BAN 5753 ADVANCED BUSINESS ANALYTICS

Mini Project #2: Deposit Opening Classification

TEAM 4: ABRACADATA

AGRAWAL, SONALI (A20406851), SONALI.AGRAWAL@OKSTATE.EDU

ALAWIYE, RHODA (A20392590), RHODA.ALAWIYE@OKSTATE.EDU

IPPILI, MURALIDHAR (A20384428), MURALIDHAR.IPPILI@OKSTATE.EDU

WOOD, JACOB, (A20385445), JACOB.WOOD10@OKSTATE.EDU

ZERMAN, ANDREA (A20387316), ANDREA.ZERMAN@OKSTATE.EDU

BUSINESS UNDERSTANDING

Goal: Identify customers likely to open a term deposit account.

- Understanding financial goals of customers to allow tailoring of products and services to meet specific needs
- Effective resource allocation and targeted marketing strategies
- Term accounts involve long-term commitments leading to longer lasting customer relationships and stability to XYZ's deposit base

The end result would lead to an enhanced overall risk management and financial planning strategy ensuring a robust and sustainable business model.

Data Set: Previous telephone calling campaign conducted between May 2008 and November 2010.

- Contains 41,188 records
- 20 attributes relating to the customers including demographic, latest contact, socio-economic and other attributes

APPROACH

Data Exploration
- Missing data analysis
- Distribution of independent variables
- Bivariate analysis between target and independent variables
- Correlation analysis

Handling Class Imbalance
- Number of customers who signed up for the term deposit are significantly lower than those who didn't
- Oversampled the minority class in the training data for better predictive model

Created Pipeline
- One-hot encoded string variables
- Further assembled and scaled one-hot encoded vectors and original numerical features
- Fitted the pipeline to the training data

Modeling
- Trained 4 models – Logistic regression, Decision tree, Random forest and Gradient boosting classifier on the preprocessed pipeline

Model Selection
- Compared key evaluation metrics for all four models
- Gradient boosting classifier selected as the champion model since it has the best AUC score and sensitivity

DETAILED PIPELINE

Converted string categorical features into numerical indices using StringIndexer

Converted the numerical indices into one-hot encoded vectors using OneHotEncoder

Converted the outcome label column 'y' into numerical labels using StringIndexer

Assembled all the one-hot encoded vectors from the string categorical features and the numeric features into a single vector column

Scaled down the assembled features into a new column named 'features' using StandardScaler

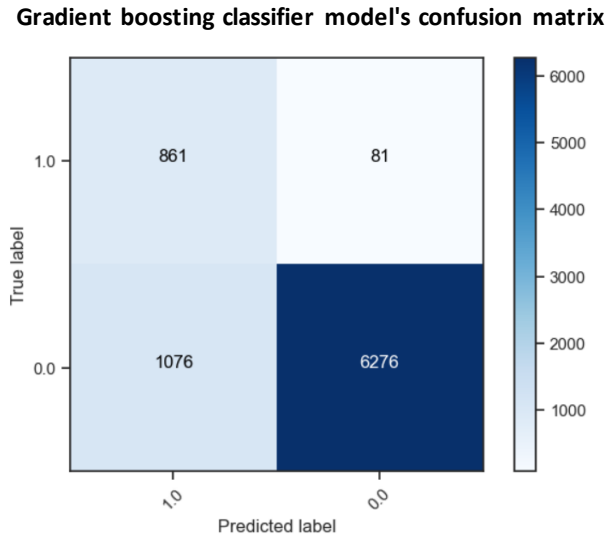
MODEL COMPARISON AND SELECTION

Model	Accuracy	Sensitivity	Specificity	Precision	AUC
Logistic Regression	0.872	0.872	0.872	0.465	0.939
Decision Tree	0.857	0.892	0.852	0.436	0.872
Random Forest	0.851	0.776	0.861	0.417	0.922
Gradient boosting classifier	0.861	0.914	0.854	0.445	0.948

- Four models were trained - Logistic regression, Decision tree, Random forest and Gradient boosting classifier
- Gradient boosting classifier performed the best on the test data in terms of sensitivity and area under the ROC curve
- Sensitivity measures the proportion of actual positives correctly predicted as positive by the model. Given our objective of correctly identifying customers likely to open a term deposit account, sensitivity is an important measure in our model selection process

Gradient boosting classifier selected as the champion model

RECOMMENDATIONS AND INSIGHTS



Adjust calling campaigns to target customers identified by Gradient boosting classifier model

- **Reduces number of customers called by 76.6% from 8294 to 1937 (861 + 1076)**
- **Success rate 44.5% versus 11.3%**

- Customers who are contacted less tend to subscribe more for term deposit accounts with the bank. Over-contacting customers might lead to irritation or a negative perception of the bank. The bank should approach customers more selectively and with personalized messages
- Admin workers and retired people are more likely to subscribe for term deposit accounts than customers of other professions. Accordingly, bank can develop marketing campaigns highlighting stability and security to specifically target them
- Also, single customers are likely to subscribe than married customers, accordingly, bank can develop targeted strategies towards singles

