Bank Marketing Campaign

(EDA and ML Model Exploration)

Contents

About the Bank Marketing Campaign

ABC Bank wants to sell it's 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).

Dataset information:

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 ('yes') or not ('no') subscribed.

The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).

Why this project matters: Improving conversion rates saves costs and boosts ROI.

Approaching the Data

The CSV file ‘*bank-additional.csv*’ had numerical and categorical columns along with the target variable ‘y’ and the data was cleaned and feature engineered with the help of *label encoding* and *one hot encoding*. Then this cleaned data was explored visually and then fed into different ML models.

The following ML models were implemented:

1. Logistic Regression (Linear Model)
2. Decision Tree (Tree-based model)
3. Random Forest (Ensemble of Trees)
4. XGBoost (Boosting Algorithm)

Cleaning the Data

This dataset was previously cleaned in Jupyter Notebook using techniques like *label encoding* and *one hot encoding* to divide the columns into relevant and more understandable data points/features, based on which the EDA has been performed. The resulting columns are displayed on the right.

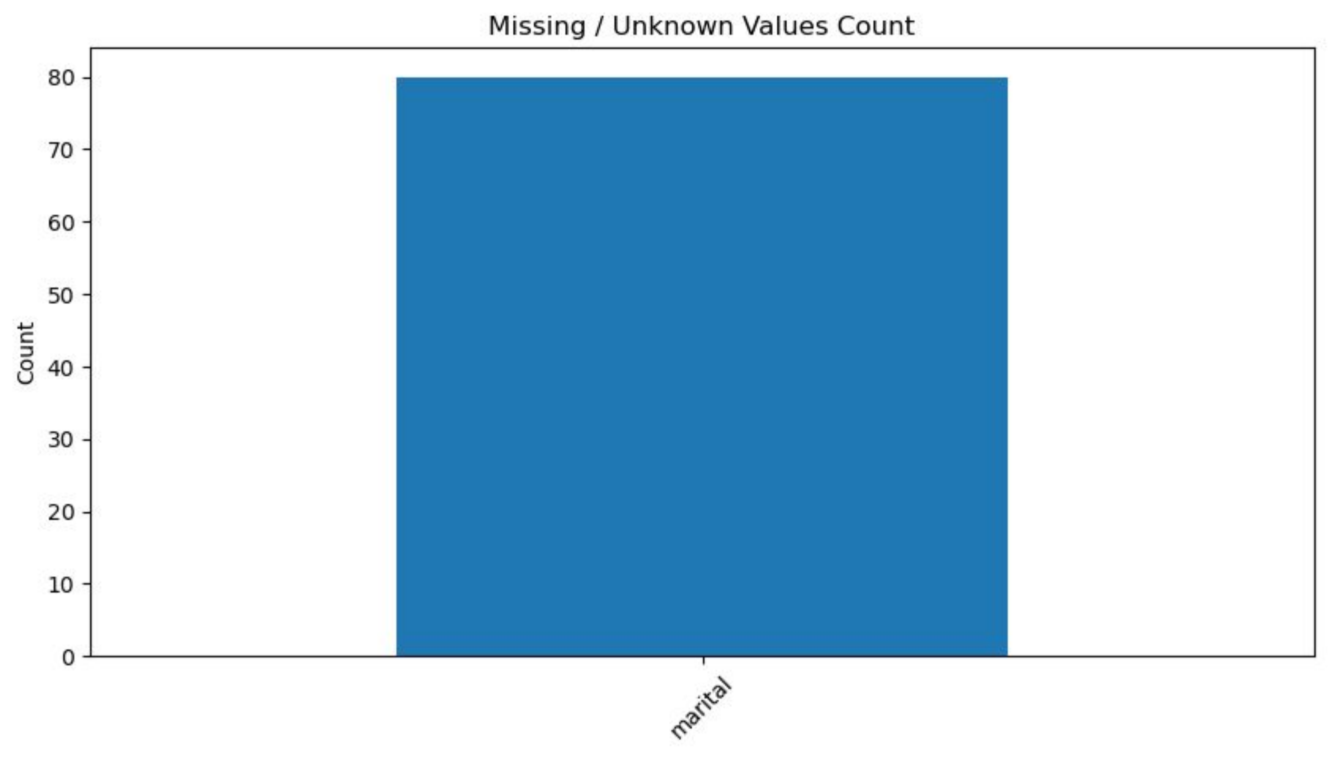
```
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   age                             41176 non-null  float64
 1   marital                         41176 non-null  object
 2   default                         41176 non-null  int64
 3   housing                         41176 non-null  int64
 4   loan                            41176 non-null  int64
 5   contact                         41176 non-null  object
 6   month                           41176 non-null  object
 7   day_of_week                     41176 non-null  object
 8   duration                        41176 non-null  float64
 9   campaign                        41176 non-null  float64
 10  pdays                           41176 non-null  float64
 11  previous                        41176 non-null  int64
 12  emp.var.rate                    41176 non-null  float64
 13  cons.price.idx                  41176 non-null  float64
 14  cons.conf.idx                   41176 non-null  float64
 15  euribor3m                       41176 non-null  float64
 16  nr.employed                     41176 non-null  float64
 17  y                               41176 non-null  int64
 18  job_blue-collar                 41176 non-null  bool
 19  job_entrepreneur                41176 non-null  bool
 20  job_housemaid                   41176 non-null  bool
 21  job_management                  41176 non-null  bool
 22  job_retired                     41176 non-null  bool
 23  job_self-employed               41176 non-null  bool
 24  job_services                    41176 non-null  bool
 25  job_student                     41176 non-null  bool
 26  job_technician                  41176 non-null  bool
 27  job_unemployed                  41176 non-null  bool
 28  job_unknown                     41176 non-null  bool
 29  education_basic.6y              41176 non-null  bool
 30  education_basic.9y              41176 non-null  bool
 31  education_high.school           41176 non-null  bool
 32  education_illiterate            41176 non-null  bool
 33  education_professional.course   41176 non-null  bool
 34  education_university.degree     41176 non-null  bool
 35  education_unknown               41176 non-null  bool
 36  poutcome_nonexistent            41176 non-null  bool
 37  poutcome_success                41176 non-null  bool
dtypes: bool(20), float64(9), int64(5), object(4)
memory usage: 6.8+ MB
```

