IBM PROFESSIONAL CERTIFICATE: Unsupervised Learning - Clustering

Ibrahim Mohamed

August 2022

MAIN OBJECTIVE

- This analysis' primary goal is to cluster bank customers into groups and validate if they are potential customer for term deposit.
- So clustering is the main target analysis for unsupervised learning
- Show the correlation between the features & the most features with impact.
- Validate the clusters with binary classification.

ABOUT THE DATA

- This dataset is downloaded from The University of California Irvine Machine Learning Repositor.
- The data is related with direct marketing campaigns of a Portuguese banking institution.
- The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be (or not) subscribed.
- This data set has 4521 records and 17 features.

DATA EXPLORATION

- Features:

age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
33	services	married	secondary	no	4789	yes	yes	cellular	11	may	220	1	339	4	failure	no
35	management	single	tertiary	no	1350	yes	no	cellular	16	apr	185	1	330	1	failure	no
30	unemployed	married	primary	no	1787	no	no	cellular	19	oct	79	1	-1	0	unknown	no
30	management	married	tertiary	no	1476	yes	yes	unknown	3	jun	199	4	-1	0	unknown	no
59	blue-collar	married	secondary	no	0	yes	no	unknown	5	may	226	1	-1	0	unknown	no

- **age:** customer age in years
- **Job:** customer job
- **marital:** customer martial status
- **education**: customer education
- **default**: has credit in default?
- **balance**: average yearly balance, in euros

DATA EXPLORATION

- Features:

- **Housing:** has housing loan?
- **loan**: has personal loan?
- **contact**: contact communication type
- **day**: last contact day of the month
- **month**: last contact month of year
- **duration**: last contact duration
- **campaign**: number of contacts performed during this campaign and for this client
- **pdays**: number of days that passed by after the client was last contacted from a previous campaign
- **previous**: number of contacts performed before this campaign and for this client
- **poutcome**: outcome of the previous marketing campaign
- **y**: has the client subscribed a term deposit? - TARGET

DATA EXPLORATION

- Description:

Mean

age:41
job: NaN
marital: NaN
education: NaN
default: NaN
balance: 1422
housing: NaN
loan: NaN
day: 15
month: NaN
duration: 263
campaign: 3
pdays: 39
previous: 0.5
poutcome: NaN
y: NaN

Std

age: 10
job: NaN
marital: NaN
education: NaN
default: NaN
balance: 3009
housing: NaN
loan: NaN
day: 8
month: NaN
duration: 259
campaign: 3
pdays: 100
previous: 1.6
poutcome: NaN
y: NaN

Max

age:87
job: NaN
marital: NaN
education: NaN
default: NaN
balance: 71188
housing: NaN
loan: NaN
day: 31
month: NaN
duration: 3025
campaign: 50
pdays: 871
previous: 25
poutcome: NaN
y: NaN

Min

age:19
job: NaN
marital: NaN
education: NaN
default: NaN
balance: -3313
housing: NaN
loan: NaN
day: 1
month: NaN
duration: 4
campaign: 1
pdays: -1
previous: 0
poutcome: NaN
y: NaN

IBM

DATA EXPLORATION

- Description:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
age	4521.0	NaN	NaN	NaN	41.170095	10.576211	19.0	33.0	39.0	49.0	87.0
job	4521	12	management	969	NaN	NaN	NaN	NaN	NaN	NaN	NaN
marital	4521	3	married	2797	NaN	NaN	NaN	NaN	NaN	NaN	NaN
education	4521	4	secondary	2306	NaN	NaN	NaN	NaN	NaN	NaN	NaN
default	4521	2	no	4445	NaN	NaN	NaN	NaN	NaN	NaN	NaN
balance	4521.0	NaN	NaN	NaN	1422.657819	3009.638142	-3313.0	69.0	444.0	1480.0	71188.0
housing	4521	2	yes	2559	NaN	NaN	NaN	NaN	NaN	NaN	NaN
loan	4521	2	no	3830	NaN	NaN	NaN	NaN	NaN	NaN	NaN
contact	4521	3	cellular	2896	NaN	NaN	NaN	NaN	NaN	NaN	NaN
day	4521.0	NaN	NaN	NaN	15.915284	8.247667	1.0	9.0	16.0	21.0	31.0
month	4521	12	may	1398	NaN	NaN	NaN	NaN	NaN	NaN	NaN
duration	4521.0	NaN	NaN	NaN	263.961292	259.856633	4.0	104.0	185.0	329.0	3025.0
campaign	4521.0	NaN	NaN	NaN	2.79363	3.109807	1.0	1.0	2.0	3.0	50.0
pdays	4521.0	NaN	NaN	NaN	39.766645	100.121124	-1.0	-1.0	-1.0	-1.0	871.0
previous	4521.0	NaN	NaN	NaN	0.542579	1.693562	0.0	0.0	0.0	0.0	25.0
poutcome	4521	4	unknown	3705	NaN	NaN	NaN	NaN	NaN	NaN	NaN
y	4521	2	no	4000	NaN	NaN	NaN	NaN	NaN	NaN	NaN