APPENDIX

DATA DICTIONARY

Attribute	Data Type	Description
Age	Numeric	Age of customer
Job	Categorical	Type of job
Marital	Categorical	Marital status
Education	Categorical	Last level of education completed
Default	Categorical	Indicates if customer has credit in default
Housing	Categorical	Indicates if customer has a housing loan
Loan	Categorical	Indicates if customer has a personal loan
Contact	Categorical	Contact communication type: cellular/telephone
Month	Categorical	Last contact month of year
Day_of_week	Categorical	Last contact day of the week
Duration	Numeric	Last contact duration
Campaign	Numeric	Number of contacts performed during this campaign for this customer
Pdays	Numeric	Number of days since last contact (999, means no previous contact)
Previous	Numeric	Number of contacts performed before this campaign
Poutcome	Categorical	Outcome of the previous marketing campaign
Emp.var.rate	Numeric	Employment variation rate - quarterly indicator
Cons.price.index	Numeric	Consumer price index - monthly indicator
Cons.conf.idx	Numeric	Consumer confidence index - monthly indicator
Euribor3m	Numeric	euribor 3-month rate - daily indicator
Nr.employed	Numeric	Number of employees - quarterly indicator

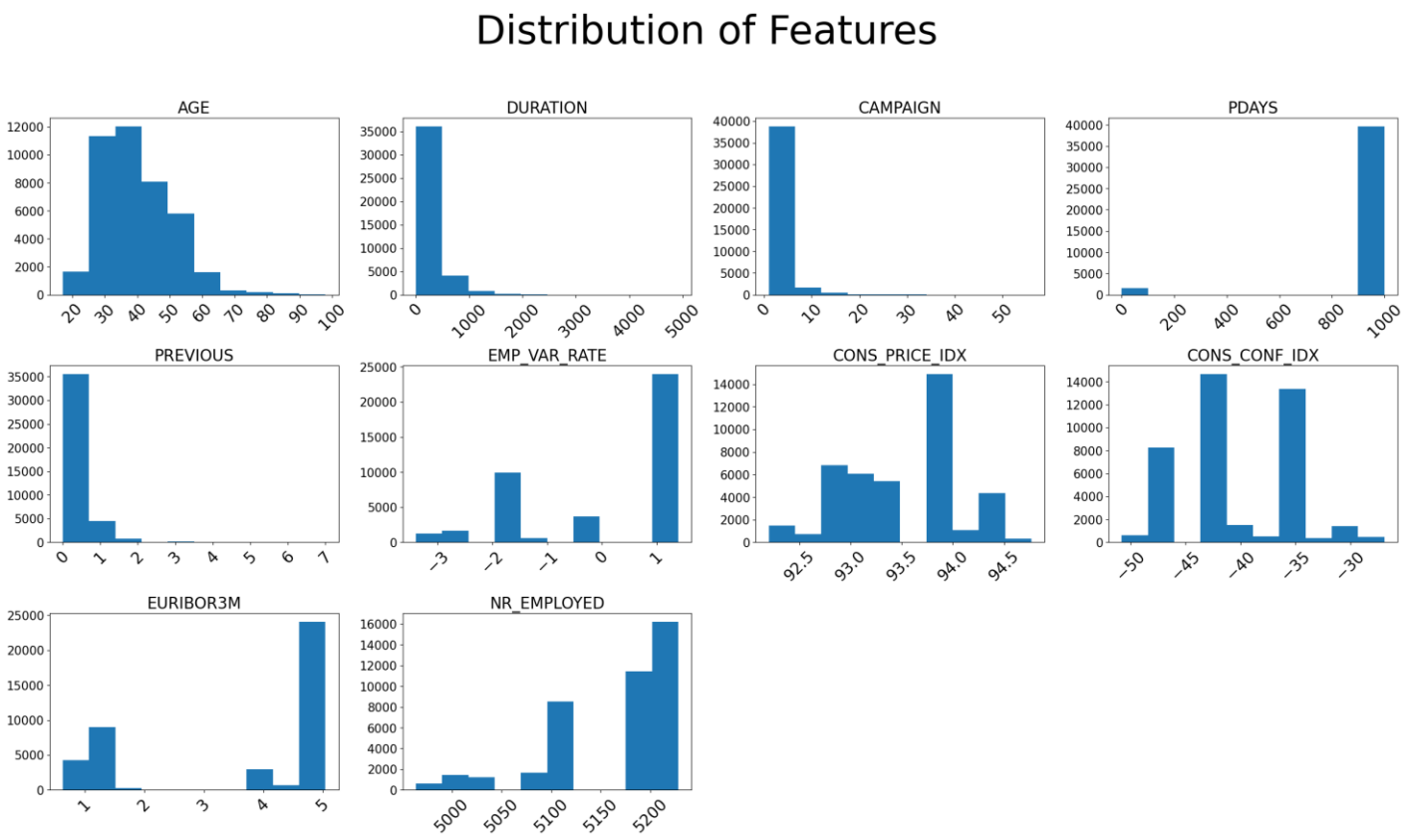
- The data set contains 41,188 observations relating to a prior direct telephone call marketing campaign
- Calling campaign occurred between May 2008 and November 2010
- Includes 20 attributes relating to the customers including demographic, latest contact, socio-economic and other attributes
- There was no missing data

