Week 11
EDA Presentation and proposed modeling technique

Data Science Intern at Data Glacier
Project: Bank Marketing (Campaign)

Name: Adrian Baysa
Email: adrianbaysa2@gmail.com
Country: Philippines
Batch Code: LISUM16

Problem Description and Github Repo

- **Problem Description:**

ABC Bank wants to sell its term deposit product to customers and before launching the product they want to develop a model which help them in understanding whether a particular customer will buy their product or not (based on customer's past interaction with bank or other Financial Institution).

- **Github Repo:** https://github.com/agbaysa/dataglacier_week11

EDA for Business Users - Overview

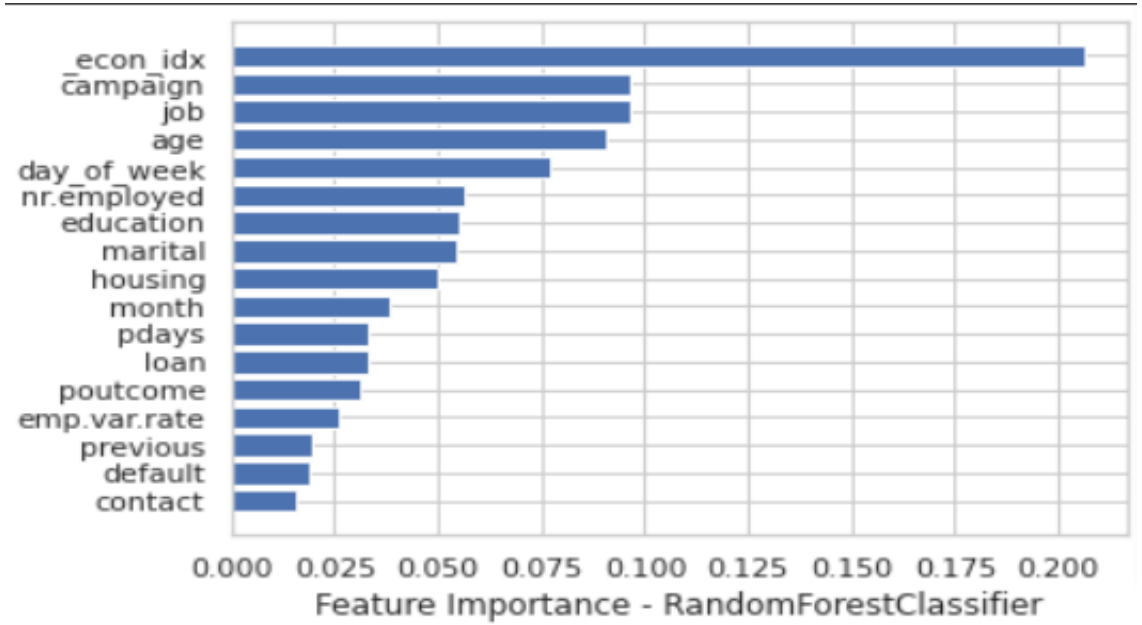
The following approach is done for Exploratory Data Analysis:

- First, a Random Forest Classifier is used to determine the feature importance
- pandas-profiling is again used on the processed data to identify skewness, cardinality, interactions, and correlations
- Formulate hypotheses based on the top features and provide key insights using EDA

EDA for Business Users – Feature Importance

The top features based on the Random Forest Classifier are as follows:

- Economic Index (factor of Consumer Price Index, Consumer Confidence Index, and Eurobor3m)
- Campaign:
- job
- age
- day_of_week
- education
- marital