DATA EXPLORATION

- Data Types & Null Values

- Our Data Types are the following
- Our data doesn't have any missing values except for balance column which has 361 null value

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4521 entries, 0 to 4520
Data columns (total 17 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   age        4521 non-null   int64
 1   job        4521 non-null   object
 2   marital    4521 non-null   object
 3   education  4521 non-null   object
 4   default    4521 non-null   object
 5   balance    4521 non-null   int64
 6   housing    4521 non-null   object
 7   loan       4521 non-null   object
 8   contact    4521 non-null   object
 9   day        4521 non-null   int64
 10  month      4521 non-null   object
 11  duration   4521 non-null   int64
 12  campaign   4521 non-null   int64
 13  pdays      4521 non-null   int64
 14  previous   4521 non-null   int64
 15  poutcome   4521 non-null   object
 16  y          4521 non-null   object
dtypes: int64(7), object(10)
memory usage: 600.6+ KB
```

	data
age	0
default	0
balance	361
housing	0
loan	0
duration	0
campaign	0
pdays	0
previous	0
y	0
contact_telephone	0
contact_unknown	0
marital_married	0
marital_single	0
job_blue-collar	0
job_entrepreneur	0
job_housemaid	0
job_management	0
job_retired	0
job_self-employed	0
job_services	0
job_student	0
job_technician	0

EXPLORATORY DATA ANALYSIS

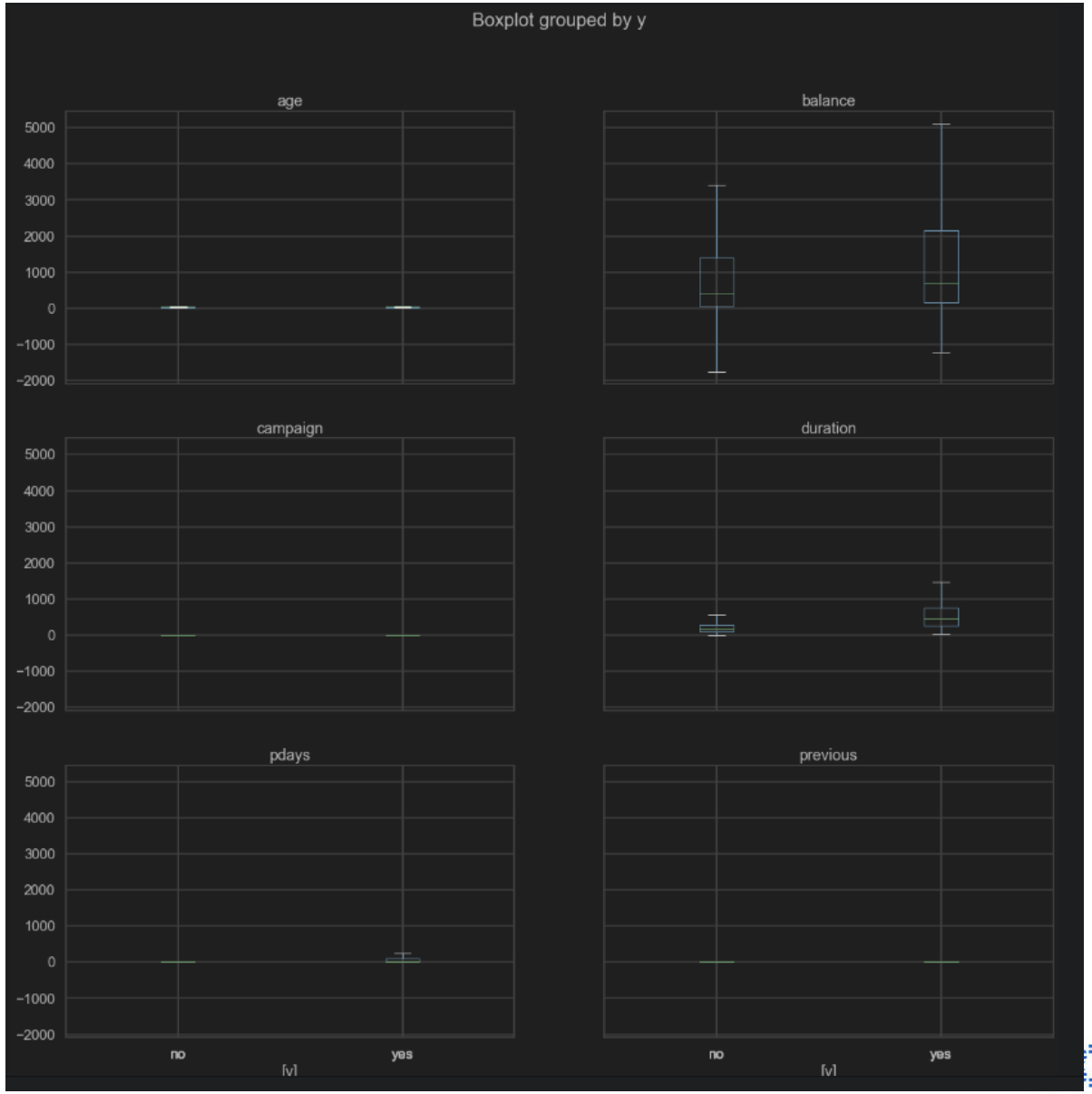
- Categorical features & Numerical features

- Our data contain both categorical features and numerical features