EXPLORATORY DATA ANALYSIS – ATTRIBUTES SUMMARY

	0	1	2	3	4	5	6	7
summary	count	mean	stddev	min	25%	50%	75%	max
age	41188	40.02406040594348	10.421249980934043	17	32	38	47	98
job	41188	None	None	admin.	None	None	None	unknown
marital	41188	None	None	divorced	None	None	None	unknown
education	41188	None	None	basic.4y	None	None	None	unknown
default	41188	None	None	no	None	None	None	yes
housing	41188	None	None	no	None	None	None	yes
loan	41188	None	None	no	None	None	None	yes
contact	41188	None	None	cellular	None	None	None	telephone
month	41188	None	None	apr	None	None	None	sep
day_of_week	41188	None	None	fri	None	None	None	wed
duration	41188	258.2850101971448	259.27924883646455	0	102	180	319	4918
campaign	41188	2.567592502670681	2.770013542902331	1	1	2	3	56
pdays	41188	962.4754540157328	186.910907344741	0	999	999	999	999
previous	41188	0.17296299893172767	0.49490107983928927	0	0	0	0	7
poutcome	41188	None	None	failure	None	None	None	success
emp_var_rate	41188	0.08188550063178966	1.57095974051703	-3.4	-1.8	1.1	1.4	1.4
cons_price_idx	41188	93.5756643682899	0.5788400489540823	92.201	93.075	93.749	93.994	94.767
cons_conf_idx	41188	-40.502600271918276	4.628197856174573	-50.8	-42.7	-41.8	-36.4	-26.9
euribor3m	41188	3.621290812858533	1.7344474048512595	0.634	1.344	4.857	4.961	5.045
nr_employed	41188	5167.035910943957	72.25152766826338	4963.6	5099.1	5191.0	5228.1	5228.1
y	41188	None	None	no	None	None	None	yes

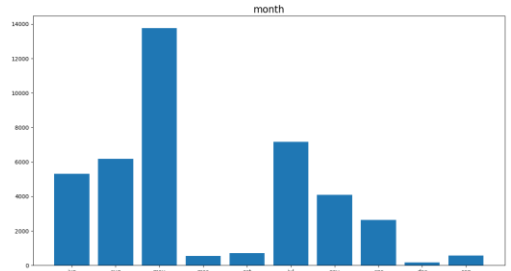
- Age IQR is 32 – 47, mean of 40
- Duration IQR is 102 – 319, mean of approx. 258, max = 4918
- Campaign IQR of 1 – 3, max = 56
- Majority customers have not been contacted previously

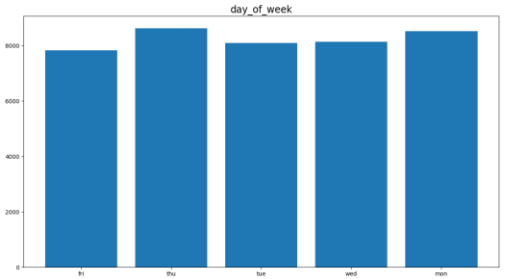
EXPLORATORY DATA ANALYSIS – DISTRIBUTION OF NUMERIC VARIABLES

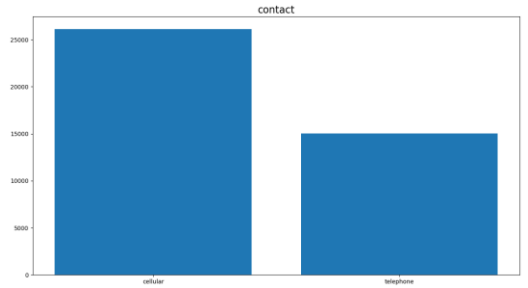


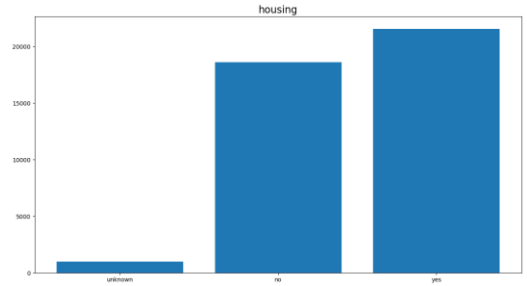
- None of the attribute is normally distributed
- Age is right skewed with thin tails
- Duration, Campaign, and Previous are leptokurtic with right tails
- Pdays shows most density at 999, which implies that customers have never been contacted before

