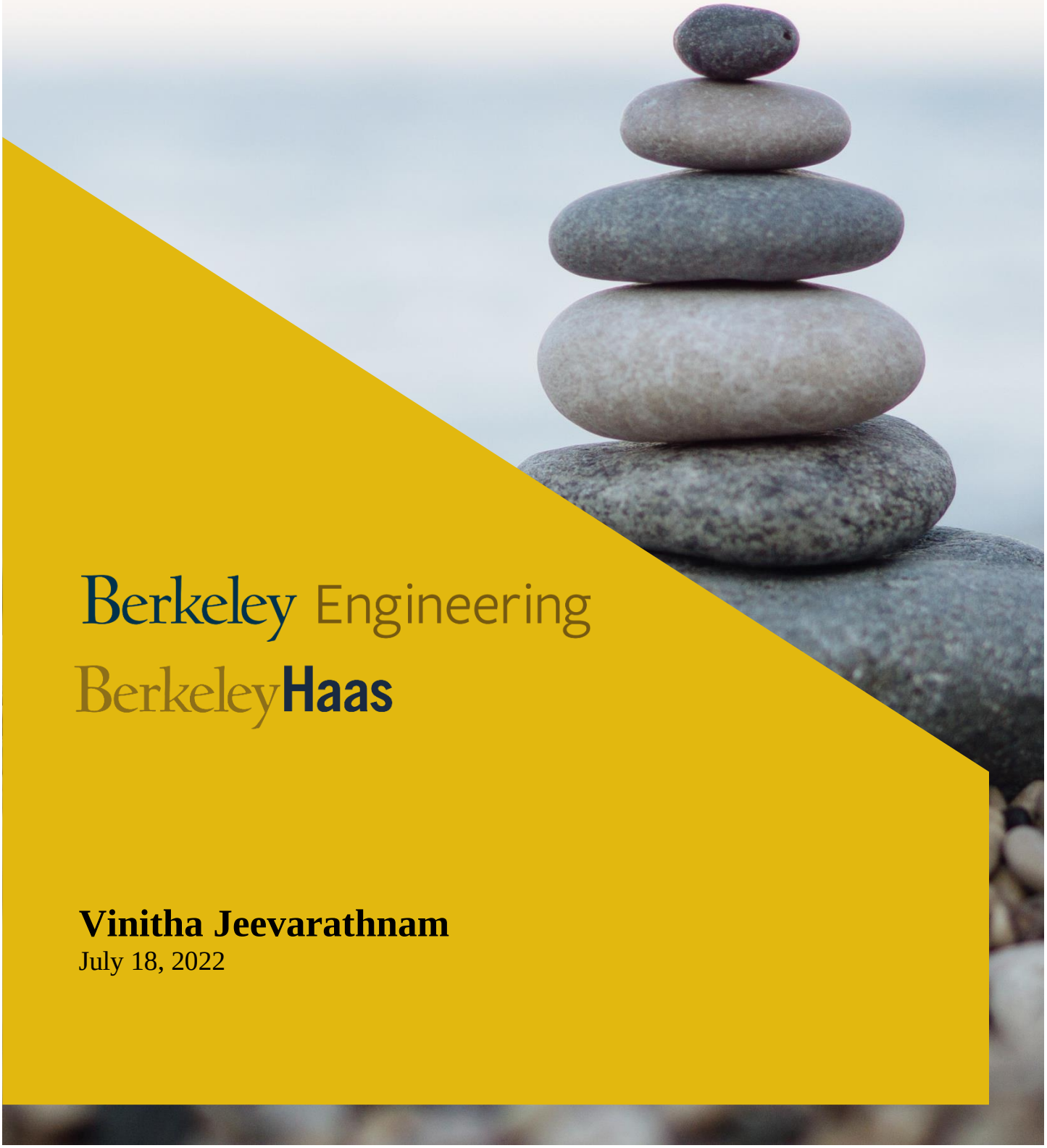
Module 17 – Practical Application 3



Berkeley Engineering

BerkeleyHaas

Vinitha Jeevarathnam

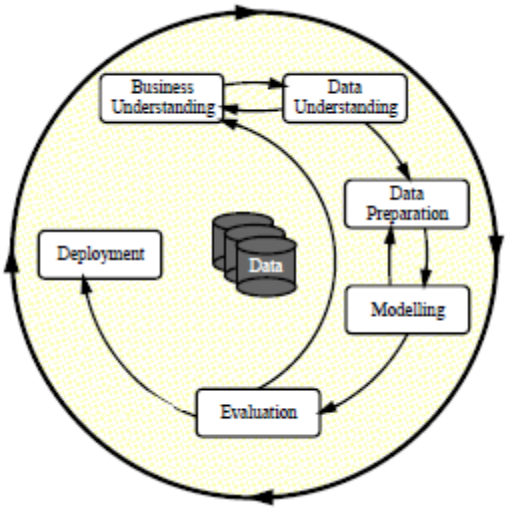
July 18, 2022

Table of Contents

1. **CRISP-DM Framework** 3
2. **Business Understanding** 4
3. **Data Understanding** 5
4. **Data Preparation** 7
5. **Modeling** 8
6. **Evaluation** 11
7. **Deployment** 18

1. CRISP-DM Framework

The CRoss-Industry Standard Process for Data Mining provides a model and framework for AI/ML data mining. For this practical application, I have adopted several steps from the CRISP-DM framework.



2. Business Understanding

2.1 Determine Business Objectives

The dataset provided for this practical application is from the UCI Machine Learning repository that contains information on approx. 41188 marketing campaign for term deposit banking product contact details (21 features). I need to use this dataset to accomplish the following objectives,

1. Identify the performance/accuracy score of Classifier Models
2. Identify the factors that affect the customer acceptance of banking products via telephone marketing
3. Summarize and provide future recommendations (if any)

2.2 Assess Situation

- **Assumptions**
  - The data provided for this application is of high quality and minimal effort is needed to cleanup
  - Availability of system resources to perform data processing of 41188 data records and perform GridSearchCV
- **Constraints**
  - None

2.3 Determine Data Mining Goals

- Maintain as much as data after data cleanup activities
- Leverage Target Encoder and StandardScaler methods to convert categorical columns to numerical columns and standardize values for better handling of data in the models.

2.4 Produce Project Plan

A project plan is not needed for this application.