- We need to change the categorical data into numerical data.
- This is done by splitting categorise into separate columns.

- **Binary variables:** ['default', 'housing', 'loan', 'y']
- **Categorical variables:** ['job', 'marital', 'education', 'contact', 'poutcome']
- **Numerical variables:** ['duration', 'previous', 'pdays', 'campaign', 'balance', 'age']

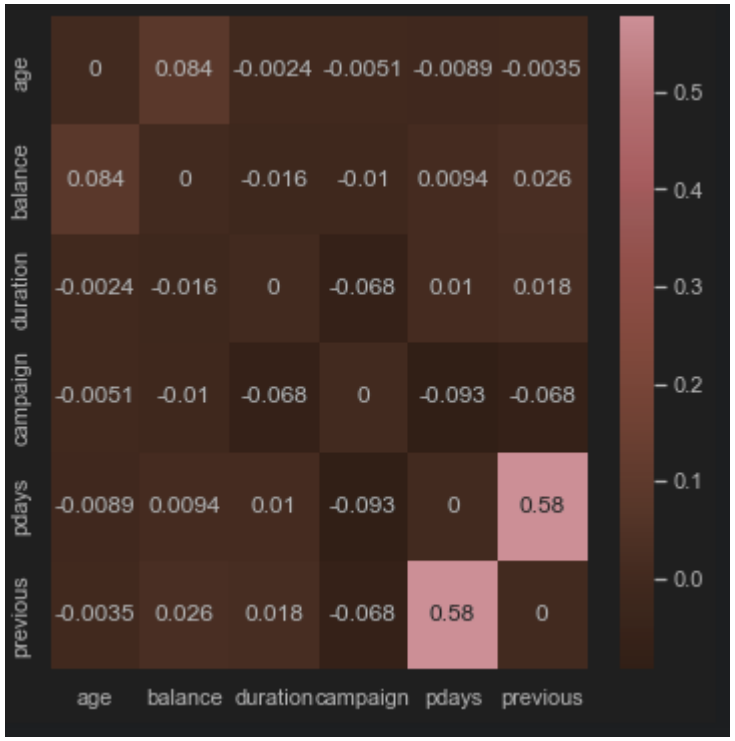
EXPLORATORY DATA ANALYSIS

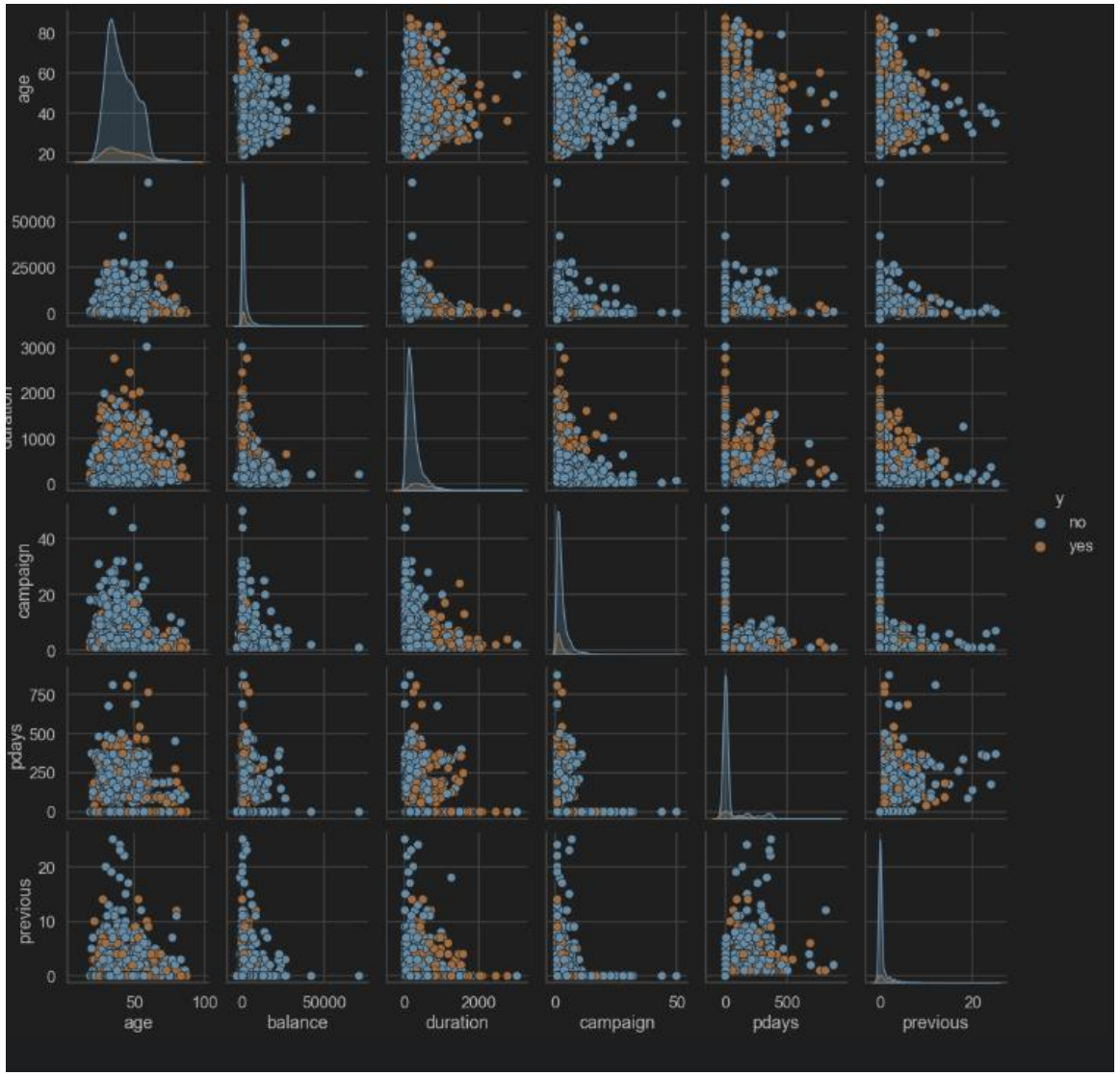
- Identify outliers in the data



EXPLORATORY DATA ANALYSIS

- Correlation between features





EXPLORATORY DATA ANALYSIS

- Correlation between features

- These features are most correlated to each other
- Age – balance
- Duration – campaign
- Campagin – pday
- Pday – previous
- Previous - pday