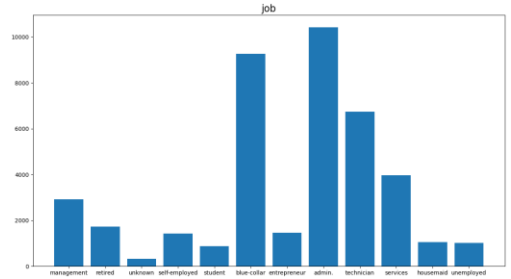
EXPLORATORY DATA ANALYSIS – DISTRIBUTION OF CATEGORICAL VARIABLES

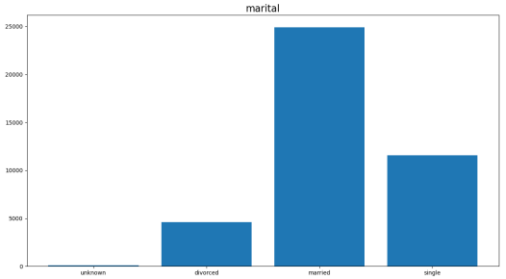


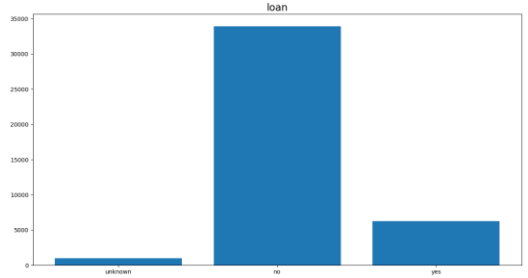


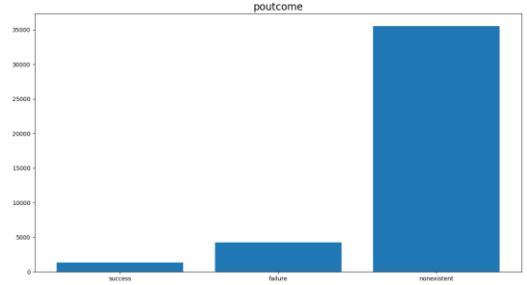


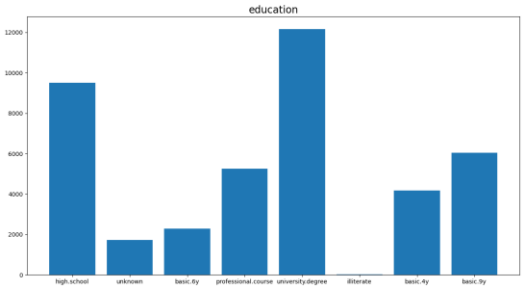


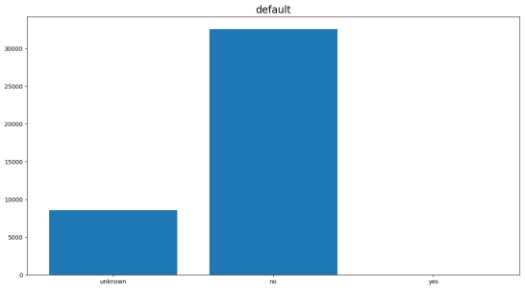


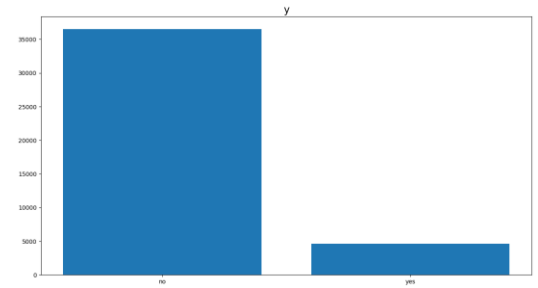






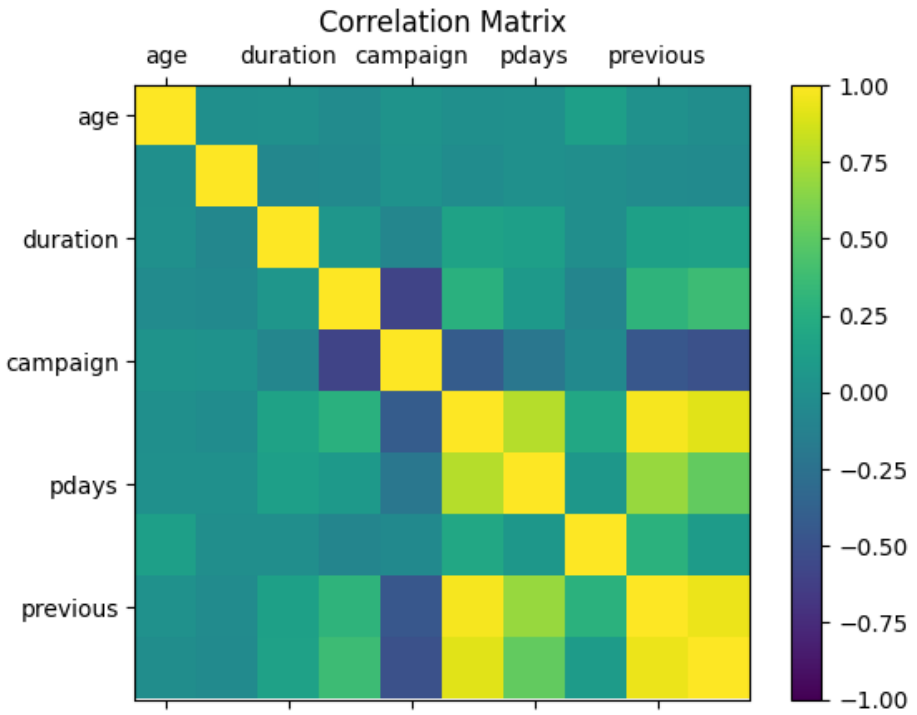