Exploratory Data Analysis: *The next few slides shall display visualizations of the following findings from the EDA of the dataset.*

- ***Null values:*** The marital column had the most amount of null values compared to all other features
- **These were the Conversion Rates based on different factors:**
    1. ***Job type:*** Students and retired folks were found to be the most likely to convert.
    2. ***Education type:*** Illiterates and people with unknown data about their education were the biggest converted groups with university educated, high school graduates and education professionals following closely behind.
    3. ***Marital Status:*** Married folks were more likely to convert into users compared to single folks and divorcees.
    4. ***Contact Style:*** People using mobile phones were way more likely to convert compared to the ones still using telephones.
    5. ***Monthly basis:*** The months from May to August were the best months in terms of achieving the highest user conversion rates.
    6. ***Weekly basis:*** All weekdays had a similar performance when it came to the highest conversion rate, with Thursday being slightly better than the others.

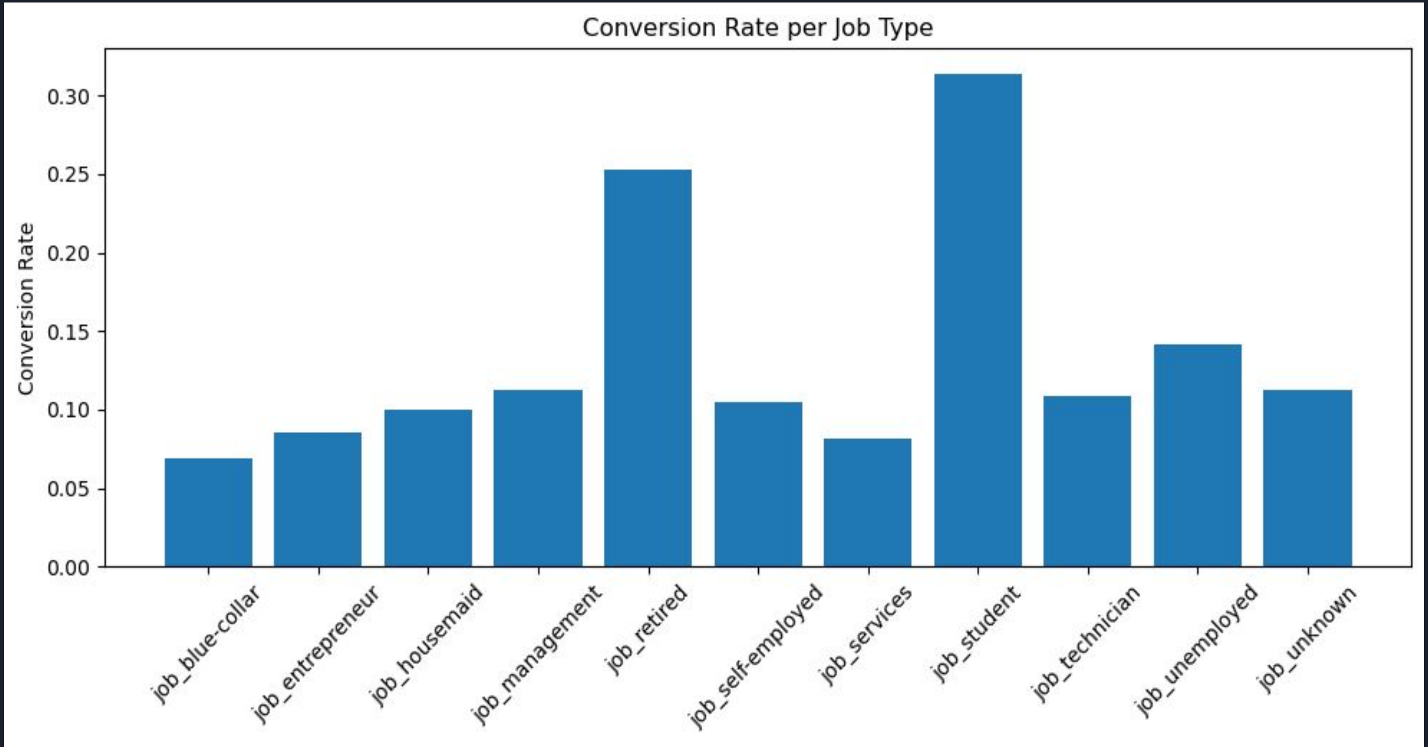
Exploratory Data Analysis

Representing Null values: The marital column had the most amount of null values out of all the other features.

