

Bank Marketing Dataset

CIND 119 – DJ0 – Final Project

Members

Ime Precious, *pime@ryerson.ca*

Sharaf Malak, *malak.sharaf@ryerson.ca*

Sharafutdinova Iuliia, *isharafutdinova@ryerson.ca*

Summary

Our team of data scientists has been tasked with developing an effective telemarketing strategy to sell term deposit accounts for a Portuguese bank. This bank has been conducting marketing campaigns, but it has not been effective. If the dataset provided by the bank is a correct the representation of their data, then it proves that their telemarketing strategy is not effective because only 12% of the customers are subscribed to their term deposit accounts.

The dataset consisted of 4521 rows and 17 attributes – 7 of which were quantitative and 10 which were qualitative including the target attribute. The dataset was prepared using R. The dataset had no missing data, but it contained a lot of outliers, and it was imbalanced.

The two machine learning methods were used in this analysis – Decision Tree and Naïve Bayes. These methods were used for all attributes and the selected attributes (*Age*, *Duration*, and *Poutcome*). The Decision Tree was created using Python and SAS, while the Naïve Bayes method was applied using Python. The performance metrics – accuracy, recall, and precision – were used to evaluate the accuracy of the models. The Decision Tree Sklearn is the best model in accurately predicting the customers who subscribed to the term deposit.

We would recommend that the bank should contact customers who are in their 40's or above, in management or a technician, married, with minimum secondary school education, and with account balance above £1600 as they are most likely to subscribe to a term deposit account.

Table of Contents

SUMMARY	II
1.0 DATA PREPARATION.....	1
1.1 DATA EXPLORATION	1
1.2 DATA FILTERING	2
1.3 DATA CLEANING	2
<i>Missing Values</i>	2
<i>Outliers</i>	2
<i>Correlation</i>	3
<i>Balance</i>	4
1.4 DATA TRANSFORMATION	5
<i>Distribution</i>	5
<i>Data Partitioning</i>	7
<i>Feature Selection</i>	8
2.0 PREDICTIVE MODELING/CLASSIFICATION	8
2.1 CLASSIFICATION USING DECISION TREE	8
<i>Decision Tree: Python – Weka Package</i>	8
<i>Decision Tree: Python – Sklearn Package</i>	9
<i>Decision Tree: SAS</i>	9
<i>Decision Tree Comparison</i>	10
2.2 CLASSIFICATION USING NAIVE BAYES	10
<i>Naïve Bayes: Python – Weka Package</i>	11
<i>Naïve Bayes: Python – Sklearn Package</i>	11
<i>Naïve Bayes Comparison</i>	12
2.3 THE BASELINE "ALL ATTRIBUTES" AND "SELECTED FEATURES" COMPARISON	12
<i>Python Weka "All Attributes" and "Selected Features" Comparison</i>	12
<i>Python Sklearn "All Attributes" and "Selected Features" Comparison</i>	13
2.4 PERFORMANCE METRICS FOR COMPARING THE TWO TECHNIQUES	13
3.0 CONCLUSIONS AND RECOMMENDATIONS.....	14
3.1 CONCLUSION	14
3.2 ANY ADDITIONAL DATA PROCESSING OR ANALYSIS SHOWN.....	15
RECOMMENDATIONS.....	18
BONUS: ADDITIONAL DATA PROCESSING OR ANALYSIS SHOWN	19
WORKLOAD DISTRIBUTION	21
REFERENCES	22

1.0 Data Preparation

The Bank Marketing Dataset was used for this project and this dataset contained both quantitative and qualitative data. Table 1 shows the attributes and the attribute type.

Table 1 - Attribute Type and Description

Column Name	Attribute type	Description
Age	quantitative	age of customers
Job	nominal	type of job, twelve categories
Marital	nominal	marital status, three categories
Education	ordinal	education level, four categories
Default	nominal	credit in default (yes or no)
Balance	quantitative	average yearly balance
Housing	nominal	housing term deposit (yes or no)
Term deposit	nominal	personal term deposit (yes or no)
Contact	nominal	method of last contact, three categories
Day	quantitative	day customer was last contacted
Month	ordinal	month the customer was last contacted, twelve categories
Duration	quantitative	last contact duration
Campaign	quantitative	number of contacts during campaign
Pdays	quantitative	number of days since last contact from a previous campaign
Previous	quantitative	number of contacts before this campaign
Poutcome	ordinal	outcome of previous campaign (yes or no)
Y	nominal	target variable, subscribed to a term deposit (yes or no)

1.1 Data Exploration

Table 2 shows the min, max, mean, median, and standard deviation of the numeric attributes. If the mean is closer to the minimum value, then the data is right-skewed, if it is closer to the maximum value then it is left-skewed, but if it is closer to the median then it is normally distributed. Based on the values, the attributes Age and Day have a mean that is close to the median so it should be normally distributed while the remaining attributes (Balance, Duration, Campaign, Pdays, Previous) are right-skewed.

Table 2 - Numerical Attributes Data Exploration

Attribute	Min	Max	Mean	Median	Standard Deviation
Age	19	87	41.14	39	10.58
Balance	-33313	71188	1422.66	444	30009.64
Day	1	31	15.92	16	8.25
Duration	4	3025	263.96	185	259.8
Campaign	1	50	2.79	2	3.11
Pdays	-1	871	39.77	-1	100.12
Previous	0	25	0.54	0	1.69

1.2 Data Filtering

The dataset does not have any duplicate records, so no rows were deleted for this reason.

1.3 Data Cleaning

Missing Values

The data was checked for missing values and there were no missing values however, some of attributes – *Job*, *Education*, *Contact*, and *Poutcome* – had some values as “unknown”. This could imply that the values are missing or not applicable for those customers. For example, the unknown values for education could mean that the options – primary, secondary, and tertiary – are not applicable to that customer. For this project, we assumed that the “unknown” values were not applicable to those customers and therefore did not treat them as missing values.

Outliers

Outliers are values that are outside the normal range or at an abnormal distance from other points in a dataset (Frost, n.d.). It is important to identify and deal with outliers because they can be errors or an anomaly which can distort the analysis of the data (Barnes, 2021). Identifying and dealing with them is a way of confirming the data quality before performing any further analysis or extracting insights (Barnes, 2021). Some of the ways to determine outliers are using graphs such as box plots, histogram or a scatter plot, and z-score (Frost, n.d.).

Outliers can be dealt with by removing the records with the outliers or transforming the data using log or square root. Removing records with outliers may not be an ideal solution in every case because of the loss of data that comes with it. Also, outliers should be investigated before deleting them from the dataset because they could be valid (Grace-Martin, n.d.). An alternative is to remove the attributes with a significant proportion of outliers but first, you must determine if the attribute is a key feature.

The box plot (univariate method) was used to determine the outliers for the quantitative attributes in this dataset. The multivariate method can also be used to determine the outliers. This method might be more accurate since it builds a model using all the attributes (Alberto Quesada, 2017).

Table 3 below shows the outliers for the quantitative attributes. From the table, *Day* has no outliers and only 0.8% of the *Age* data are outliers. However, 18% of *Pdays* and *Previous* are outliers while 11% of

Balance are outliers. Since data manipulation is out of the scope of this project and to prevent loss of data, no records with outliers were removed.

Table 3 - Numerical Attribute Outlier Analysis

Attribute	# of Outliers	% Outliers
Age	38	0.8%
Balance	506	11.2%
Day	0	0%
Duration	330	7.3%
Campaign	318	7%
Pdays	816	18%
Previous	816	18%

Correlation

Correlation indicates the relationship between 2 or more variables. A positive correlation means that as an independent variable increases, the dependent variable increases while a negative correlation means that the dependent and independent variables have an inverse relationship so as the independent variable increases the dependent variable decreases. However, the correlation coefficient must be statistically significant for this to be applicable – that means the correlation must be strong.