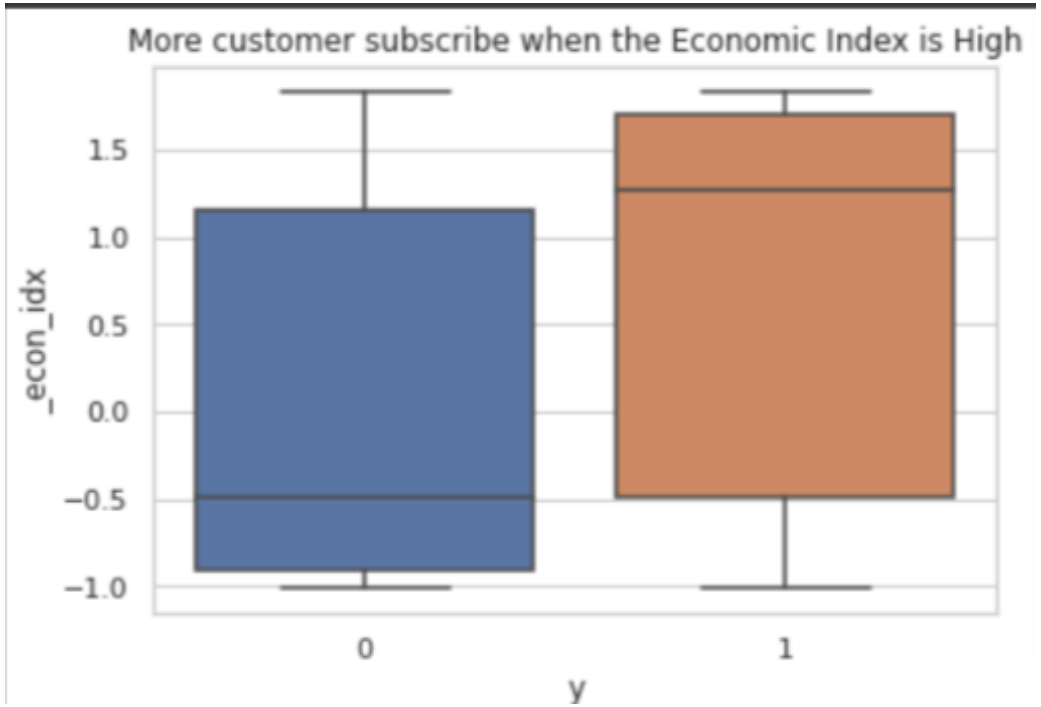
EDA for Business Users – Hypotheses

The following hypothesis were derived based on the top features:

- The target variable and _econ_idx has a linear relationship
- The lower the campaign, the more likely the customer will subscribe to a term deposit
- Those with white-collared jobs are more likely to subscribe
- Senior customers are more like to subscribe
- Clients mostly subscribe on the first day of the week (Monday)
- Customers with a higher education are more likely to avail
- Married individuals are more likely to avail

The target variable and _econ_idx has a linear relationship

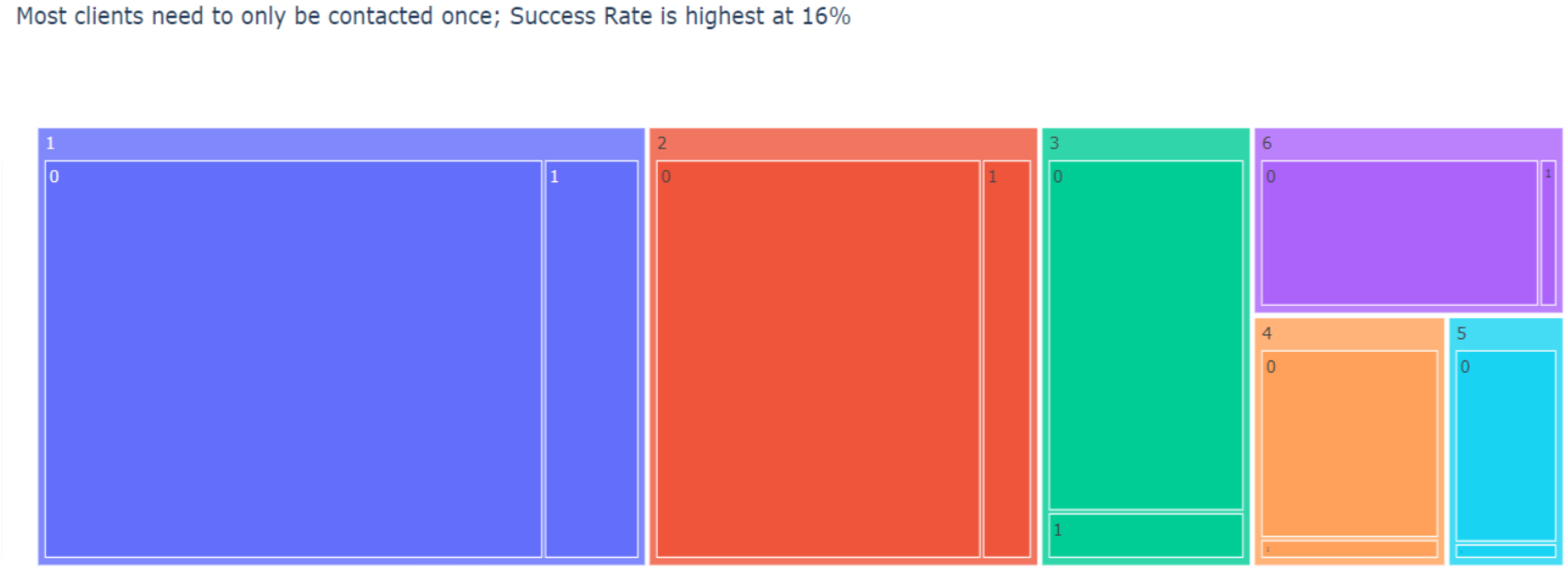
The Economic Index is the most important feature. The graph shows that clients subscribe to term deposit when the Economic Index is higher (i.e. with an average of above 1.0).



