# DSGA 1017 Final Project Report

Chitvan Goyal
cg4174@nyu.edu

Ilina Tripathi
it2189@nyu.edu

May 2023

## 1 Background

Banks and financial institutions spend a lot of resources on finding clients and getting them to open a term deposit with the bank. The primary aim of a marketing campaign is to provide value to the customers and also, to build strong relationships which will generate more business from clients. A good marketing campaign can also generate good word-of-mouth publicity for the bank and can allow it to enter new markets.
There are various factors that have to be considered to design an effective marketing campaign such as the price offered, the target demographic, the distribution channels etc. The ADS we are auditing aims to analyse the past marketing campaign data of a Portuguese financial institution and identify what drives a good marketing strategy that results in clients opening a term deposit. The ADS conducts an in-depth analysis of all these attributes in the dataset, performs exploratory data analysis, data preprocessing and then trains various classifiers on the data to obtain the best results. There are various features that could be considered sensitive in this dataset such as age, education and marital status.

Often banks target certain kind of people because they feel that it might be more profitable to conduct business with them. This means they also tend to avoid doing business with people who are a credit risk, or seem like they are vulnerable financially. People who are students might have education loans and that puts them in debt and therefore affects them financially. Also, people's marital status might have a significant impact on their bank balance. These biases can also reflect in the marketing strategies designed by the bank, based on their past experiences. Unfortunately, this might result in the bank missing out on some potential clients, and it also might show that the bank is biased against certain groups and it might affect their relationship with existing clients because they feel they are not being catered to properly. In order to avoid these kinds of issues from arising, it is necessary to ensure that the ADS is fair and performs similarly across different subgroups such as different age groups and different marital status.
**Kaggle Solution for ADS audit-** https://www.kaggle.com/code/janiobachmann/bank-marketing-campaign-opening-a-term-deposit.([1])

## 2 Input and Output

The dataset used in this study was obtained from Kaggle [2], where it was sourced from UCI machine learning repository[6]. The dataset contains information on the direct marketing campaigns of a Portuguese banking institution, with a focus on the outcomes of the previous marketing campaign and whether or not a deposit was made. Specifically, it consists of 11,162 rows and 17 columns. The dataset does not contain any missing values, but for some variables such as poutcome, pdays, duration, and balance, we observe default values when the data is not available. In these cases, the default value has a specific meaning, such as no contact being made to the customer or the default value, as in the case of duration=0, being true.

1. **age** (numeric)( age less than 35- young, age greater than 55- old and remaining middle-aged)
2. **job**: type of job (categorical)
3. **marital**: marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)

4. **education** (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unk
5. **default**: has credit in default? (categorical: 'no','yes','unknown')
6. **housing**: has housing loan? (categorical: 'no','yes','unknown')
7. **loan**: has personal loan? (categorical: 'no','yes','unknown')
8. **contact**: contact communication type (categorical: 'cellular','telephone')
9. **month**: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
10. `day_of_week`: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
11. **duration**: last contact duration, in seconds (numeric).
12. **campaign**: number of contacts performed during this campaign and for this client (numeric, includes last contact)
13. **pdays**: number of days that passed since last contact(numeric; 999 means client was not previously contacted)
14. **previous**: number of contacts performed before this campaign and for this client (numeric)
15. **poutcome**: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

Table 1: Data types of various attributes

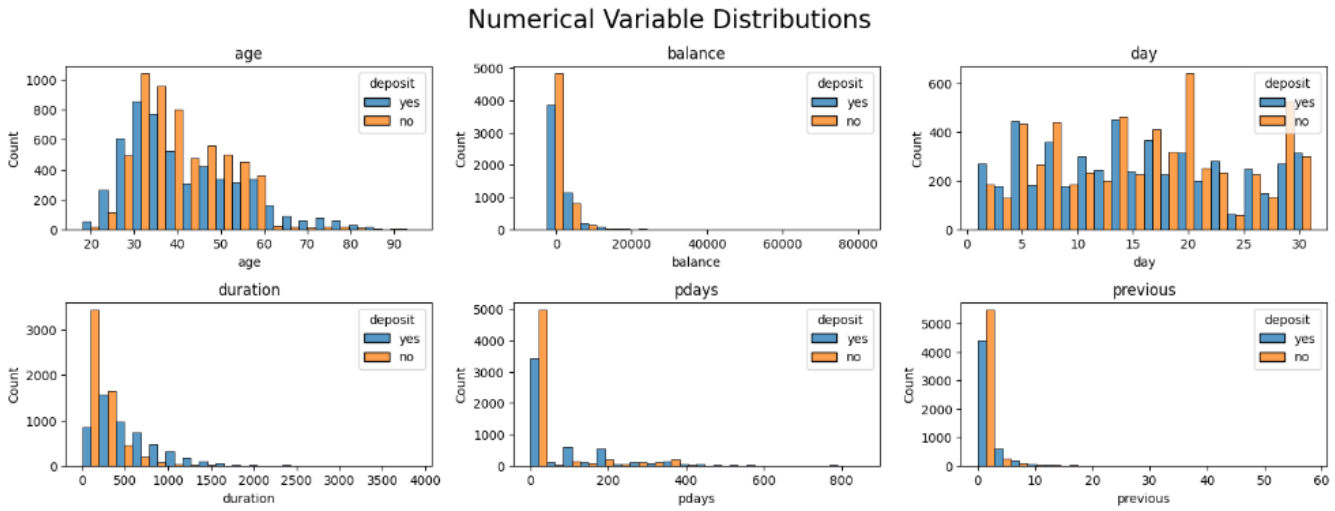
| Variable | Type |
|---|---|
| Age | int64 |
| Job | object |
| Marital Status | object |
| Education | int64 |
| Default | object |
| Housing | object |
| Loan | object |
| Contact | object |
| Month | int64 |
| Day of week | object |
| Duration | int64 |
| Campaign | int64 |
| pdays | int64 |
| previous | int64 |
| poutcome | object |

The **target variable** in the dataset is 'deposit', which is a binary categorical variable that indicates whether or not the client has made a deposit with the bank. It has two possible values: 'yes' and 'no'. The goal of this dataset is to build a model that can predict whether or not a client will make a deposit based on the other variables in the dataset.