Correlation coefficients can assume any value between -1 and 1 and the closer the coefficient is to 0, the weaker the correlation, the closer it is to 1 the stronger the correlation, and a correlation coefficient of 0 indicates that there is no correlation between the values. The Pearson correlation is used to examine the correlation coefficient between the numeric values while the Phik correlation is used to compare the correlation coefficient between the categorical and numerical data.

Table 4 shows the Pearson correlation matrix for the numerical attributes. The Pearson correlation matrix shows that there is no correlation between *Age* and *Duration* and between *Age* and *Previous* because the correlation coefficient is 0. The result shows that there is a positive correlation between the *Pdays* and *Previous* since the correlation coefficient is 0.58. The result also shows that the maximum correlation, apart from 0.58 is 0.16 between *Campaign* and *Day*. Since the correlation for the other variables are low, we can conclude that the remaining attributes are not correlated to each other.

Table 4 - Pearson Correlation Matrix for Numeric Attributes

	Age	Balance	Day	Duration	Campaign	Pdays	Previous
Age	1.00	0.08	-0.02	0.00	-0.01	-0.01	0.00
Balance	0.08	1.00	-0.01	-0.02	-0.01	0.01	0.03
Day	-0.02	-0.01	1.00	-0.02	0.16	0.09	-0.06
Duration	0.00	-0.02	-0.02	1.00	-0.07	0.01	0.02
Campaign	-0.01	-0.01	0.16	-0.07	1.00	-0.09	-0.07
Pdays	-0.01	0.01	-0.09	0.01	-0.09	1.00	0.58
Previous	0.00	0.03	-0.06	0.02	-0.07	0.58	1.00

Figure 1 shows that *Poutcome*, *Duration* and *Month* have the highest correlation to the target attribute, *Y*. Based on the figure, we can conclude that *Poutcome*, *Duration* and *Month* have correlation coefficients of ~ 0.6 , ~ 0.4 , and ~ 0.3 with respect to *Y*.

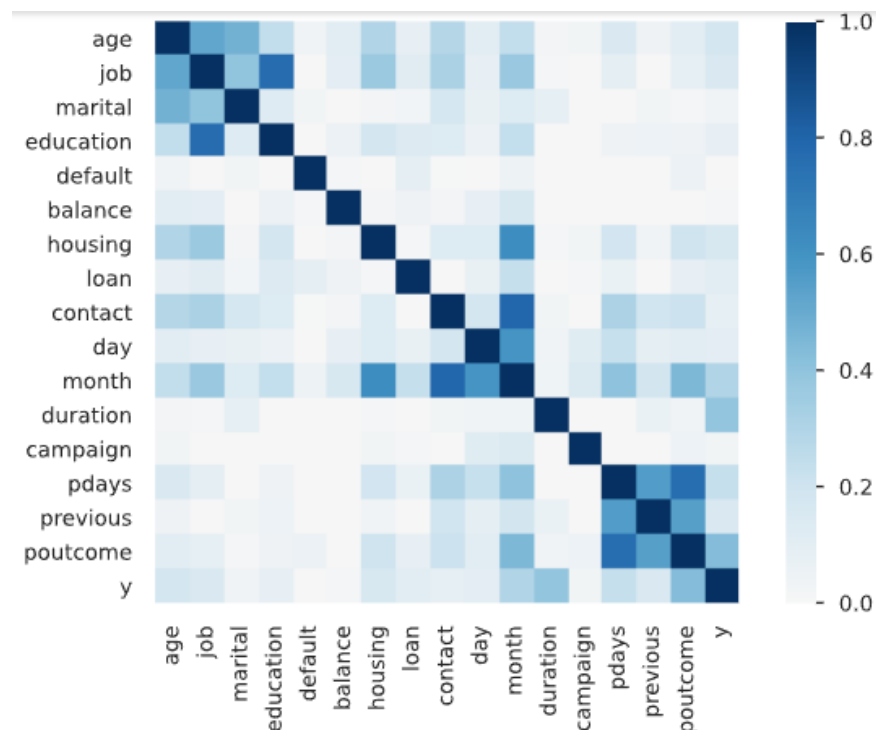


Figure 1 - Phik's Correlations Matrix showing categorical Attributes

Some other attributes with a strong positive phik correlation coefficients are shown below:

- *Job & Education* ~ 0.8
- *Month & Contact* ~ 0.8
- *Month & Housing* ~ 0.6
- *Pdays & Poutcome* ~ 0.8
- *Pdays & Previous* ~ 0.6