The lower the campaign, the more likely the customer will subscribe to a term deposit

Frequency of calls do not necessarily translate to deposit subscriptions. Those contacted only once has a Success Rate of 16%.

y	campaign	no	yes	percent
0	1	11635	2235	16.11
1	2	7711	1191	13.38
2	3	4248	572	11.87
3	4	2269	248	9.85
5	6	2862	186	6.10
4	5	1411	119	7.78



Those with white-collared jobs are more likely to subscribe

Admin, Technicians, and either Retired/Self-employed have higher Success Rates.

y	job	no	yes	percent
3	3	7550	1318	14.86
5	5	5865	1246	17.52
2	2	5035	720	12.51
0	0	6199	622	9.12
1	1	2286	323	12.38
4	4	3201	322	9.14

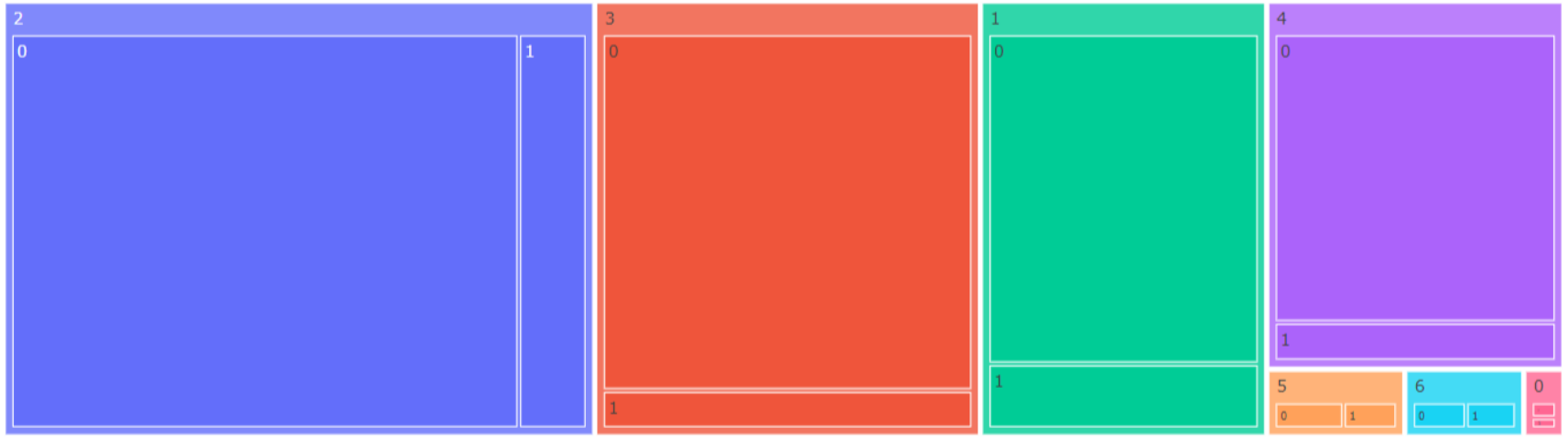
where: 'blue-collar': 0, 'management': 1, 'technician': 2, 'admin.': 3, 'services': 4, 'entrepreneur': 5, 'unknown': 5, 'retired': 5, 'self-employed': 5, 'unemployed': 5, 'housemaid': 5, 'student': 5

Senior customers are more like to subscribe

Both young and senior customers at the ends of the age spectrum are mostly likely to avail a term deposit with Success Rates that are above 40%

y	age	no	yes	percent
2	2	11589	1556	11.84
1	1	5320	1044	16.40
3	3	7746	828	9.66
4	4	4920	660	11.83
5	5	268	208	43.70
6	6	212	198	48.29
0	0	81	57	41.30

Customers ages 61 and above and those below 21 years old have higher Success Rates (above 40%)