**Sensitive features :** Age, Education, Marital-status

Table 2: Descriptive statistics for numerical variables

| Variable | Count | Mean | St. Dev. | Min | Max | 25th Perc. | 50th Perc. | 75th Perc. |
|---|---|---|---|---|---|---|---|---|
| Age | 11162 | 41.231 | 11.913 | 18 | 95 | 32 | 39 | 49 |
| Balance | 11162 | 1528.53 | 3225.41 | -6847 | 81204 | 122 | 550 | 1708 |
| Day | 11162 | 15.65 | 8.42 | 1 | 31 | 8 | 15 | 22 |
| Duration | 11162 | 371.99 | 347.12 | 2 | 3881 | 138 | 255 | 496 |
| Campaign | 11162 | 2.51 | 2.72 | 1 | 63 | 1 | 2 | 3 |
| Pdays | 11162 | 51.33 | 108.76 | -1 | 854 | -1 | -1 | 20.75 |
| Previous | 11162 | 0.83 | 2.29 | 0 | 58 | 0 | 0 | 1 |



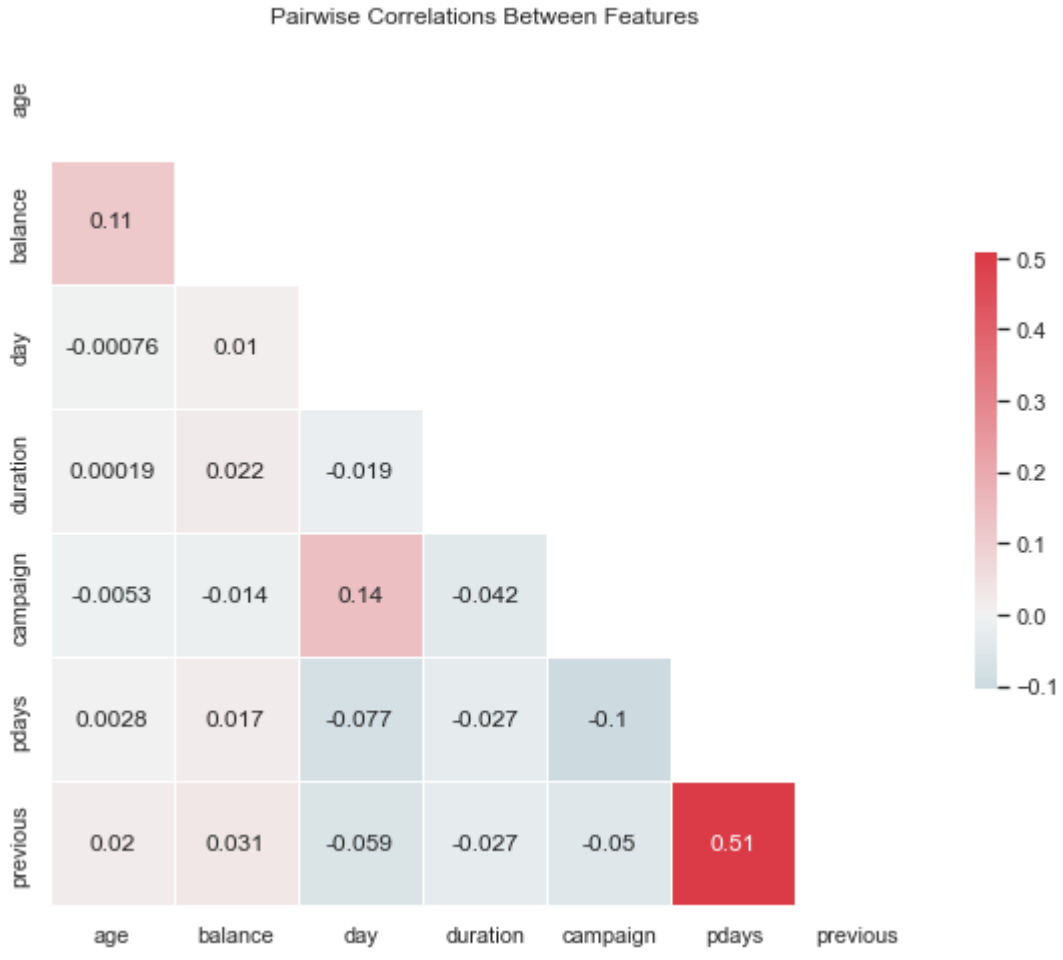




Figure 1: pairwise correlation of numeric variables

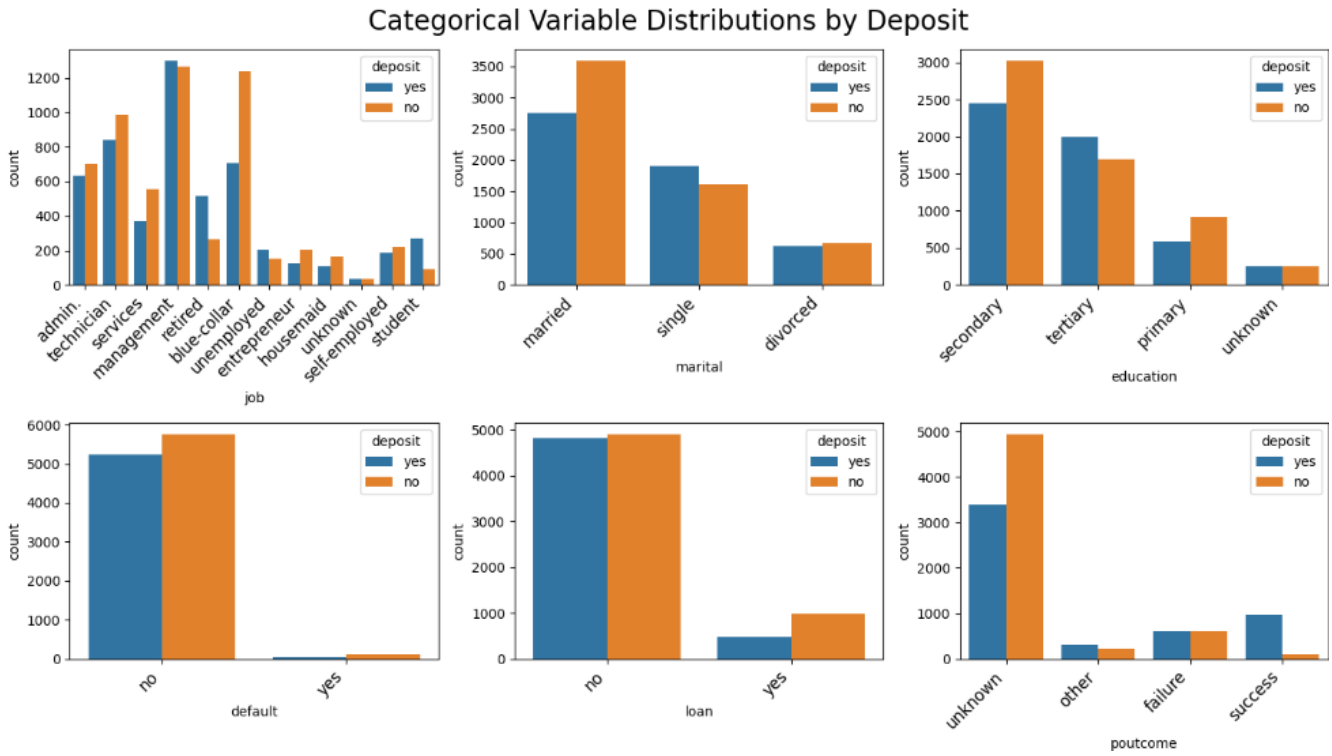



Figure 2: categorical variable distribution by target variable

# 3 Implementation and validation

- Data cleaning and pre-processing- There were no missing values in the dataset so no rows were dropped. Much of the information is in the categorical attributes, which had to first be label encoded. There was no scaling performed because the authore used the Gradient Boost Classifier which is a tree-based ensemble machine learning algorithm.

- The author then performs comprehensive exploratory data analysis to check how the various features are distributed and what features might be correlated. They also checked if combination of two features like education and marital status has any major impact on the outcome.