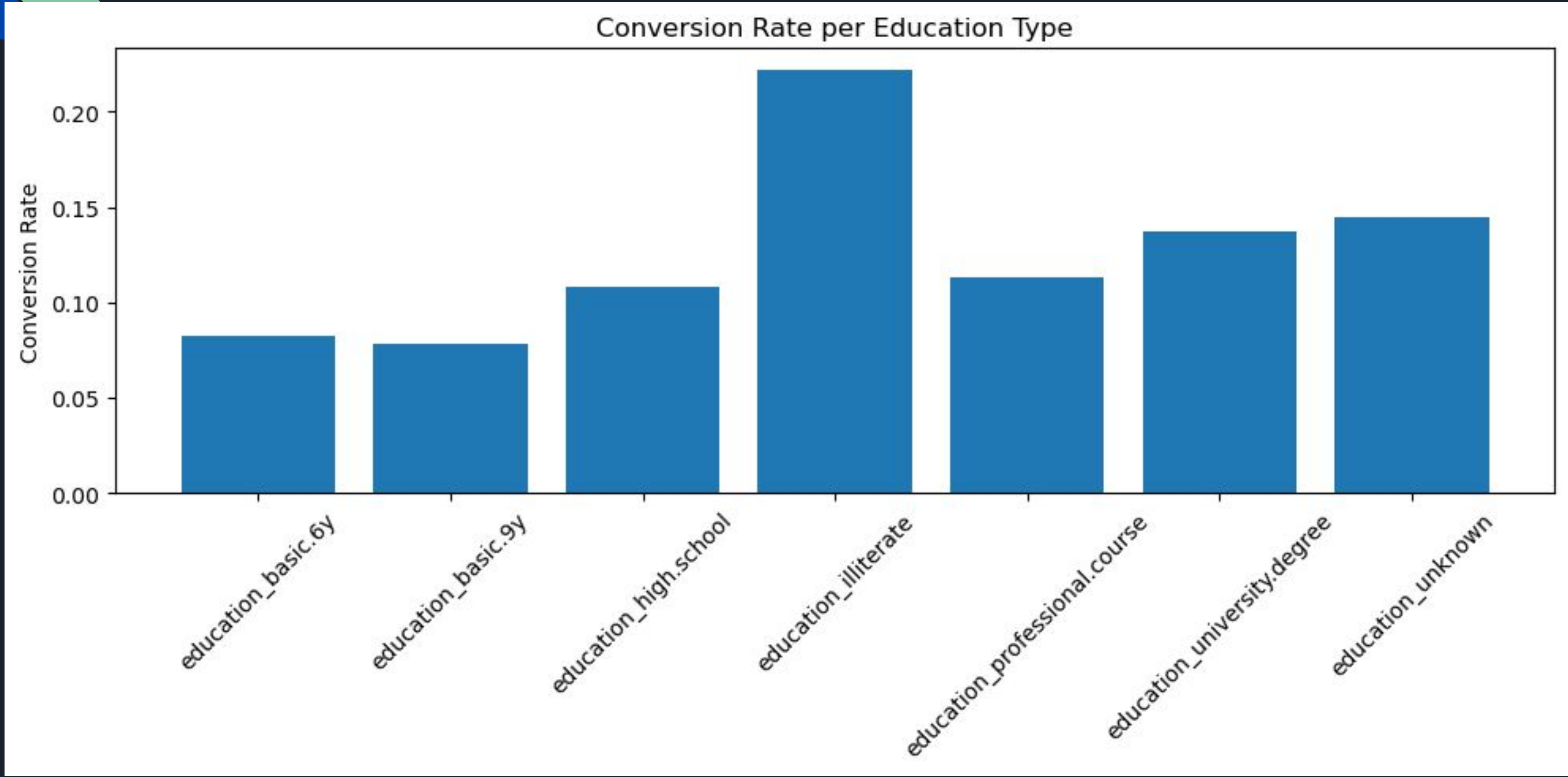


Conversion rate based on Job Type

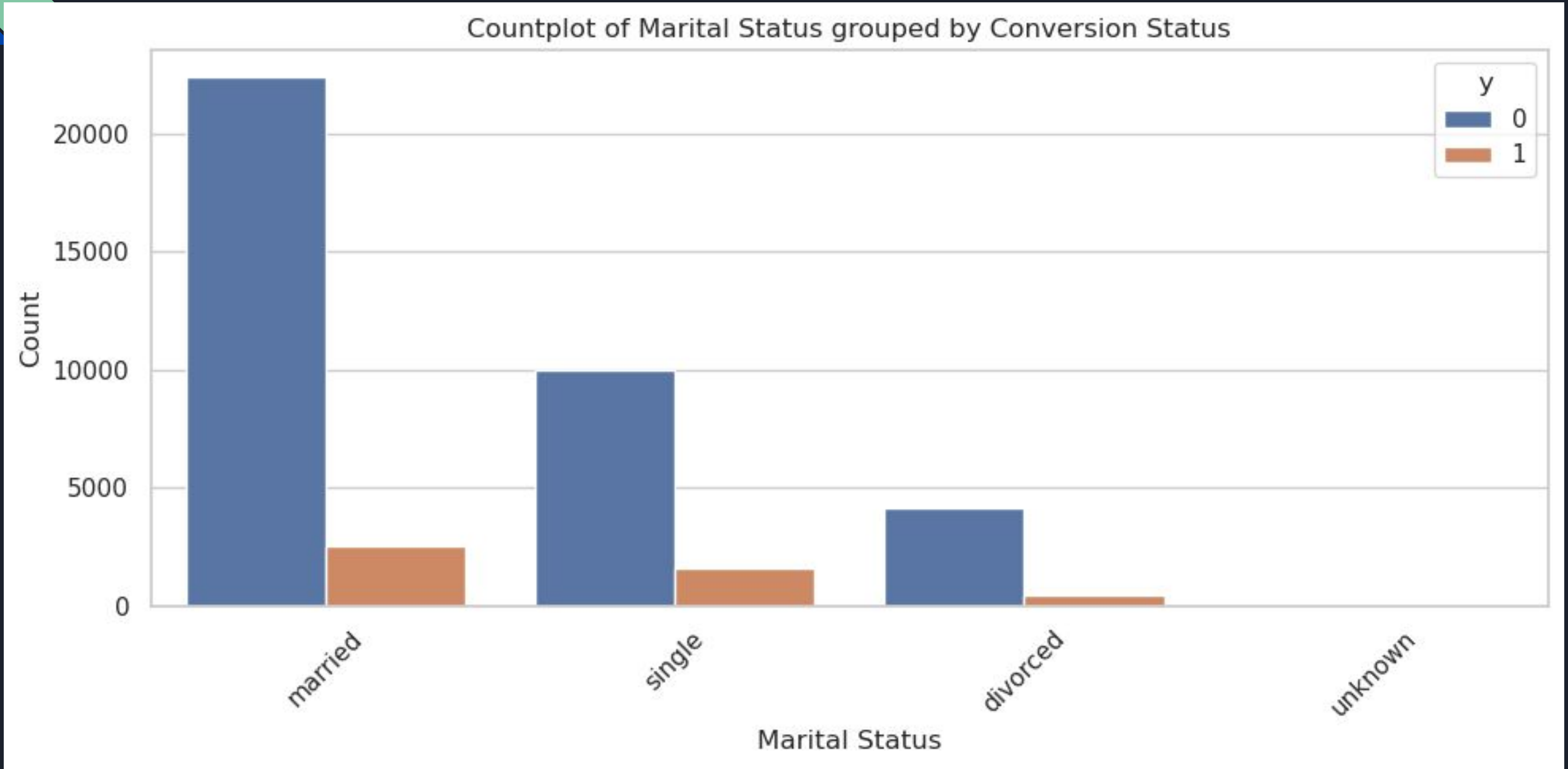
Students and retired folks were found to be the most likely to convert.



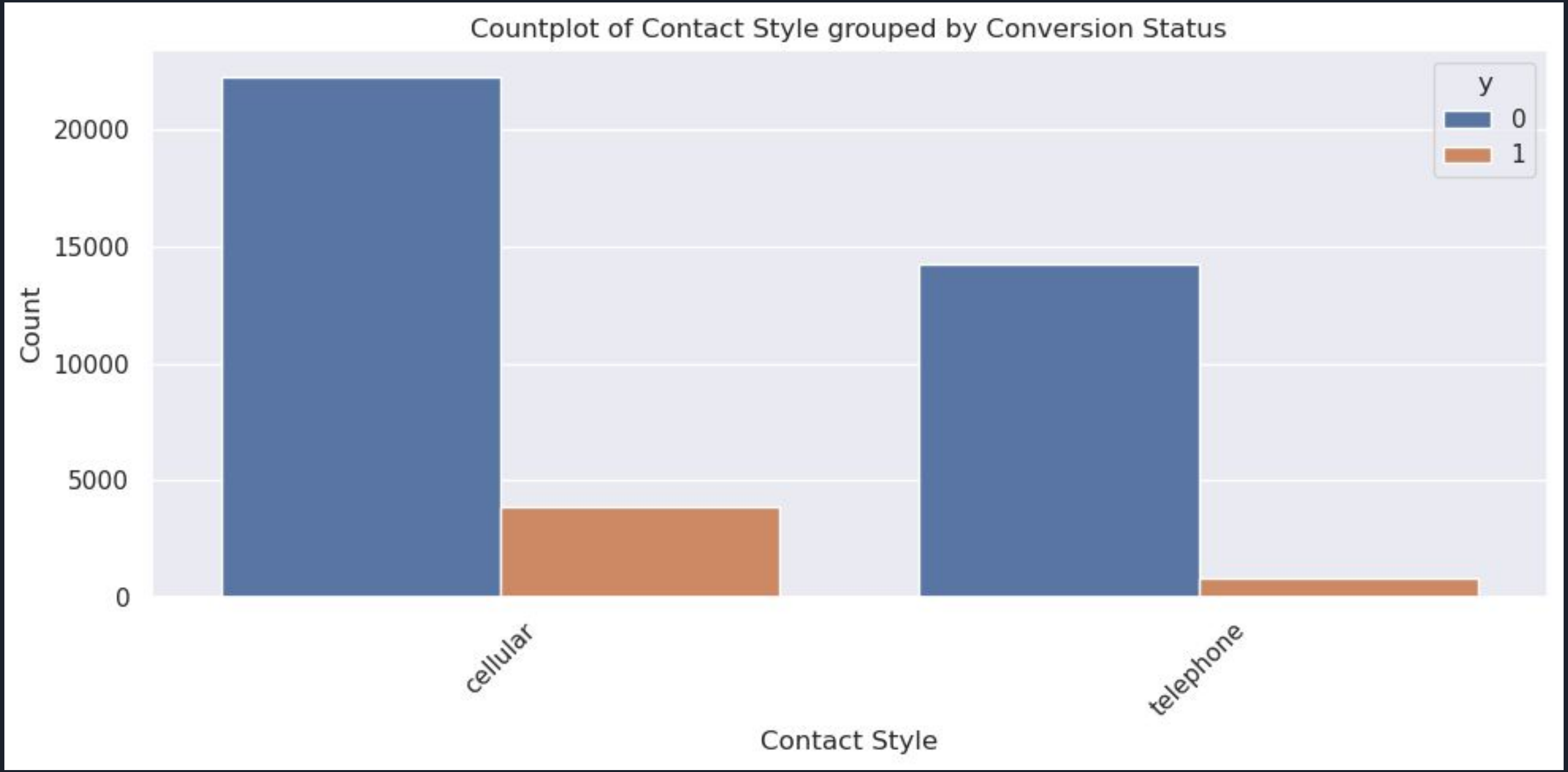
Illiterates and people with unknown data about their education were the biggest converted groups with university educated, high school graduates and education professionals following closely behind.