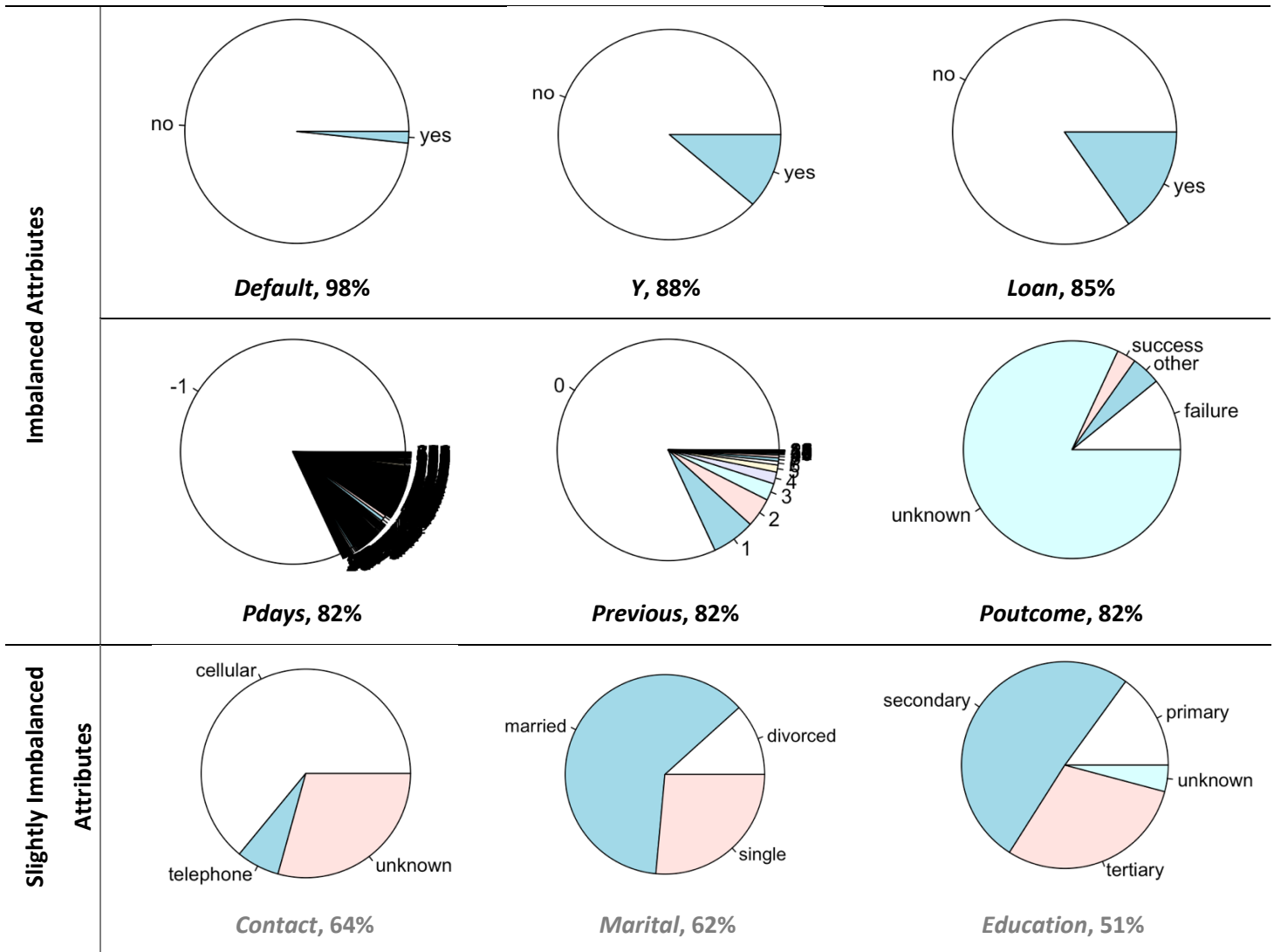
Balance

A dataset is balanced when there is relatively equal representation of all variables for each attribute. Based on the pie charts in Table 5, we can conclude that the dataset is not balanced. The results from the target attribute, *Y* shows that 88% of the dataset represent customers who did not subscribe to the term deposit so this dataset will be biased towards accurately predicting these customers since most of the data represents this segment of customers.

The effect of imbalance in a dataset can be analyzed using evaluation metrics such as accuracy, precision, and recall in evaluating the performance of the model. Imbalance can be dealt with by over sampling which is simply adding duplicate records of the lesser represented variables to the dataset or under sampling which is removing some records from the overrepresented variables from the dataset

(Tripathi, 2019). Another way to deal with imbalance is feature selection using statistical methods. This leads to a more accurate classification model and reduces imbalance in the dataset (Tripathi, 2019).

Table 5 - Pie Charts for Imbalanced Attributes



1.4 Data Transformation

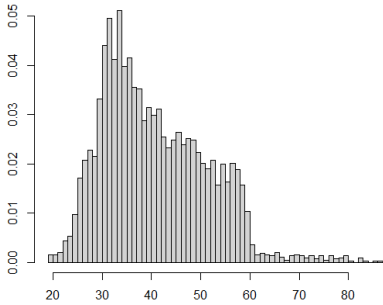
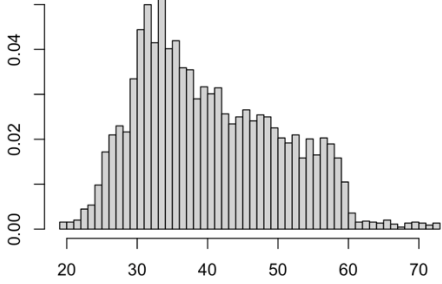
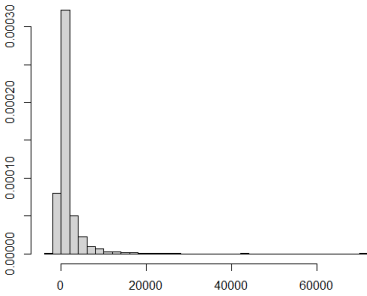
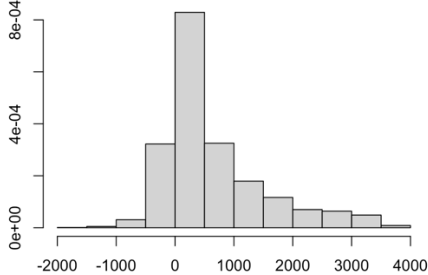
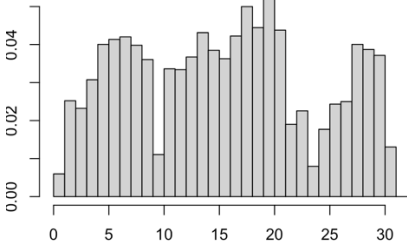
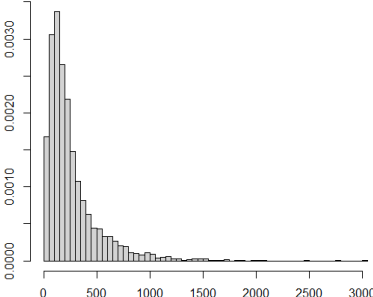
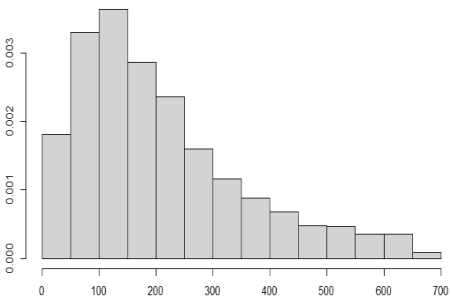
Distribution

Distribution shows all the possible values and how often they occur for an attribute in a dataset. It tells us about the shape and spread of data. Distribution can be normal, left or right skewed, bimodal, or multimodal. A histogram was used to show the distribution of the numeric attributes.

From the histograms in Table 6, *Age* and *Balance* are normally distributed while *Duration* and *Campaign* are right-skewed. Skewed distributions can be dealt with by normalizing the values using square root, log, or reciprocal on every value for each skewed attribute (Wade, 2018). However, the skewed attributes were not normalized because it is out of the scope of this project.

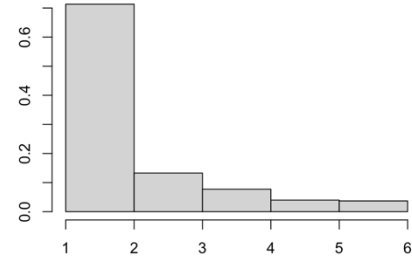
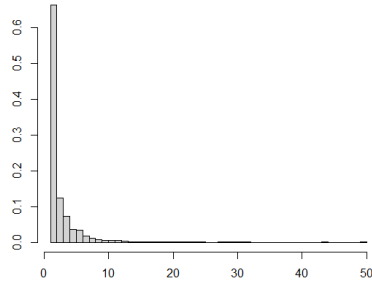
The *Duration* and *Campaign* attributes are of concern because of the outliers so the histogram was plotted without the outliers. The *Previous* and *Pdays* attributes have a very high spike at -1 and 0 because 82% of the customers were not contacted previously. *Pdays* was plotted without the customers that were previously contacted and without the outliers. The plots without the outliers showed a clearer shape of the distribution.

Table 6 - Histogram Distribution for Numerical Attributes

Attribute	Distribution	Distribution with Outliers	Distribution without Outliers
Age	Normal Distribution		
Balance	Normal Distribution		
Day	Multimodal Distribution		NA – no outliers
Duration	Right Skewed		

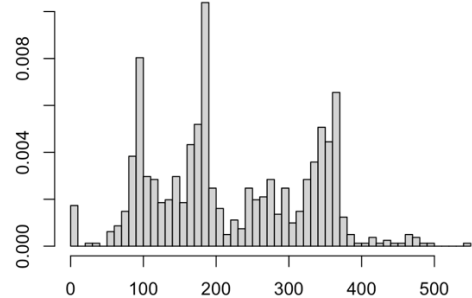
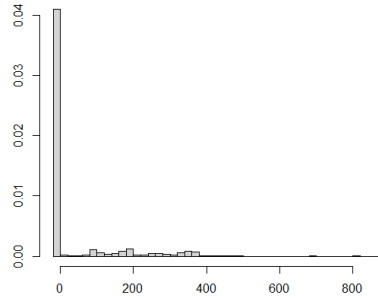
Campaign

Right
Skewed



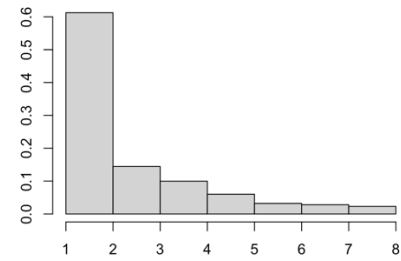
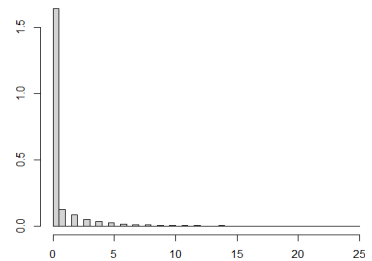
Pdays

Multimodal
Distribution



Previous

Right
Skewed



Data Partitioning

Data partitioning is the process of splitting data into different subsets for machine learning – training, validation, and testing (Data Partition, n.d.). In this project, the data was split into training and testing subsets using Python Weka and Sklearn. The training dataset is used for training the model, the validation dataset is used to evaluate the performance of the model and adjust the model parameters, while the test dataset is used for the final evaluation of the model (Brownlee, 2017).