- Stratified sampling was used to get the training and testing sets. Stratified sampling can ensure that each subgroup is represented in the sample. If a subgroup is underrepresented in the sample, then the statistical inference for that subgroup will be less accurate. This also prevents overfitting. Stratified sampling was done before cross validation after selecting 3 features of importance to distribute the data.

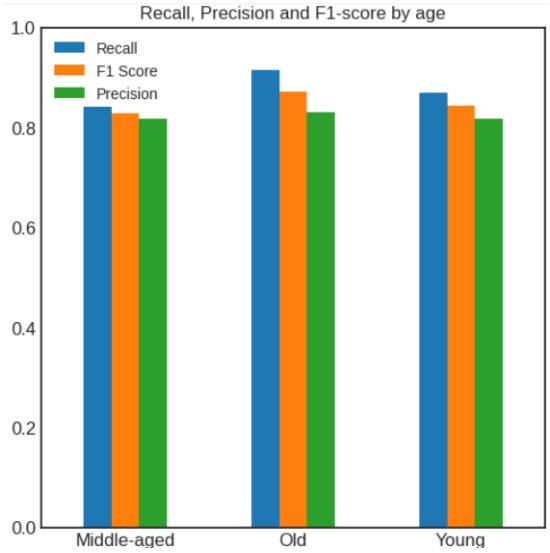
- In this ADS, multiple classifiers were trained and the performance of each was analysed. Based on that, one of the top performing models was used to predict the outcome. While all models had good performance, the owner of the ADS selected Gradient Boosting Classifier and hence, we have analysed the results from that model.

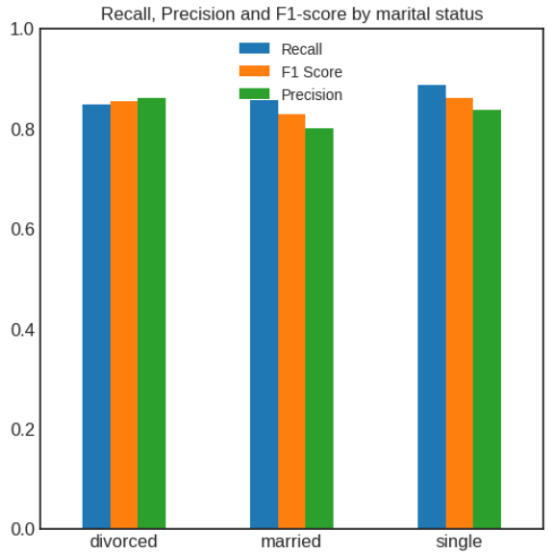
- The ADS used the ROC curve on the training dataset to analyse performance of the Gradient Boost Classifier. It obtained an ROC score of 91.73%. The test accuracy obtained by the model is 84.639%. The ADS was able to predict the outcome of whether a user will open a term deposit or not, with good accuracy.
  **Note :** The user has considered feature duration for model development which is incorrect since [6] UCI repository has stated that is will bias the result because we will have duration data once the customer is contacted. In this model our goal is to identify customers to contact so we will not have that info at hand.

