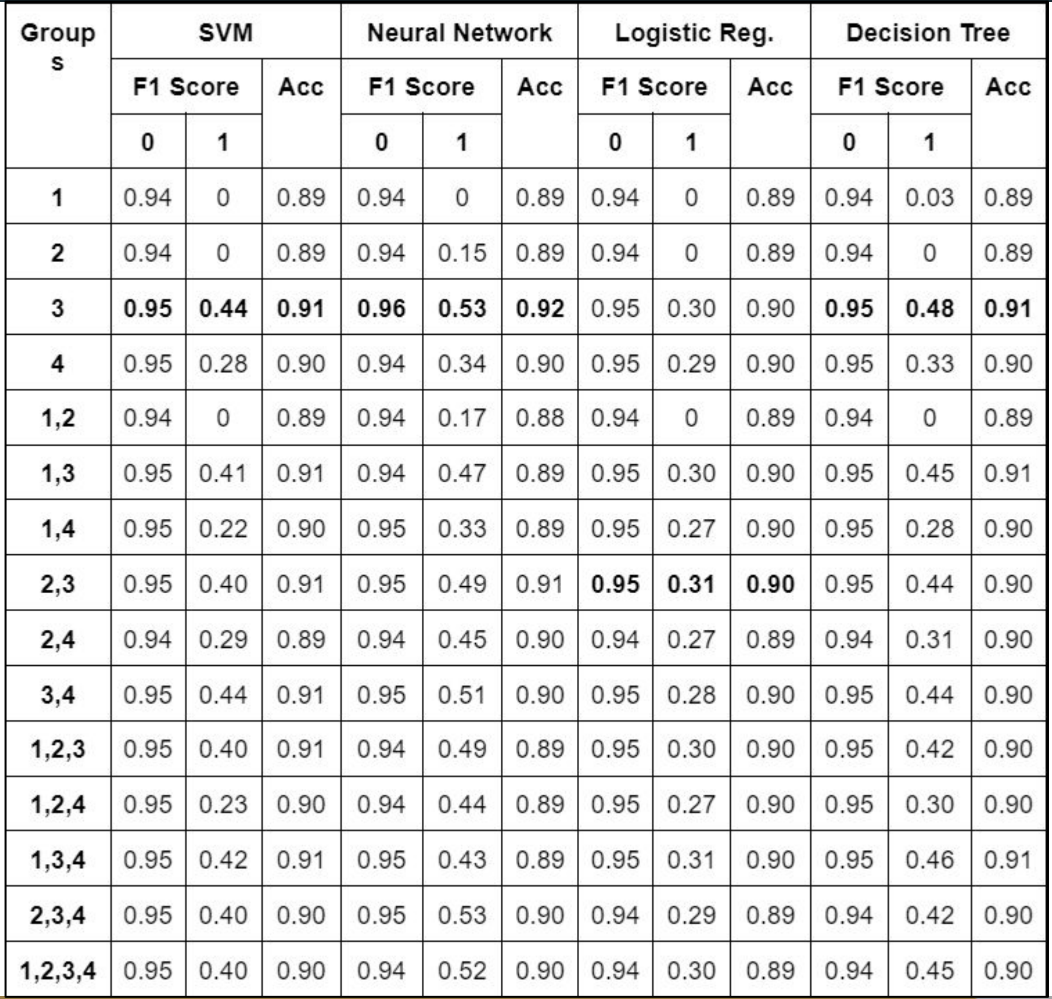
We divided the set of 16 features into 4 groups:

Group 1: Client Information related features---age, job, marital, education, default

Group 2: Banking related attributes---housing, loan, balance, month_int

Group 3: Related to the last contact made---day, month, duration, campaign, pdays, previous, poutcome

Group 4: Most important 5 features as can be seen from Feature Importance---duration, balance, day, age



Group s	SVM			Neural Network			Logistic Reg.			Decision Tree		
	F1 Score		Acc	F1 Score		Acc	F1 Score		Acc	F1 Score		Acc
	0	1		0	1		0	1		0	1	
1	0.94	0	0.89	0.94	0	0.89	0.94	0	0.89	0.94	0.03	0.89
2	0.94	0	0.89	0.94	0.15	0.89	0.94	0	0.89	0.94	0	0.89
3	**0.95**	**0.44**	**0.91**	**0.96**	**0.53**	**0.92**	0.95	0.30	0.90	**0.95**	**0.48**	**0.91**
4	0.95	0.28	0.90	0.94	0.34	0.90	0.95	0.29	0.90	0.95	0.33	0.90
1,2	0.94	0	0.89	0.94	0.17	0.88	0.94	0	0.89	0.94	0	0.89
1,3	0.95	0.41	0.91	0.94	0.47	0.89	0.95	0.30	0.90	0.95	0.45	0.91
1,4	0.95	0.22	0.90	0.95	0.33	0.89	0.95	0.27	0.90	0.95	0.28	0.90
2,3	0.95	0.40	0.91	0.95	0.49	0.91	**0.95**	**0.31**	**0.90**	0.95	0.44	0.90
2,4	0.94	0.29	0.89	0.94	0.45	0.90	0.94	0.27	0.89	0.94	0.31	0.90
3,4	0.95	0.44	0.91	0.95	0.51	0.90	0.95	0.28	0.90	0.95	0.44	0.90
1,2,3	0.95	0.40	0.91	0.94	0.49	0.89	0.95	0.30	0.90	0.95	0.42	0.90
1,2,4	0.95	0.23	0.90	0.94	0.44	0.89	0.95	0.27	0.90	0.95	0.30	0.90
1,3,4	0.95	0.42	0.91	0.95	0.43	0.89	0.95	0.31	0.90	0.95	0.46	0.91
2,3,4	0.95	0.40	0.90	0.95	0.53	0.90	0.94	0.29	0.89	0.94	0.42	0.90
1,2,3,4	0.95	0.40	0.90	0.94	0.52	0.90	0.94	0.30	0.89	0.94	0.45	0.90

Conclusion

1. SVM’s have the best precision score
2. Neural Networks provide the best AUC

Contributions

1. Prishat Bachhar [213050078] - SVM, Data Balancing, EDA
2. Anmol Namdev [213050044] - NN, data Balancing, Ablation Study
3. Shivam Gautam [213050003] - Data Cleaning
4. Subodh Lathkar [213050047] - Decision Trees
5. Kalpankur Pandey [204277002]- Logistic Regression

Code sources

1. https://www.kaggle.com/henriqueyamahata/bank-marketing-classification-roc-f1-recall
2. SKlearn Website
3. https://towardsdatascience.com/methods-for-dealing-with-imbalanced-data-5b761be45a18

Thank you