3. Data Understanding

3.1 Collect Initial Data

- Import all the necessary packages needed to perform functions related to pandas, modeling (sklearn), and plotting
- Import the data using read_csv
- The dataset contains 41188 records and 21 columns

3.2 Describe and Explore Data

Input variables:

 # bank client data:

- age (numeric)
- job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
- marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
- education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
- default: has credit in default? (categorical: 'no','yes','unknown')
- housing: has housing loan? (categorical: 'no','yes','unknown')
- loan: has personal loan? (categorical: 'no','yes','unknown')

 # related with the last contact of the current campaign:

- contact: contact communication type (categorical: 'cellular','telephone')
- month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
- day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
- duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

 # other attributes:

- campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
- 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)
- previous: number of contacts performed before this campaign and for this client (numeric)
- poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

 # social and economic context attributes

- emp.var.rate: employment variation rate - quarterly indicator (numeric)
- cons.price.idx: consumer price index - monthly indicator (numeric)
- cons.conf.idx: consumer confidence index - monthly indicator (numeric)
- euribor3m: euribor 3 month rate - daily indicator (numeric)
- nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):

- y - has the client subscribed a term deposit? (binary: 'yes','no')

3.3 Verify Data Quality

- None of the columns were reporting any missing values (NaN)

```
age               0
job               0
marital           0
education         0
default           0
housing           0
loan              0
contact           0
month             0
day_of_week       0
duration          0
campaign          0
pdays             0
previous          0
poutcome          0
emp.var.rate      0
cons.price.idx    0
cons.conf.idx     0
euribor3m         0
nr.employed       0
y                 0
dtype: int64
```

- The numerical features – Age, Duration, and Campaign have outliers

	age	duration	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
count	41188.00000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000
mean	40.02406	258.285010	2.567593	962.475454	0.172963	0.081886	93.575664	-40.502600	3.621291	5167.035911
std	10.42125	259.279249	2.770014	186.910907	0.494901	1.570960	0.578840	4.628198	1.734447	72.251528
min	17.00000	0.000000	1.000000	0.000000	0.000000	-3.400000	92.201000	-50.800000	0.634000	4963.600000
25%	32.00000	102.000000	1.000000	999.000000	0.000000	-1.800000	93.075000	-42.700000	1.344000	5099.100000
50%	38.00000	180.000000	2.000000	999.000000	0.000000	1.100000	93.749000	-41.800000	4.857000	5191.000000
75%	47.00000	319.000000	3.000000	999.000000	0.000000	1.400000	93.994000	-36.400000	4.961000	5228.100000
max	98.00000	4918.000000	56.000000	999.000000	7.000000	1.400000	94.767000	-26.900000	5.045000	5228.100000

4. Data Preparation

4.1 Select and Clean Data

- Used **IQR** to remove outliers from **Age, Duration, and Campaign**
- Since **Duration** feature should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model. Hence, this feature is dropped for our model analysis.

4.2 Construct and Format Data

- Using numerical **Label Encoder** and **StandardScalar**, all the categorical columns and values are converted to numerical features.

	age	job	marital	education	default	housing	loan	contact	month	day_of_week	campaign	pdays	previous	poutcome	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed	y
0	56	3	1	0	0	0	0	1	6	1	1	999	0	1	1.1	93.994	-36.4	4.857	5191.0	0
1	57	7	1	3	1	0	0	1	6	1	1	999	0	1	1.1	93.994	-36.4	4.857	5191.0	0
2	37	7	1	3	0	2	0	1	6	1	1	999	0	1	1.1	93.994	-36.4	4.857	5191.0	0
3	40	0	1	1	0	0	0	1	6	1	1	999	0	1	1.1	93.994	-36.4	4.857	5191.0	0
4	56	7	1	3	0	0	2	1	6	1	1	999	0	1	1.1	93.994	-36.4	4.857	5191.0	0

	age	pdays	cons.price.idx	cons.conf.idx	euribor3m	nr.employed	campaign	previous	emp.var.rate
count	3.463400e+04	3.463400e+04	3.463400e+04	3.463400e+04	3.463400e+04	3.463400e+04	34634.000000	34634.000000	34634.000000
mean	-1.308905e-16	2.232114e-16	-6.453437e-15	5.514636e-16	-5.252034e-17	7.799270e-15	1.941156	0.176156	0.058298
std	1.000014e+00	1.000014e+00	1.000014e+00	1.000014e+00	1.000014e+00	1.000014e+00	1.128839	0.494951	1.569052
min	-2.332039e+00	-5.198117e+00	-2.362075e+00	-2.220096e+00	-1.713958e+00	-2.830377e+00	1.000000	0.000000	-3.400000
25%	-7.833798e-01	1.935809e-01	-8.478155e-01	-4.727208e-01	-1.304199e+00	-9.390798e-01	1.000000	0.000000	-1.800000
50%	-1.639162e-01	1.935809e-01	-2.085000e-01	-2.785681e-01	7.232415e-01	3.436527e-01	2.000000	0.000000	1.100000
75%	7.652792e-01	1.935809e-01	7.444093e-01	8.863486e-01	7.832625e-01	8.614913e-01	3.000000	0.000000	1.400000
max	3.036646e+00	1.935809e-01	2.083680e+00	2.935739e+00	8.317410e-01	8.614913e-01	5.000000	7.000000	1.400000

	age	pdays	cons.price.idx	cons.conf.idx	euribor3m	nr.employed	job	marital	education	default	housing	loan	contact	month	day_of_week	campaign	previous	poutcome	emp.var.rate	y
0	1.694475	0.193581	0.744409	0.886349	0.723241	0.343653	3	1	0	0	0	0	1	6	1	1	0	1	1.1	0
1	1.797719	0.193581	0.744409	0.886349	0.723241	0.343653	7	1	3	1	0	0	1	6	1	1	0	1	1.1	0
2	-0.267160	0.193581	0.744409	0.886349	0.723241	0.343653	7	1	3	0	2	0	1	6	1	1	0	1	1.1	0
3	0.042572	0.193581	0.744409	0.886349	0.723241	0.343653	0	1	1	0	0	0	1	6	1	1	0	1	1.1	0
4	1.694475	0.193581	0.744409	0.886349	0.723241	0.343653	7	1	3	0	0	2	1	6	1	1	0	1	1.1	0

4.3 Integrate Data

This step is not needed as we are not integrating with other datasources or datasets.