# 4 Outcomes

1. Upon evaluating the model performance on subpopulations of sensitive features, it appears that the model outperforms the baseline accuracy. While there is no significant difference between the model accuracy and baseline accuracy for the subpopulations, the model demonstrates balance in terms of accuracy. Overall, these findings suggest that the model is performing well in accurately predicting outcomes within the subpopulations.

   The evaluation of the model's performance on subpopulations of sensitive features has recall, precision, and F1 scores all within the range of 0.8-0.9 for each subpopulation. These high scores indicate that the model is performing consistently well in accurately identifying and classifying sensitive features across all subpopulations. A high precision score implies that model is making accurate positive predictions, while a high recall score indicates that the model is able to identify most of the positive instances. Overall, the model is beneficial for bank to maximize profits based on accuracy metrics.



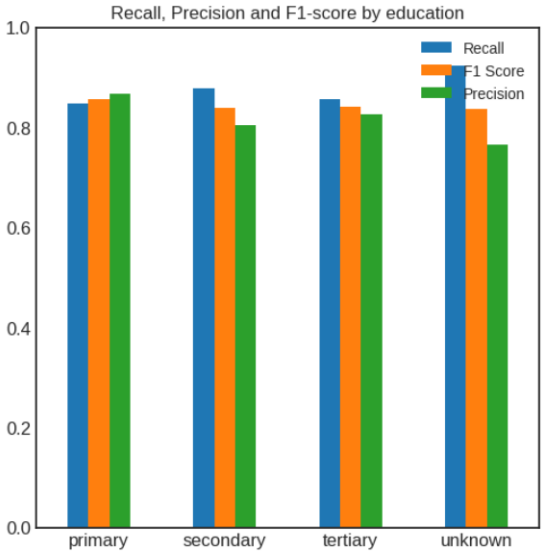



Table 3: Baseline accuracy and model accuracy over different subgroups for the sensitive features

| Sensitive features | Subpopulation | Baseline Accuracy | Model accuracy |
|---|---|---|---|
| Education | Primary | 60.60% | 88.29% |
| | Secondary | 55.26% | 81.69% |
| | Tertiary | 54.11% | 85.70% |
| Age | Young | 50.08% | 83.44% |
| | Middle-aged | 58.66% | 86.14% |
| | Old | 61.05% | 82.41% |
| Marital-status | Single | 54.35% | 84.27% |
| | Marries | 56.62% | 84.62% |
| | Divorced | 51.89% | 85.66% |

2. We used Fairlearn metrics to evaluate the fairness of the classifier [3].

Table 4: Fairness metrics over different subgroups for the sensitive features

| Sensitive feature | Subpopulation | Selection Rate | FNR | FPR |
|---|---|---|---|---|
| Education | Primary | 0.4047 | 0.1532 | 0.0914 |
| | Secondary | 0.5956 | 0.1225 | 0.2560 |
| | Tertiary | 0.4584 | 0.1434 | 0.1426 |
| Age | Young | 0.5423 | 0.1301 | 0.2025 |
| | Middle-aged | 0.4104 | 0.1599 | 0.1244 |
| | Old | 0.7134 | 0.0854 | 0.3426 |
| Marital Status | Single | 0.5809 | 0.1133 | 0.2107 |
| | Married | 0.4614 | 0.1436 | 0.1615 |
| | Divorced | 0.4906 | 0.1515 | 0.1353 |

(a) **Fairness evaluation for education -** [3]

**Selection rate difference** is 0.1909 which suggests that selection rates across the groups(Primary, secondary and tertiary) do not vary significantly and is not leading to bias against customer on the basis of their education. Higher selection rate is observed for people with secondary education while customers with primary education have lower selection rate.

**Equalized odds ratio** is 0.3571 which suggests that false positive rates and false negative rates are not balanced across the subgroups based on education. Customers with secondary education have highest false positive rate while also having lowest false negative rate.

After evaluating the fairness metrics for education feature, it is evident that the subpopulation with primary education is at a disadvantage due to the low positive outcome of the model. This implies that the model is less likely to predict positive outcomes for customers with primary education, causing them to miss out on the benefits of the marketing campaign. This can potentially lead to a loss in revenue for the bank, as deserving customers are not targeted by the campaign.

Additionally, the subpopulation with secondary education has a high false positive rate, indicating that the model is more likely to falsely predict positive outcomes for this group. As a result, the bank may end up spending more on the campaign for this group, leading to a higher cost without a proportional increase in revenue.

In summary, the bank's marketing campaign may not be fair to customers with primary education, as they are less likely to be targeted despite being deserving of the campaign benefits. On the other hand, customers with secondary education may be unfairly targeted, leading to higher campaign costs without a proportional increase in revenue.

(b) **Fairness evaluation for age -**[3]

**Selection rate difference** is 0.3029 which suggests that selection rates across the groups(young, middle-aged and old) varies significantly and is leading to bias against customer on the basis of their age. Higher selection rate is observed for older age group while middle-aged customers have lower selection rate.

**Equalized odds ratio** is 0.3632 which suggests that false positive rates and false negative rates are not balanced across the subgroups based on education. Older customers have significantly high false positive rate while they also having lowest false negative rate.

Upon evaluating the fairness metrics for the age feature, we found that the middle-aged population is at a disadvantage due to the model's low positive outcome. This means that the marketing campaign is less likely to target middle-aged customers who deserve to benefit from it, potentially leading to a loss in revenue for the bank.

Furthermore, the older population has a significantly high false positive rate, which suggests that the model is more likely to wrongly predict positive outcomes for this group. As a result, the bank may end up spending more on the campaign for this group, without a proportional increase in revenue.

To sum up, the marketing campaign may not be fair to middle-aged customers, as they are less likely to be targeted despite being deserving of the campaign benefits. Meanwhile, the campaign may be unfairly targeting older customers, resulting in higher campaign costs without a proportional increase in revenue.

