Wednesday, April 19, 2023

CS-354N, IIT Indore

SUPERVISOR

Dr. Aruni Tiwari

STUDENTS

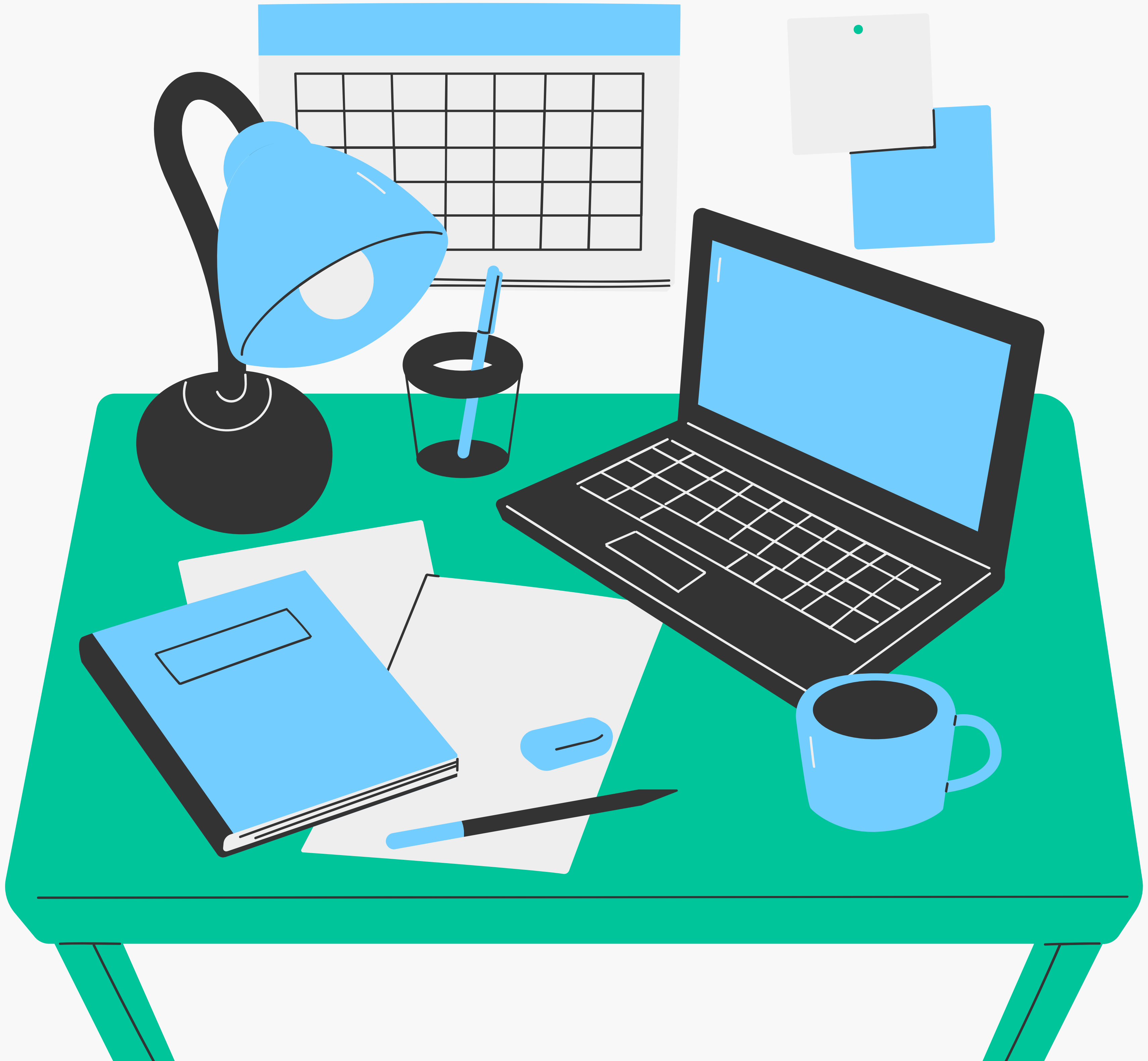
Mukul Jain(200001050)
Nilay Ganvit(200001053)

Term Deposit Prediction

- This problem involves building a binary classification model to predict the probability of a client subscribing to a term deposit.

- The goal is to minimise the calls by eliminating those who will not subscribe to the Term Deposit.

We both want to pursue our Master's in Business Administration and chose a topic that would benefit our Resume.



Finding Data

For the following problem, only a single type of Dataset is present all over the internet with a different number of data points.

A balanced one with undersampled data and an imbalanced one.

We chose the imbalanced data set with more data.

Choosing one

The dataset used to predict Term Deposit subscriptions is related to direct telemarketing campaigns of a Portugese Banking Institution, collected from 2008 to 2013. The Dataset consists of 16 predictors (9 categorical and 7 numeric) and 1 output.

III Preprocessing

Preprocessing involves the following:

- Data Cleaning: The dataset does not contain missing values or invalid entries.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 45211 entries, 0 to 45210
Data columns (total 17 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   age         45211 non-null  int64 
 1   job         45211 non-null  object
 2   marital     45211 non-null  object
 3   education   45211 non-null  object
 4   default     45211 non-null  object
 5   balance     45211 non-null  int64 
 6   housing     45211 non-null  object
 7   loan        45211 non-null  object
 8   contact     45211 non-null  object
 9   day         45211 non-null  int64 
 10  month       45211 non-null  object
 11  duration    45211 non-null  int64 
 12  campaign    45211 non-null  int64 
 13  pdays       45211 non-null  int64 
 14  previous    45211 non-null  int64 
 15  poutcome    45211 non-null  object
 16  subscribed  45211 non-null  object
dtypes: int64(7), object(10)
memory usage: 6.2+ MB
```

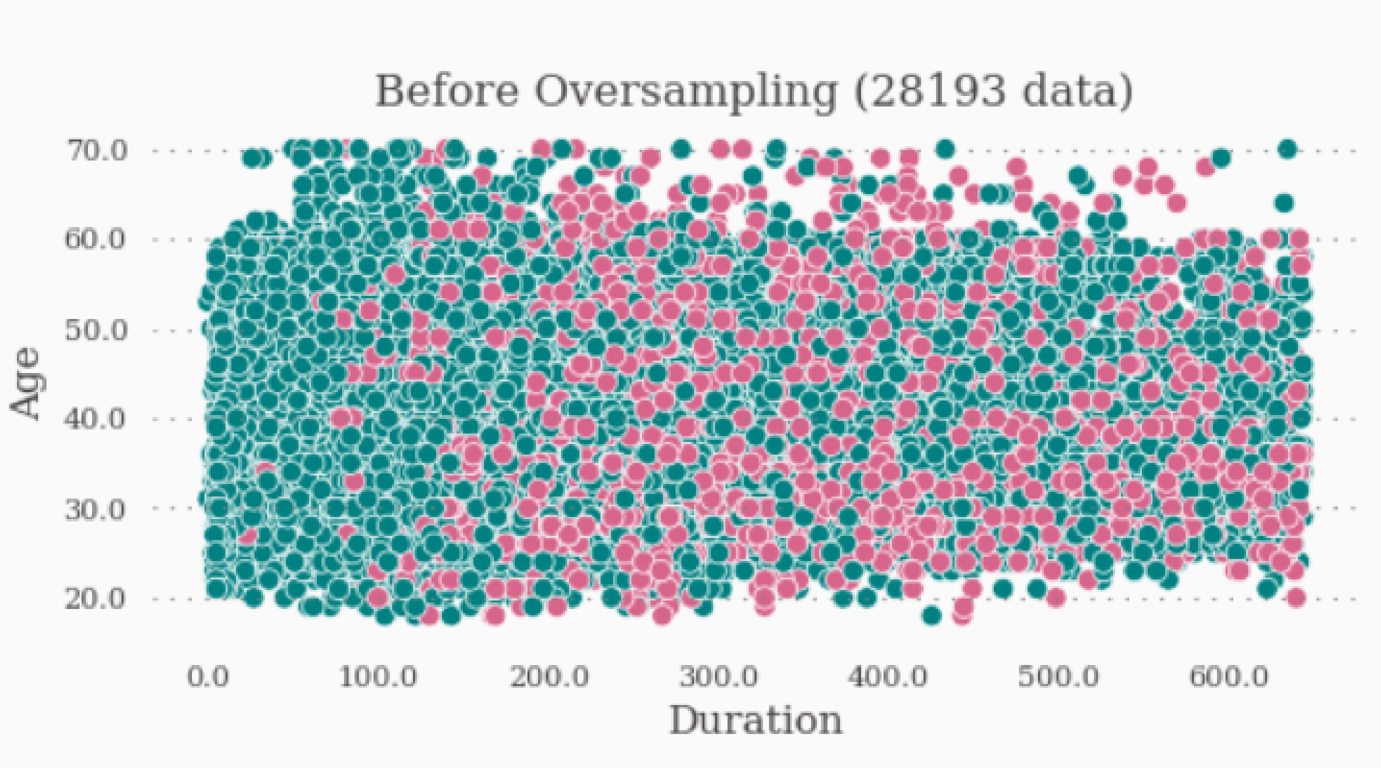
- Data Transformation: Categorical data were transformed into numerical data using One Hot Encoding
- Normalisation: The data was scaled down to mean zero and unit variance.

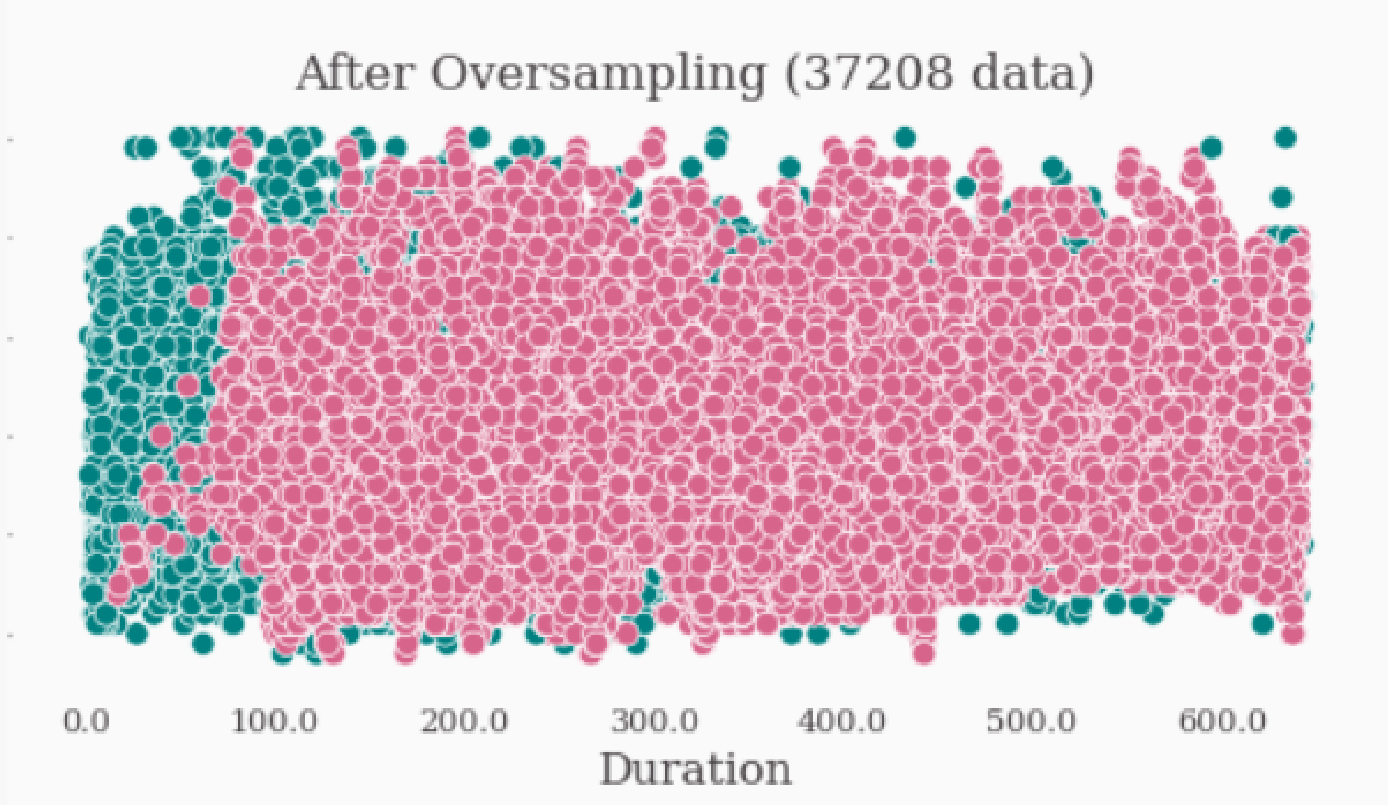
```
df2 = encoding.fetch()
df2.head()
```

	age	job	marital	education	default	balance	housing	loan	contact	duration	campaign	subscribed