5. Modeling

5.1 Select Modeling

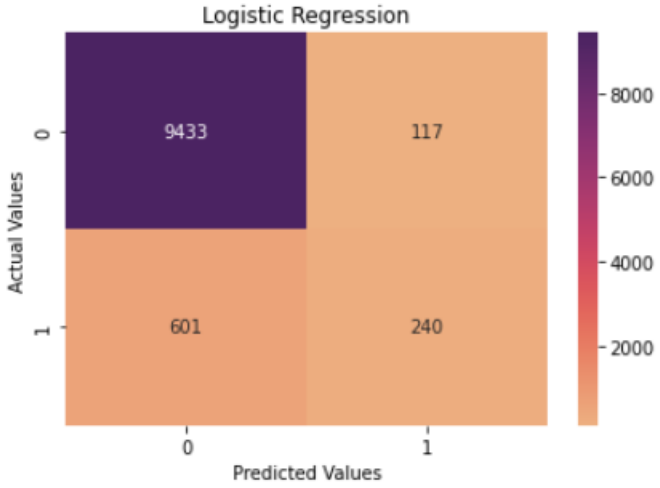
- The following Supervised Classification models are leveraged for this application
  - DummyClassifier() - for baseline prediction
  - LogisticRegression()
  - KNeighborClassifier()
  - DecisionTreeClassifier()
  - RandomForestClassifier()
  - SVC()
- Once the baseline accuracy scores are identified (models used with default or no hyperparameters), the performance of these models were improved by adding the best hyperparameters using GridSearchCV()

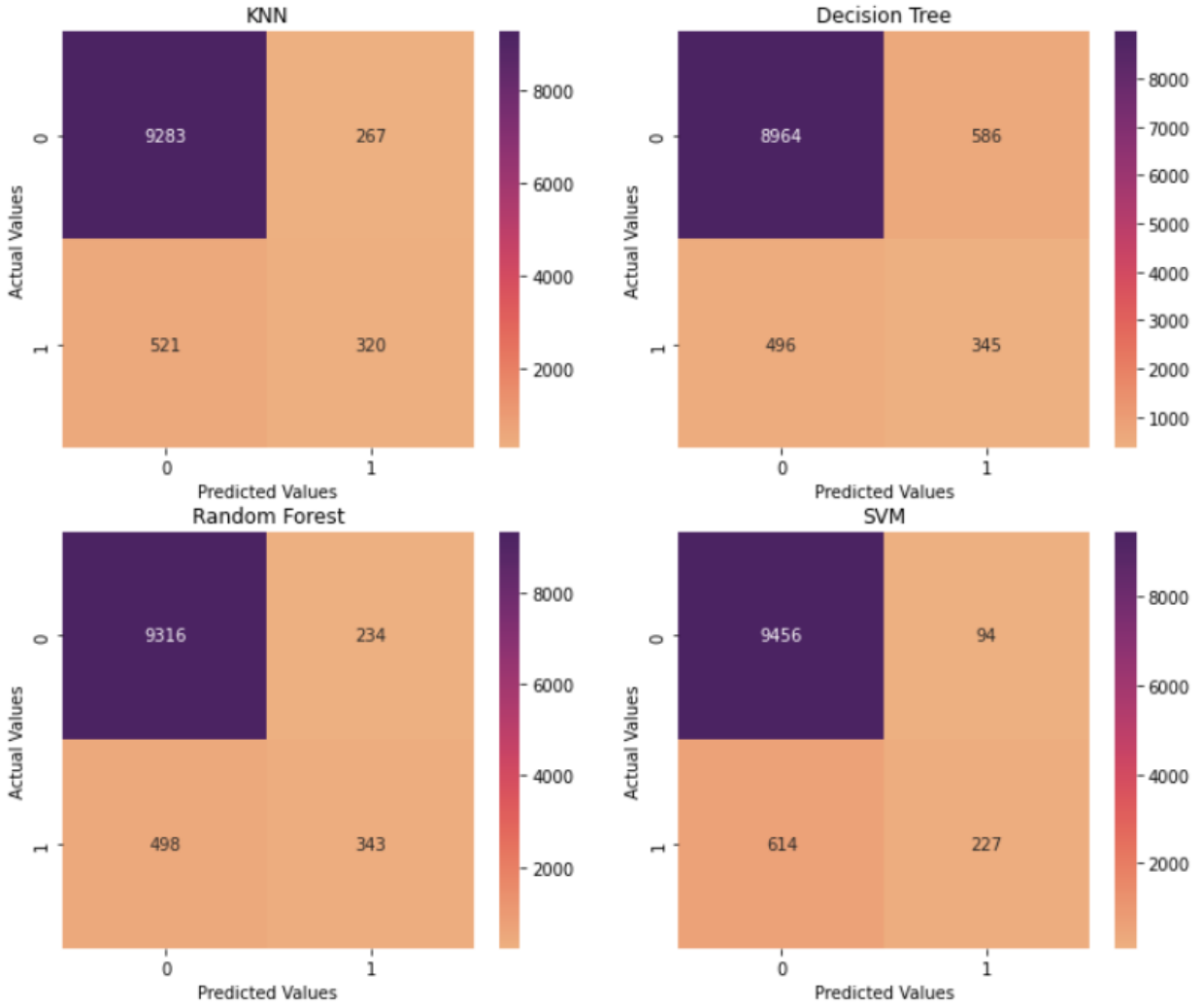
5.2 Generate Test Design

- Using **train_test_split** method, the data is split into train set (70%) and test set (30%)
- X_Train set shape – (**24243, 19)**
- X_Test set shape – (**10391, 19)**

5.3 Build and Assess Model

Initial analysis and Confusion Matrix of the selected supervised classification models,





A summary of the modeling results is saved in a dataframe,

	Model	Train_Time	Train_Score	Test_Score	Accuracy_Score	AUC
0	Baseline	0.004537	0.498041	0.500241	N/A	N/A
1	Logistic Regression	0.250259	0.929010	0.930902	0.930902	0.636562
2	KNN	0.038941	0.939075	0.924165	0.924165	0.676271
3	Decision Tree	0.141615	0.997195	0.895871	0.895871	0.674432
4	Random Forest	2.197746	0.997195	0.929554	0.929554	0.691673
5	SVM	14.651222	0.928845	0.931864	0.931864	0.630037

Based on the initial analysis, In the initial analysis, **KNeighborClassifier()** and **LogisticalRegression()** models outperformed the other models.