(c) **Fairness evaluation for marital status -**[3]

**Selection rate difference** is 0.1196 which suggests that selection rates across the groups(single, married and divorced) does not vary significantly and is not leading to bias against customer on the basis of their marital status. Slightly higher selection rate is observed for single population compared to other groups.

**Equalized odds ratio** is 0.6423 which suggests that false positive rates and false negative rates are not balanced across the subgroups based on marital status. Single customers have significantly high false positive rate while also having lowest false negative rate.

After examining the fairness metrics for the marital-status feature, it is evident that the single subpopulation is receiving an advantage due to the model's high positive outcome. This indicates that non-single customers who could also benefit from the marketing campaign may be overlooked, leading to a potential loss in revenue for the bank.

Furthermore, the single population has a significantly high false positive rate, suggesting that the model is more likely to falsely predict positive outcomes for this group. Consequently, the bank may end up spending more on the campaign for the single population, without a proportional increase in revenue.

In summary, the marketing campaign may be biased towards single customers, as they are more likely to be targeted despite high false positive rates. At the same time, non-single customers may be unfairly excluded from the campaign, resulting in higher campaign costs without a proportional increase in revenue.

3. We used SHAP[4][5] for model interpretability to gain insight into features that are driving model outcome. The summary_plot was used to summarize the most important features where each dot in the plot is an instance and the colour gives the magnitude of the values of that word. Then SHAP values (on x-axis) give us the contribution of each feature on that prediction. The colour is indicating the level of impact and the value on x-axis indicates positive or negative impact. In Figure3, we see that duration, pdays have high positive impact, and the contact feature has high negative impact on the outcome.

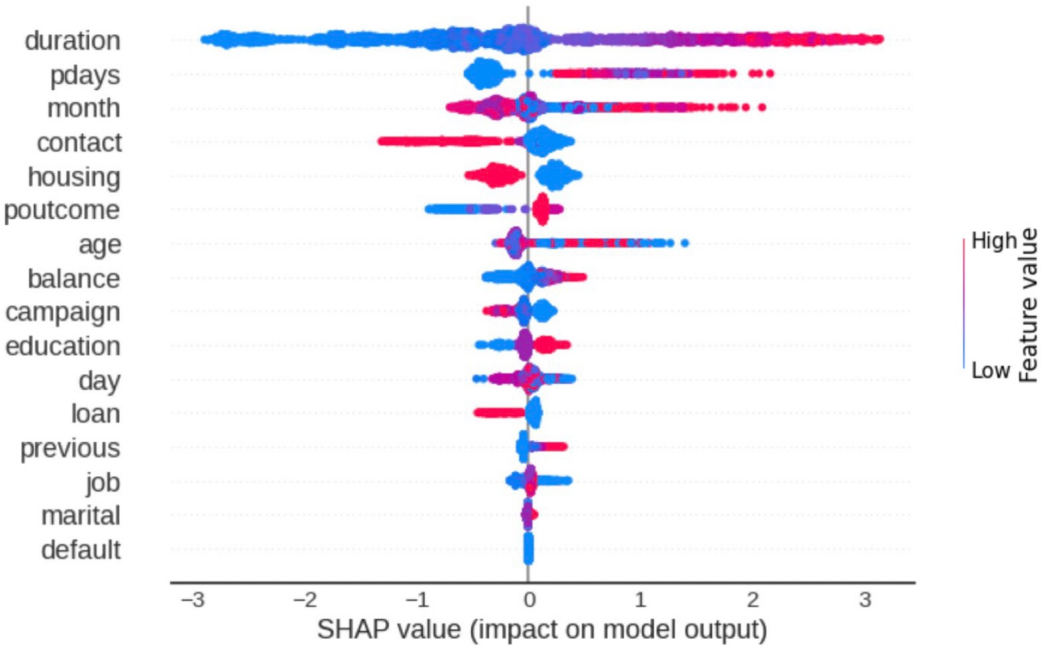



Figure 3: SHAP summary plot indicating the contribution of each feature to the outcome

Figure 4 shows the SHAP explanation for a correctly classified instance. We can see that duration has the highest impact for the negative outcome. Other factors like pdays, month, age also affect the outcome and contribute to the negative prediction. Attributes like balance and housing are contributing to the positive outcome a little, but ultimately that final SHAP value is -2.33 which means it is strongly predicting 'No' for this instance.
Figure 5 shows the SHAP explanation for an incorrectly classified instance. This is an example of a false positive. As seen from our analysis, the old age group(age greater than 55) has a high selection rate and a high FPR. As seen in the figure, age is contributing to the positive outcome strongly. Duration has an impact too but age has a stronger impact and therefore, the classifier is predicting 'Yes' for this outcome which is the incorrect prediction. This shows that the classifier is biased towards older age groups.