```
corr_mat.abs().idxmax()
```

	data
age	balance
balance	age
duration	campaign
campaign	pdays
pdays	previous
previous	pdays

EXPLORATORY DATA ANALYSIS

- Feature Engineering

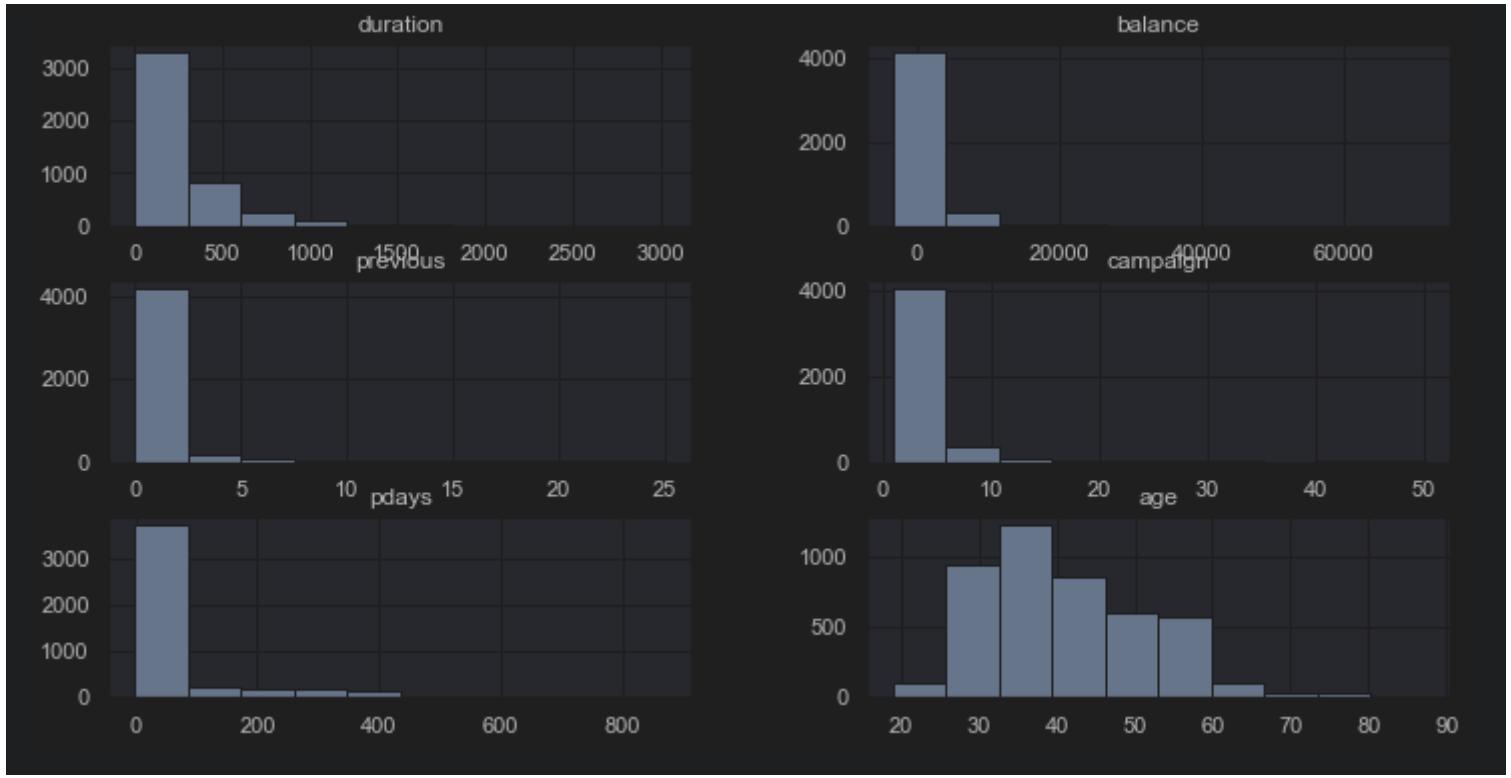
- We dropped the un needed columns like day, month.
- Get the number of unique values

Variable	Unique Values
age	67
job	12
marital	3
education	4
default	2
balance	2353
housing	2
loan	2
contact	3
duration	875
campaign	32