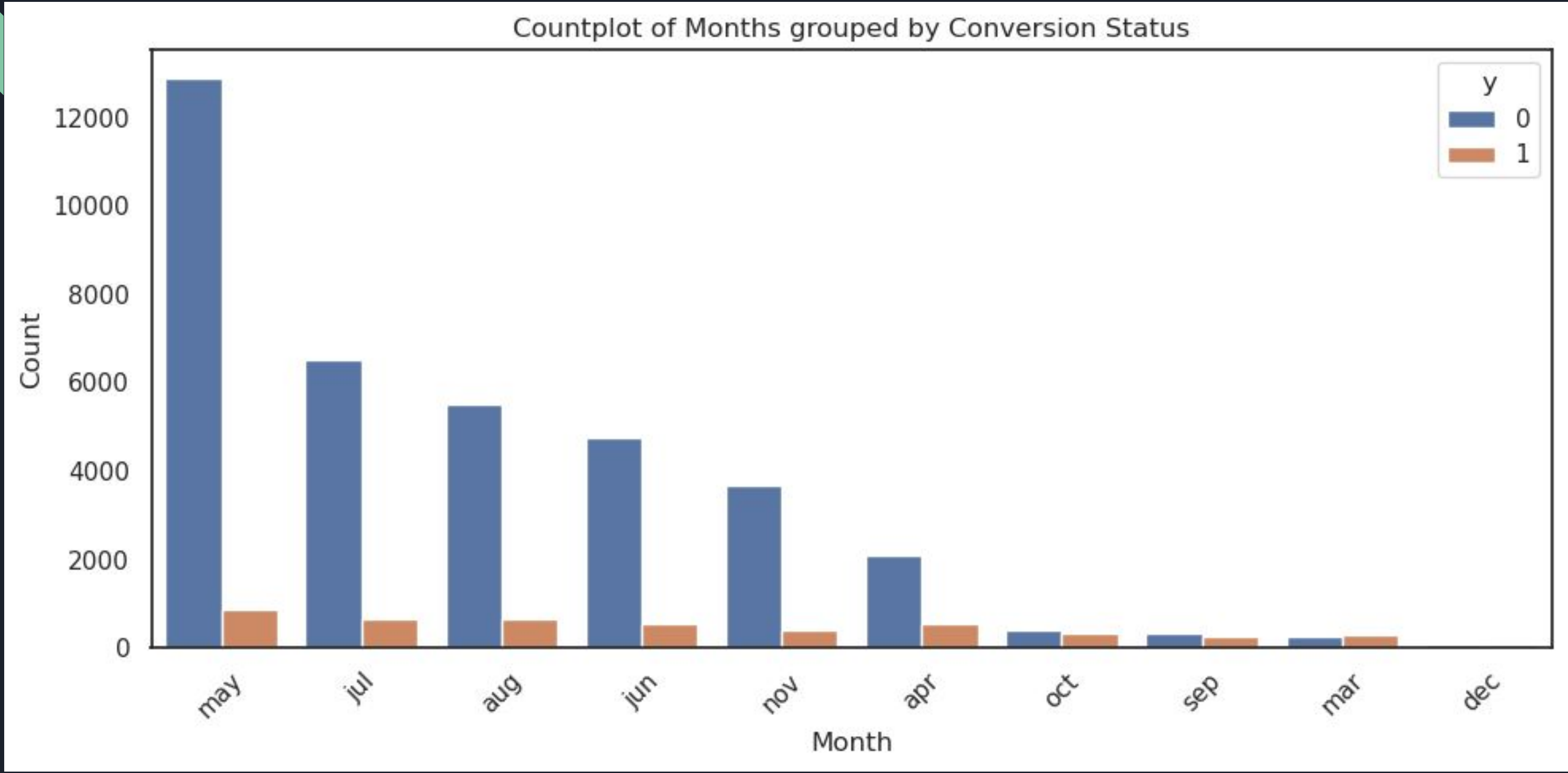
Married folks were more likely to convert into users compared to single folks and divorcees.