- Imbalanced target variable, 4640 = y, 36548 = n
- Large spike of last contact month in May, low in December
- Day of week relatively evenly distributed
- Majority contacted via cell phone, were married, have housing, and no defaults
- Most contacted job is admin followed by blue collar

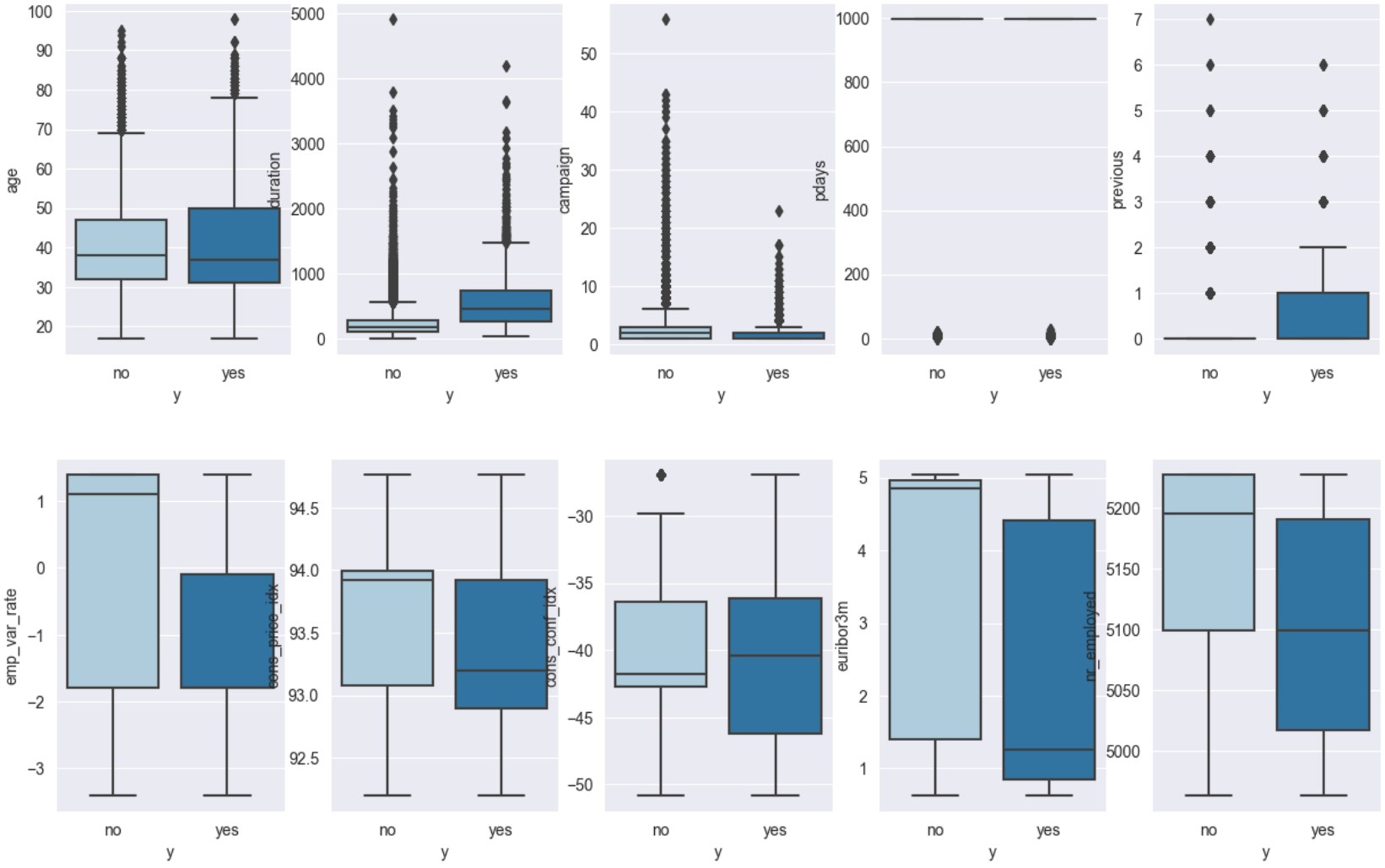
EXPLORATORY DATA ANALYSIS – CORRELATION MATRIX