Improving the Models

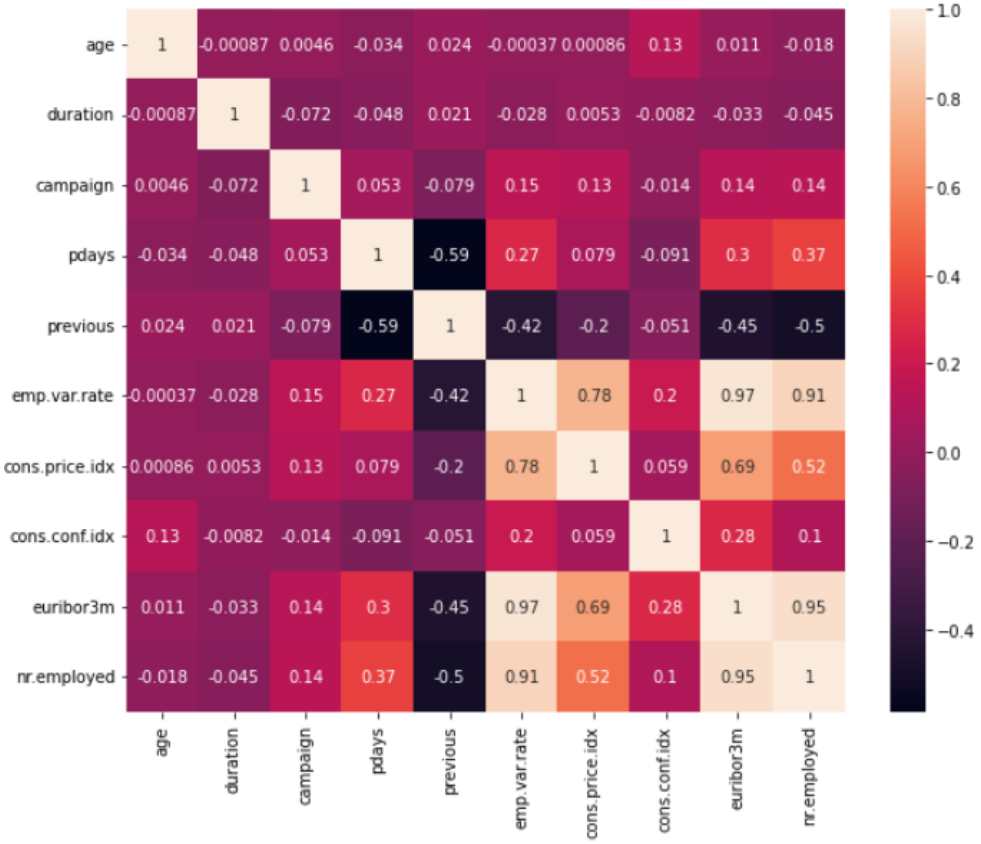
After the initial analysis, the GridSearchCV() is applied to identify the best hyperparamaters and improved the performance of these Classifier Models.

Among the improved models, **KNeighborClassifier()** and **DecisionTreeClassifier()** model outperformed the other improved classifier models.

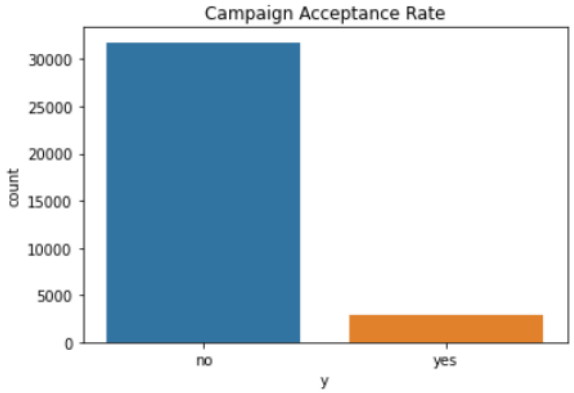
6. Evaluation

6.1 Evaluate Results

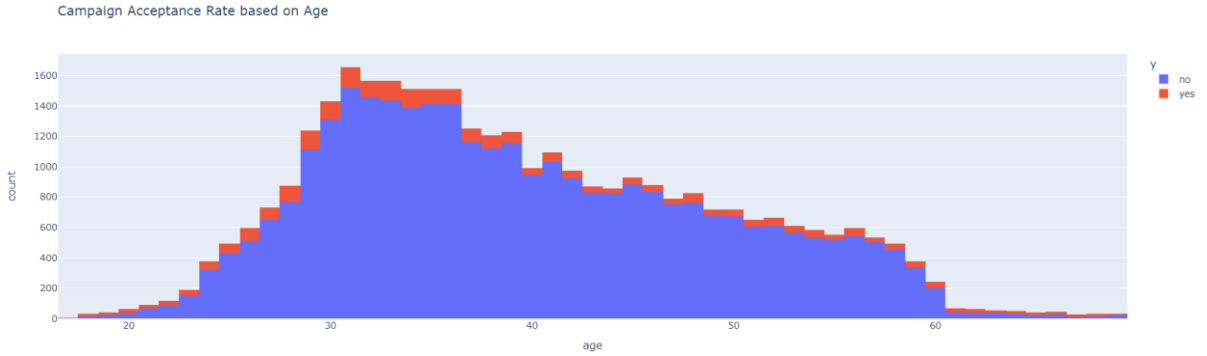
- Both in the baseline model analysis and improved best hyperparameter analysis, **KNeighborClassifier()** outperformed the other classification models for the selected data.
- Based on all the model analysis, the following features exhibited the highest feature importance and influences the acceptance of the banking product via marketing campaign over the telephone.
    - 1. **euribor3m**
    - 2. **nr.employed**
- Apart from the modeling, the following is inferred from plotting the dataframe using plotly, seaborn and matplotlib,
    - There is a high correlation between **Euribor rate** and **Employment Variation Rate** (0.97). Followed by **Euribor** and **Employee Count** (0.95) and then by **Employment Variation Rate** and **Employee Count** (0.91).



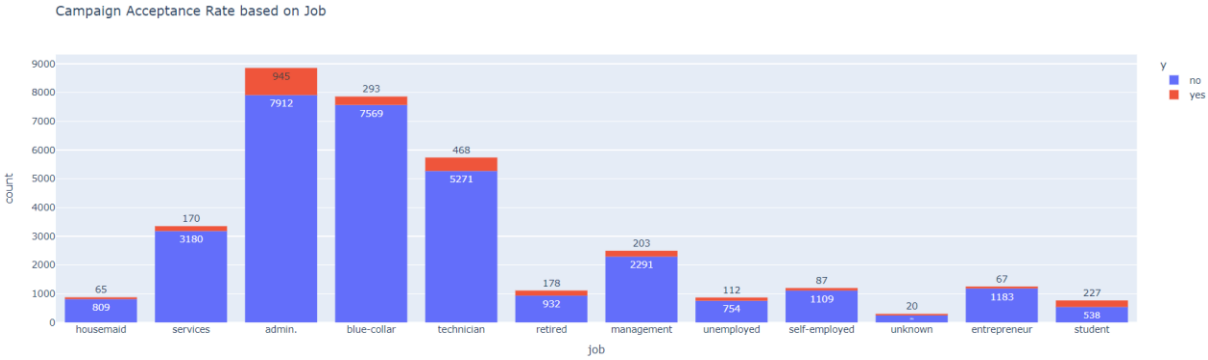
- Out of the 34634 campaign responses (reduced outliers) recorded in this dataset, only 8.19% subscribed to a term deposit.