0	58	4	1	2	0	2143	1	0	0	261	1	0
1	44	9	2	1	0	29	1	0	0	151	1	0
2	33	2	1	1	0	2	1	1	0	76	1	0
3	47	1	1	1	0	1506	1	0	0	92	1	0
4	33	1	2	1	0	1	0	0	0	198	1	0

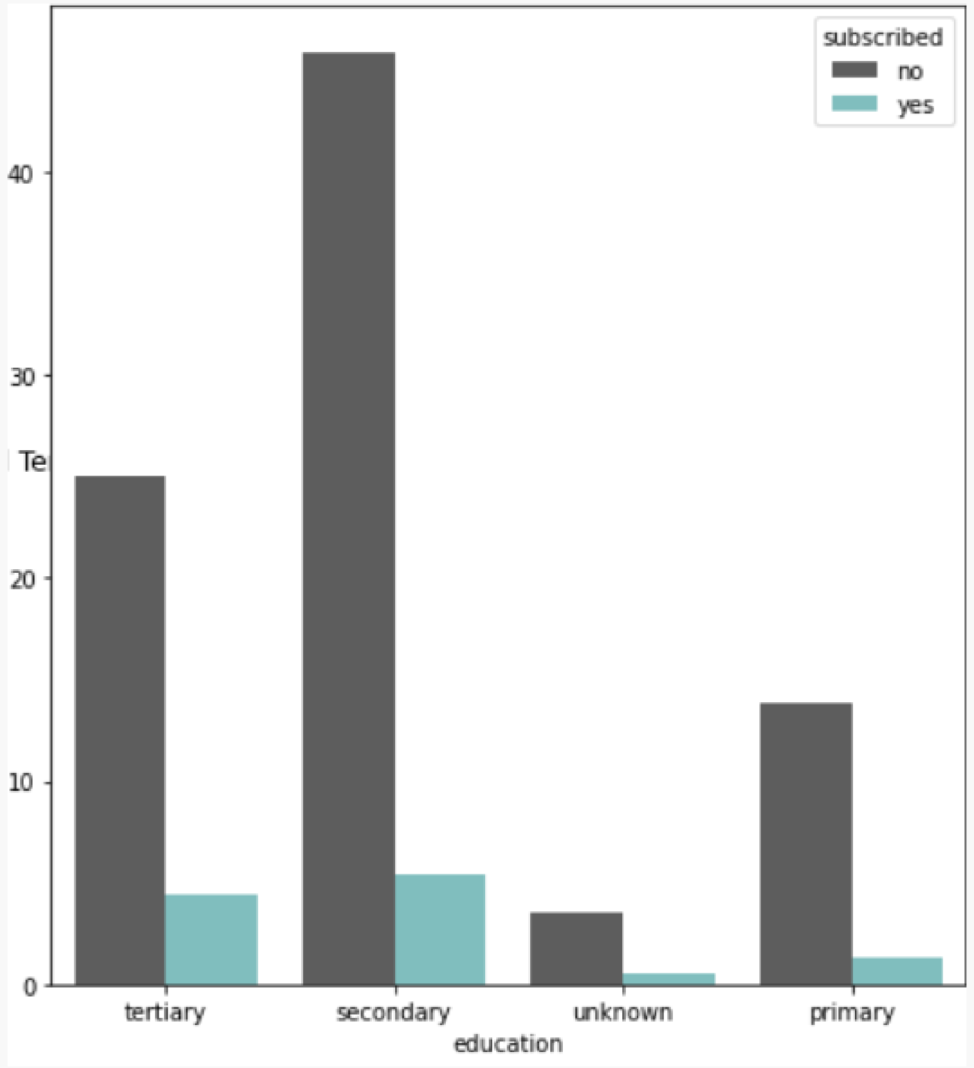
Preprocessing involves the following:

- Data Imbalance: Since the dataset was imbalanced, we applied Oversampling using SMOTE
- Splitting: Data was split using Stratified Sampling into training validation and test sets.



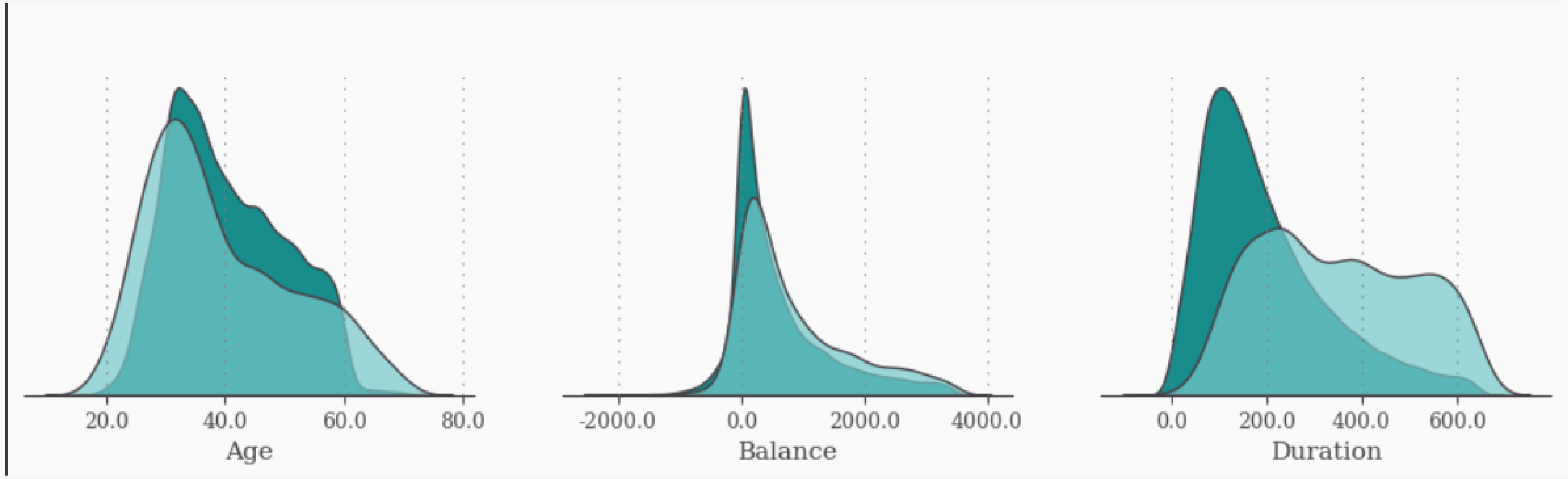


Bar Chart



We looked at the variance of Long Term Subscriptions over the type of education spreading over tertiary, secondary, primary...

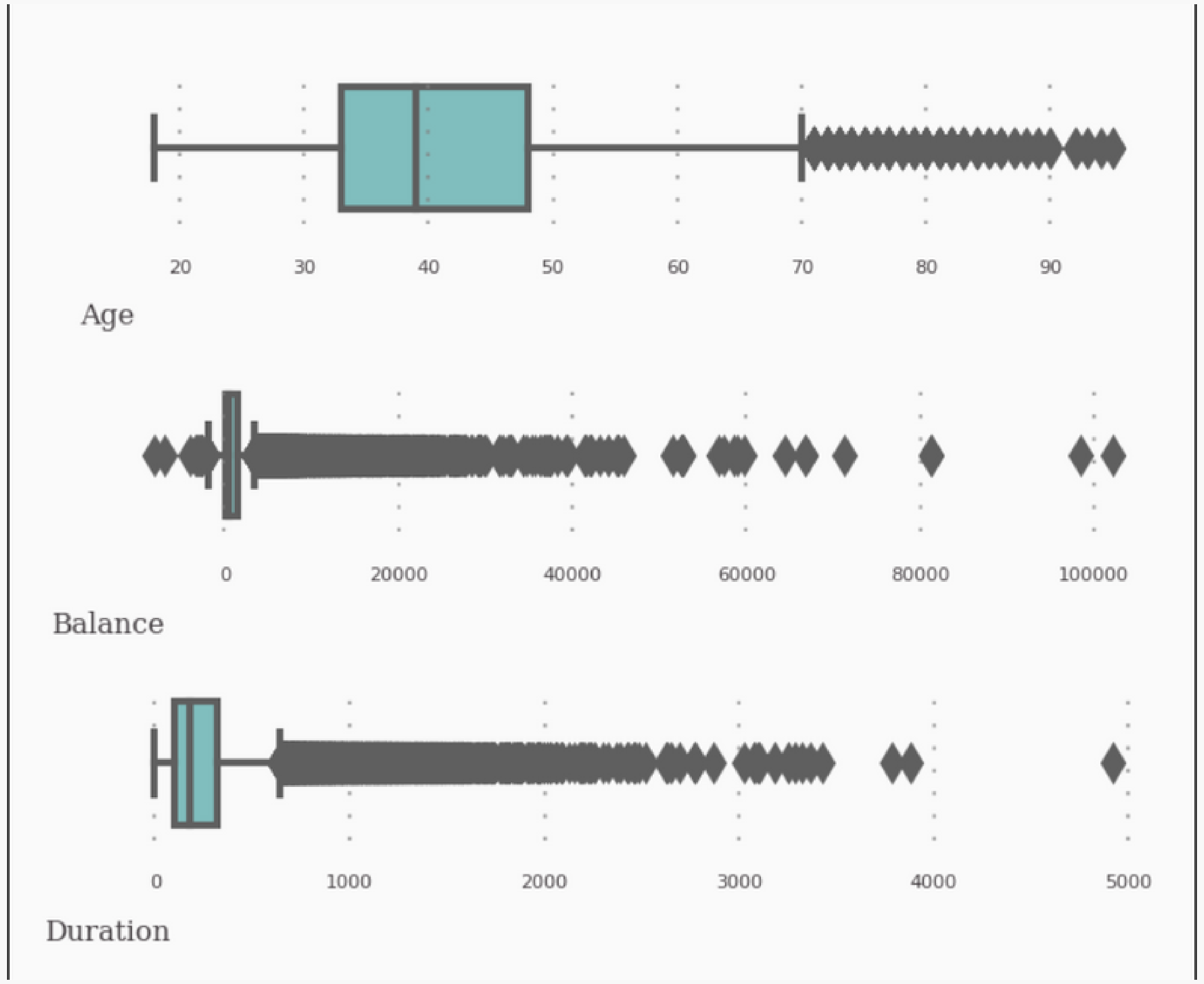
Factors



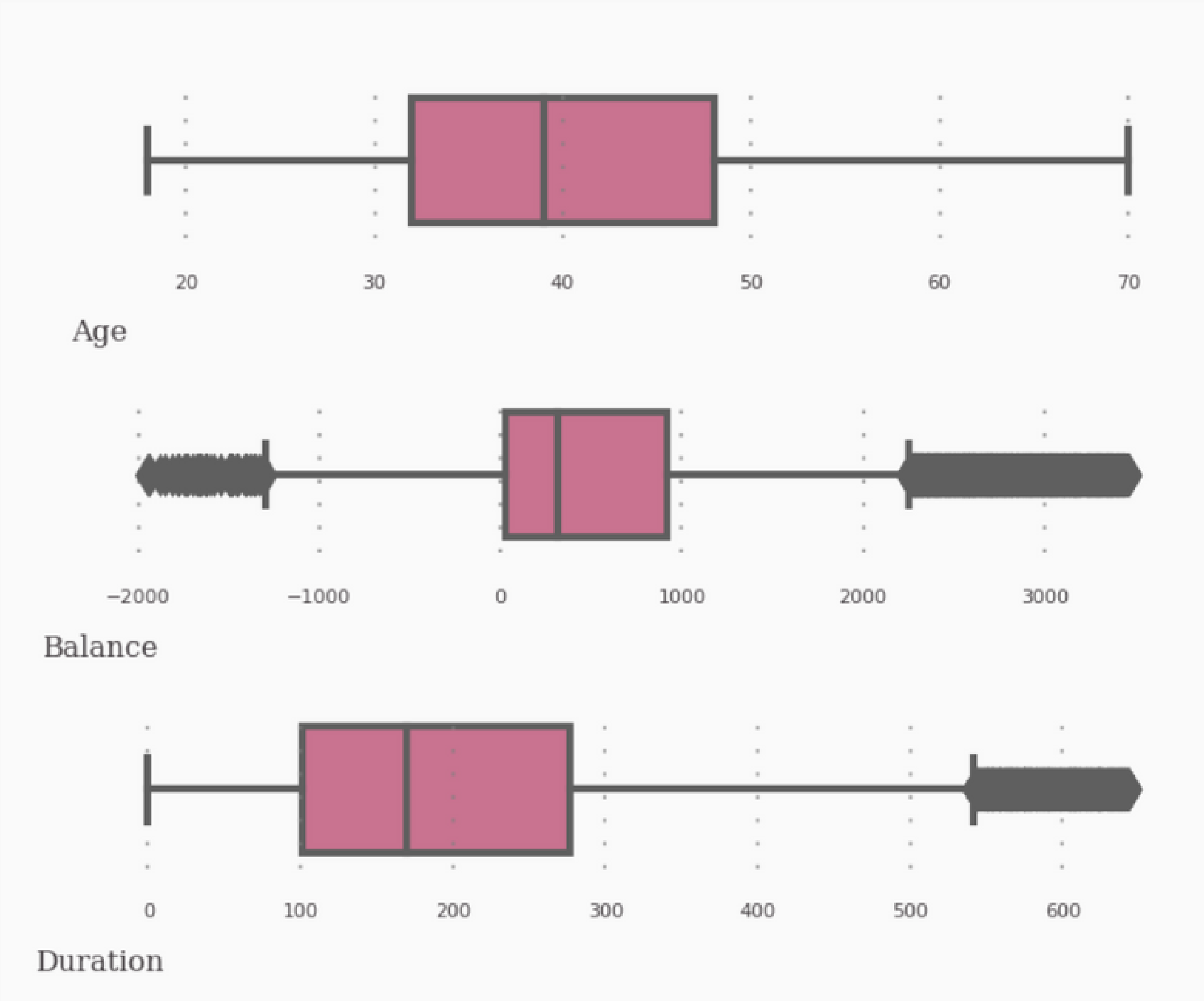
As seen in the graph above the area of Age and Balnce overlap more than Duration, it concludes that Duration can be a differentiating factor when it comes to Long therm deposits

IV Analysis

Outliers



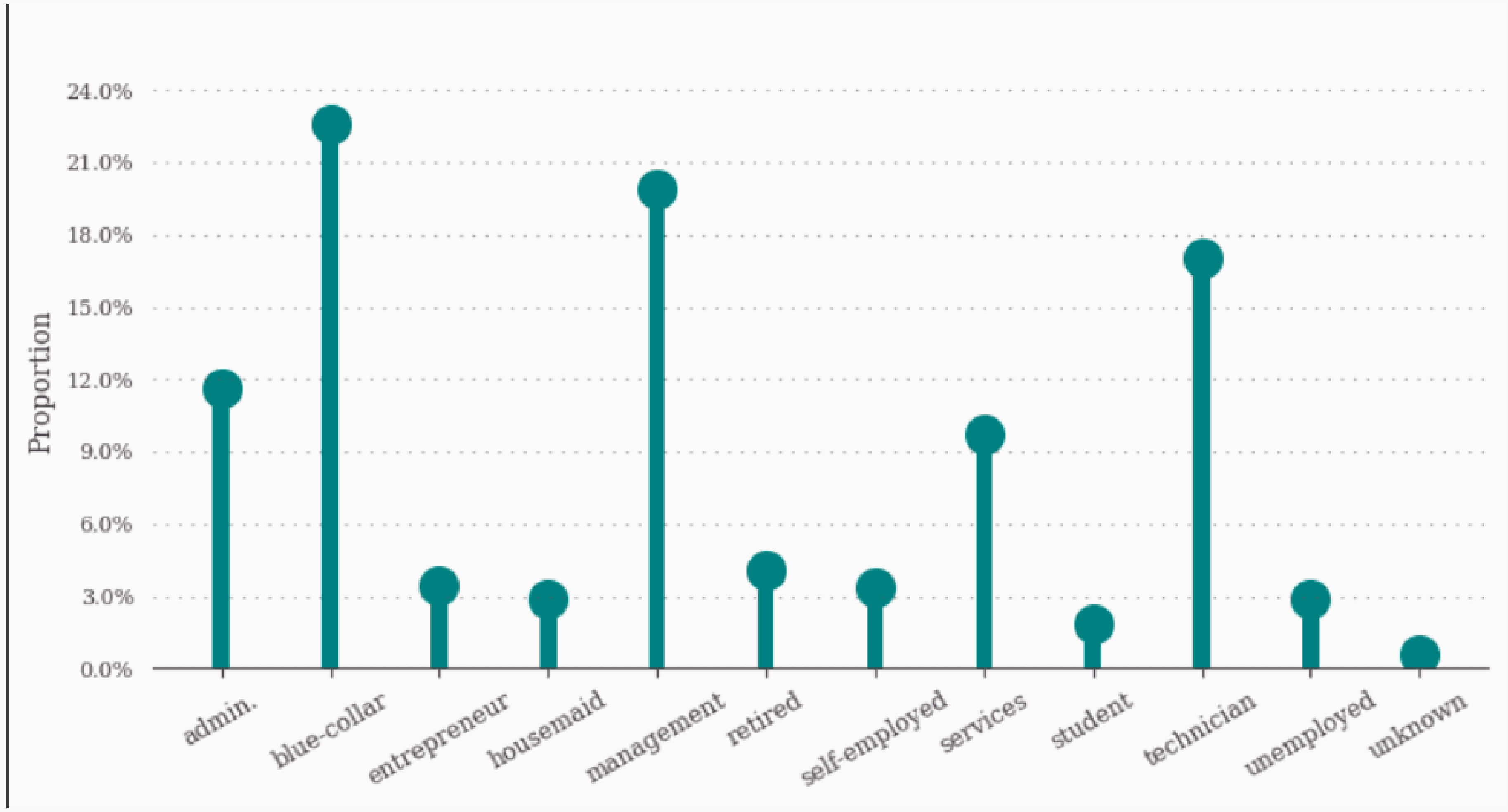
Trimming the Outliers



We found relations of Long term Deposits with Age, Balance and Duration while also being acknowledged of the outliers present in them

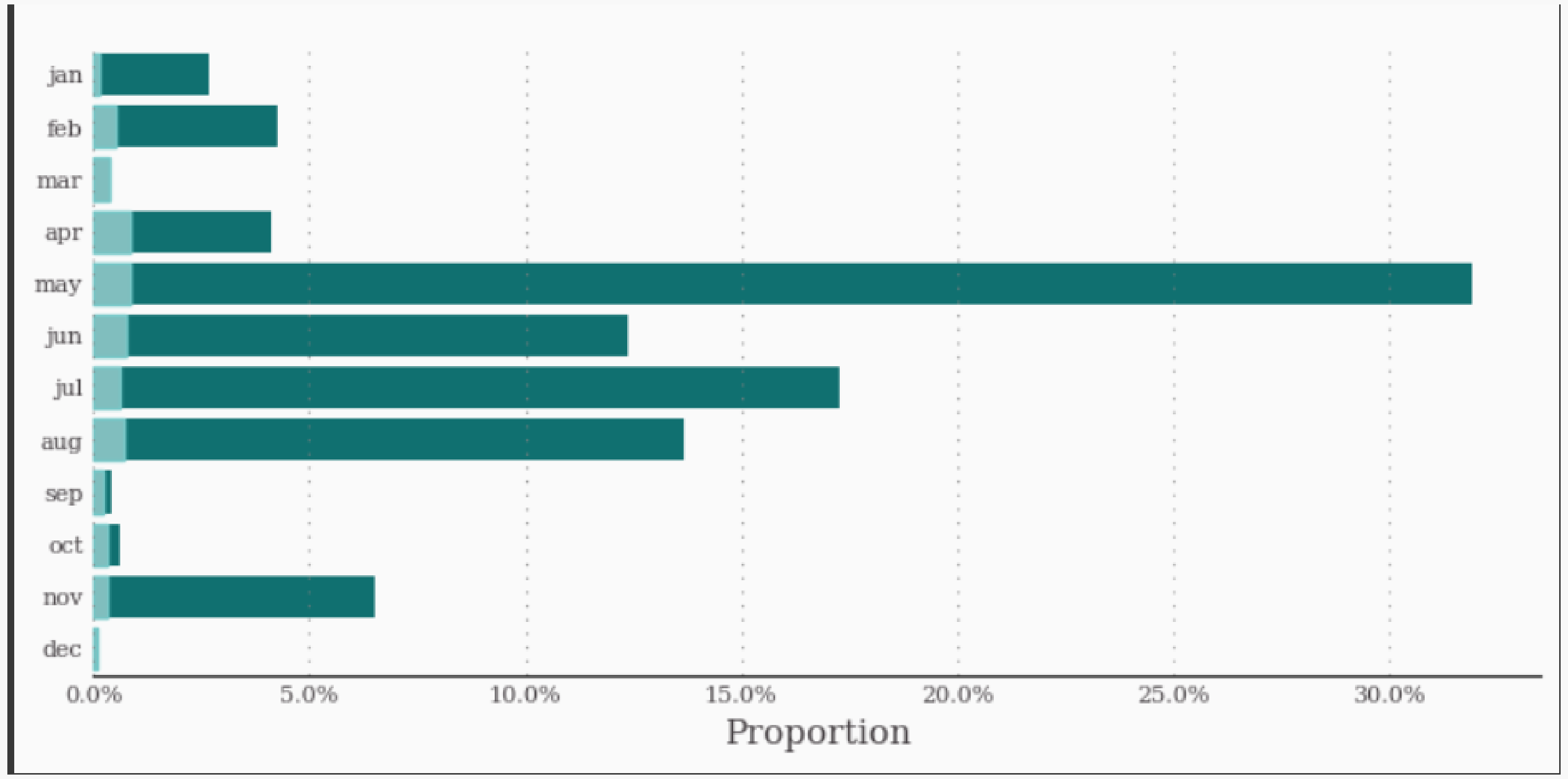
The Outliers were trimmed as a step to clean the data as outliers can lead to worse effects on the study and trimming them refines the scope of study

Type of employment



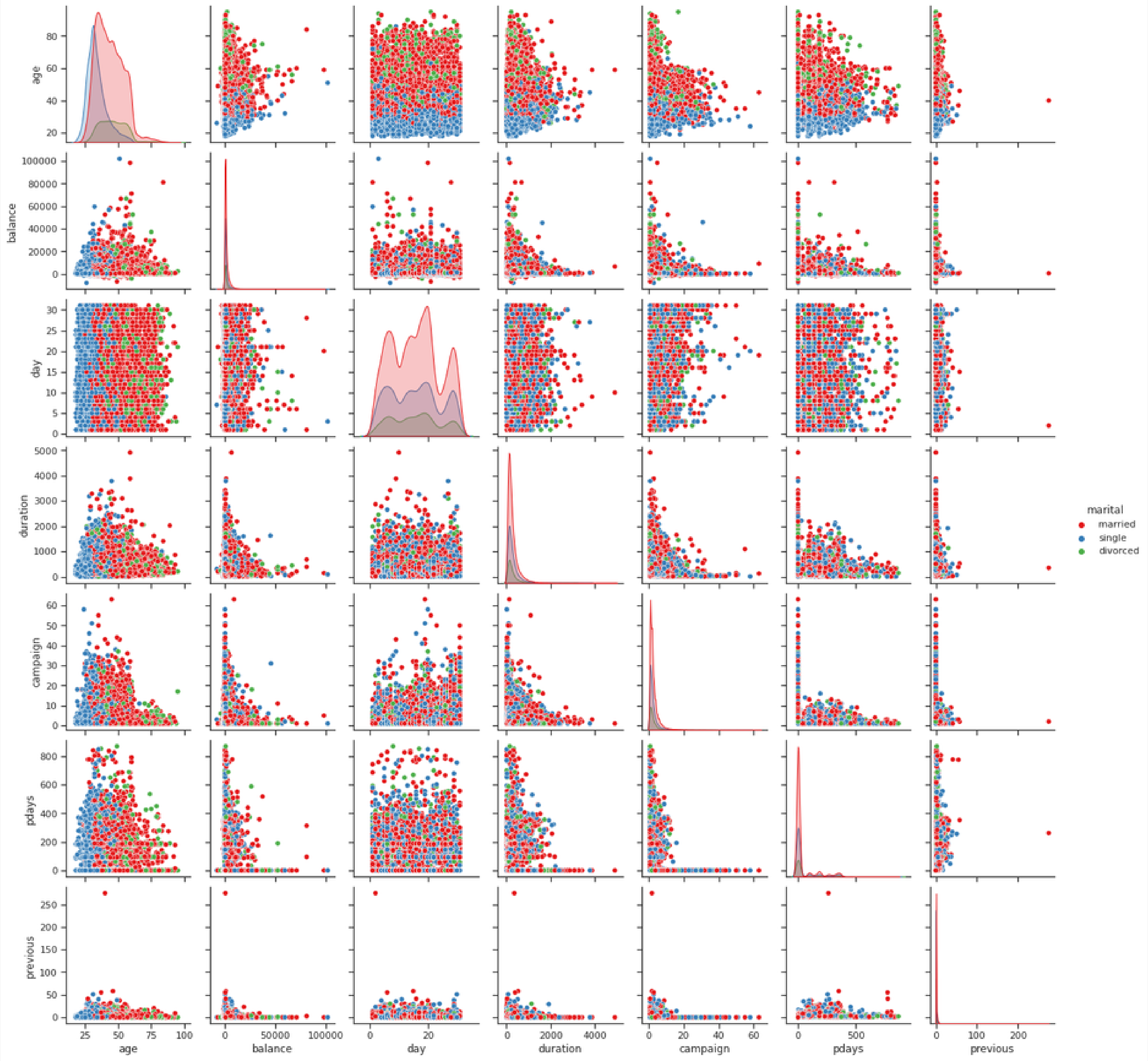