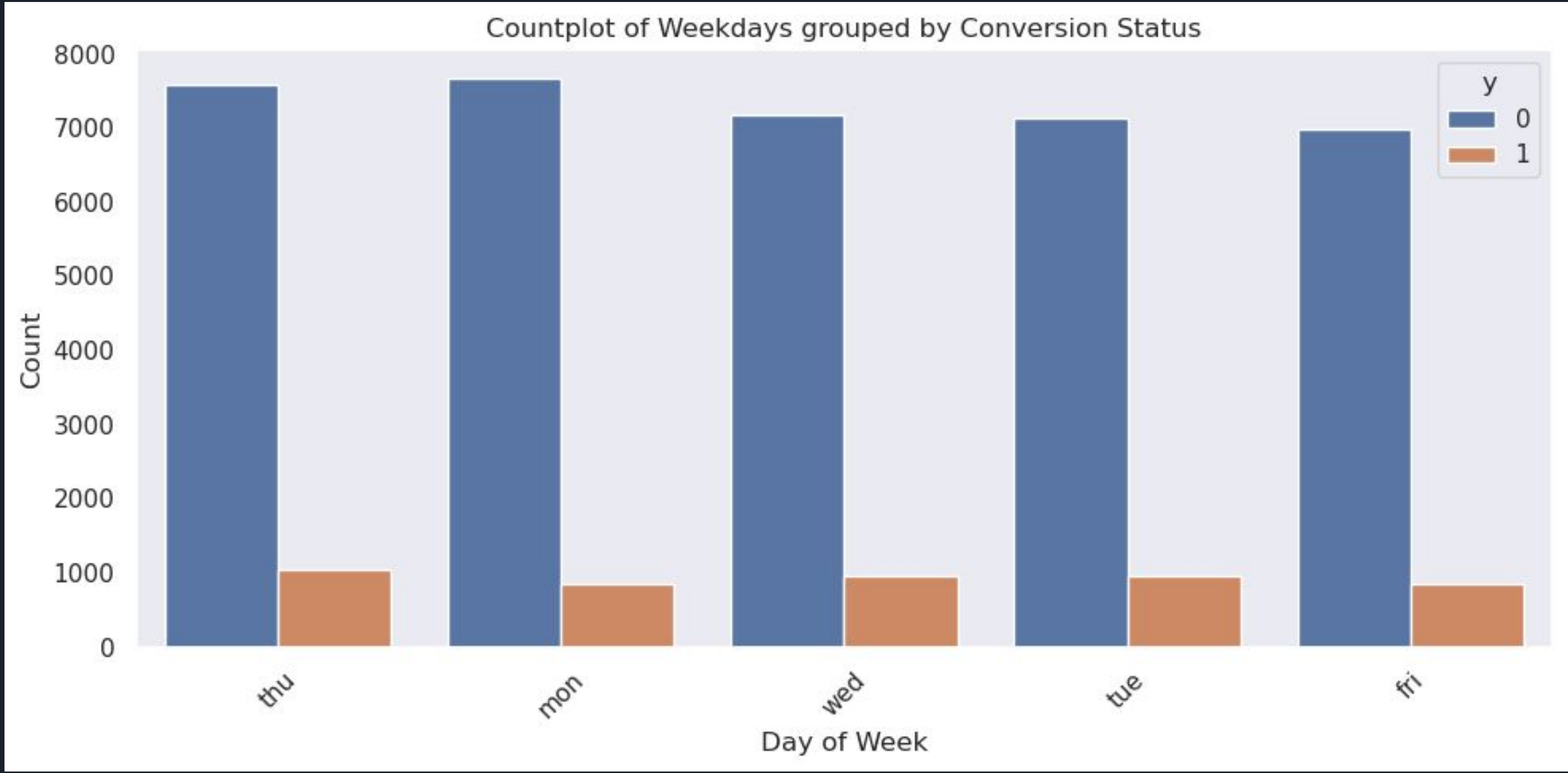
People using mobile phones were way more likely to convert compared to the ones still using telephones.



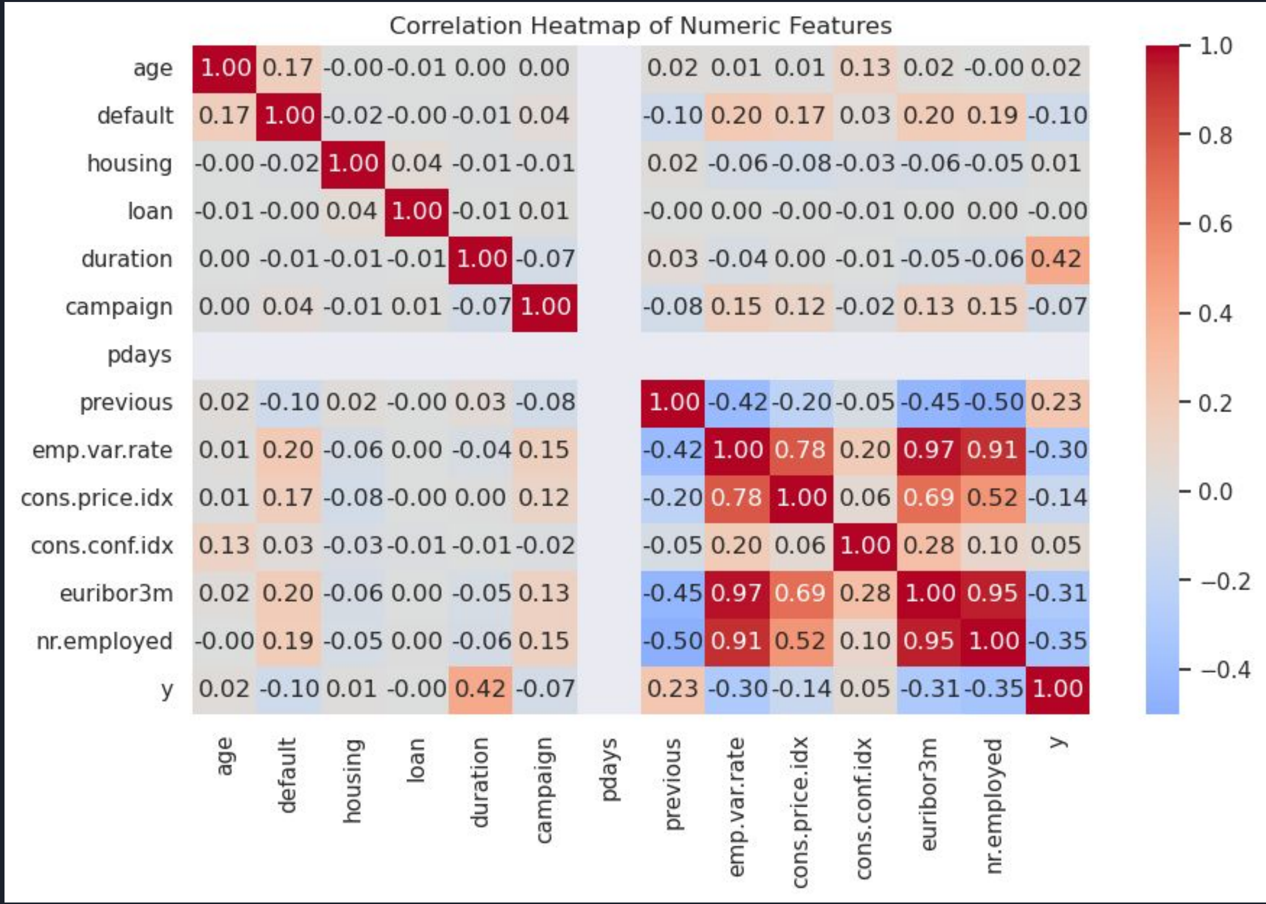
The months from May to August were the best months in terms of achieving the highest user conversion rates.