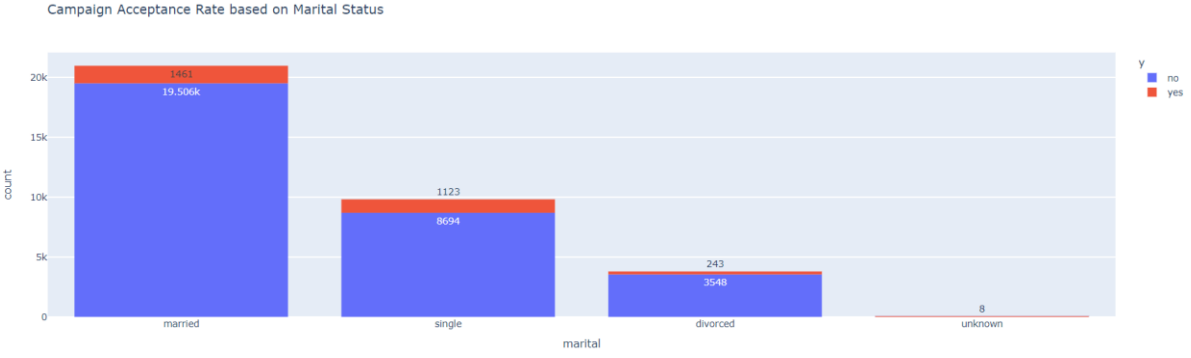
- A high Campaign was targeted to the **age group between 28 and 40**, and the acceptance rate is also higher in this group



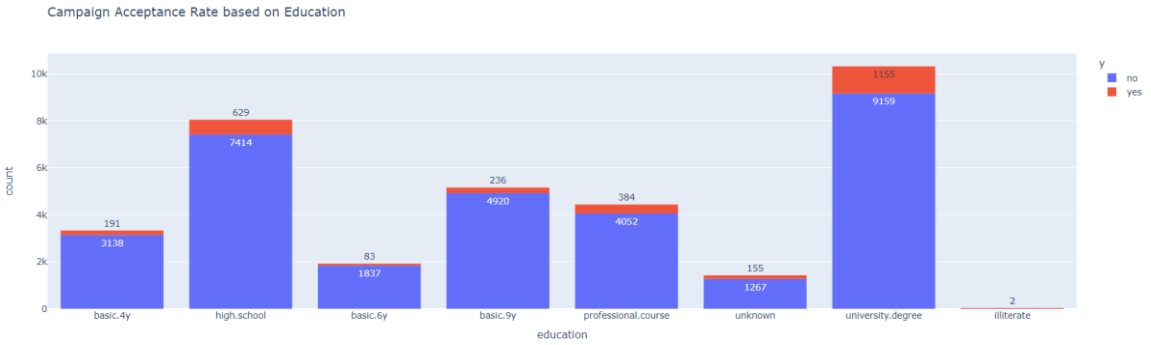
- A high campaign was targeted to people with Admin and Blue-Collar jobs. People with **Admin jobs** shows a higher acceptance to term-deposit.



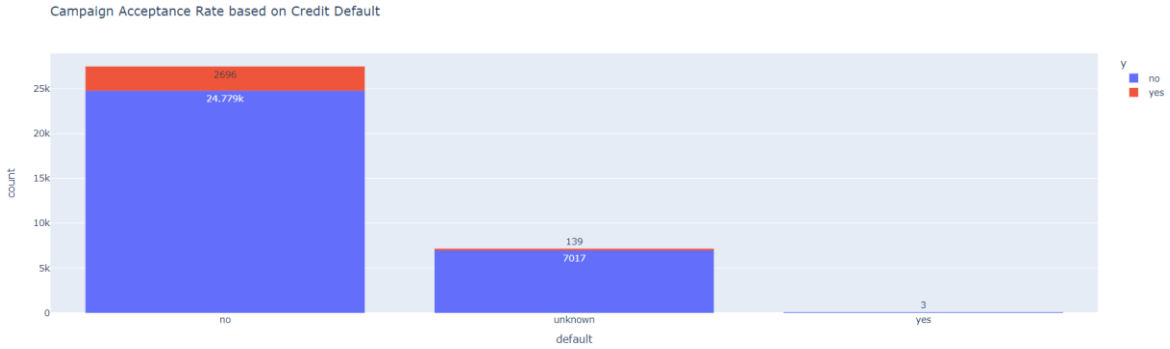
- A high campaign was targeted to people with '**Married**' **Marital Status** and also shows a higher acceptance of term-deposit.



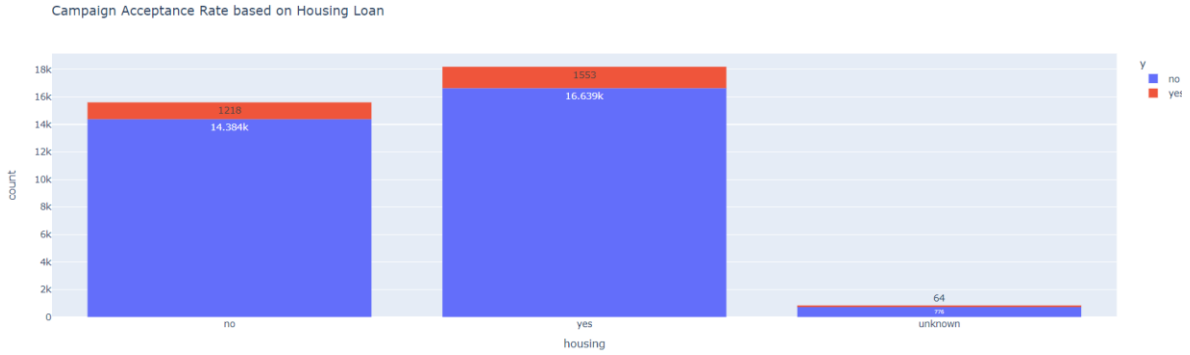
- A high campaign was targeted to people with University and High-School degrees. People with **University degree** shows a higher acceptance to term-deposit.



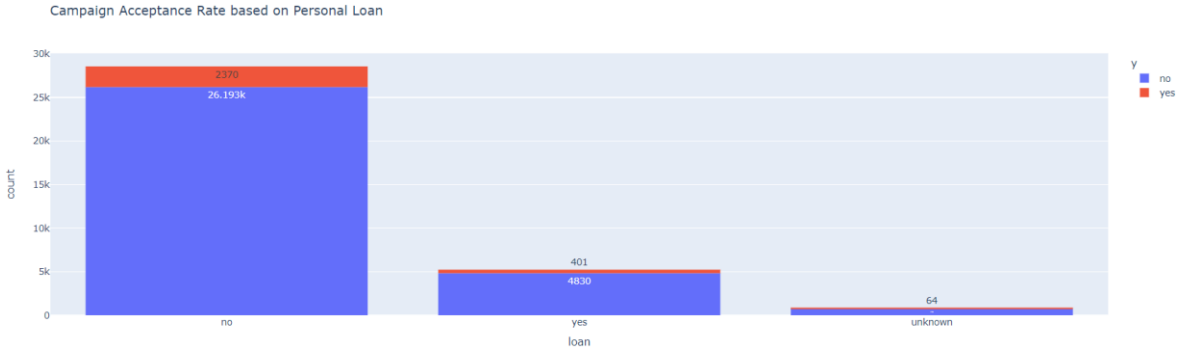
- A high campaign was targeted to people with **'No' credit default** status and also shows a higher acceptance to term-deposit.