It is apperent from the above chart that middle level type of employment plays a role in factoring if someone will subscribe or not

Variance over months



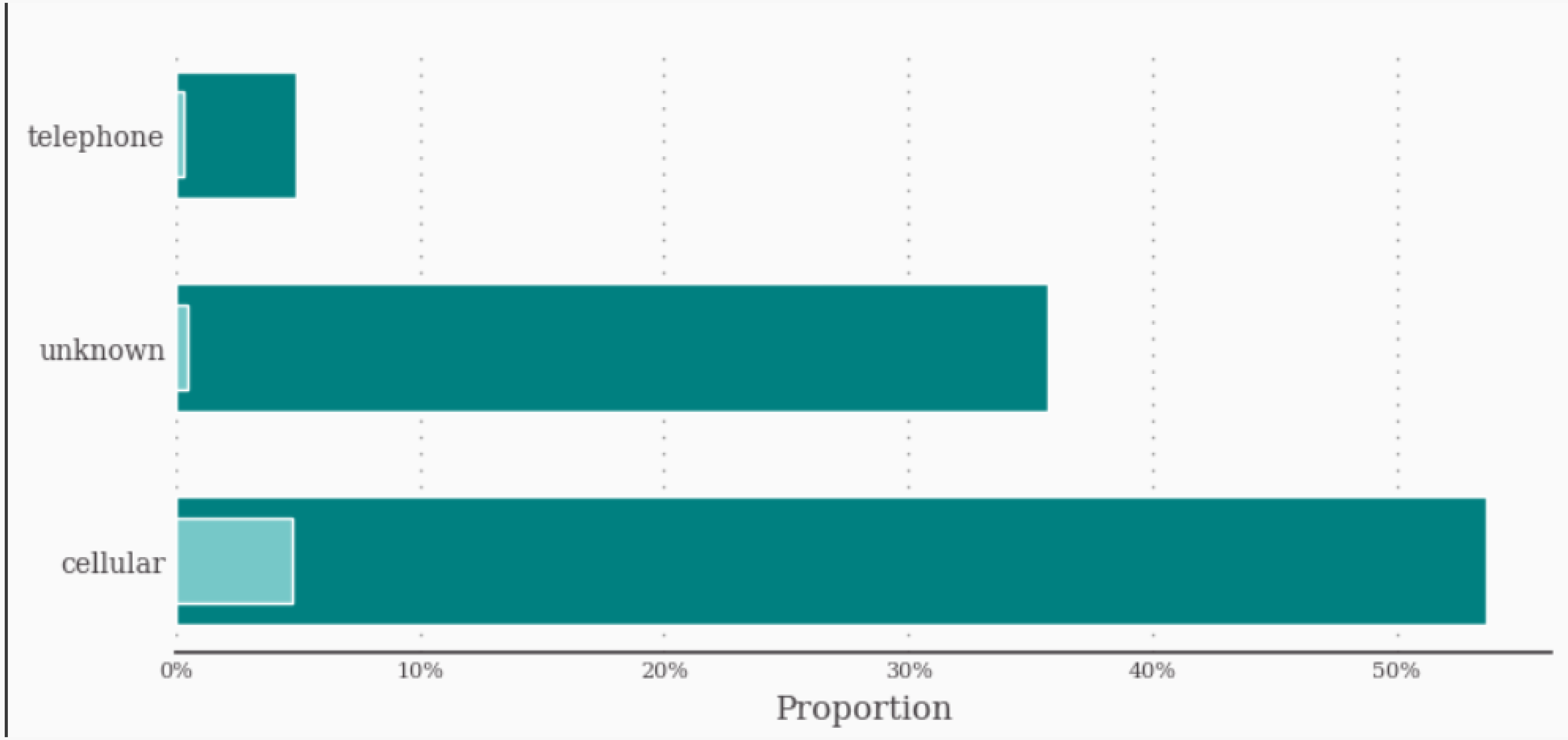
Doing a bivariate analysis over the deposits over months in a year, we can see how the annual nature of fiscal cycle affects the probability of contacts and conversions

IV Analysis



Plotting the graphs between marital status and all other attributes, we can see how it affects not only the subscription rate but also most of the other attributes.

This signifies the effect of someone being single, married, or divorced over the finances of individuals or families and the psychological changes or implications on decisions made while it has a say on other factors, it indirectly induces its weightage on term deposits.

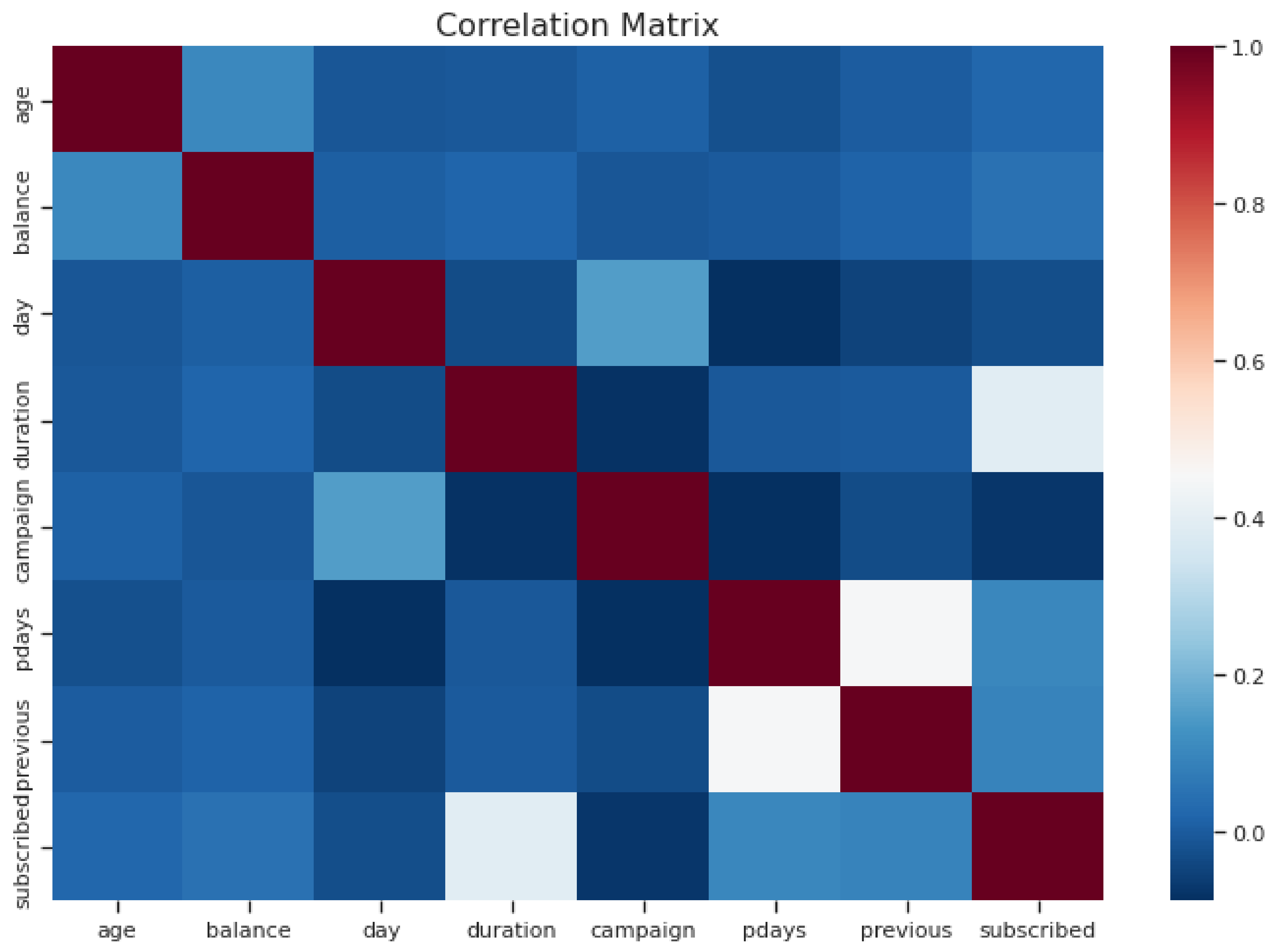


Last but not the least, means of contact doesn't imply much on conversions but we gains insight on which modes are popular and thus will help banks to efficienty allocate resources

IV Analysis

There is no better way to conclude research results than a correlation matrix heatmap, which demonstrates how strongly attributes weigh on each other, going from colder to warmer colours and the diagonal being warmest as characteristics are compared to themselves.