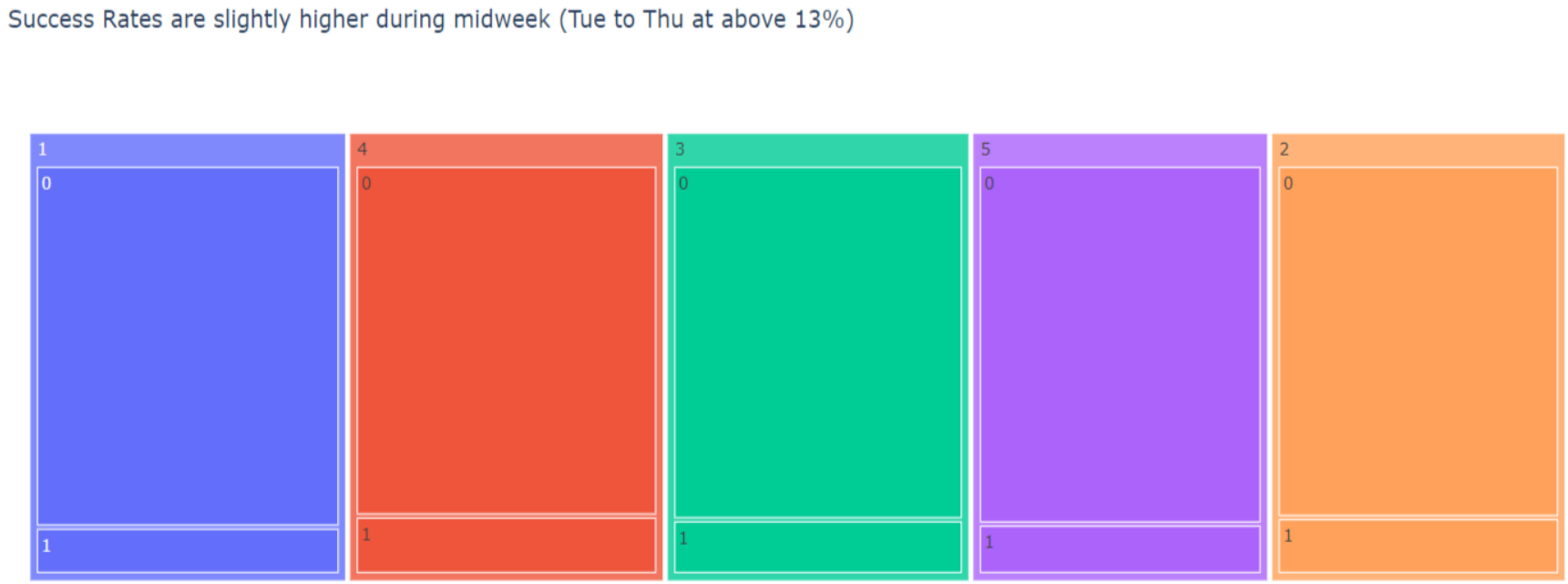


where: 20 and below: 0, 21 to 30: 1, 31 to 40: 2, 41 to 50: 3, 51 to 60: 4, 61 to 70: 5, 71 and above: 6

Clients mostly subscribe on the first day of the week (Monday)

The graph shows that deposit subscription rate is higher during the mid week (Tue to Thur) with an average Success Rate of above 13%.

y	day_of_week	no	yes	percent
3	4	6132	1024	14.31
1	2	5768	935	13.95
2	3	5965	923	13.40
0	1	6370	837	11.61
4	5	5901	832	12.36

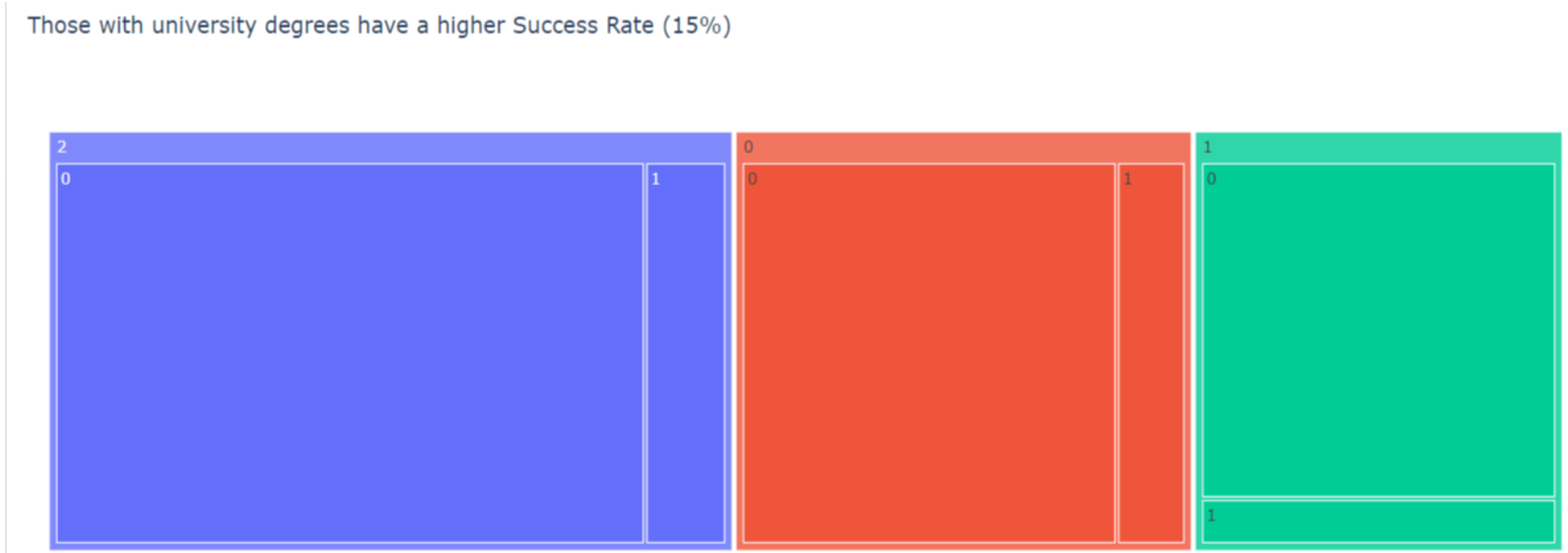


where: Mon: 1, Tue: 2, Wed: 3, Thu: 4, Fri: 5

Customers with a higher education are more likely to avail

Those with university degrees have a significantly higher Success Rate of deposit subscription at 14%.