- # A high campaign was targeted to people with an **existing Housing Loan** and the same shows a higher acceptance to term-deposit

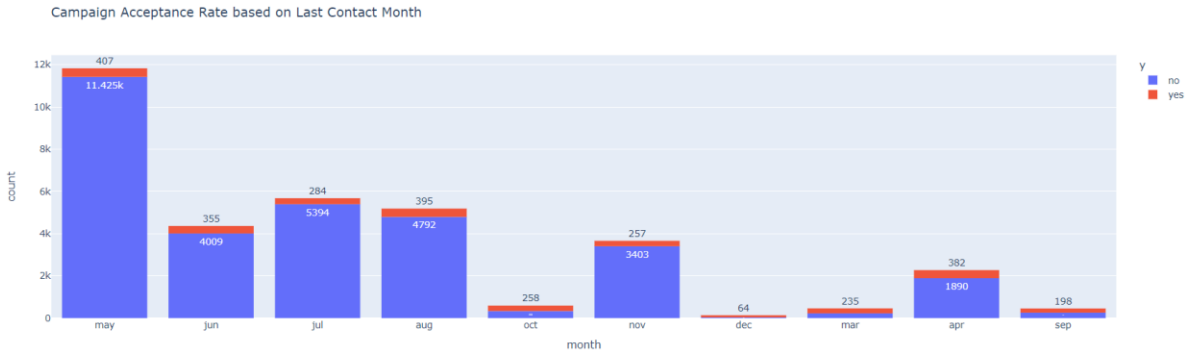


- A high campaign was targeted to people with **no existing Personal Loan** and the same shows a higher acceptance to term-deposit.

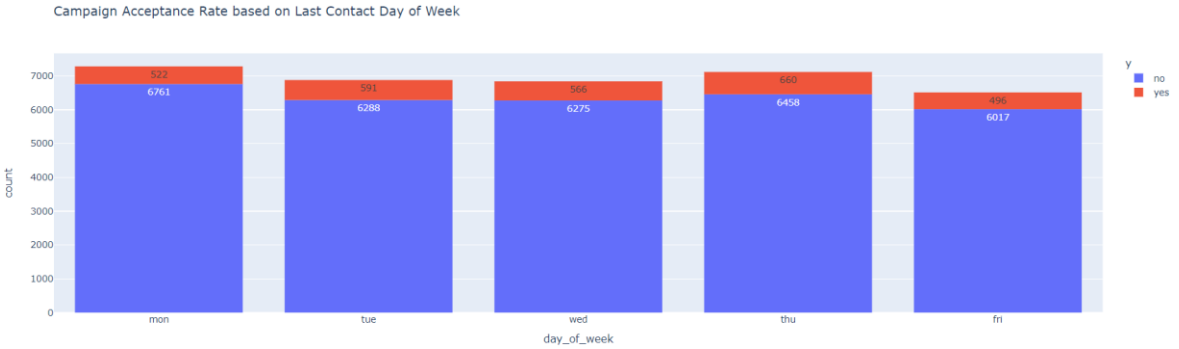


- A high **campaign was targeted to people via cellphones** and the same shows a higher acceptance to term-deposit.

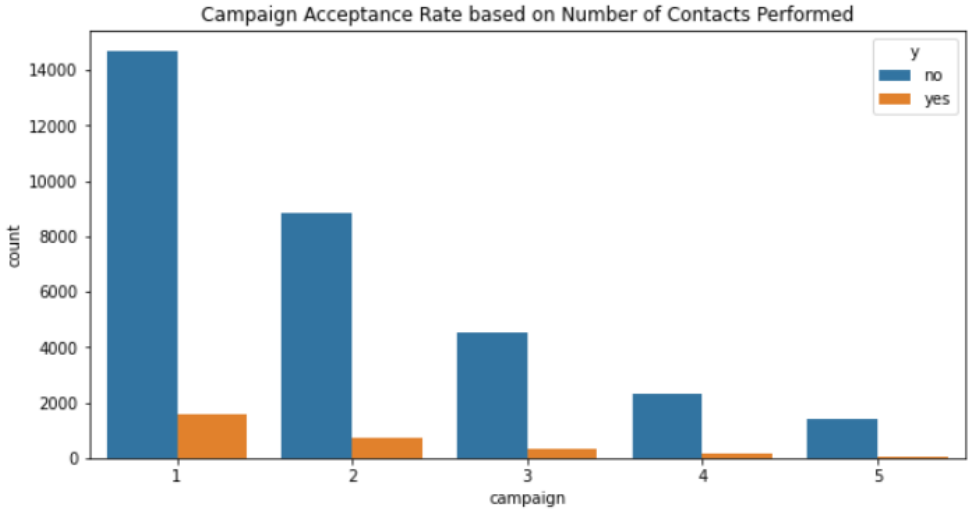
- A high campaign was targeted to people during **the month of 'May'** and the same shows a higher acceptance to term-deposit.



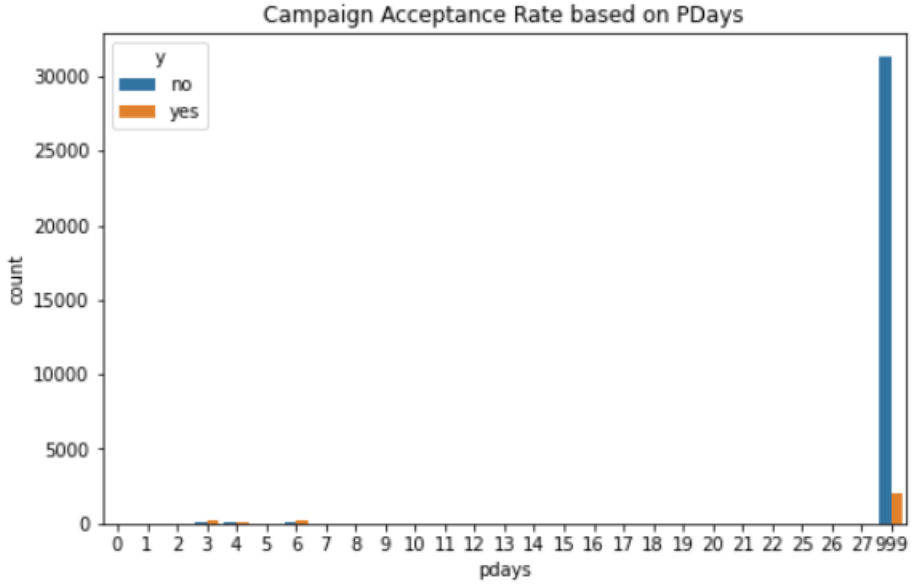
- A high campaign was targeted **on 'Mondays'**. But, the acceptance rate for term-deposit is higher on **'Thursdays'.**