- Here we can see age correlating with balance, which is obvious
- also seen is the relation between day and campaign
- Duration and subscription are fundamentally related
- lastly, pdays and previous attributes have about half a correlation



Precision

Precision is the ratio of true positives to the total number of positive predictions. In other words, it measures how many of the positive predictions made by the model were actually correct.

Recall

Recall is the ratio of true positives to the total number of actual positive cases. In other words, it measures how many of the positive cases the model was able to identify correctly.

Accuracy

Accuracy is the ratio of correct predictions to the total number of predictions. In other words, it measures how many of the predictions made by the model were correct.

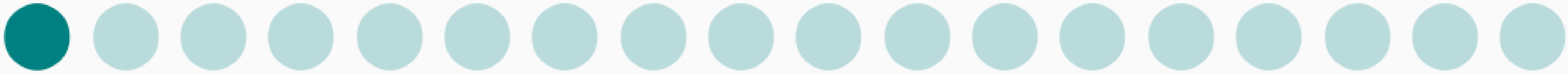
F1 Score

F1 score is the harmonic mean of precision and recall. It combines the precision and recall into a single metric that balances both measures.

V Our Metric

Client Decided To Subscribed

Only 1 in 18 people [1761 out of 28193]



- Only 1 out of 18 people decide to make a Term Deposit. The goal is to minimise the calls by eliminating those who will not subscribe to the Term Deposit. This way, a company can save vast amounts of money and redirect to where necessary.