In data partitioning, the records in the training data set are different from the records in the test or validation sets. This is to better evaluate the performance of the model. If the complete dataset is used for training and testing, then the model might predict the data accurately but might not predict unseen data as expected (Data Partition, n.d.). This process is used in supervised learning techniques like classification trees as seen in this project.

For this project, the Bank Marketing Dataset, was randomly divided into a training and test set to prevent bias. The Sklearn package used 70% for training the model and 30% for testing the model while the Weka ML package used 66% for training and 34% for testing.

Feature Selection

Feature selection is the process of using the most important attributes in the dataset for developing the model thereby reducing the number of attributes and possibly improving the performance of the method (Brownee, 2020). This process was applied in this project using SAS and Python Weka. Based on Weka the selected features were **Age**, **Duration**, and **Poutcome** and SAS had **Age**, **Duration**, **Poutcome**, and **Day**.

Table 7 - Variable Importance Table Output from SAS

Variable	Relative	Importance
Duration	1.0000	12.3938
Poutcome	0.6353	7.8737
Age	0.2841	3.5213
Day	0.2499	3.0972

2.0 Predictive Modeling/Classification

Classification is a supervised machine learning technique used to train models which predict classes based on the training from the training dataset. In this project, the Decision Tree and Naïve Bayes classification techniques were used to train, test, and evaluate the performance of the dataset. The **train-test split** method was used for the classification.

2.1 Classification using Decision Tree

A Decision Tree is a classification technique that uses attributes in a dataset to predict a class based on decisions. Decision trees are referred to as classification for qualitative attributes or regression trees for continuous variables (Gupta, 2017). For this project SAS and Python (Sklearn and Weka packages) were used to create the decision tree.

Decision Tree: Python – Weka Package

The Weka package split the data randomly into two subsets – the train (66%) and test (34%) subset and produced a J48 pruned tree using the C4.5 algorithm. The evaluation of this model was done on the test subset which was 1537 records (34% of the total dataset). The accuracy of this model is 88.5% which means 88.5% of the test data was predicted correctly.

Table 8 - Confusion Matrix for Python Weka Package (Decision Tree)

		Predicted Values	
		Negative - No	Positive - Yes
Actual Values	Negative - No	TN = 1300	FP = 57
	Positive - Yes	FN = 120	TP = 60

Table 9 shows that the True Positive rate is 0.333 which means that the model correctly predicts customers who subscribe for the deposit term 33% of the time and it incorrectly predicts the customers

who subscribe for the deposit term 4.2% of the time. The precision shows that 51.3% of the customers who were predicted as subscribers were correctly predicted.

Table 9 - Evaluation Metrics for Weka Package (Decision Tree)

Target Variable, Y	Precision	Recall	True Positive	False Positive
Class 1 - Yes	0.513	0.333	0.333	0.042
Class 0 - No	0.915	0.958	0.958	0.667

Decision Tree: Python – Sklearn Package

The Sklearn package was also used to create the decision tree and the dataset was randomly split into the training (70%) and test (30%) subset dataset. The precision for the customers who subscribed is 59% which means 59% of the customers who were predicted as subscribed were classified correctly. The recall for this model 29% which means that only 29% of the customers that subscribed for the term deposit were correctly classified. This indicates that the model is not reliable in correctly classifying customers who subscribed for the term deposit. The accuracy from this model is 87% which is slightly less than the accuracy from the Weka Package.

Table 10 - Evaluation Metrics for Sklearn Package (Decision Tree)

	Precision	Recall	True Positive
Positive – Yes	0.59	0.29	0.29
Negative – No	0.91	0.97	0.97
Accuracy	0.87	0.89	0.89

Decision Tree: SAS

SAS was also used to create a decision tree and to evaluate the decision tree classification model. The dataset was used for training and testing the algorithm because from the confusion matrix, there are 4521 total values which means all the data is being evaluated unlike the Python Weka and Sklearn packages which had only 34% and 30% of the dataset respectively for the testing.

From the confusion matrix in Table 11, the accuracy is 90.8% which is a good estimate. This means that 90.8% of all the customers were correctly classified. The accuracy is higher than the results from the Python models. This can be attributed to the complete dataset being used for training and testing which can lead to a more optimistic but not exactly accurate model.

The table also shows a low error rate of 2.1% for the customers who did not subscribe and a high error rate of 63.34% for the customers who subscribed. Table 12 shows that the precision for the customers who subscribed is 70% and the recall is 56%. This means that 70% of the customers who were predicted as subscribing customers were classified correctly and 56% of the actual customers who subscribed to the term deposit were classified correctly.

Table 11 - Confusion Matrix for SAS (Decision Tree)

		Predicted Values		Error Rate
		Negative - No	Positive - Yes	
Actual Values	Negative - No	3916	84	0.0210
	Positive - Yes	330	191	0.6334

Table 12 - Evaluation Metrics for SAS (Decision Tree)

	Precision	Recall	True Positive	False Positive
Positive – Yes	0.70	0.56	0.70	0.27
Negative – No	0.92	0.98	0.93	0.021

The sensitivity is 97.9% which means that the model has a high rate of correctly classifying customers that do not subscribe. Since the specificity is 27%, it implies that the model has a low rate of correctly classifying customers that subscribe for the term deposit. Table 13 shows a GINI index of 0.15 indicating this decision tree has a relatively good split since the GINI index is close to 0. The entropy level of 38% shows that the model is good because it indicates low levels of impurity.

Table 13 - Model-Based Fit Statistics from SAS (Decision Tree)

N Leaves	Mis-class	Sensitivity	Specificity	Entropy	Gini
6	0.0916	0.9790	0.3666	0.3849	0.1514

Decision Tree Comparison

Only the results from Weka and Sklearn will be compared for this project. The Weka is a better model because it has a higher accuracy (88.5% vs 87%) and true positive rate (33% vs 29%). This means that the Weka model correctly predicts those who subscribe to the term deposit 33% of the time while the Sklearn model predicts it correctly 29% of the time.

2.2 Classification using Naive Bayes

The Naïve Bayes model is an old method of classification and predictor selection and is known for its simplicity and stability. The Naïve Bayes considers all attributes to be of equal importance as opposed to the decision tree in determining the more important attributes. The evaluation metrics yield similar results for the TP rate, FP rate, Recall, and Precision.

Naïve Bayes: Python – Weka Package

The accuracy of this model is 87.57% which means 87.57% of the test data was predicted correctly.