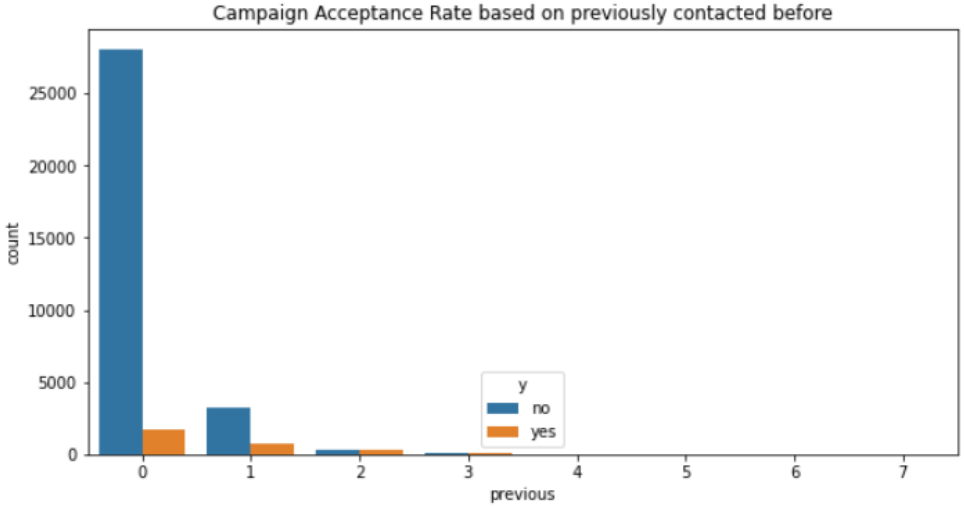
- High campaign contacts led to very minimum acceptance to term-deposit.



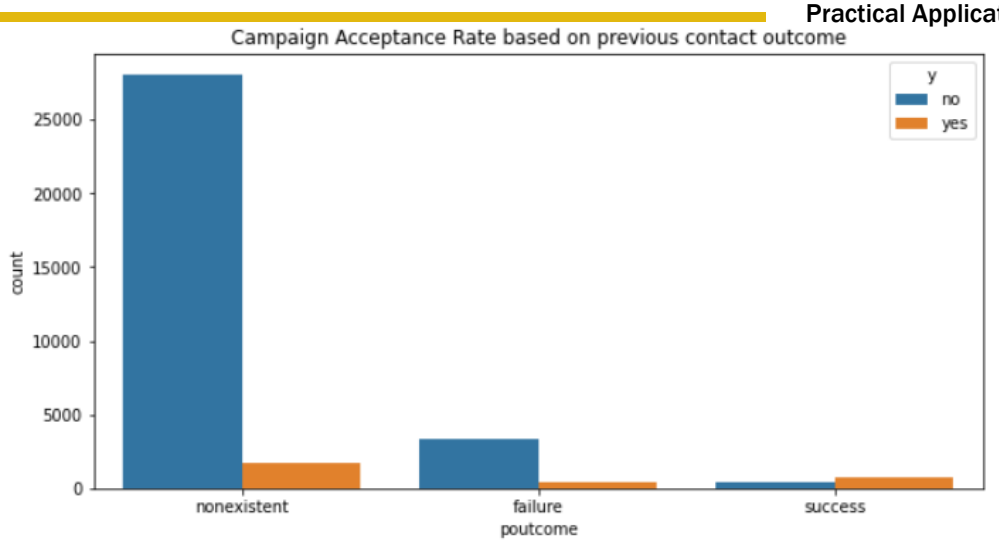
- Most of the campaign outreach happened to customers that were never been contacted before



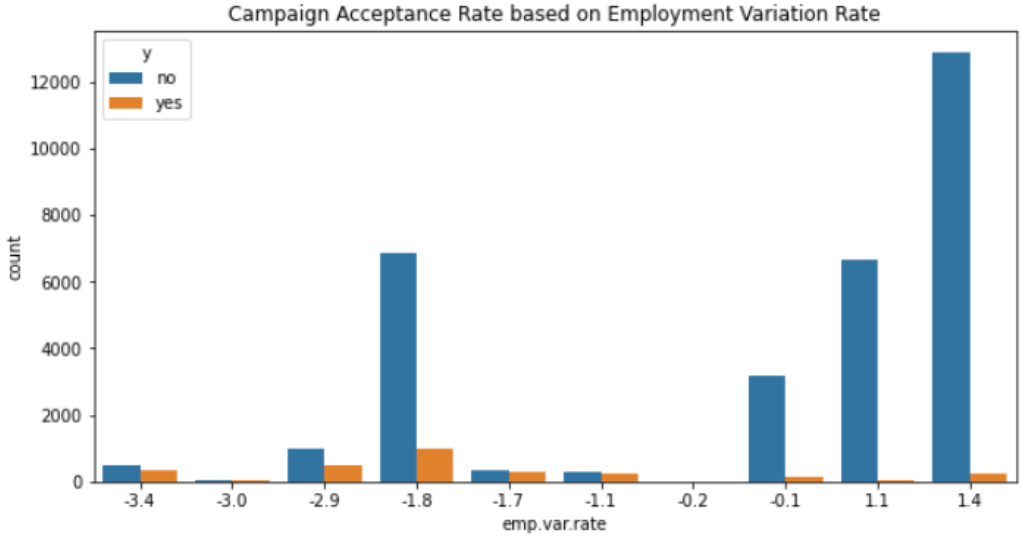
- Most of the campaign outreach happened to customers that were never been contacted before



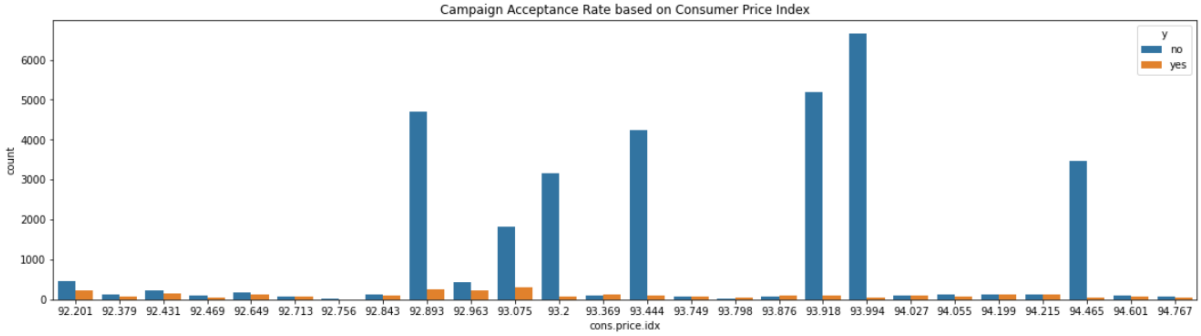
- Most of the campaign outreach happend to customers that were never contacted before and also shows a better acceptance rate to term-deposit.