- Most correlations are weak
- Some Strong Positive Correlations
  - Emp_var_rate and cons_price_idx = 0.77
  - Emp_var_rate and euribor3m = 0.97
  - Emp_var_rate and nr_employed = 0.91
  - Euribor3m and nr_employed = 0.95

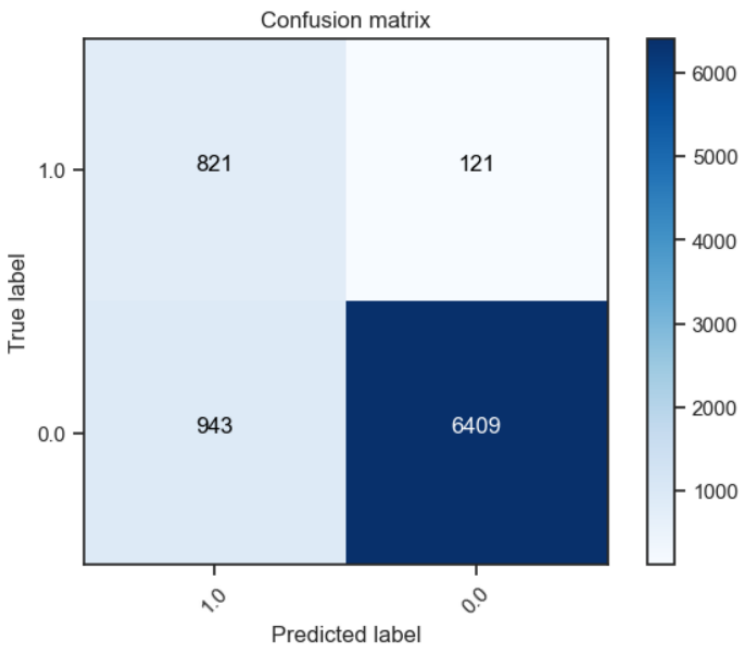
age	duration	campaign	pdays	previous	emp_var_rate	cons_price_idx	cons_conf_idx	euribor3m	nr_employed
1.0	-8.65705010140913...	0.004593580493413432	-0.03436895116685...	0.02436474093611654	-3.70685467441012...	8.567149710785426E-4	0.12937161424620508	0.010767429541674797	-0.01772513191192...
-8.65705010140913...	1.0	-0.0716992262641536	-0.0475770154456121	0.020640350701749122	-0.02796788448933175	0.005312267762748574	-0.00817287281392...	-0.03289665570187576	-0.04470322316241789
0.004593580493413432	-0.0716992262641536	1.0	0.052583573385026956	-0.07914147244884145	0.15075380555786647	0.1278359116094557 3	-0.01373309874190...	0.13513251080435904	0.14409489484472365
-0.03436895116685...	-0.0475770154456121	0.052583573385026956	1.0	-0.587513856136789	0.27100417426183293	0.0788891087159522	-0.09134235397835197	0.29689911239700334	0.37260474218583123
0.02436474093611654	0.020640350701749122	-0.07914147244884145	-0.587513856136789	1.0	-0.42048910941333256	-0.2031299674503254	-0.05093635090673017	-0.45449365360773475	-0.5013329290362544
-3.70685467441012...	-0.02796788448933175	0.15075380555786647	0.27100417426183293	-0.42048910941333256	1.0	0.7753341708352004	0.19604126813197817	0.9722446711516908	0.9069701012559852
8.567149710785426E-4	0.005312267762748574	0.12783591160945573	0.0788891087159522	-0.2031299674503254	0.7753341708352004	1.0	0.05898618174887866	0.6882301070378115	0.5220339770133208
0.12937161424620508	-0.00817287281392...	-0.01373309874190...	-0.09134235397835197	-0.05093635090673017	0.19604126813197817	0.05898618174887866	1.0	0.27768621966372714	0.10051343183754938
0.010767429541674797	-0.03289665570187576	0.13513251080435904	0.29689911239700334	-0.45449365360773475	0.9722446711516908	0.6882301070378115	0.27768621966372714	1.0	0.9451544313981852
-0.01772513191192...	-0.04470322316241789	0.14409489484472365	0.37260474218583123	-0.5013329290362544	0.9069701012559852	0.5220339770133208	0.10051343183754938	0.9451544313981852	1.0