All weekdays had a similar performance when it came to the highest conversion rate, with Thursday being slightly better than the others.



A quick comparison of the numerical features of the dataset:



Summary of the heatmap

- euribor3m and nr.employed have strong positive correlations with features like emp.var.rate and cons.price.idx (values above 0.9).
- euribor3m and nr.employed have moderate negative correlation with y (around -0.3 to -0.35), meaning as they increase, success rate (y=1) tends to decrease.
- duration has a moderate positive correlation with y (0.42), meaning longer calls are linked to higher conversion.
- Most other features like age, loan, housing, etc., have very weak or no correlation with y (values near 0).

So, to summarise, **we shall focus more on *duration*, *euribor3m*, *nr.employed***, and ***emp.var.rate*** as they seem ***more relevant to y.***

Model Exploration: Logistic Regression

These were the Performance metrics for the Logistic Regression model:

```
Training Accuracy: 0.8576
Test Accuracy: 0.8616
Cross-validation Scores: [0.85761991 0.85686096 0.85458409 0.85777171 0.85913783]
Mean CV Accuracy: 0.8572

Confusion Matrix:
[[6252 1056]
 [  84  844]]

Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.86      0.92      7308
           1       0.44      0.91      0.60       928

    accuracy                           0.86      8236
   macro avg       0.72      0.88      0.76      8236
weighted avg       0.93      0.86      0.88      8236
```

Model Exploration: Logistic Regression

Logistic Regression Results

- Training Accuracy: **85.7%**, Test Accuracy: **86.1%**
- Strong consistency in **cross-validation (CV Avg: 85.7%)**
- High precision for non-subscribers (0.99), but lower for subscribers (0.44)
- High recall for subscribers (0.91), ensuring fewer missed positives
- Overall Accuracy: **86%**, Weighted F1-Score: **0.88**

The model performs well overall, but has challenges with **subscriber precision**, meaning it predicts more false positives.

Model Exploration: Decision Tree Classifier

Performance Metrics for the Decision Tree Classifier model:

```
Best Parameters: {'criterion': 'gini', 'max_depth': 5, 'min_samples_leaf': 4, 'min_samples_split': 2}
Accuracy: 0.9123360854783875

Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.96      0.95      7308
           1       0.62      0.56      0.59       928

    accuracy                           0.91      8236
   macro avg       0.78      0.76      0.77      8236
weighted avg       0.91      0.91      0.91      8236


Confusion Matrix:
 [[6990  318]
 [ 404  524]]
```

Feature Importance metrics of the Decision Tree Classifier model

The model relies heavily on just some features and the rest are almost negligible, and some are completely unused (importance = 0).

Feature	Importance
duration	0.508516
nr.employed	0.349553
poutcome_success	0.043254
euribor3m	0.035085
cons.conf.idx	0.033143
month_oct	0.012255
cons.price.idx	0.007934
contact_telephone	0.004231
month_may	0.004226
day_of_week_mon	0.001013
job_unknown	0.000790
month_aug	0.000000
poutcome_nonexistent	0.000000
education_unknown	0.000000
marital_married	0.000000
marital_single	0.000000
marital_unknown	0.000000

Model Exploration: Random Forest

Performance Metrics

MSE = mean squared error

```
Random Forest R² Score: 0.4151
Random Forest RMSE: 0.2466
```

Top 10 Important Features:

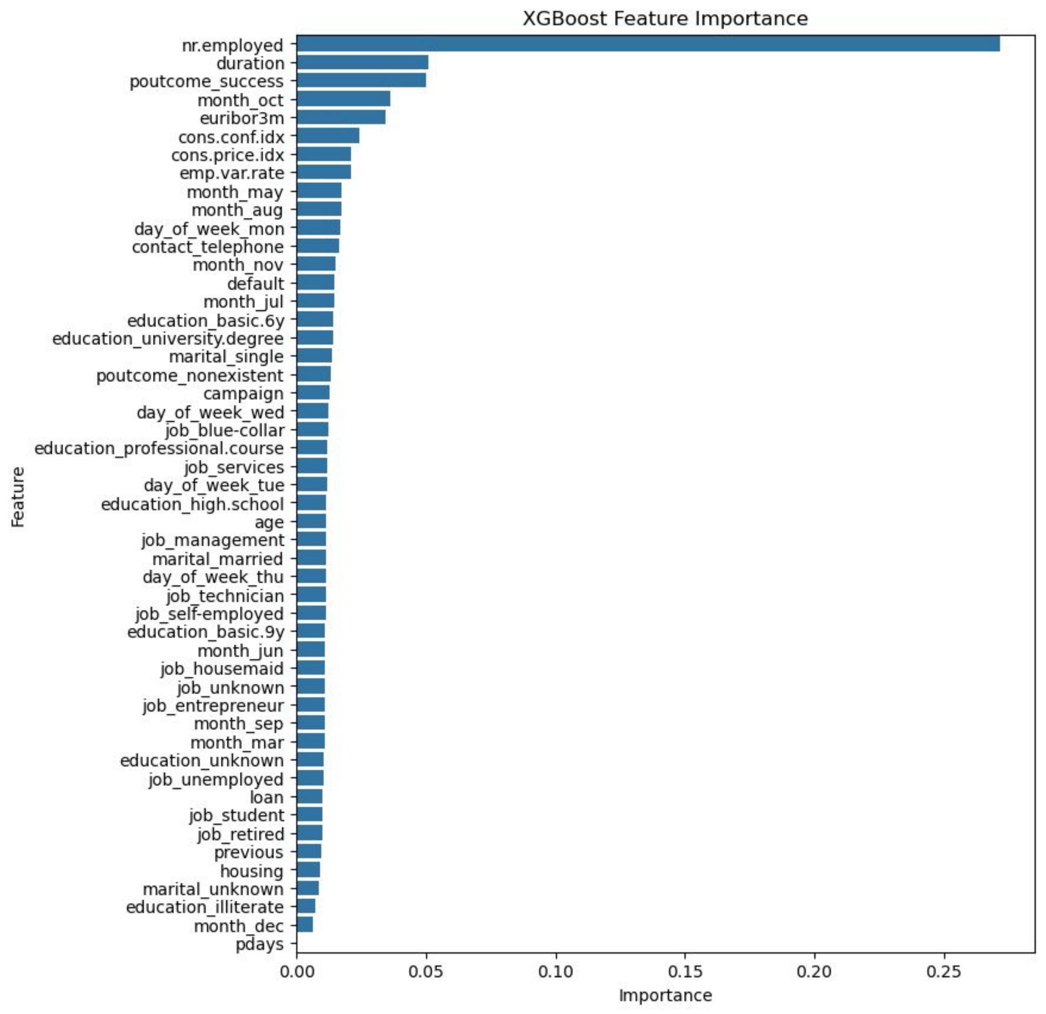
	Feature	Importance
4	duration	0.326841
12	nr.employed	0.157336
0	age	0.089557
11	euribor3m	0.082881
5	campaign	0.035177
10	cons.conf.idx	0.024657
32	poutcome_success	0.020333
2	housing	0.015659
9	cons.price.idx	0.013559
3	loan	0.012479

Model Exploration: Random Forest

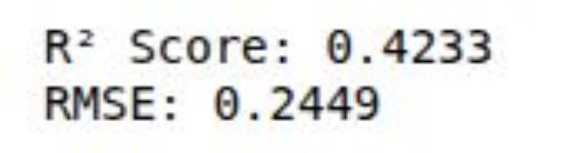
From the aforementioned results we can deduce that:

- The model explains ~41.5% of the variance in the target variable.
- The low RMSE means prediction errors are relatively small compared to the scale of the target.
- Duration has the highest importance (0.326), meaning the model relies on it heavily for predictions.

Model Exploration: XGBoost



Performance Metrics:


R² Score: 0.4233
RMSE: 0.2449

Model Exploration: XGBoost

From the aforementioned performance metrics and plot we can deduce that:

- The top features driving the model are: **nr.employed** (employment-related indicator), **duration** (length of call), **poutcome_success** (previous campaign outcome), **month_oct, euribor3m** (macroeconomic indicators and seasonality)
- **R^2 score = 0.4233**: The model explains about 42.3% of the variance, which is moderate.
- **RMSE = 0.2449:** shows the average error in predictions; the smaller, the better (interpretation depends on your target variable's scale).

To summarise, the XGBoost model is picking up employment, call duration, campaign history, and economic factors as the strongest predictors.

Insights from Modeling

- **Linear vs. Non-Linear:** Logistic Regression underperformed compared to tree-based methods, highlighting non-linear patterns in the data.
- **Tree Models:** Decision Tree was interpretable but unstable; Random Forest improved reliability by averaging multiple trees.
- **Boosting (XGBoost):** Delivered the best trade-off between accuracy and generalization.
    - ***Feature Importance:*** Employment levels, call duration, past campaign success, and macroeconomic indicators drive campaign outcomes most strongly.

Final Takeaway 🚀

Best Model: *XGBoost* delivered the strongest and most reliable performance.

Key Drivers Identified:

1. Employment indicators
2. Call duration
3. Previous campaign success
4. Economic conditions (e.g., euribor3m, nr.employed)

Actionable Insights for the Bank:

1. Prioritize high-potential customers for targeted outreach.
2. Optimize call duration & timing to maximize conversions.
3. Align campaigns with favorable economic conditions.

Impact: *More efficient campaigns → reduced costs → higher conversion rates.*

Thank You!