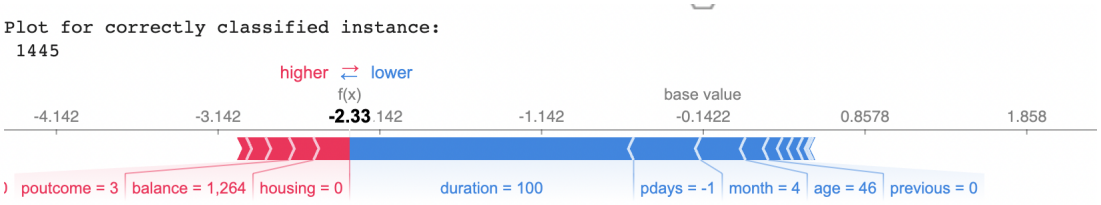



Figure 4: SHAP explanation for a correctly classified instance

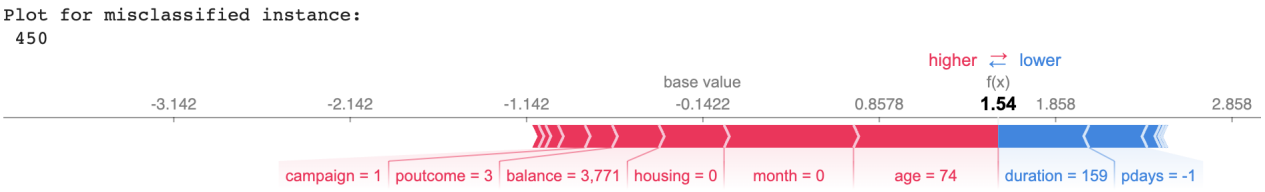



Figure 5: SHAP explanation for a misclassified instance

# 5 Summary

- The goal of the ADS is to predict whether a customer who has been targeted by the bank's marketing campaign will open a term deposit or not. The Bank Marketing data set that the ADS is using is a widely used data set for this task or similar tasks, and has been used in many studies. The data set contains information about the bank's marketing campaigns, which includes various attributes relevant to the such as age, job, education, and housing. The dataset is reasonably large (11162 rows) and has a mix of numerical and categorical features which are very informative. The data set also has no missing values.

- Based on our analysis, we believe that the ADS is not very robust as it is highly dependent on the 'duration' feature. This feature is highly correlated with the target output. However, the duration of a call is not known before the customer is contacted and so this feature should not be included if the aim is to create a realistic predictive model. As seen in our SHAP analysis, it appears that duration has a strong impact on the prediction.
  To test the robustness of this classifier, we implemented the classifier and used it on the dataset with 'duration' feature equated to zero. The classifier achieved an accuracy of 54.32% on the test set. This is much lower than the accuracy of the model when trained on data with the 'duration' feature. Hence, we can conclude that the ADS is not very robust.
  Furthermore, the model has some bias when it comes to the sensitive features. By analysing the fairness metrics like FPR, FNR differences among subgroups and checking the selection rates and equalised odds ratio, we can determine if there is any bias in the classifier and whether the bank might be spending resources on the wrong targets because of it.

  - Education attribute- The model often predicts Yes for customers with secondary education even though many times they don't open a term deposit. This can be problematic for the bank as they are spending time, money and effort of the campaign on individuals who might not open a term deposit. The individuals who have primary or tertiary education tend to then miss out on the the bank's services. This is also problematic for the bank because the dissatisfaction among customers can cause them to lose clients.
  - Age attribute- The model often predicts Yes for customers who are either in the older category or younger category even though many of them don't open a term deposit with the bank. It has a very strong bias towards the older age group which is shown by the high FPR and selection rate. This can be problematic for the bank as they are using resources targeting the wrong people. Middle-aged individuals tend to miss out on the advertisements from the bank because they are often classified as 'not likely to open a term deposit'. This can cause dissatisfaction with the services among the middle-aged individuals as they will not be targeted and hence, get subpar service. This is problematic for the bank also because they might lose out on potential clients among the middle-aged individuals.
  - Marital status attribute- The model often tends to predict Yes for customers who are single even though many of them don't open a term deposit with the bank. There is a significant difference in the FPR and FNR for this subgroup so the equalised odds ratio indicates that the error rates are not balanced. This can be problematic for the bank as they are using resources targeting the wrong people. The classifier often predicts No for the divorced individuals which means that they are not being approached even though they could possibly open a term deposit with the bank which leads to dissatisfaction with the bank's customer service.

Based on the above analysis it can be seen that the classifier is unfair towards certain subgroups particularly the single, the old and the secondary educated populations. If the bank decided to use predictions of this model to design their campaign and target customers, it could lead to some dissatisfaction among customers and the bank will end up wasting resources and revenue on the the wrong targets.

- We would recommend that this ADS not be deployed in the industry as it is not very robust. The high dependence on duration indicates that the predictions are not very reliable as duration would would not be available to use, before the campaign has begun and clients have been contacted. Furthermore, the differences in FPR, FNR and selection rates indicate that the model is particularly biased towards certain demographics and using the predictions from this model can cause the bank to waste resources, time and effort on incorrect targets. This can also cause dissatisfaction among people who feel that the bank is not approaching them with relevant services and products.

# References


[1] Bank marketing campaign: Opening a term deposit. `https://www.kaggle.com/code/janiobachmann/bank-marketing-campaign-opening-a-term-deposit`.

[2] Bank marketing dataset. `https://www.kaggle.com/datasets/janiobachmann/bank-marketing-dataset`.

[3] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in AI. Technical Report MSR-TR-2020-32, Microsoft, May 2020.

[4] Julia Stoyanovich. Ds-ga 1017 responsible data science, 2023.

[5] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017)*, pages 4765–4774, 2017.

[6] S Moro, P Cortez, and P Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, June 2014.