Table 14 - Confusion Matrix for Python Weka Package (Naive Bayes)

		Predicted Values	
		Negative - No	Positive - Yes
Actual Values	Negative - No	TN = 1261	FP = 96
	Positive - Yes	FN = 95	TP = 85

The precision for the customers who subscribed is 47.0% which means 47.0% of the customers who were predicted as subscribed were classified correctly. The recall for this model 47.2% which means that only 47.2% of the customers that subscribed for the term deposit were correctly classified. This indicates that the model is not reliable in correctly classifying customers who subscribed for the term deposit.

Table 16 shows that the True Positive rate is 47.0% which means that the customers who subscribed are correctly predicted 47.0% of the time and the False Positive rate of 7.1% indicates that the model incorrectly predicts the subscribed customers 7.1% of the time.

Table 15 - Evaluation Metrics for Weka Package (Naive Bayes)

Target Attribute, Y	Precision	Recall	True Positive	False Positive
Class 1 – Yes	0.470	0.472	0.470	0.071
Class 0 – No	0.930	0.930	0.930	0.527

Naïve Bayes: Python – Sklearn Package

The accuracy from this package using cross-validation is 47% which is significantly lower than the Python Weka package and the decision tree results.

Table 16 - Confusion Matrix for Python Sklearn Package (Naive Bayes)

		Predicted Values	
		Negative - No	Positive - Yes
Actual Values	Negative - No	TN = 516	FP = 672
	Positive - Yes	FN = 46	TP = 123

The precision for the customers who subscribed is 15% which means only 15% of the customers who were predicted as subscribed to term deposits were classified correctly. The recall and true positive rate for this model 73% which means that 73% of the customers that subscribed for the term deposit were correctly classified. Table 17 shows that the False Positive rate is 56% which indicates that the model incorrectly predicts the subscribed customers 56% of the time.

Table 17 - Evaluation Metrics for Sklearn Package (Naïve Bayes)

	Precision	Recall	True Positive	False Positive
Positive – Yes	0.15	0.73	0.73	0.56
Negative – No	0.92	0.43	0.43	0.27

Naïve Bayes Comparison

Based on the results from Weka and Sklearn, the Sklearn is a better model because it has a higher true positive rate (73% vs 47%). This means that the Sklearn model correctly predicts those who subscribe to the term deposit 73% of the time while the Weka model predicts it correctly 47% of the time.

2.3 The baseline "all attributes" and "selected features" comparison

Python Weka "All Attributes" and "Selected Features" Comparison

Weka attribute selection module selected *Age*, *Duration*, and *Poutcome* using the Weka Attribute Selector module. Shows the results from using all attributes and selected features for the Weka Decision Tree and Naïve Bayes classification models.

The decision tree from all attributes has 104 leaves with a size of 142 while the decision tree from the selected features has 10 leaves and has a size of 14. The tree is much smaller for the selected features because there are less attributes used in the classification.

Table 18 - Selected Features Confusion Matrix for Python Weka Package (Decision Tree) 0 – 88.87% Accuracy

		Predicted Values	
		Negative - No	Positive - Yes
Actual Values	Negative - No	TN = 1284	FP = 73
	Positive - Yes	FN = 98	TP = 82

Table 19 - Selected Features Confusion Matrix for Python Weka Package (Naïve Bayes) – 88.93% Accuracy

		Predicted Values	
		Negative - No	Positive - Yes
Actual Values	Negative - No	TN = 1310	FP = 47
	Positive - Yes	FN = 123	TP = 57

Error! Reference source not found. shows that the accuracy and precision of the selected feature ui slightly higher than the "All Attributes" dataset.

Table 20 - "All Attributes" and "Selected Features" Comparison for Python Weka Package

Evaluation	Decision Tree		Naïve Bayes	
	All	Selected	All	Selected

Metrics	Attributes	Features	Attributes	Features
TP	0.33	0.46	0.47	0.32
FP	0.04	0.05	0.07	0.03
Accuracy	0.885	0.8887	0.88	0.8893
Precision	0.51	0.53	0.47	0.55
Recall	0.33	0.47	0.47	0.32

Python Sklearn “All Attributes” and “Selected Features” Comparison

The decision tree from all attributes has 104 leaves with a size of 142 while the decision tree from the selected features has 10 leaves and has a size of 14. The tree is much smaller for the selected features because there are less attributes used in the classification.

Table 21 - Selected Features Confusion Matrix for Python Sklearn Package (Naïve Bayes) – 79% Accuracy

		Predicted Values	
		Negative - No	Positive - Yes
Actual Values	Negative - No	TN = 968	FP = 220
	Positive - Yes	FN = 66	TP = 103

Error! Reference source not found. shows that the accuracy and precision of the selected feature is lower than the “All Attributes” dataset therefore, the selected features do not produce a better model for the Sklearn model.

Table 22 - “All Attributes” and “Selected Features” Comparison for Python Sklearn Package

Evaluation Metrics	Decision Tree		Naïve Bayes	
	All Attributes	Selected Features	All Attributes	Selected Features
Accuracy	0.88	0.79	0.89	0.87
Precision	0.51	0.32	0.15	0.64
Recall	0.26	0.61	0.73	0.21

2.4 Performance metrics for comparing the two techniques

The confusion matrix, accuracy, precision, recall and false positive rate were used to compare the performance of the two techniques.

Accuracy shows the percentage of the data that was predicted correctly.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision is used to measure what percentage of the positive class was predicted correctly.

$$Precision = \frac{TP}{TP + FP}$$

Recall calculates the percentage of the predicted positive attributes that were correctly predicted as positive. This is the same measure as the True Positive rate.

$$Recall = \frac{TP}{TP + FN}$$

From Table 23, we can see that recall from the Naïve Bayes for Sklearn has the highest recall of 73% while the Decision Tree for Sklearn has the lowest recall of 29%. Also, Decision Tree for Sklearn has the highest precision of 59% while Naïve Bayes for Sklearn has the lowest precision of 15%.

Overall, the Decision Tree for Sklearn performs the best because it has the highest accuracy, so it predicts customers that subscribe and that do not subscribe with 89% correctness. It also has the highest precision of 59% which indicates that out of the customers predicted to subscribe for the term deposit, 59% are correctly predicted.

However, it might be useful to consider the recall for this project since the aim is to develop a more effective telemarketing strategy that will target the customers that will subscribe. The precision shows how many of the predicted subscribed customers are actually subscribed customers while the recall shows how many actual subscribed customers were not predicted as subscribed customers. It might be useful to have a higher recall since it predicts more of the actual positives (subscribed customers) correctly so this way, less potential subscribing customers are lost. If this logic is followed, the Naïve Bayes WEKA method might be best model since Sklearn has a low accuracy. Naïve Bayes Sklearn is not chosen because of the low accuracy of 47% and low precision of 15%.

Table 23 - Evaluation Metrics Comparison for the Class Yes

Evaluation Metrics	Decision Trees		Naïve Bayes	
	J48 WEKA	One-Hot encoded (Sklearn)	WEKA	Sklearn
Overall Accuracy	0.88	0.89	0.87	0.47
Recall	0.33	0.29	0.47	0.73
Precision	0.51	0.59	0.46	0.15

3.0 Conclusions and Recommendations

3.1 Conclusion

The dataset contained 16 attributes and 1 target attribute for binary classification used to determine if a customer will subscribe to a Portuguese Bank's term deposit or not. The selected attributes for this dataset from the Weka Attribute Selector module were *Age*, *Duration*, and *Poutcome*. Based on the model comparison, Decision Tree Sklearn was chosen as the most accurate model with an accuracy of 89% while the Naïve Bayes for Weka was recommended because of the recall of 47% to prevent loss of potential subscribing customers.