y	education	no	yes	percent
2	2	13802	1903	12.12
0	0	8875	1629	15.51
1	1	7459	1019	12.02

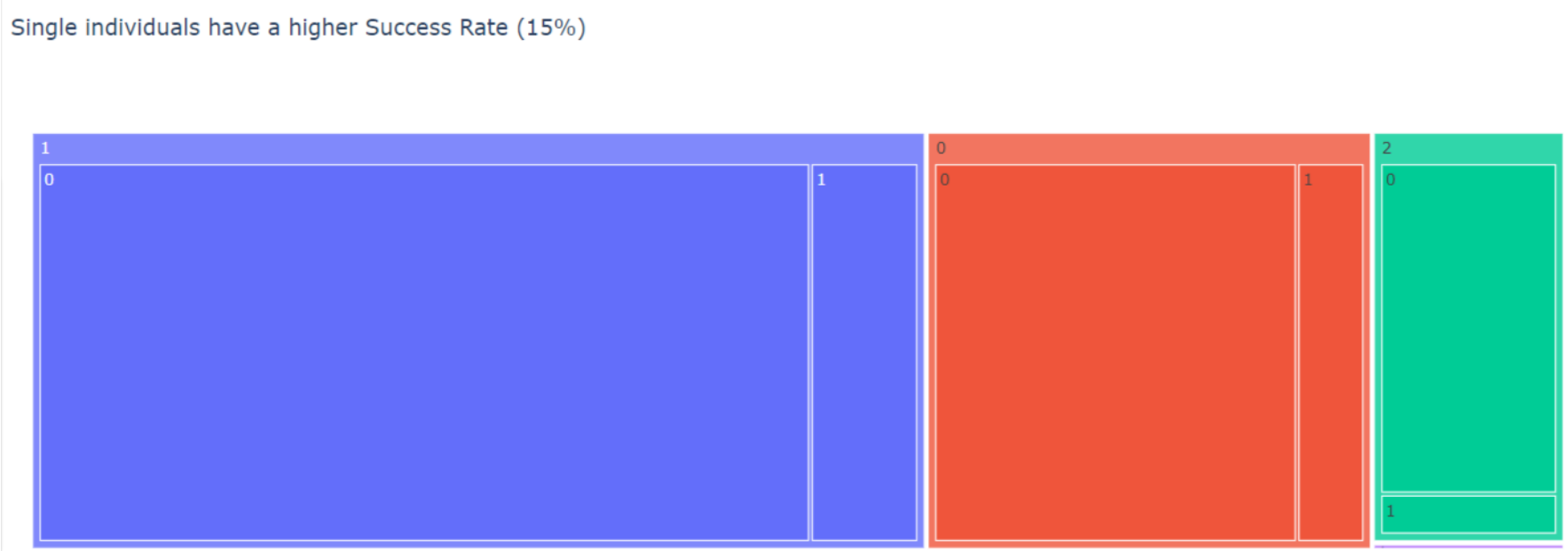


where:'university.degree': 0, 'high.school': 1, 'basic.9y': 2, 'professional.course': 2, 'basic.4y': 2, 'basic.6y': 2, 'unknown': 2, 'illiterate': 2

Married individuals are more likely to avail

Single individuals have the highest Success Rate at 14%. Note the 15% Success Rate for customers with unknown marital status.

y	marital	no	yes	percent
1	1	17758	2487	12.28
0	0	8504	1582	15.69
2	2	3807	471	11.01
3	4	67	11	14.10



where:'single': 0, 'married':1, 'divorced':2, 'unknown': 4

EDA: Data Profiling Highlights

Dataset statistics

Number of variables	18
Number of observations	34687
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	6.0 MiB
Average record size in memory	182.5 B

Variable types

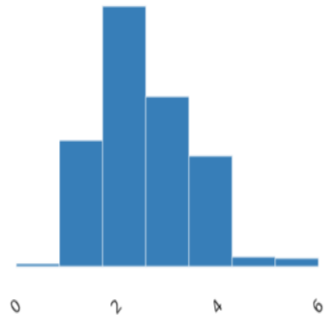
Numeric	8
Categorical	10

EDA: Data Profiling Highlights

age

Real number (ℝ)