# A Appendix

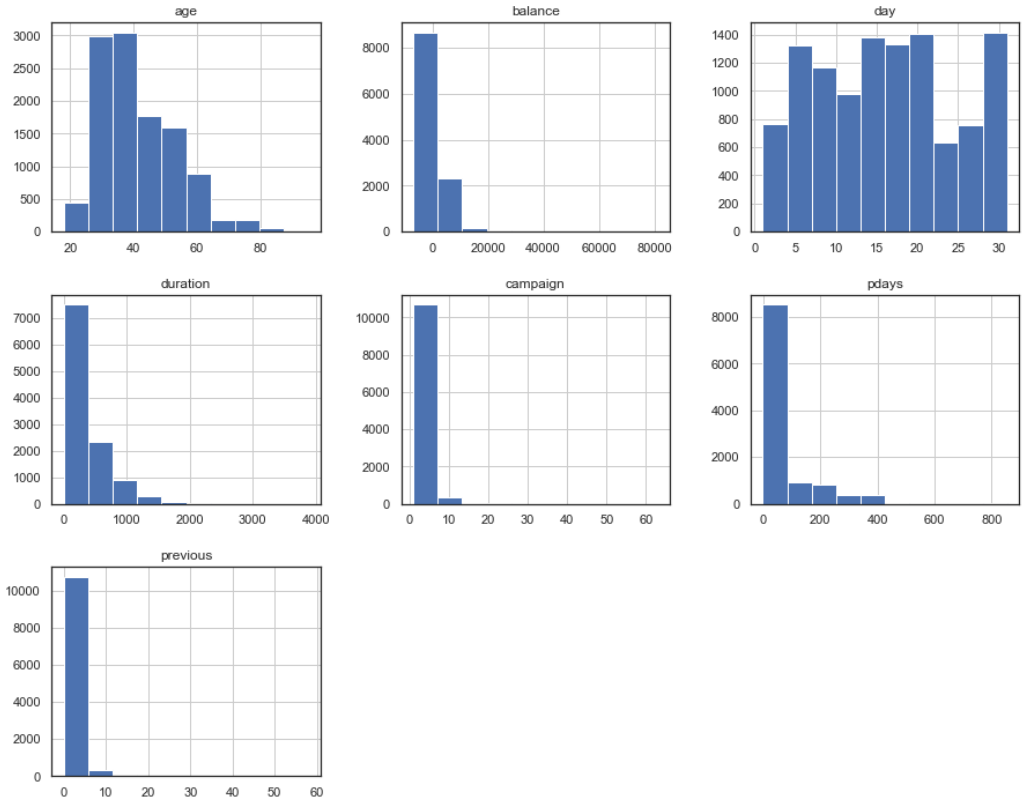



Figure 6: Distribution for numerical variables

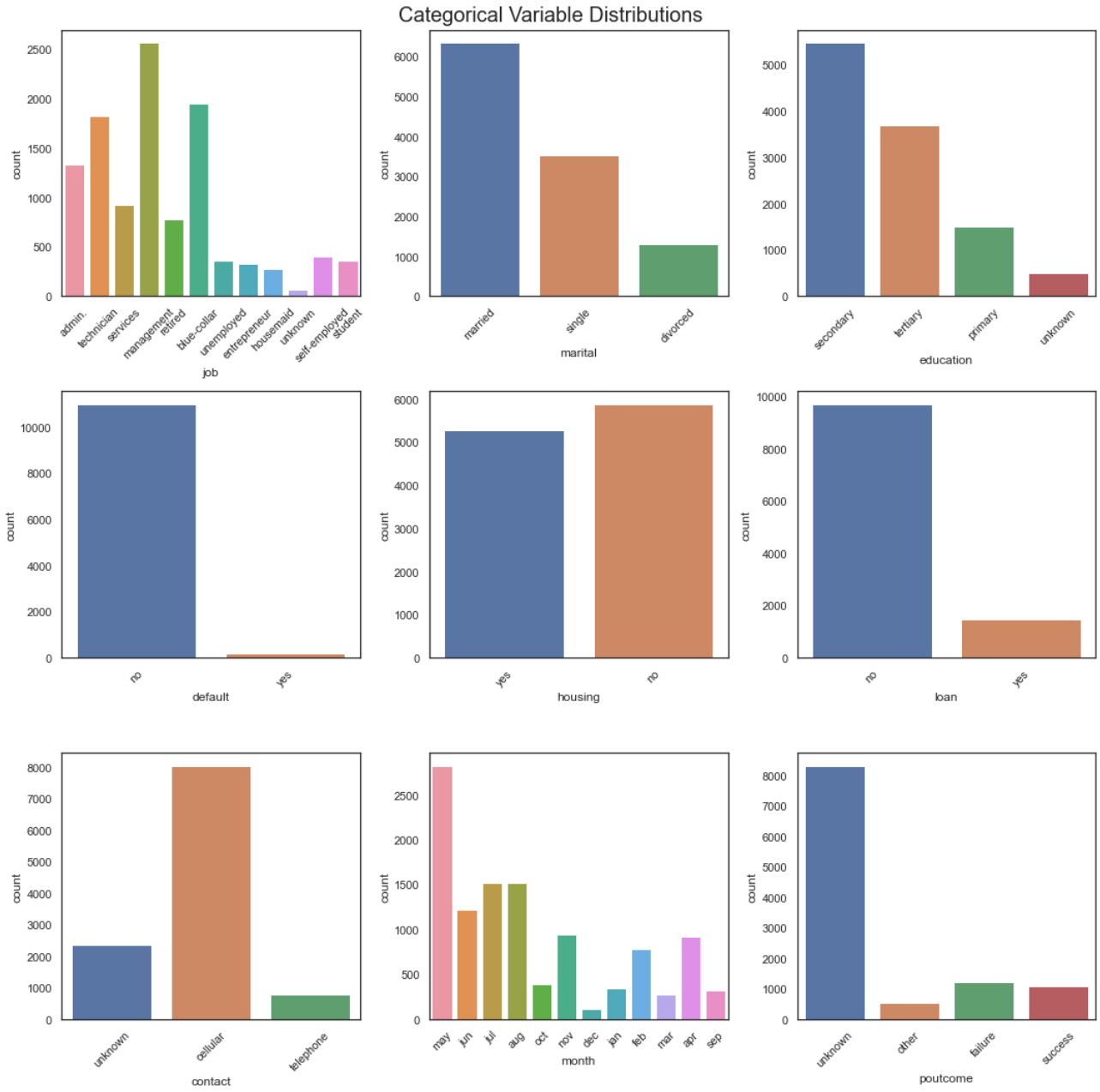



Figure 7: pairwise correlation of numeric variables

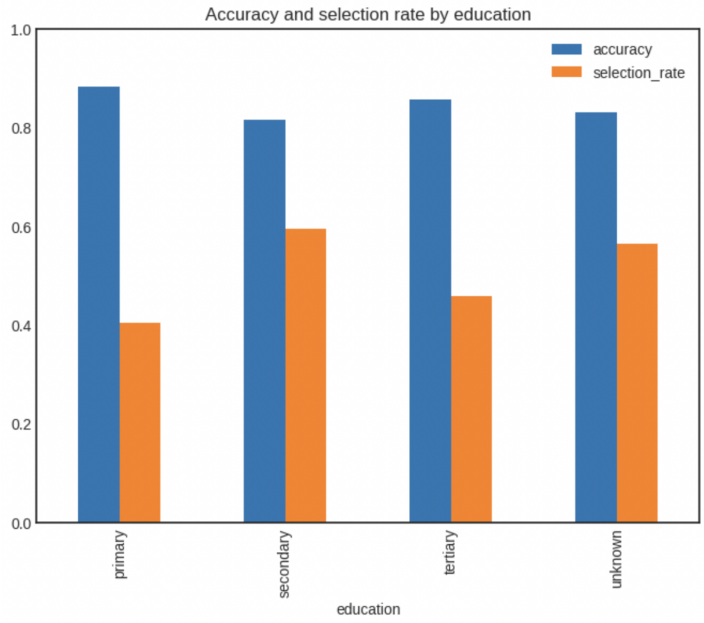



Figure 8: Accuracy and selection rate across subgroups in the 'education' attribute