- Therefore, Recall is more important because a company does not want to miss out its potential customers.
- However, if we optimise only Recall, Precision would be affected badly. Thus, we use F2 Score
- **F2 Score** gives more weight to Recall and less weight to Precision. Thus, Recall would be much improved without affecting Precision too much.

$$F2 = 5\left(\frac{(precision)\,(recall)}{4\;precision + recall}\right)$$

Naive Bayes

A probabilistic algorithm that makes predictions based on the probabilities of each feature and how they relate to the outcome.

Logistic Regression

A statistical method that models the probability of a binary outcome using a linear combination of the input features.

Decision Tree

A tree-like model that uses a series of binary decisions to classify or predict an outcome based on the input features.

Random Forest

An ensemble learning method that combines multiple decision trees to improve the accuracy and robustness of predictions.

Gradient Boosting

Gradient boosting combines weak models to create a strong one, by correcting the errors of the previous models through iterative addition of new models.

XGBoost

XGBoost is a fast and powerful machine learning algorithm that uses gradient boosting and regularization to improve accuracy and prevent overfitting.

Grid Search CV

Grid search CV searches for the best combination of hyperparameters for a machine learning model by evaluating the model's performance with each combination using cross-validation.

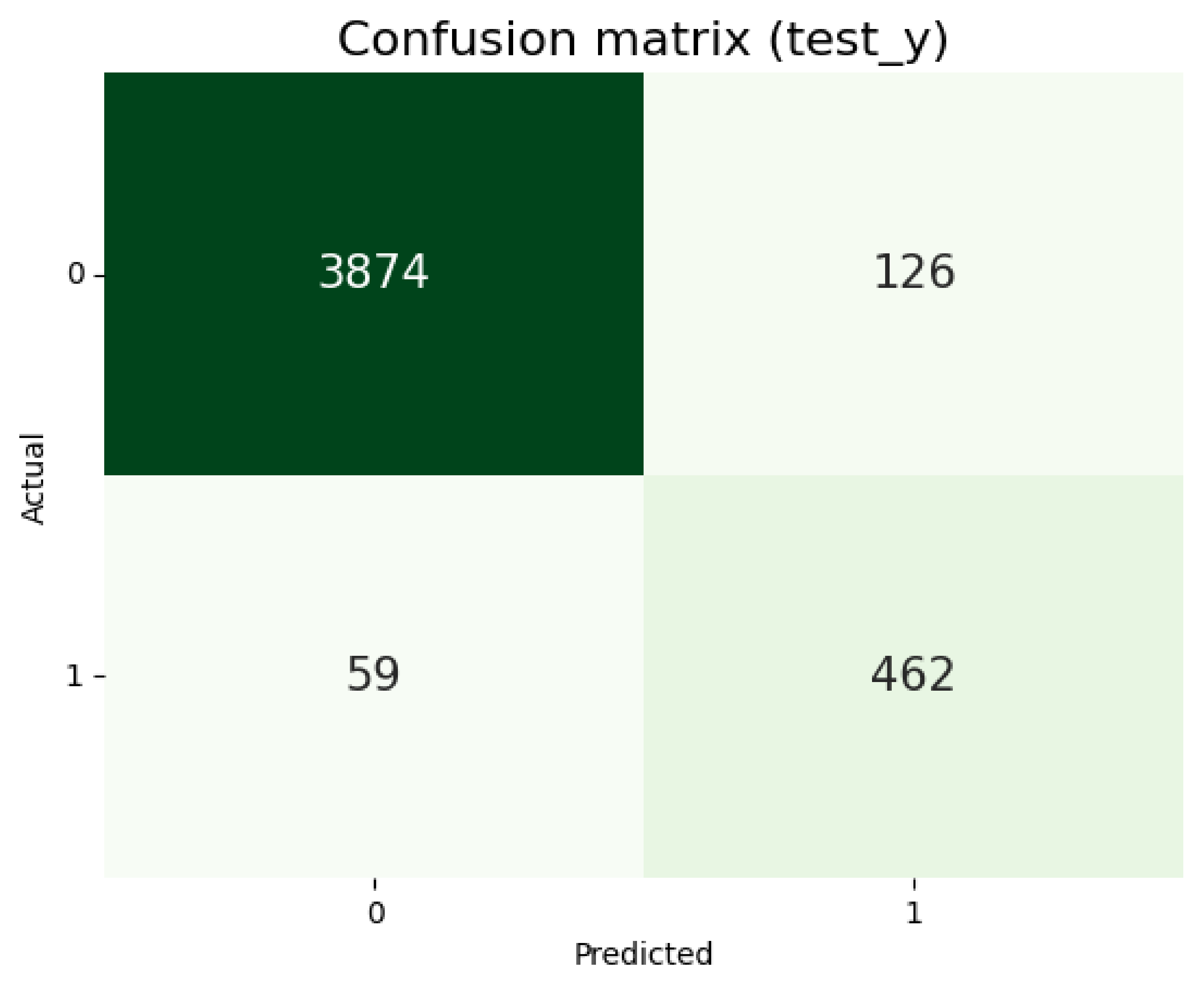
XGBoost over Grid Search CV

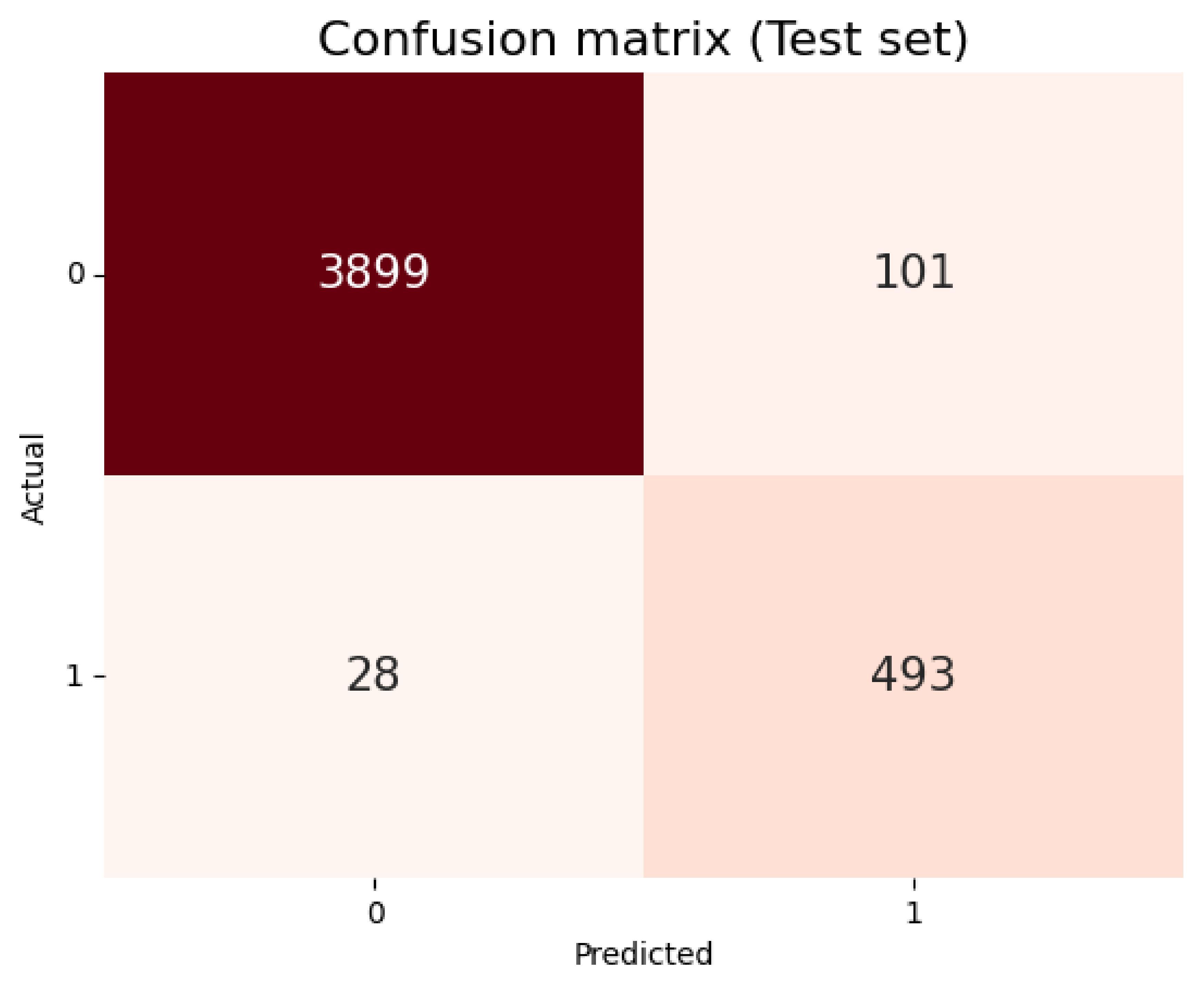
Grid search CV can be used with XGBoost to optimize its hyperparameters, leading to better accuracy and generalization to new data.

Method	Accuracy	F2 Score
Naive Bayes	70.34%	0.37
Logistic Regression	81.19%	0.41
Decision Tree	82.85%	0.31
Random Forest	97.15%	0.92
XGBoost	88.50%	0.41

The results achieved through these methods were not quite good in comparison to the results achieved by other papers.

We were also advised to use ensemble methods and used Random Forest which gave great results.


Confusion matrix (test_y)
3874
126
59
462
0
1
0
1
Actual
Predicted


Confusion matrix (Test set)
3899
101
28
493
0
1
0
1
Actual
Predicted

Accuracy = 97%

S. No.	Classification Models	Evolution Metrics of Classification Models				
		Accuracy	Precision	Recall	F1-Score	AUROC-score
1.	LR	92.72	70.58	54.54	61.53	93.62
2.	GNB	85.0	38.55	72.72	50.39	84.76
3.	RF	91	59.37	43.18	50.0	93.53
4.	SVM	90.53	58.62	38.63	46.57	91.4
5.	DT	91.0	56.75	47.72	51.85	71.69

https://journal.stic.ac.th/index.php/sjhs/article/view/296/85

VIII Results of other Research Papers

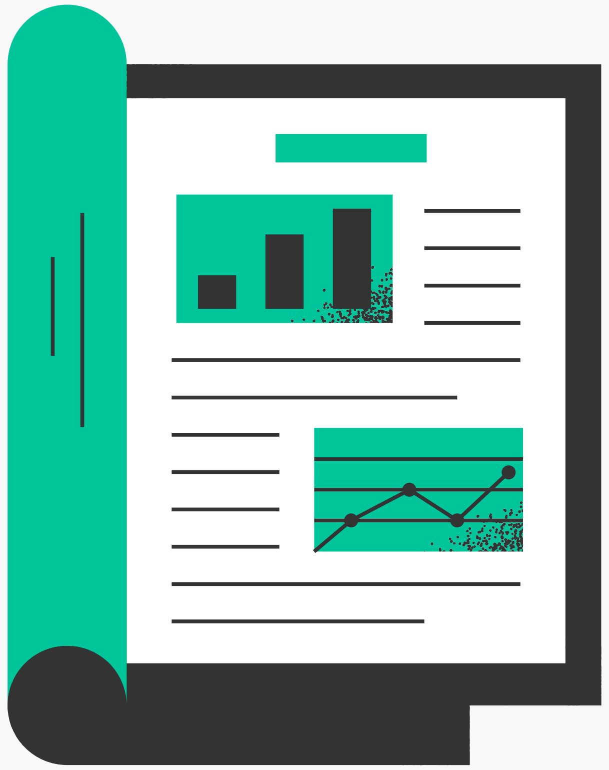