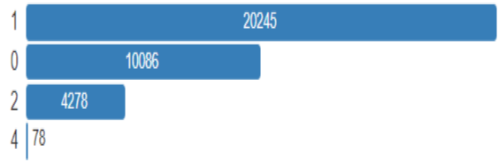
Distinct	7	Minimum	0
Distinct (%)	< 0.1%	Maximum	6
Missing	0	Zeros	138
Missing (%)	0.0%	Zeros (%)	0.4%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	2.4659382	Memory size	1.5 MiB



marital

Categorical

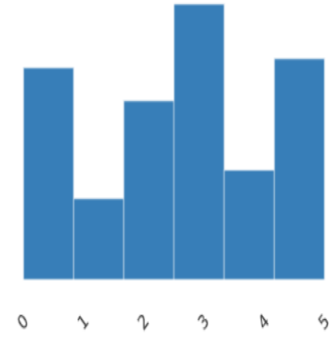
Distinct	4
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	1.5 MiB



job

Real number (ℝ)

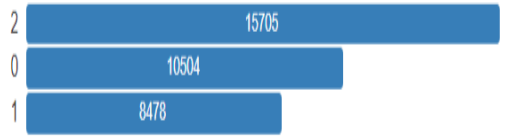
Distinct	6	Minimum	0
Distinct (%)	< 0.1%	Maximum	5
Missing	0	Zeros	6821
Missing (%)	0.0%	Zeros (%)	19.7%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	2.6052988	Memory size	1.5 MiB



education

Categorical

Distinct	3
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	1.5 MiB

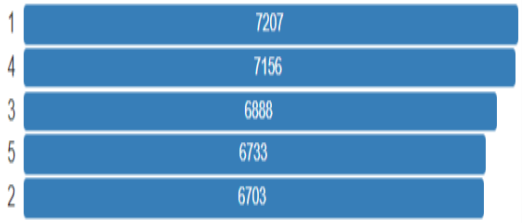


EDA: Data Profiling Highlights

day_of_week

Categorical

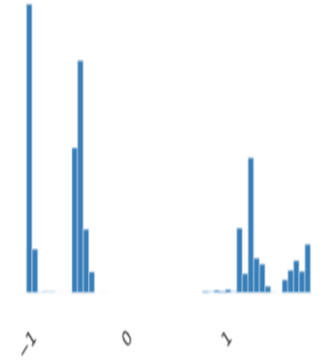
Distinct	5
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	1.5 MiB



_econ_idx

Real number (ℝ)

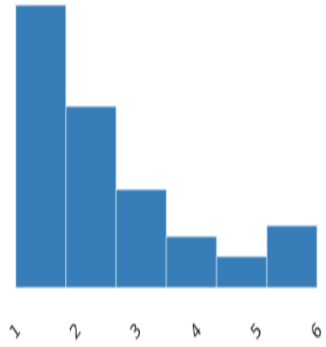
Distinct	375	Minimum	-1.0145415
Distinct (%)	1.1%	Maximum	1.8367855
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	22436
Infinite (%)	0.0%	Negative (%)	64.7%
Mean	0.039892994	Memory size	1.5 MiB



campaign

Real number (ℝ)

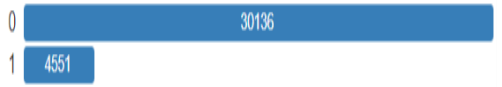
Distinct	6	Minimum	1
Distinct (%)	< 0.1%	Maximum	6
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	2.3680341	Memory size	1.5 MiB