The total number of customers who opened a term deposit using the Decision Tree for Weka and Sklearn and the Naive Bayes for Weka is 89% and the number that did not open term deposit is 11%.

From our analysis, these are the criteria for the people who are more likely to subscribe for the term deposit.

1. Mean age should be 41 years old
2. Jobs should be Management or Technician
3. Marital status should be Married
4. Should be in Secondary or Tertiary Education
5. Balance should be at least £1600
6. Mode of contact should be Cellular
7. No housing term deposit or default

3.2 Any additional data processing or analysis shown

Figure 2 shows that people with a secondary or tertiary education are more likely to subscribe for the term deposit.

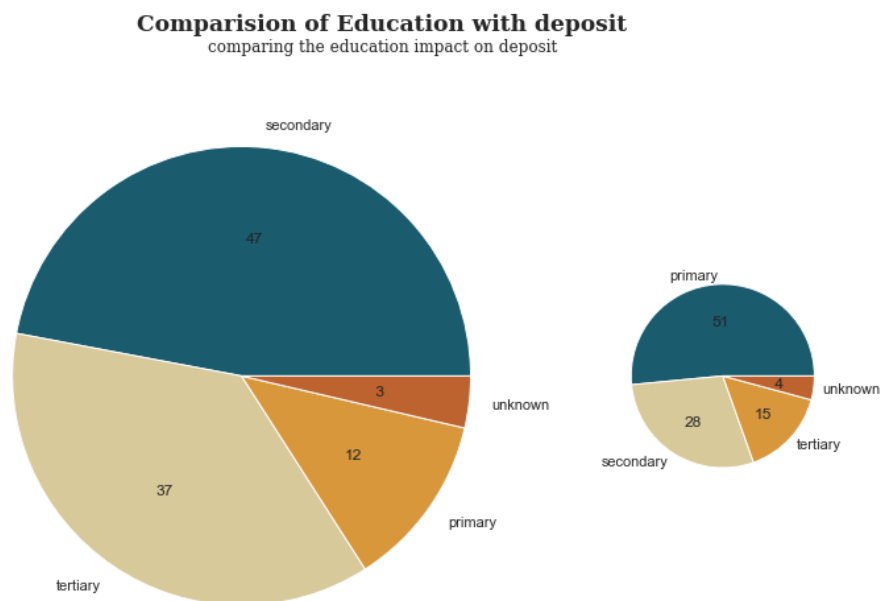


Figure 2 - Comparing the education impact on deposit

Figure 3 shows that people with a secondary and tertiary education have the most balance so to have people with a higher balance, we recommend contacting people with secondary and tertiary education since one of the criteria requires a balance of at least £1600.

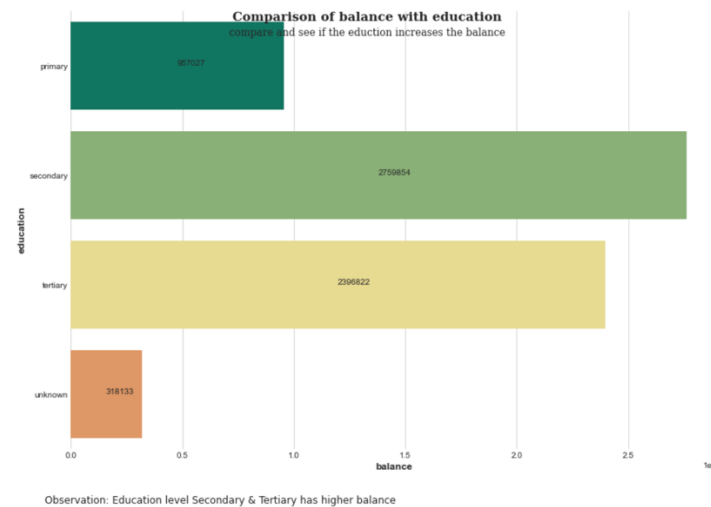


Figure 3 - Comparison of Balance and Education

Figure 4 shows that people's balance increase with age so the older someone is the more likely they are to subscribe to the term deposit.

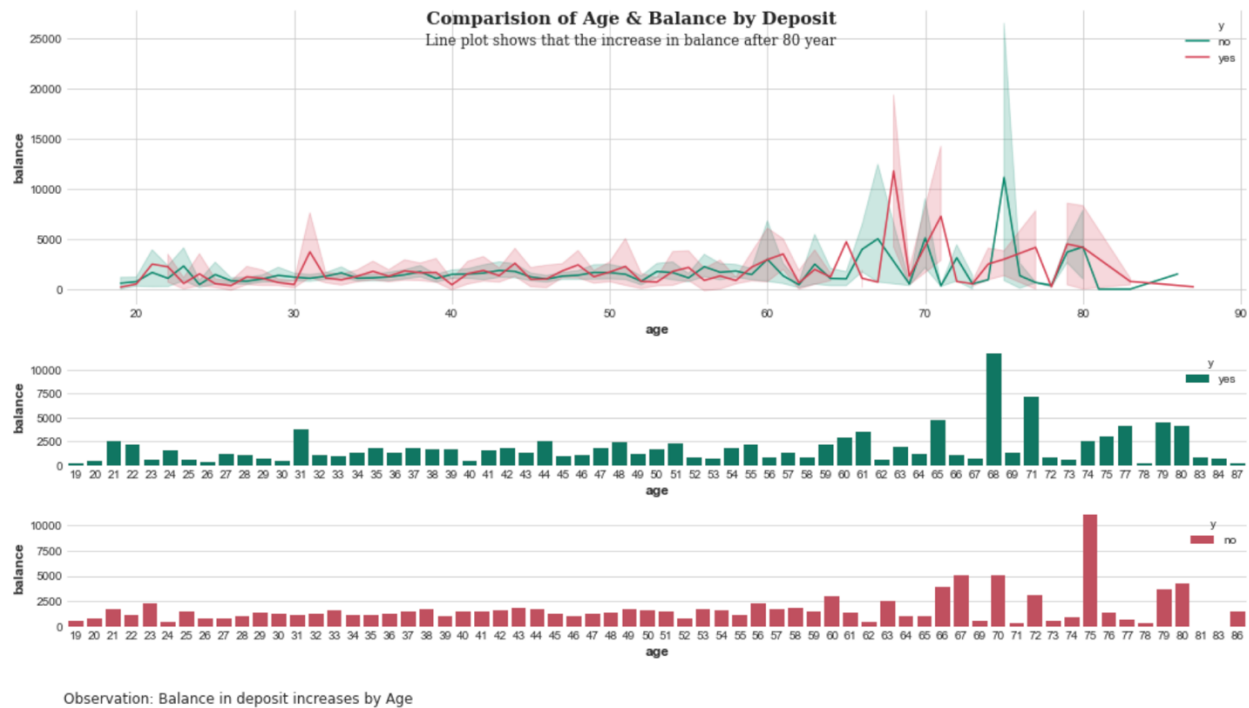


Figure 4 - Comparison of Age and Balance by Deposit

Figure 5 shows that the people are more likely to subscribe to the term deposit if they do not have a loan.