Table 1. Comparison result of Neural Network, SVM, Naïve Bayes and Stacking with Ensemble methods

		Accuracy (%)	Error-rate (%)	Sensitivity (%)	Specificity (%)
Neural Network	NN	94.8684	5.1316	95.5175	98.225
	NN (Bagging)	96.6158	3.3842	97.14	99.075
	NN (Boosting)	94.8684	5.1316	94.21	98.275
SVM	SVM	89.7589	10.2411	85.20	96.875
	SVM (Bagging)	89.8031	10.1969	89.84	96.95
	SVM (Boosting)	90.1198	9.8872	90.50	97.00
Naive Bayes	NB	88.2327	11.7673	84.32	94.2
	NB (Bagging)	88.3654	11.6346	87.40	94.325
	NB (Boosting)	88.7193	11.2807	89.70	95.025
Stacking	Classifier (SVM, NN) Meta classifier (SVM)	91.3294	8.6706	89.45	98.975

https://ojs.wiserpub.com/index.php/AIE/article/view/880/591

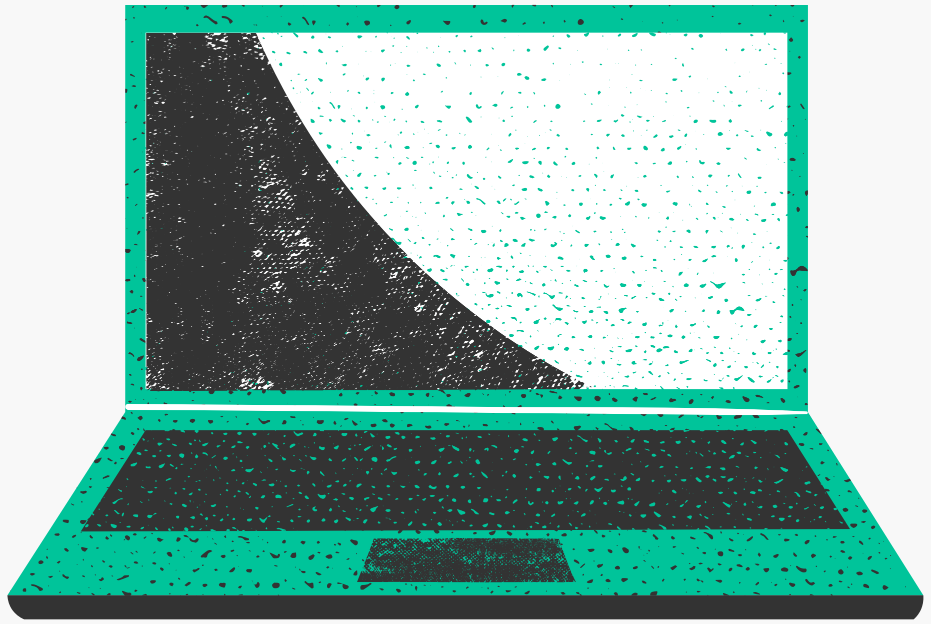
Ethical Considerations

We need to ensure that the model does not discriminate against any particular group of customers based on their age, gender, ethnicity, or other protected characteristics.



Feature Engineering

Explore different data sources and incorporate additional features that can help us better predict the likelihood of a client subscribing to a term deposit.



Deployment

We can deploy it in a real-world setting and integrate it with the bank's existing systems.

Wednesday, April 19, 2023

CS-354N, IIT Indore

SUPREVISOR

Dr. Aruna Tiwari

STUDENTS

Mukul Jain(200001050)
Nilay Ganvit(200001053)

Thank you for listening!