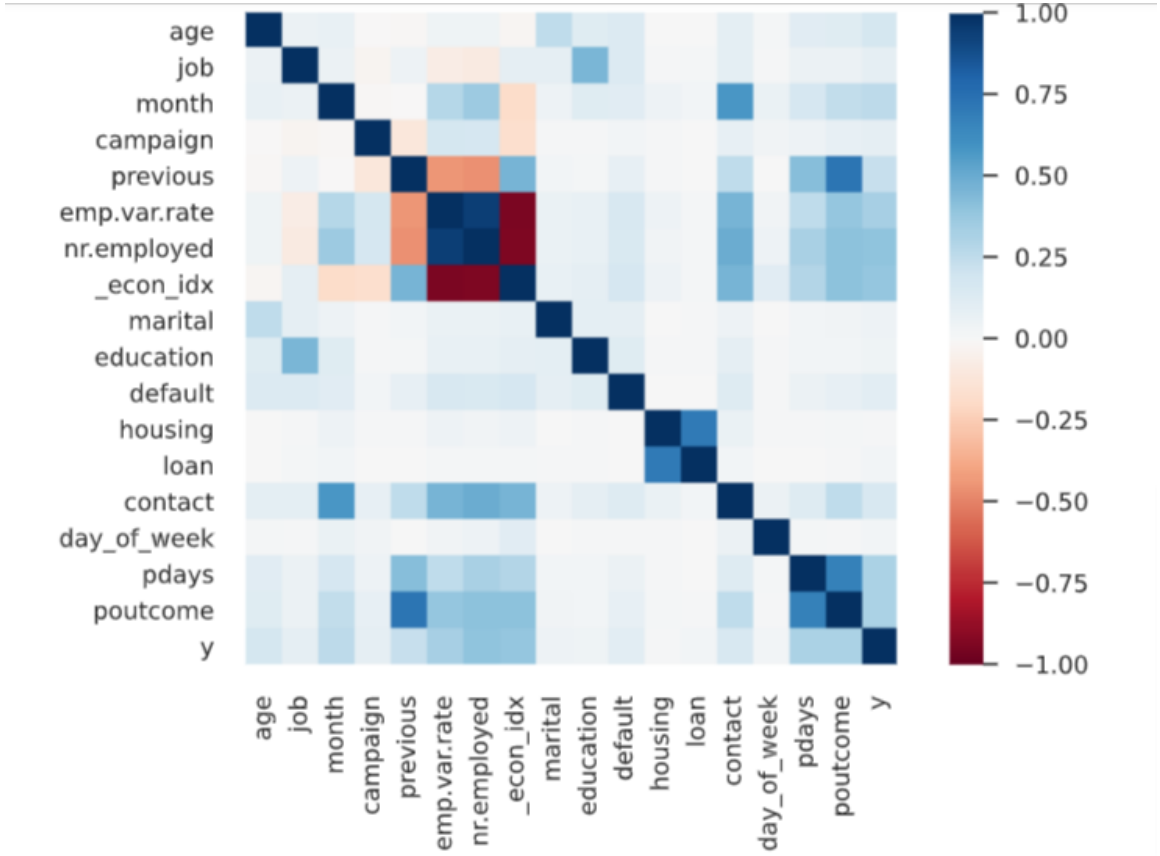
y

Categorical

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	1.5 MiB



EDA: Correlations



Recommended Model

Comparing Random Forest Classifier and Linear Discriminant Analysis, the latter performed better in terms of ROC-AUC and Accuracy based on cross-validation results. LDA has an ROC-AUC of 0.7720 and and an accurate of 0.8780.

```
# Compare RFC and LDA

models = []
models.append(('RFC', RandomForestClassifier()))
models.append(('LDA', LinearDiscriminantAnalysis()))

scoring = 'roc_auc'
results = []
names = []

for name, model in models:
  kfold = KFold(n_splits=num_folds)
  cv_results = cross_val_score(model, X_train, y_train, cv=kfold, scoring=scoring)
  results.append(cv_results)
  names.append(name)
  msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
  print(msg)
```

```
RFC: 0.743589 (0.014258)
LDA: 0.772081 (0.012481)
```

```
[423]  scoring = 'accuracy'
       results = []
       names = []

       for name, model in models:
         kfold = KFold(n_splits=num_folds)
         cv_results = cross_val_score(model, X_train, y_train, cv=kfold, scoring=scoring)
         results.append(cv_results)
         names.append(name)
         msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
         print(msg)
```

```
RFC: 0.868493 (0.004901)
LDA: 0.878048 (0.005732)
```