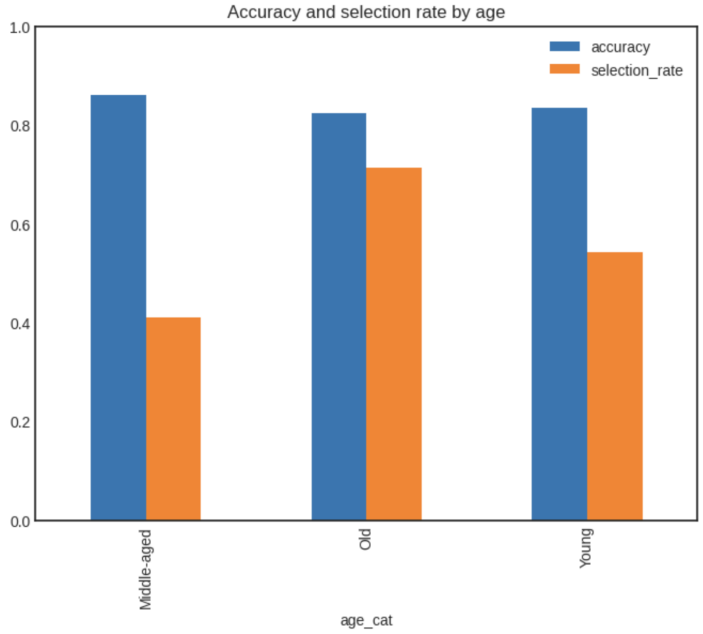



Figure 9: Accuracy and selection rate across subgroups in the 'age' attribute

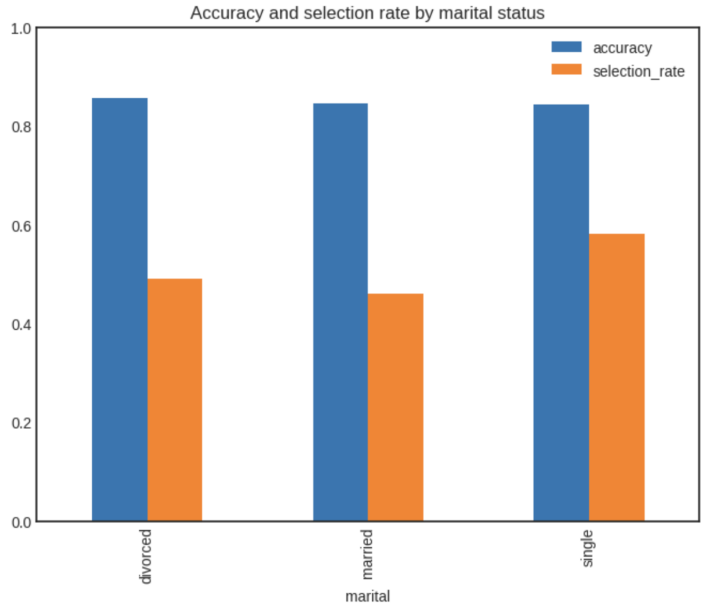



Figure 10: Accuracy and selection rate across subgroups in the 'marital' attribute

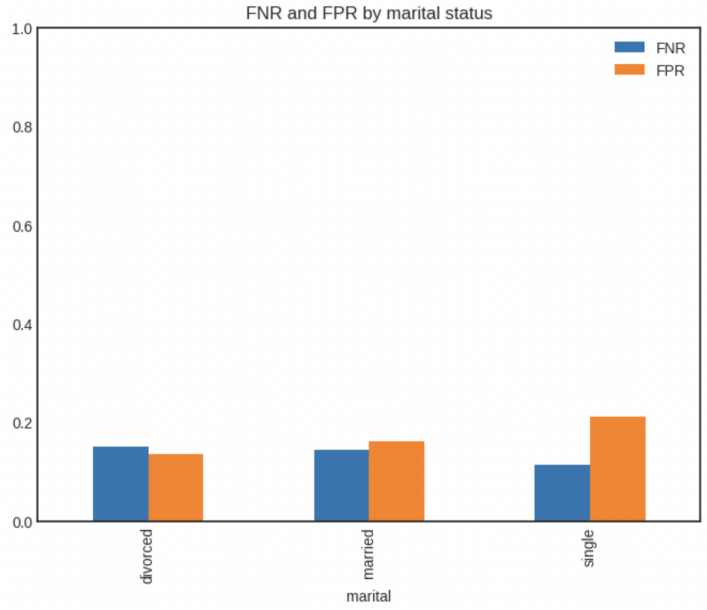



Figure 11: FPR and FNR across subgroups in the 'marital' attribute