pdays	292
previous	24
poutcome	4
y	2

EXPLORATORY DATA ANALYSIS

- Feature Engineering

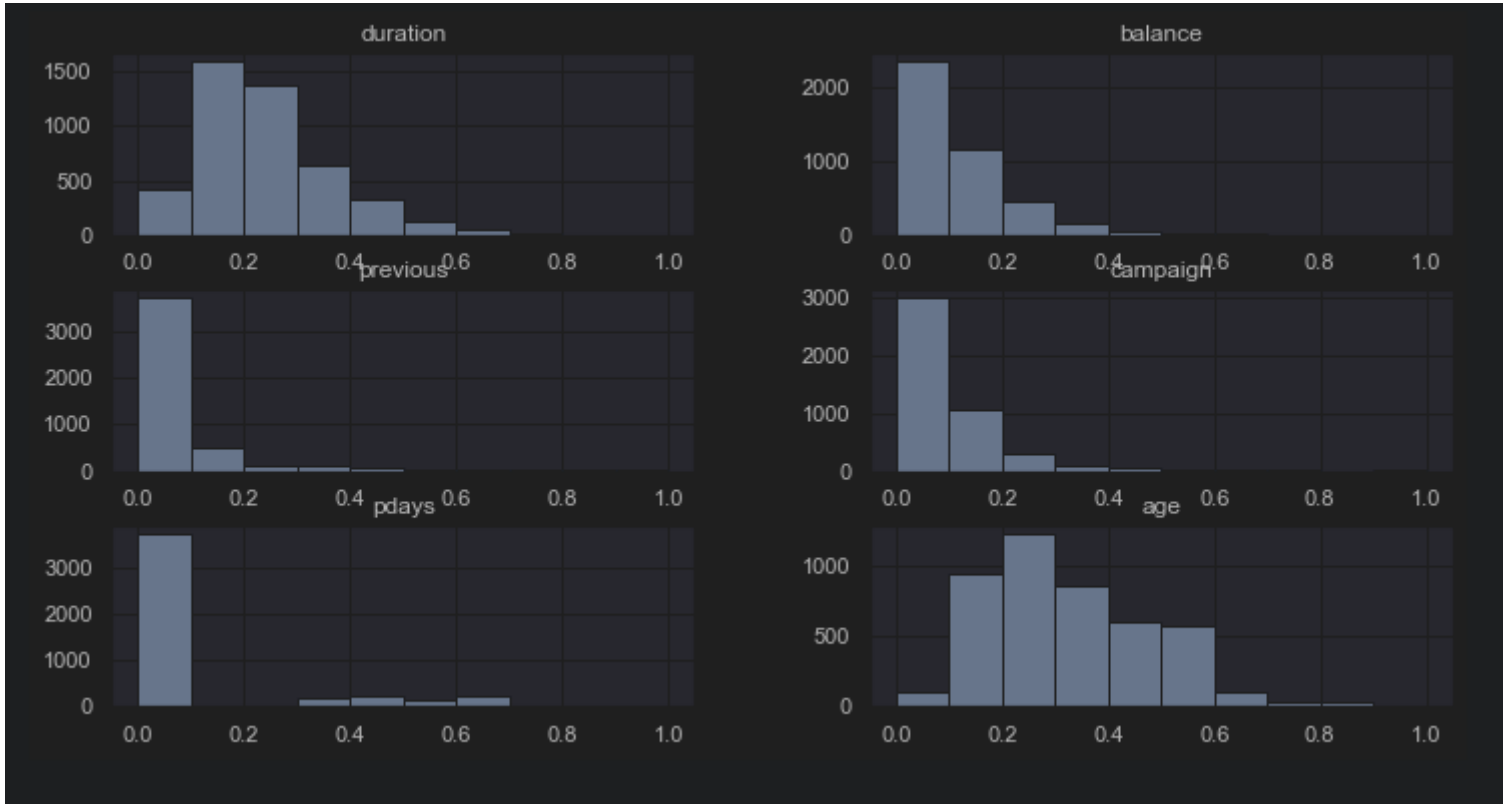
- Plot of the numerical values



EXPLORATORY DATA ANALYSIS

- Feature Engineering

- Remove Skew with boxcox1p transformation



MACHINE LEARNING ANALYSIS

- K-Means Clustering

Model Features and Parameters:

- Model = KMeans()
- N_clusters = 2
- Init = ‘k-means++’
- N_init = 10
- Max_iter = 300
- Random_state = 0