Recommended Model

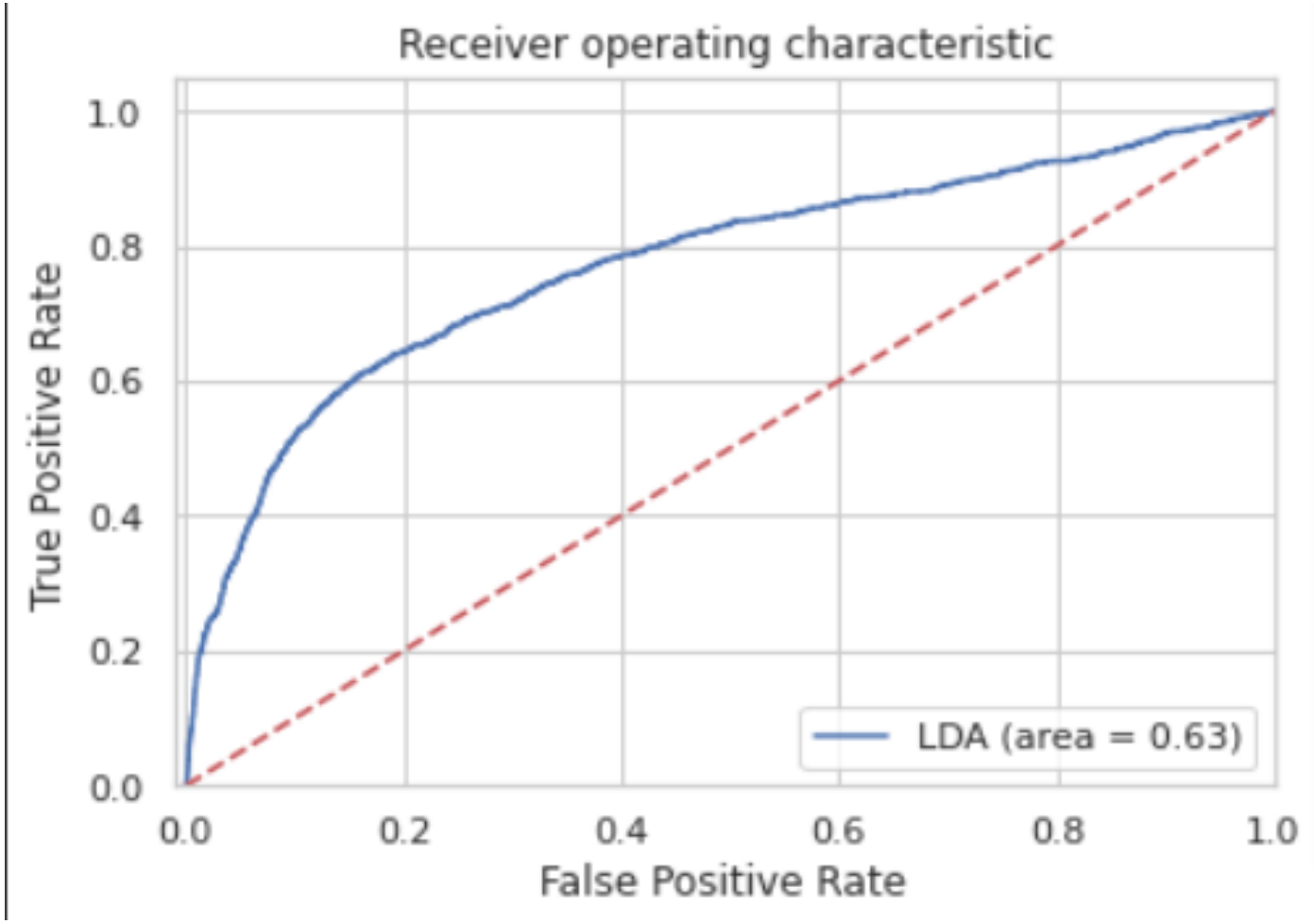
Based on the Confusion Matrix, several items are highlighted below:

- We have correctly identified clients who will surely avail of the term deposit (i.e. low hanging fruit) consisting of 223 customers. These accounts should be prioritized.
- We have also correctly identified 8,944 customers who will surely be unlikely to avail of the subscription (i.e. poison fruit). These accounts should be avoided.
- For those not identified by the model, the campaign should focus on the demographics that were highlighted.

```
[424]  lda = LinearDiscriminantAnalysis()
       lda.fit(X_train, y_train)
       y_pred = logreg.predict(X_valid)
       print('Accuracy of LDA: {:.2f}'.format(lda.score(X_valid, y_valid)))

       from sklearn.metrics import confusion_matrix
       confusion_matrix = confusion_matrix(y_valid, y_pred)
       print("Confusion Matrix for LDA :\n",confusion_matrix)
       print("Classification Report for LDA:\n",classification_report(y_valid, y_pred))
```

```
Accuracy of LDA: 0.88
Confusion Matrix for LDA :
 [[8944   98]
 [1142  223]]
Classification Report for LDA:
              precision    recall  f1-score   support

           0       0.89      0.99      0.94      9042
           1       0.69      0.16      0.26      1365

    accuracy                           0.88     10407
   macro avg       0.79      0.58      0.60     10407
weighted avg       0.86      0.88      0.85     10407
```

Recommended Model



Business Recommendations

Leverage on the LDA Model

- Run the LDA model in order to determine the customers most likely to avail of the deposit (223 customers or 2% of the total). These clients should be prioritized in terms of the campaign. **Impact: Translates to cheaper cost of acquisition due to the high probability of deposit subscription.**
- Similarly, avoid customers that are most likely to not avail of the subscription (8,944 customers or 85% of the total). **Impact: Translates to campaign cost saves of 85% as these clients will be deprioritized.**

Actions for Identified Fence-Sitters

- For the remaining clients (1,240 or 13%), rules-of-thumb in terms of campaign strategy (e.g. who to reach out first, etc.) can rely on the demographics of the customer in terms of the following:
  - Campaign calls can only be one and not necessarily repetitive.
  - Priotize those with jobs in Admin Technicians, Retired/Self-Employed
  - Prioritize clients 20 years old and below and those 61 and above.
  - Campaigns are more effective on Tuesdays to Thursdays
  - Prioritize those with University Degrees
  - Single and Married individuals can be prioritized
  - Adjust the campaign based on economic conditions. The Economic Index can measure the economic environment (Consumer Price Index * Consumer Confidence Index * Euribor3m). An index of 1.0 and above signifies the likelihood of deposit subscriptions.

Thank You