EXPLORATORY DATA ANALYSIS – BOXPLOT, NUMERIC VARIABLES

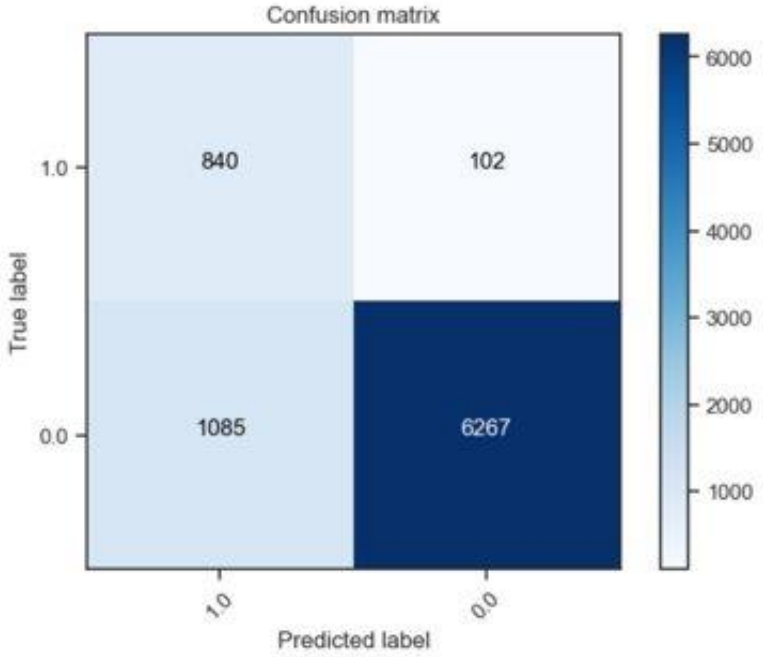


- The boxplot provides insights into the variable distribution across the target variable categories which shows the presence of several outliers
- "Age" and "Duration" have large number of outliers above the upper whisker
- "Campaign" has fewer outliers compared to the above
- "Previous" plot has large number of outliers present in both categories.

MODELING – LOGISTIC REGRESSION AND DECISION TREE



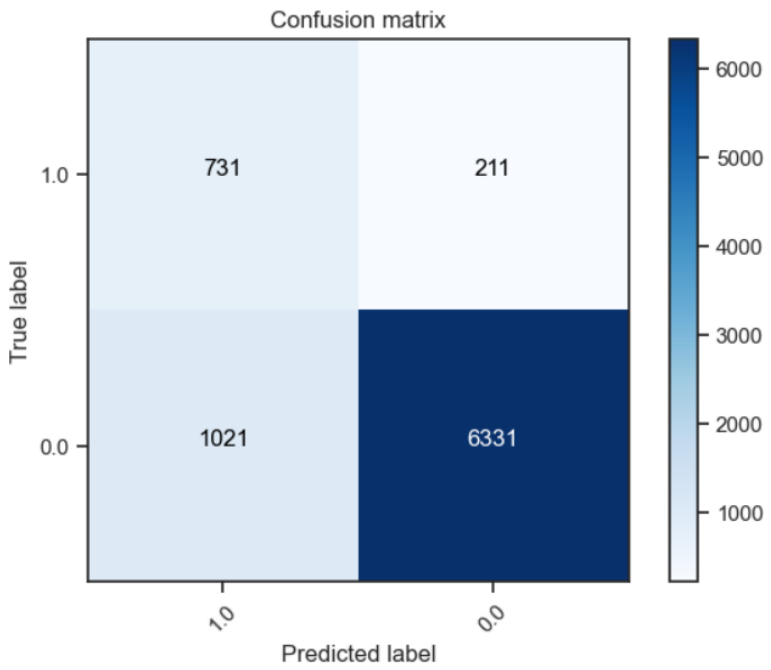
Logistic Regression Evaluation Metrics
Accuracy = 0.872
Sensitivity = 0.872
Specificity = 0.872
Precision = 0.465
Test Area Under ROC: 0.939