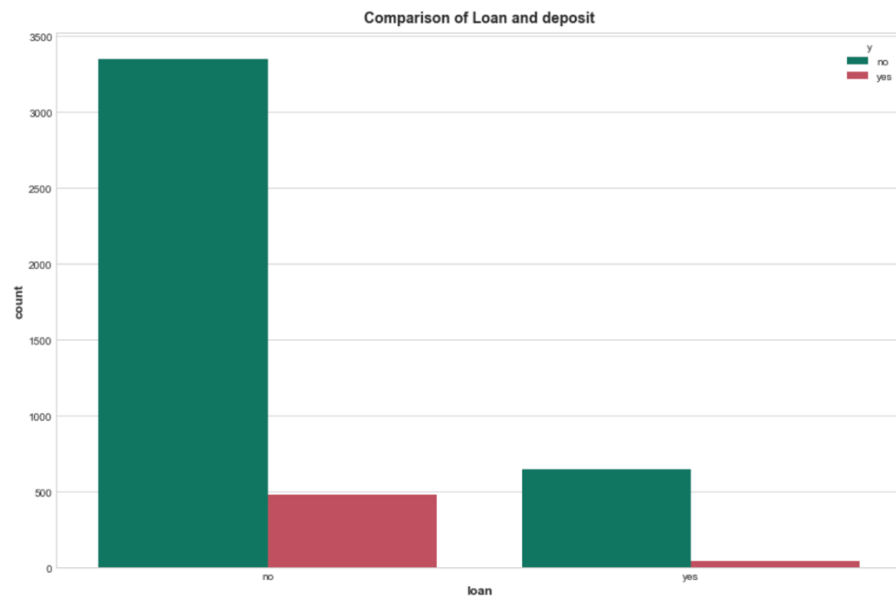


Figure 5 - Comparison of Loan and Deposit

Figure 6 shows that married and divorced people are more likely to subscribe for the term deposit than singles. The mean age of singles, married and divorced is 34, 43 and 45 respectively. This proves that the recommendation to contact people over 40 is valid since they are married, divorced and have more balance.

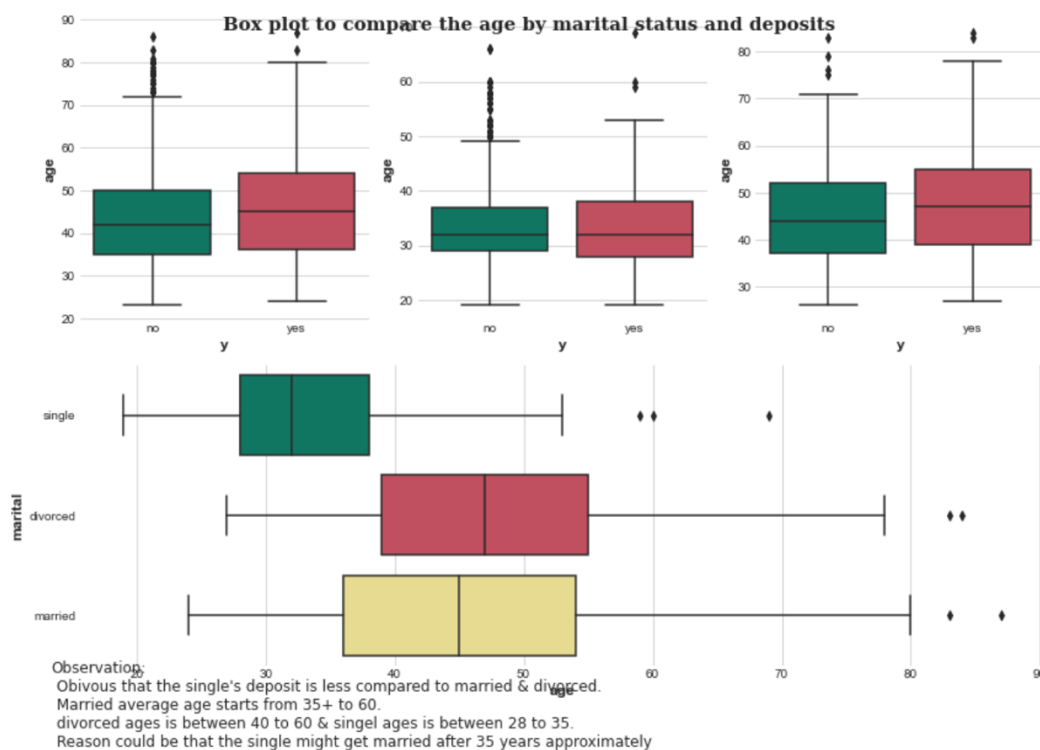


Figure 6 - Comparison of Age by Marital Status and Deposit

Recommendations

Key outcomes of the analysis are the recommendations for future marketing campaigns:

The three major attributes from the analysis are *Age*, *Duration*, and *Poutcome*. Drawing from our analysis, we recommend that the bank should contact customers who are in their 40's or above, in management or a technician, married, with minimum secondary school education, and with account balance above £1600 as they are most likely to subscribe to a term deposit account.

Another recommendation is to take into consideration is to contact customers who have been contacted before because 65% of the customers who subscribed were contacted before. Also, the marketing team should engage in longer conversations with the customers because the data shows that the higher the duration, the more likely it is that the customer will subscribe to the term deposit.

The customer's account has a huge influence on the campaign's outcome. People with an account balance above £1600 are more likely to subscribe for term deposit. The customer's age, job, and mode of contact affects campaign outcome as well. Future campaigns should contact customers from 40 years old working in management or as technicians via cellular phone. Number of contacts with the customer during the campaign is also very important as it increases the likelihood of a customer to subscribe for the term deposit.