```
# Lets take the optimal number of clusters as 6
km = KMeans(n_clusters=num_clusters, init='k-means++', n_init=10, max_iter=300, random_state=0)
km.fit(X_train)
# predict kmeamns labels
y_pred_kmeans = km.predict(X_test)
```

IBM

MACHINE LEARNING ANALYSIS

- Mean Shift Clustering

Model Features and Parameters:

- Model = MeanShift()
- Min_bin_freq = 10
- Max_iter = 1000

```
clustering = MeanShift(min_bin_freq=10,max_iter=1000).fit(X_train)
y_pred_meanshift = clustering.predict(X_test)
```

MACHINE LEARNING ANALYSIS

- Validation with classification model

Model Features and Parameters:

- Model = DecisionTreeClassifer()
- Random_state = 42

```
# Lets fit a simple Decision Tree for the classification
dt = DecisionTreeClassifier(random_state=42)
dt = dt.fit(X_train, y_train)
y_test_pred = dt.predict(X_test)
```

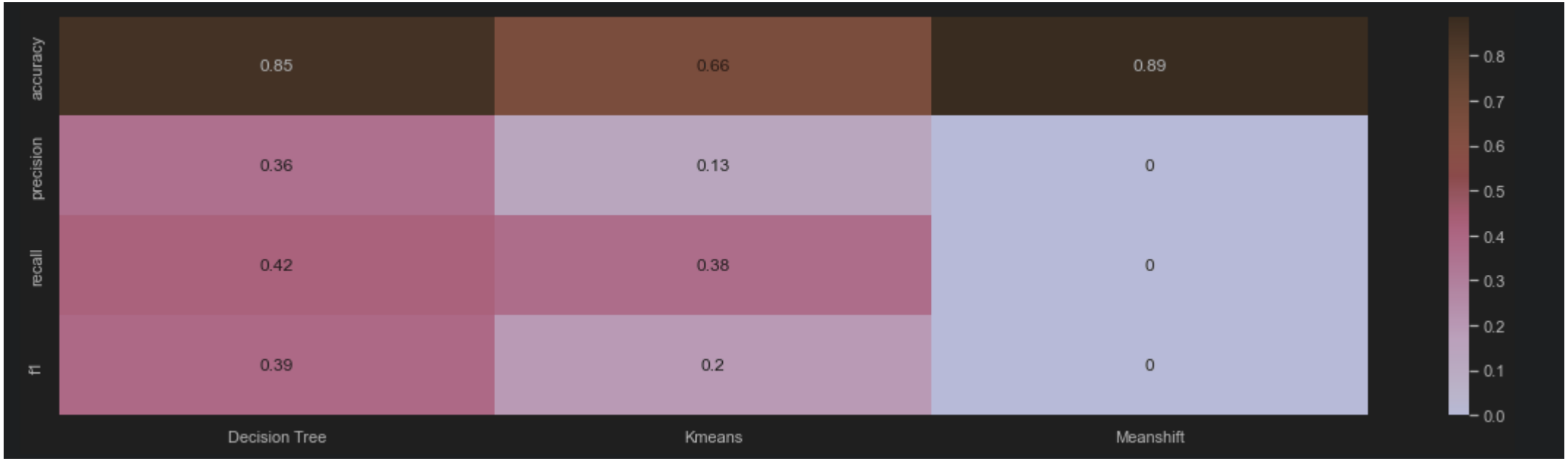
MACHINE LEARNING ANALYSIS

- Comparing Models

Actual	DecisionTree	Kmeans	Meanshift	number
0	0	0	0	774
		1	0	318
	1	0	0	62
		1	0	51
1	0	0	0	62
		1	0	26
	1	0	0	33
		1	0	31

MACHINE LEARNING ANALYSIS

- Comparing Models



MACHINE LEARNING ANALYSIS

- Models Comparison

- With the above analysis for Unsupervised Learning, K-Means Clustering perform better than MeanShift Clustering.
- MeanShift clustering predicted all the customers into single cluster
- Compared to Supervised "Decision Tree" model, K-Means Clustering model performs with the accuracy.
- We will conclude, K-Means Clustering model is a good model for unsupervised learning for this dataset.

	RMSE	R2	RMSE-SGD	R2-SGD
Linear	4496.560111	0.862103	4531.504262	0.859951
Lasso	4496.577652	0.862102	4570.227510	0.857548
Ridge	4494.682980	0.862218	4512.691171	0.861112
ElasticNet	4494.417701	0.862234	4528.496874	0.860137

ANALYSIS NEXT STEPS

- Models Flaws and Strength and further suggestions

- We have used only small dataset for this assignment. To proceed further, we will use large dataset.
- Apart from K-Means clustering model, we can explore other clustering models like DBSCAN, Agglomerative Clustering, etc.
- We can use Decision Tree model for classification and run in parallel to K-Means Clustering model in the long run to get better results.

IBM