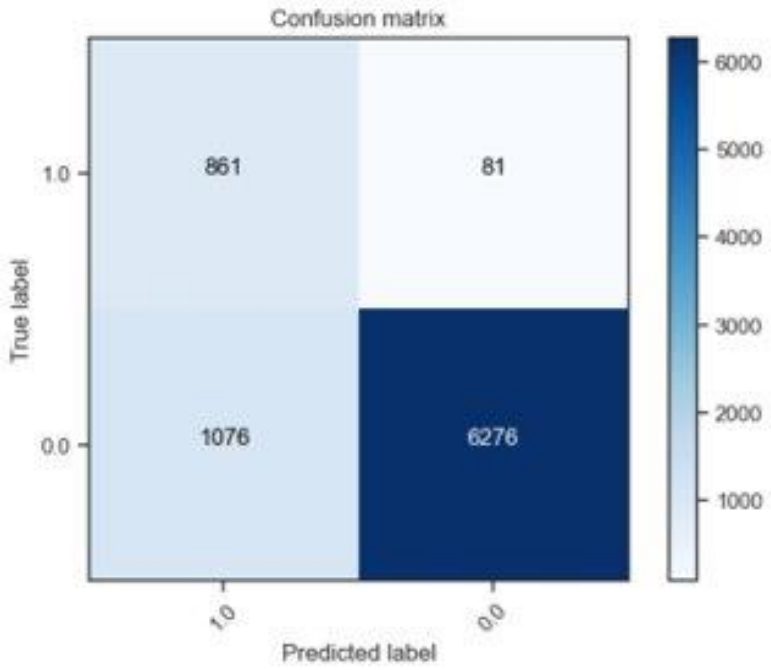


Decision Tree Evaluation Metrics
Accuracy = 0.857
Sensitivity = 0.892
Specificity = 0.852
Precision = 0.436
Test Area Under ROC: 0.872

- Both models performed well with .872 and .857 accuracy as well as .939 and .872 AUC.
- Decision Tree had a slightly better sensitivity with predicting the target outcome at .892.

MODELING – RANDOM FOREST AND GRADIENT BOOSTED TREES



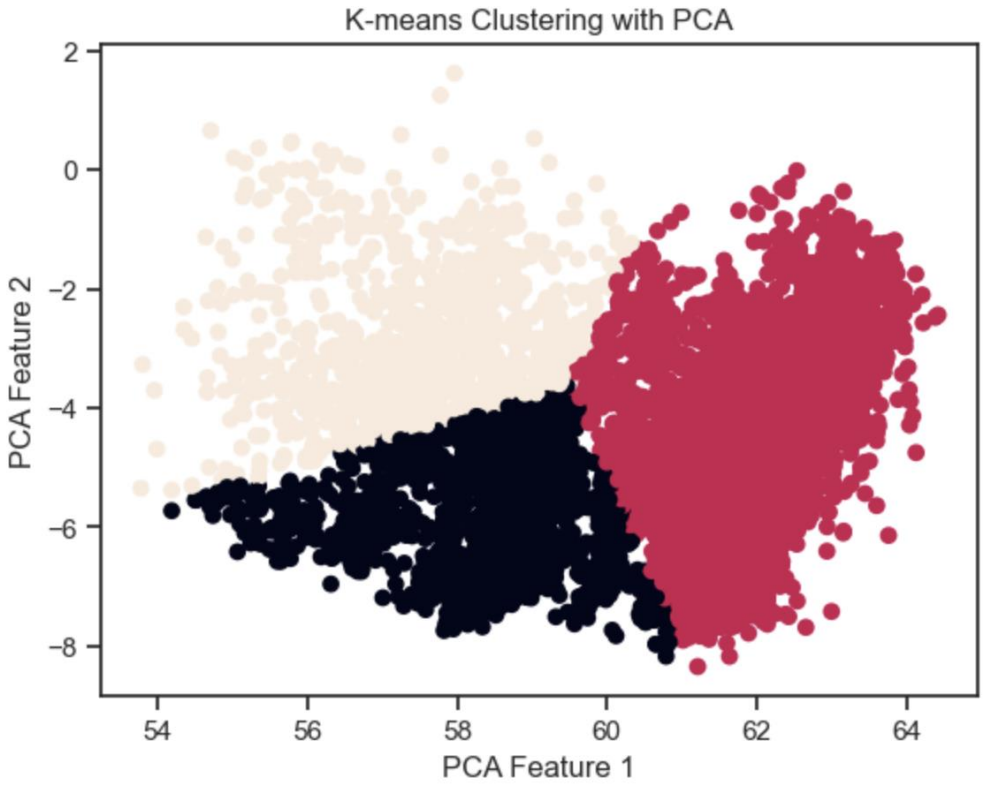



```
Random Forset Evaluation Metrics
Accuracy = 0.851
Sensitivity = 0.776
Specificity = 0.861
Precision = 0.417
Test Area Under ROC: 0.922
```

```
Gradient Boosting Trees Evaluation Metrics
Accuracy = 0.861
Sensitivity = 0.914
Specificity = 0.854
Precision = 0.445
Test Area Under ROC: 0.948
```

- Both models performed well with .851 and .861 accuracy scores as well as .922 and .949 AUC.
- Gradient Boosted Trees outperformed all models in terms of sensitivity as .914.
- It more accurately predicts the customer who will open a term deposit account.

K-MEANS CLUSTERING



Silhouette score of 0.568

- With a Silhouette score of 0.568, the clustering is moderately good, indicating some separation between the clusters
- The group on the right is distinctly separate from the other two, suggesting it has unique characteristics that set it apart in the higher-dimensional space
- There is some overlap between the two clusters means that there are some similarities in the features that aren't captured by first two principal components
- There are a few outliers in the plot indicates that there is variability in the dataset
- Overall, the K-means model has clustered the data into meaningful groups