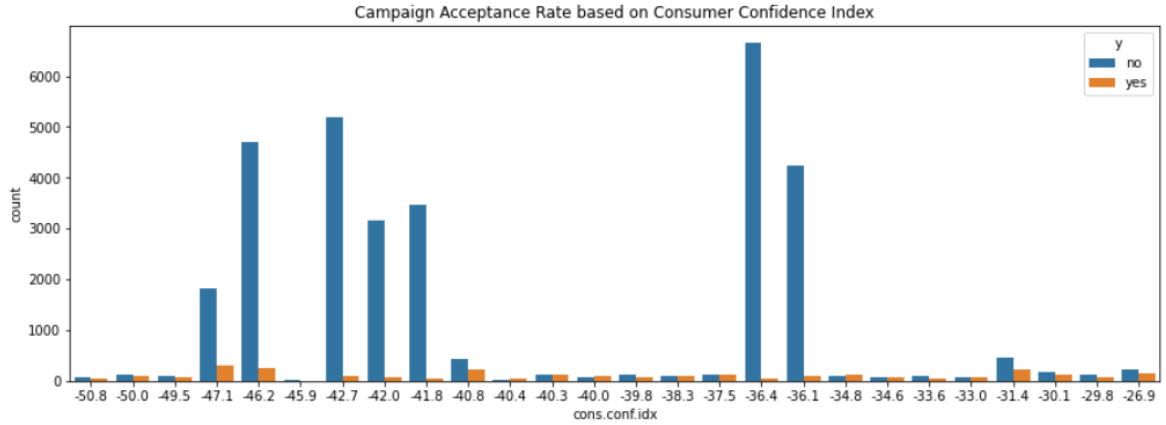
- Most of the campaign outreach happens when the EVR is high. But, the acceptance and non-acceptance of term-deposits are equal when the EVR is between -1.7 and -1.1.



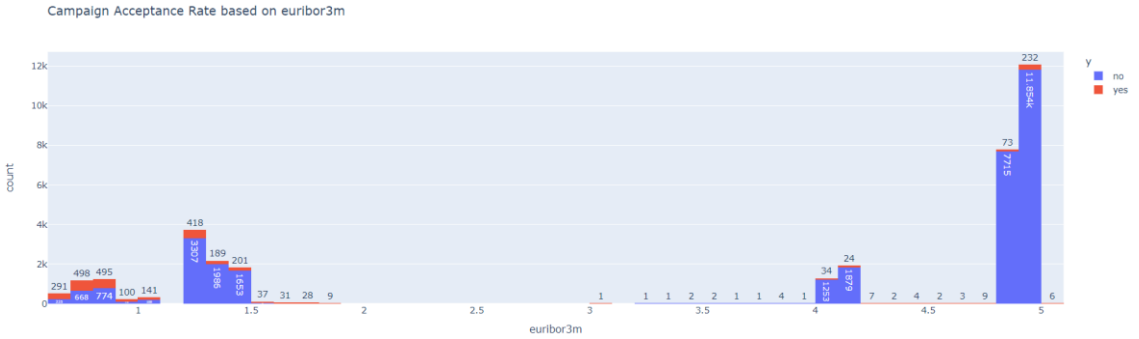
- ost of the campaign outreach happened when the CPI is over 92.89. But, the acceptance and non-acceptance of term-deposits become equal when the CPI is between 94.0 and 94.2.



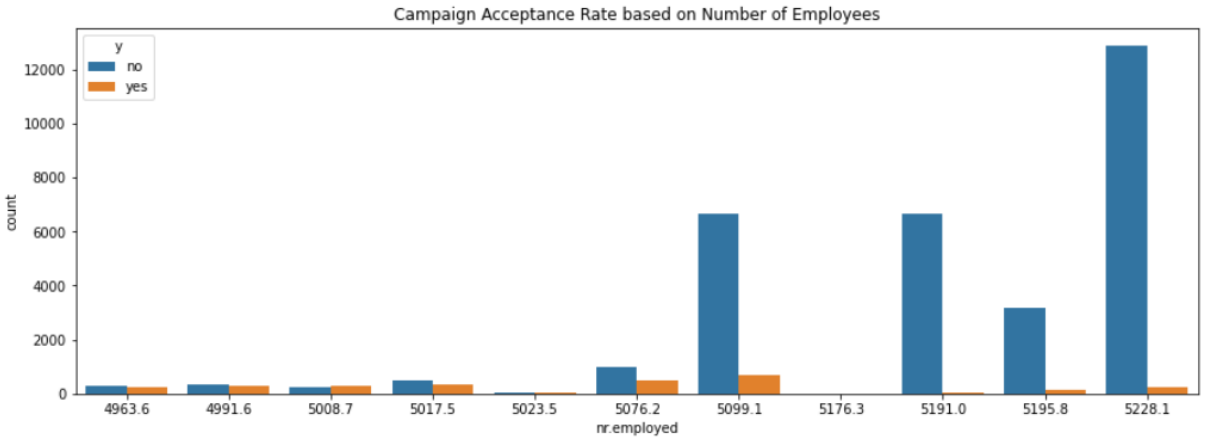
- Most of the campaign outreach happened when the CCI is low. But, the acceptance and non-acceptance of term-deposits are equal when the CCI is between -38.3 and -37.5.



- Most of the campaign outreach happened when the Euribor rates were high. But, the acceptance rate for term-deposits is high when the Euribor rates were low.



- Most of the campaign outreach happened when the Employee Count is high. But, the acceptance and non-acceptance of term-deposits are equal when the employee count is minimal.



6.2 Determine Next Steps

- Explore additional classification models like Naive Bayes, Stochastic Gradient Descent and measure the performance based on modeling metrics.
- since euribor3m is a socio-economic feature, performing *time-series* based modeling will help forecast and predict rates.
- By focusing on feature importance – euribor3 and nr.employed features, will enable the marketing team for better acceptance of bank products to its customers.

7. Deployment

This step is not needed for this practical application and can be skipped.