Bonus: Additional data processing or analysis shown

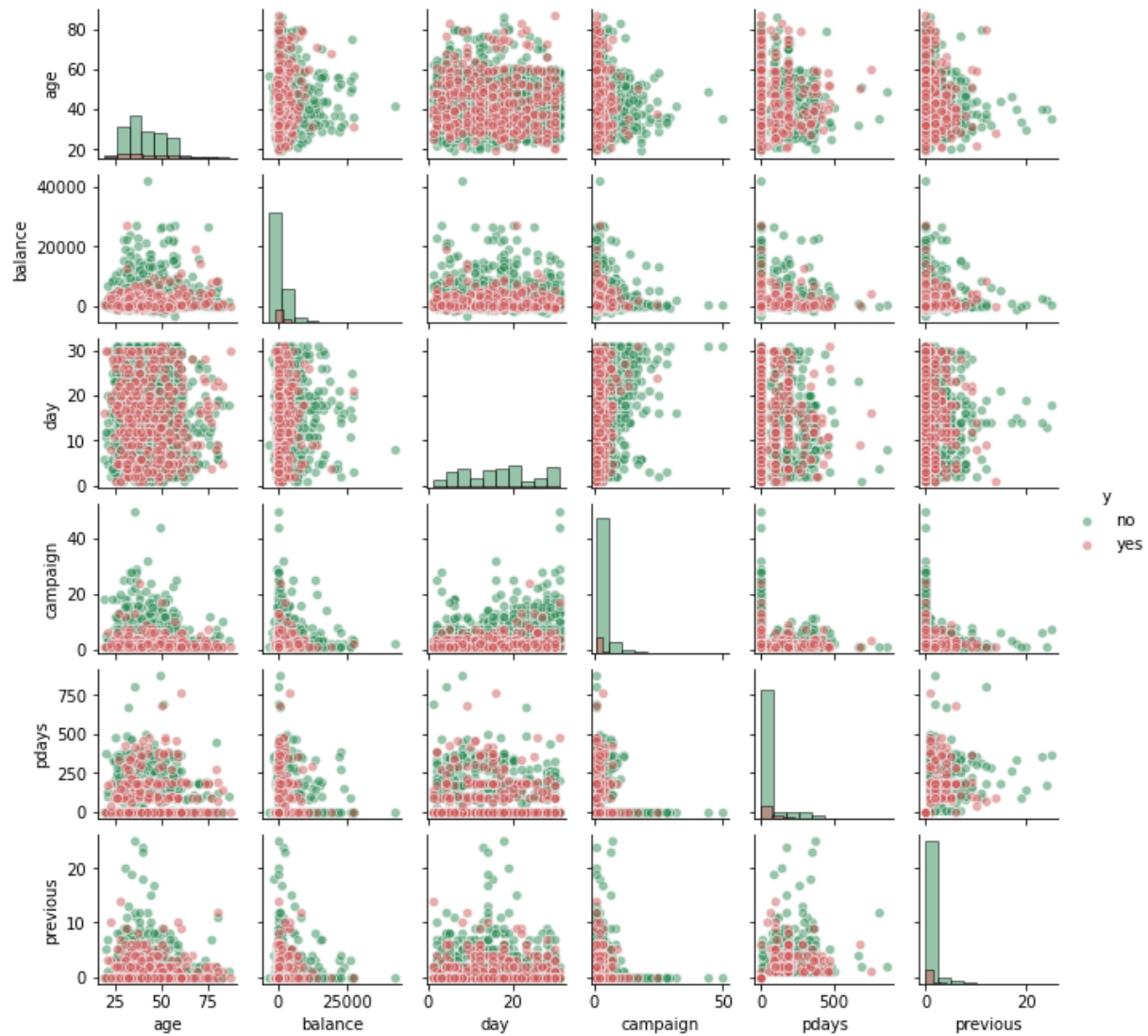


Figure 7 - Scatterplots to search for Linear and non-Linear Relationships and Histogram

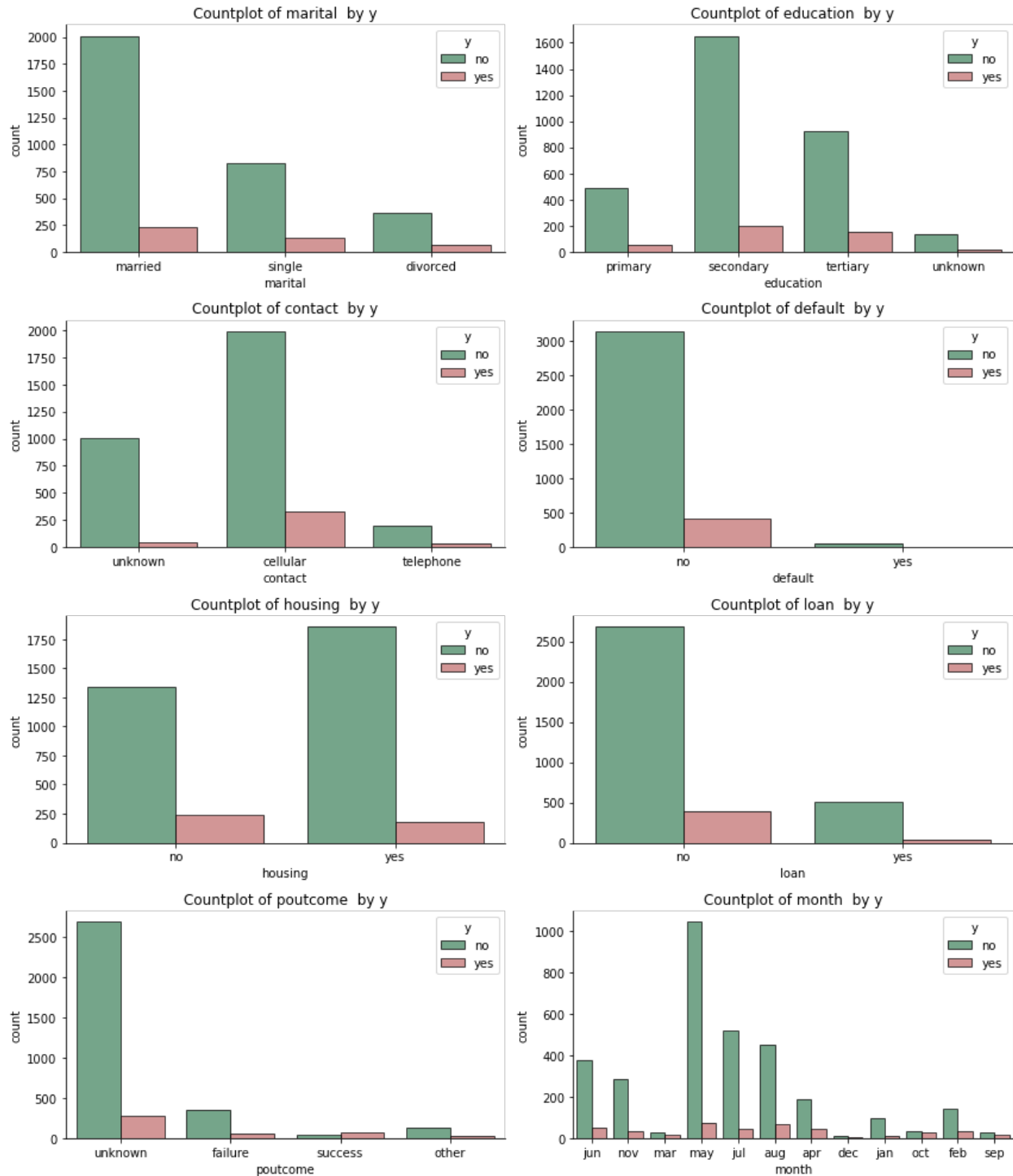


Figure 8 - Bar Charts showing the Frequency of each Attribute

Workload Distribution

All members assisted and shared their thoughts and findings on every aspect of the project. The report was equally written by all members and collated into one single report.

Member Name	List of Tasks Performed
Ime Precious	Data Preparation
Sharaf Malak	Predictive Modelling
Sharafutdinova Iuliia	Summary and Conclusion/Recommendation

References

- Alberto Quesada, A. (2017, January). *3 methods to deal with outliers*. Retrieved from KD Nuggets: <https://www.kdnuggets.com/2017/01/3-methods-deal-outliers.html>
- Barnes, R. (2021, August 9). *What is an Outlier?* Retrieved from The Data School: <https://dataschool.com/fundamentals-of-analysis/what-is-an-outlier/>
- Brownlee, J. (2017, July 14). *Machine Learning Process: What is the Difference Between Test and Validation Datasets?* Retrieved from Machine Learning Mastery: <https://machinelearningmastery.com/difference-test-validation-datasets/>
- Data Partition*. (n.d.). Retrieved from Statistics: <https://www.statistics.com/glossary/data-partition/#:~:text=Data%20partitioning%20in%20data%20mining,is%20selected%20for%20the%20partitions.>
- Frost, J. (n.d.). *5 Ways to Find Outliers in Your Data*. Retrieved from Statistics By Jim: <https://statisticsbyjim.com/basics/outliers/>
- Grace-Martin, K. (n.d.). *Outliers: To Drop or Not to Drop*. Retrieved from The Analysis Factor: <https://www.theanalysisfactor.com/outliers-to-drop-or-not-to-drop/>
- Gupta, P. (2017, May 17). *Decision Trees in Machine Learning*. Retrieved from Towards Data Science: <https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052>
- Tripathi, H. (2019, September 24). *What Is Balanced And Imbalanced Dataset?* Retrieved from Medium: <https://medium.com/analytics-vidhya/what-is-balance-and-imbalance-dataset-89e8d7f46bc5>
- Wade, C. (2018, August 21). *Transforming Skewed Data*. Retrieved from Towards Data Science: <https://towardsdatascience.com/transforming-skewed-data-73da4c2d0d16>
- Zach. (2020, January 22). *What is Considered to Be a “Strong” Correlation?* Retrieved from Statology: <https://www.statology.org/what-is-a-strong-correlation/>